

Openness in Higher Education: Open Source, Open Standards, Open Access

Brian Kelly¹; Scott Wilson²; Randy Metcalfe³

¹ UKOLN, University of Bath, Bath, BA2 7AY, United Kingdom

e-mail: b.kelly@ukoln.ac.uk

² CETIS, University of Bolton, Deane Road, Bolton BL3 5AB, United Kingdom

email: scott.bradley.wilson@gmail.com

³ OSS Watch, University of Oxford, 13 Banbury Road, Oxford OX2 6NN, United Kingdom

e-mail: randolph.metcalfe@oucs.ox.ac.uk

Abstract

For national advisory services in the UK (UKOLN, CETIS, and OSS Watch), varieties of openness (open source software, open standards, and open access to research publications and data) present an interesting challenge. Higher education is often keen to embrace openness, including new tools such as blogs and wikis for students and staff. For advisory services, the goal is to achieve the best solution for any individual institution's needs, balancing its enthusiasm with its own internal constraints and long term commitments. For example, open standards are a genuine good, but they may fail to gain market acceptance. Rushing headlong to standardize on open standards may not be the best approach. Instead a healthy dose of pragmatism is required. Similarly, open source software is an excellent choice when it best meets the needs of an institution, but not perhaps without reference to those needs. Providing open access to data owned by museums sounds like the right thing to do, but progress towards open access needs to also consider the sustainability plan for the service. Regrettably institutional policies and practices may not be in step with the possibilities that present themselves. Often a period of reflection on the implications of such activity is what is needed. Advisory services can help to provide this reflective moment. UKOLN, for example, has developed a Quality Assurance (QA) model for making use of open standards. Originally developed to support the Joint Information Systems Committee's (JISC) digital library development programmes, it has subsequently been extended across other programmes areas. Another example is provided by OSS Watch's contribution to the development of JISC's own policy on open source software for its projects and services. The JISC policy does not mandate the use of open source, but instead guides development projects through a series of steps dealing with IPR issues, code management, and community development, which serve to enhance any JISC-funded project that takes up an open source development methodology. CETIS has provided a range of services to support community awareness and capability to make effective decisions about open standards in e-learning, and has informed the JISC policy and practices in relation to open standards in e-learning development. Again, rather than a mandate, the policy requires development projects to become involved in a community of practice relevant to their domain where there is a contextualised understanding of open standards.

Keywords: open standards; open source; open access; quality assurance; advisory services

1 Introduction

The Joint Information Systems Committee (JISC) of the UK education funding councils has been engaged in a long-running process of engaging with the concept of 'openness' in educational technology and digital content. This engagement has moved through several phases, from initial evangelism into today's more pragmatic stance, and effected through the agency of three services:

- UKOLN has been charged with the development of the JISC information environment, formerly known as the Distributed National Electronic Resource (DNER), the UK education sector framework for the distribution of published digital content;
- CETIS is the Centre For Educational Technology & Interoperability Standards, and has responsibility for the development of open standards to support e-learning;
- OSSWatch provides advice and guidance on the use of open-source technologies in education.

Together these three services offer the JISC support at the policy and strategy level on the three strands of 'openness' in technology discourse - namely, open content, open standards, and open source. In each area the emergence of widespread use of social software and distributed systems (the 'Web 2.0' phenomena) has provided a disruption affecting each service and its strategy on 'openness'.

2 Definition of Open Standards

In a paper on open standards it is important to have a clear definition of the meaning of the term. In practice, however, it can be difficult to reach an agreed definition. Rather than attempting to produce a formal definition the following list of the characteristics of open standards is given:

- The development of open standards is the responsibility of a trusted neutral organisation;
- The responsibility for the ongoing maintenance and development of open standards is taken by a trusted neutral organisation;
- Involvement in the development of open standards is open to all;
- There are no discriminatory barriers to use of open standards;
- Access to open standards is available to all, without any financial barriers.

It should be noted, however, that such characteristics do not necessarily apply to all organisations with a responsibility for open standards. For example within organisations such as W3C (the World Wide Web Consortium) discussions on areas in which standardisation will occur are decided by member organisations who have paid the required membership fee. Similarly the initial discussions and agreements on the preferred approaches to the standardisation work may be determined by such member organisations. Also standards produced by organisation such as the BSI (British Standards Institution) are not necessarily available free-of-charge.

3 Why Use Open Standards?

Open standards are important in the development of networked services for several reasons. They aim to:

Support interoperability: Interoperability is often critical to those creating digital services. There will be a need to ensure that services and data can be used not only within a correct environment, but also across other digital services and across other application areas. A prime purpose of open standards is to provide such interoperability.

Maximise access: Cultural heritage services normally seek to maximise access to their resources and services. Ideally access will not be limited by constraints such as the device used by the end user; their physical location; their location on the network; etc. or personal factors such as disabilities;

Provide application- and device-independence: The dangers of lock-in to particular applications or hardware platforms are widely acknowledged;

Ensure architectural integrity: Unlike proprietary solutions, for which the development and intended usage is likely to be constrained by commercial and competitive factors, open standards which are developed within a wider context can help to ensure architectural integrity across a wide range of scenarios;

Provide long-term access to resources and services: Long term access to scholarly resources and cultural heritage resources is of particular importance for public sector organisations.

The authors of this paper feel that an understanding of such benefits is widely accepted within the development community. What, therefore, are the barriers to an implementation of a vision based on this approach?

4 The Complexities of Open Standards

The reality is that despite the widespread acceptance of the importance of open standards and the feeling among some that use of open standards should be mandatory in the development of networked services in practice, many organisations fail to implement open standards in their provision of access to digital resources. This may be due to several factors:

Disagreements over the Meaning: There are many complex issues involved when selecting and encouraging use of open standards. Firstly there are disagreements over the definition of open standards. For example Java, Flash and PDF are considered by some to be open standards, although they are, in fact, owned by Sun, Macromedia and Adobe, respectively, who, despite documenting the formats and perhaps having open processes for the evolution of the formats, still have the rights to change the licence conditions governing their use (perhaps due to changes in the business environment, company takeovers, etc.) Similarly there are questions regarding the governance of apparent open standards, with the control of RSS 1.0 and RSS 2.0 providing an interesting example; this lightweight but powerful syndication format for Web context has a complex history plagued by disagreements over governance and the roadmap for future developments;

Difficulties in Mandating and Enforcing Compliance: There are also issues with the mandating of open standards. For example: What exactly does 'must' mean? When told you must comply with HTML standards a developer working on a project might first ask what if I don't? Then what if nobody does? They might also ask what if I use PDF instead of HTML? There is a need to clarify the meaning of must and for an understandable, realistic and reasonable compliance regime;

Failure in the Market Place: It also needs to be recognised that open standards do not always succeed in gaining acceptance in the market place: they are often regarded as too complex to be deployed and the user community may be content to use existing closed solutions and reluctant to make the investment needed to make changes to existing working practices;

Failure to Satisfy User Needs and Expectations: There is a danger that a development approach over-emphasises the importance of open standards to the detriment of the end user and the end user's needs and expectations. It is often tempting to look only at the benefits of open standards for the developer or the provider of a service. We can see the temptation to develop a service based on a rich standard which can address a wide variety of use case scenarios. The danger would be that the end user rejects the service in preference to a simpler one.

Despite such reservations, in reality many IT development programmes are successful. The success may be based on the deployment of agreed and well-defined open standards. However in other cases development work may adopt a more pragmatic approach, making use of mature open standards, but having a more flexible approach to newer standards, for which there has been no time to reflect on the strengths and weaknesses and the experiences gained in their use.

5 Experiences in the UK

The Joint Information Systems Committee (JISC) (<http://www.jisc.ac.uk/>) who provide leadership in the innovative use of Information and Communications Technology to support education and research in the UK, have traditionally based their funding of development programmes around the use of open standards. Technical development for JISC's eLib programme, which was launched in 1996, was based on a standards document (eLib, 1996). The document formed the basis of a revised standards document which was produced to support JISC's Distributed National Electronic Resource (DNER) programme (which was later renamed the JISC Information Environment). Standards document (JISC, 2001). This work in turn influenced the NOF-digitise Technical Standards document (NOF, 2001) which was used by the national NOF-digitisation programme, which was responsible for digitisation projects across the cultural heritage sector.

The authors have been involved in providing technical advice and a support infrastructure for JISC-funded development programmes.

Experiences of the QA Focus Project

Although projects funded by the eLib programme were expected to comply with the eLib standards document, in practice compliance was never formally checked. It was probably sensible at the time (the mid 1990s) to avoid mandating a formal technical architecture and corresponding open standards – that could easily have led to mandating use of Gopher! In those early days of the Web, we were seeing rapid developments in the variety of services which were being provided on the Web and many new open standards being developed. However over time, and as the Web matured and the rate of innovation slowed, there was an increasing realisation of the need to provide a more stable environment for technical developments and the corresponding need to address the issue of compliance.

In 2000 JISC funded the QA Focus project (<http://www.ukoln.ac.uk/qa-focus/>) to develop a quality assurance framework, which would help ensure that future projects would comply with standards and recommendations and deploy best practices (Kelly, 2003). The project's aim was to develop a quality assurance (QA) methodology which would help to ensure that projects funded by JISC digital library programmes were functional, widely accessible and interoperable; to provide support materials to accompany the QA framework and to help to embed the QA methodology in projects' working practices. Liaison with a number of projects provided feedback on the current approach to use of standards. The feedback indicated: (a) a lack of awareness of the standards document; (b) difficulties in seeing how the standards could be applied to projects' particular needs; (c) concerns that the standards would change during the project lifetime; (d) lack of technical expertise and time to implement appropriate standards; (e) concerns that standards may not be sufficiently mature to be used; (f) concerns that the mainstream browsers may not support appropriate standards and (g) concerns that projects were not always starting from scratch but may be building on existing work and in such cases it would be difficult to deploy appropriate standards. Many of these were legitimate concerns, which needed to be addressed in future programmes.

This feedback was very valuable and provided a counter-balance to views which suggested the need for a heavyweight compliance regime which forced projects to comply fully with a technical architecture and corresponding open standards. The feedback led to the development of a contextual framework which is described later.

6 Open Standards: The CETIS Experience

In the late 1990s CETIS began life as the UK IMS Centre, a project funded by JISC to engage in the new IMS (instructional management systems) specification consortium. IMS began developing a series of specifications for XML data and content interoperability for elearning following the emerging paradigm of 'Learning Objects'. CETIS engaged in the development of the specifications, while also engaging with the UK education community to disseminate information about open standards, promoting a message that placed open standards as the key mechanism for preventing vendor lock-in and supporting long-term sustainability for the newly emerging 'Virtual Learning Environment' technology sector. As the sector developed, CETIS expanded to engage in a wide range of open standards work at a UK, European, and international level.

This message proved very attractive for policy-makers, who were keen to find a new procurement strategy following the unpopularity of the 'single primary vendor' approach that had been used previously within the schools sector, but still needed to provide some form of strategic co-ordination to prevent resources being wasted. Open standards seemed an ideal tool for this policy task: standards could be mandated such that the choice of systems were restricted to those that could conform; these conforming systems could then be more easily replaced by institutions using the interoperability effected by open standards if they were no longer the optimal choice. This style of procurement policy was adopted in various ways by the Learning and Skills Council, JISC, BECTa, and the DfES, and continues to be the key approach of agencies in the UK education sector to this day through initiatives such as the e-Framework¹, the BECTa Learning Platform Framework, and DfES Information Standards Board.

While the overall message has been an attractive one at the policy level, the experience of open standards at a practical level has proved less clear-cut. In particular, the intended effect of interoperability and reduced opportunity for vendor lock-in has not always been well served by the means of open standards. There have been influences from the political, business and technology context of the development and application of open

¹ See <http://www.e-framework.org>

standards that in some cases have served to either reduce or completely reverse the effect of standards on interoperability.

The process of standardisation can be a difficult one for those concerned. For example, the specification process itself was being driven largely by the vendors themselves, for whom it may be argued the interests are not served best by the agenda of open standards. A good example of this is the first attempt by IMS at a standards framework for Learning Management Systems. This was implemented by the company now known as BlackBoard as a 'reference implementation' of the APIs defined by IMS. However, this reference implementation formed the basis of a product (the BlackBoard LMS) that competed with the other consortium offerings, resulting in the collapse of the first standards agreement.

IMS reorganised its efforts and offered a second set of standards based on XML document transfer rather than system APIs. These new standards had their own problems, however. Many of the new specifications offered little real interoperability as practically all aspects of the specification had become optional to accommodate the diverse capabilities of consortium members. Customers attempting to use the specifications to interoperate systems found that their vendors had implemented incompatible subsets of the specification that resulted in data and content transfer requiring costly manual transformation; the very thing standards had sought to eliminate. In response a number of application profiles were developed, the most well-known today being SCORM², to improve interoperability for particular purposes.

In some cases interoperability in practice did not match customer expectations. For example, the early implementations of IMS Content Packaging, the specification for open transfer of content by e-learning systems, used an approach one of the authors of this paper calls the 'white screen of lock-in' approach. This involves inserting between the open content manifest, and open (typically HTML) content a layer of proprietary XML metadata containing instructions to a specific system on how to load the content. Other systems importing the content see the table of contents, but as users click on items in the table all they see is a blank screen as the system renders the proprietary metadata instead of the content. This approach was used by both WebCT and Blackboard in their initial implementations; it may be the case here that neither company expected the specifications to be actually used for interoperability purposes, but simply wanted to assert 'conformance'. At this point in the development of the market it is also highly likely that most customers had just taken delivery of systems and were probably not very interested in ensuring they had a clear exit strategy, and were quite happy to take a conformance statement as sufficient evidence of goodwill in terms of future interoperability.

The issue of standards conformance and compliance has been a difficult one within the e-learning community, particularly with the number of competing application profiles developed. The general approach CETIS took was to take the pragmatic step of inviting vendors to demonstrate working interoperability with other partners within a closed environment, giving developers the opportunity to identify and fix issues before exposing interoperability problems to customers. An alternative approach was to take a more rigorous approach to the definition of application profiles with the intent of producing formal conformance tests, which was the subject of the TELCERT project. CETIS was also involved in the development of the RELOAD³ tool to implement the IMS content specifications in a rigorous fashion to help users overcome interoperability issues. Today, many institutions use RELOAD to fix errors in standards-conformant content, or convert between incompatible implementations.

There has also been the claim from many smaller vendors that the standards developed by consortia (which often require annual membership fees for access) are themselves a form of lock-in. By releasing complex specifications that are difficult to implement, a barrier to entry is raised that only the largest vendors can afford to cross. This accusation has been levelled at a range of standards, most notably the Web Services specification stack promoted by Microsoft, Sun and BEA, which has swelled to an enormous volume of standards weighing in at thousands of pages. Whether this is a result of deliberate conspiracy or a rather monolithic development approach is moot; the overall effect has been that some developers have found WS-* excessively cumbersome and instead embraced various forms of simpler web services based on HTTP and XML (e.g. REST⁴) using simple proprietary API definitions. These proprietary lightweight APIs are the basis of many of the services considered part of "Web 2.0", such as del.icio.us, Blogger, and Google. It should be noted, however, that in another case, IMS QTI, smaller vendors were actually more able to implement a complex specification than the major vendors, so the argument that standards can be raised as a barrier to entry needs to be looked at critically.

² Sharable Content Object Reference Model. See <http://www.adlnet.org/>

³ Reusable Learning Object Authoring and Delivery. See <http://www.reload.ac.uk>

⁴ Representation State Transfer. An architectural model for web resources. See Fielding, 2000.

Another twist in the open standards story has been the issue of patents and IPR claims. While open standards are generally thought of as being free to use, this is conditional on the licensing of appropriate patents by contributing companies and the copyright policy of the standards organisation. In two recent cases, this has resulted in the 'encumbering' of open standards with patent issues. The first case involved the company ContentGuard, who were granted US patents for a range of technologies concerned with Digital Rights Management (DRM). ContentGuard actively engaged in the standardisation process through IEEE, developing the Open Digital Rights language (ODRL) in competition with their own XML Rights Management Language (XrML). However, they did this knowing that whichever technology customers used, they would still have to pay a license fee to ContentGuard, even if they chose to use the 'open' standard. The ContentGuard DRM patent situation has been the ongoing subject of legal disputes and commercial negotiations (Rosenblatt, 2005).

The second case involved the infamous '44 claims' of the Blackboard patent (see Feldstein, 2006, and Geist, 2006), which covers many of the features of modern e-learning systems, many of which were implemented by Blackboard at its inception as a result of implementing the first IMS specifications. Ironically, this then created the situation where vendors and open-source projects were then unsure whether adopting IMS specifications would also result in patent infringement. The patent issue, combined with the merger of Blackboard and WebCT into a single dominant vendor, have increased the pressure on institutions to create an exit strategy from their existing platform. Open standards should have made this far easier to accomplish this type of technology switch, which will be costly to implement for many institutions involving a large amount of content and data migration.

The use of patents as bargaining power, leverage, and influencer in open standards has been considered in other sectors, for example, Henrik Glimstedt's work on analysing the open standards process within the mobile telephony market (Glimstedt, 2001 & 2000). However, in educational technology patents in open standards have only recently become an important factor as a result of the Blackboard case.

While standards are a technology artifact, the process of constructing a standards involves an interplay of political and economic motives and is not simply a quest for an optimal technical solution. Where efforts on a particular axis are stalled or meet with opposition, a common tactic is for the proponents to find a new venue to pursue standardisation goals; a useful analysis of how the standards process involves the interplay of personal and organisation motives is given in zur Muehlen et al (2005) in their description of the evolution of open standards for workflow, and how various standards bodies have engaged in a sort of dance with various key players moving between organisations to pursue particular goals. In e-learning a similar interplay has been seen with new standards organisation proposed or created in response to the changing political or business context, such as HEKATE and LETSI. Krechmer (2005) set out a set of criteria for openness in standards, covering the areas of participation (open meetings, consensus, due process), dissemination (open IPR, open change, open documents) and usage (one world, open interface, ongoing support) which in practice are hard to reconcile with the practices of standardisation as seen in the organisations CETIS works with. While most specification bodies have due process, an open IPR policy of some sort, and a one world (i.e. single international standard rather than regionalisation) approach, most do not support open meetings and instead favour a membership payment model. IMS, for example, decided in 2006 to delay releasing draft documents for public scrutiny to provide a competitive advantage for subscribing members; while understandable in terms of marketing membership fees, this violates Krechmer's 'Open documents' principle. Taken together, Krechmer's principles, applied in practice, show there is a great deal of interpretation possible for the meaning of 'open' in an 'open standard'.

To date, a substantial part of the effort of CETIS has been influencing the prevention of unnecessary or conflicting standards rather than the creation of desirable standards. An example of the type of case where standards prevention is necessary is where standardisation is initiated very early in the development of a technology, in a situation where adoption of a standard would genuinely impact on innovation (this is unusual; mostly, the opposite is true, as standards unlock opportunities to innovate). While early standardisation can be very tempting as a 'land grab' technique by pioneers in new types of applications, it can ultimately be damaging to the healthy diversity of solutions on offer as it sets, rather than an interoperability specification, a de jure dominant design which prevents entry into the market by alternatives (Abernathy and Utterbeck, 1978). The e-learning standards area was dominated early on by what Baskin, Krechmer and Sherif (1998) call anticipatory standards: "standards that must be created before widespread acceptance of the device or services", rather than responsive or participatory standards. This can be interpreted as "whoever defines the standard designs the future", and provides a temptation to develop standards prematurely.

While there are known caveats and issues in the area of open standards, there have also been some remarkable successes achieved as a result; understanding the critical success factors involved in open standards is an

ongoing effort by CETIS. Tim Bray, one of the original developers of XML, considers that the number of successful XML-based standards is very small, and that 5 critical standards (HTML, DocBook, ODF, UBL, and Atom) form the core of achievement in XML standards to date (Bray, 2006). In other sectors, such as mobile telephony, there has been a considerable body of research on the standards process and its contribution to the mobile telephony market (see, e.g., Glimstedt, 2000, 2001; Pfannes provides an excellent overview of sources).

This complex story has informed the evolution of the approach to open standards taken by CETIS, which since procurement as a JISC service in 2006 (as the JISC-CETIS Service) has moved away from promoting adoption of open standards in a fairly unambiguous way to explicitly supporting a more complex message on interoperability. While the goal of interoperability has remained the same, and is at the heart of the strategy of the JISC-CETIS Service, the means by which interoperability is achieved is now seen by JISC-CETIS as having a number of strands and strategies, only some of which involve the use of open standards. The new multi-faceted approach sees a role for a range of technology interventions to achieve interoperability:

- adoption of open standards to exchange data and content;
- adoption of common infrastructure, such as the emergence of de-facto common libraries and open-source platforms;
- common implementation patterns and conventions that make it easier to engineer interoperating solutions;
- post-hoc interoperability achieved using latent semantic analysis and other techniques to analyse proprietary systems and their data;
- proprietary but publicly-documented interfaces;
- open processes and communications that support a dialogue about interoperability;
- adoption of emerging standards and patterns from communities of practice.

Some of these new strands have been added to the JISC-CETIS strategy as a result of observing the development of working interoperability within Web 2.0, where the standards process has, if anything, been even more convoluted and compromised than in the education sector (there are, for example, somewhere from 7 to 9 known variants of 'RSS'; see Pilgrim, 2004). The interoperability that has been achieved using the basic approach of 'Simple, Sloppy, and Scalable' (as Google's Adam Bosworth puts it; see Steinberg, 2005) has been highly successful and enabled large numbers of new services and initiatives. By contrast, the e-learning sector has seen a long period of consolidation with relatively little innovation but increasing costs. In some cases it may be argued that the prevalence of open standards may have actually reduced practical interoperability; for example, the existence of learning object specifications such as SCORM and IEEE Learning Object Metadata, and their place in mandated conformance and procurement regimes, may have negatively impacted the uptake of content syndication formats (RSS, Atom) in education.

This broader approach to interoperability seems to offer a much greater prospect of lasting impact than a purely standards-based approach, as it enables JISC-CETIS to engage with a wider range of communities and stakeholders and to try different strategies to meet particular needs. For example, it allows JISC-CETIS to engage with open-source initiatives such as Moodle and LAMS in a more balanced way in terms of their overall impact and value, rather than keeping a standards-conformance scorecard as a simplistic measure of positive impact. It also offers a more pragmatic basis to look at the role of Web 2.0 services, and wholly proprietary developments such as Second Life, in the evolving picture of e-learning technology.

The CETIS/JISC-CETIS experience represents an evolution in the organisation's understanding of the concept of 'openness' in terms of interoperability as an interplay of many factors. The net result of this new pragmatism is to focus attention on the desired state and the role in which 'openness', in various forms, contributes to progression towards it. Rather than recommending that organisations mandate open standards and enforce their conformance, JISC-CETIS instead encourages interoperability conversations and convergence on common approaches, backed up by simple functional evaluations of interoperability in practice.

7 Open Source: The OSSWatch Experience

Early evidence demonstrated that open source software was used in UK higher and further education institutions in advance of any advisory service being set up by JISC⁵. Much as expected, OSS Watch's initial scoping study in 2003 revealed a mixed economy. No institution was maintaining an exclusively proprietary nor exclusively open source environment. That raised a number of questions for a new advisory service.

Why were institutions turning to open source solutions? Institutional policy in this area notoriously lags behind practice. OSS Watch's 2006 survey, for example, found that less than 25% of institutions had any mention of open source in their IT policies (OSS-Watch, 2006). Yet more than 75% investigated open source solutions at every viable opportunity.

The top three reasons that institutions gave for considering open source software (in 2003) were: interoperability, cost, and security. Interoperability, in particular, was a surprise. However, in retrospect it seems clear that the tendency for open source software to conform to open standards was already beginning to reap benefits with the infrastructure IT stack. This connection between open source and open standards needed elaboration if unbiased advice and guidance was to be provided to universities and colleges in the UK.

One challenge that we face initially is purely definitional. What is 'open source software'? There are competing useful guides. The earliest and more philosophically driven movement is free software movement, led by the Free Software Foundation and its Free Software Definition (Free Software Foundation, 2005). A related but less ideologically motivated option is the Open Source Initiative's Open Source Definition (OSI, 2006). The latter is, of course, based on the Debian Free Software Guidelines (Debian Project, 2004). In addition to these there are numerous other more local variations. However, since OSS Watch was established by its funders as an open source software advisory service, it seemed most sensible to accept the OSI's definition of open source software. Thus in numerous places on the OSS Watch site you will find a clear statement that, "For OSS Watch open source software is always software released under an Open Source Initiative (OSI) certified licence" (OSS-Watch, 2005).

A clear and consistent statement of what open source software is, however, does not require suppression of alternate characterisations of free software. OSS Watch regularly makes reference especially to the Free Software Definition and encourages institutions to become familiar with the differences in language and intent between the significant groups in this space. Universities and colleges engaging with free and open source software in a sensible fashion cannot be shielded from the complexities of their own engagement. On the other hand, dealing with these complexities head on can alleviate some of the anxiety they may generate for those less certain of their grounding here.

In some respects open source software is better placed, definitionally, than open standards. There appears to be universal agreement that the Open Source Initiative is the maintainer of the Open Source Definition, even if some vendors do not feel bound by the need to pursue OSI certification for licences they describe as "open source" under which they release some or all of their software. For a time this practice can weaken the clarity that an advisory service can provide in its advice and guidance. Fortunately, the open source community is such that most high-profile vendors flouting the norms of the "open source" appellation find the negative public relations it generates to be counter-productive. Recently a number of such companies have reformed their practices and can now be acknowledged as open source companies even by OSS Watch⁶.

Whether a project is using an OSI-certified licence is important. It underwrites what can usefully be said about its licensing conditions in the absence of additional paid legal advice. Institutions involved in procurement exercises are not typically interested in software that requires additional legal advice to know what can and cannot be done with it or how it can be further developed, in the case where the source code is provided. This might be one explanation for the slow take up in the UK of the Bodington virtual learning environment (VLE) as against Moodle. Although Bodington was "open sourced" by its home development institution, the University of Leeds, the licence placed upon it was not OSI-certified. This created a challenge since it could not be proclaimed as open source software by those with a strict adherence to OSI-certification as the key marker of open source software. It took some years for the Bodington community to sort this licensing issue satisfactorily (OSS-Watch,

⁵ See OSS Watch Scoping Study, 2003, <http://www.oss-watch.ac.uk/studies/scoping/>

⁶ Notable here is Alfresco's move to a GNU GPL release of its principal codebase (see <http://www.alfresco.com/legal/licensing/whitepaper/>). A smaller example, but one prominent in the university web content management market is Squiz.net's MySource Matrix (see <http://matrix.squiz.net/evaluations/licence/choosing-gpl-or-ssv>).

2006b). In the interim Moodle, which is released under the GNU GPL, was able to increase its market share in UK further education colleges to approximately 56% (OSS-Watch, 2006).

However, although an OSI-certified licence is important, it is not the sole determining of software suitability. OSS Watch therefore avoids making specific software recommendations. Instead the principal task is to help universities and colleges understand legal, social, technical and economic issues that arise when they engage with free and open source software. The goal is not the promotion of open source software for its own sake. Indeed, for OSS Watch the choice of proprietary or open source solutions is immaterial. What matters is that institutions have the resources to think through their procurement, deployment, or development IT concerns in a sensible and rational fashion. The best solution for any single institution will depend upon local conditions and individual needs.

This pragmatic approach to advice and guidance is consistent with that employed by UKOLN in its work on standards. It is also a guiding principle in the JISC Policy on Open source software for JISC projects and services (JISC, 2005). This policy is based on the UK government policy in this area and should be seen as an implementation of that policy⁷. Neither the government policy nor the JISC policy mandate open source software for deployment or open source licensing for release of development outputs. Rather, both policies draw attention to open source as one possible exploitation route for software which has been developed with government funds. The JISC policy goes further, providing useful guidance notes for those projects wishing to take up an open source development methodology (see, e.g., Raymond, 1997, and Fogel, 2005).

OSS Watch works closely with JISC-funded development project to aid their understanding of open source development methodologies. Since the JISC policy essentially urges projects to “get their IPR house in order at the earliest possible time”, early consultation meetings using involve discussions around licence choice. Again a pragmatic approach rises to the top. Licence choice for software development project can be a fraught affair. The tendency to simply choose the licence you have heard most often mentioned is disconcerting. Without presuming to provide legal advice, OSS Watch helps projects think through the options available to them. In the end the choice will remain entirely in their hands, but issues such as compatibility with other code, potential for developing a community around the project, and an initial long term sustainability plan will certainly be explored.

8 A Contextual Approach

We have described some of the limitations of open standards and the feedback we have received from those seeking to make use of open standards in their development work. We have also described the experience of using open source. However, this need not mean an abandonment of a commitment to seek to exploit the benefits of open standards or open source. Nor should it mean imposing a stricter regime for ensuring compliance. Experience has made it clear that there is a need to adopt a culture, which is supportive of use of open standards and open source but provides flexibility to cater for the difficulties in achieving this.

This culture and approach is based on:

- A contextual model which recognizes the diversity and complexities of the technical, development and funding environments;
- A process of learning and refinement from patterns of successful and unsuccessful experiences;
- A support infrastructure based on openness, such as use of Creative Commons to encourage take-up of support materials and address the maintenance and sustainability of such resources.

It is apparent that there is a need to recognise the contextual nature to this problem; i.e. there is not a universal solution, but we should try to recognise local, regional and cultural factors, which will inform the selection and use of open standards.

Over time, in response to the problems outlined, the authors and others have developed a layered approach towards open standards intended for use in development work (Kelly, 2005). This approach is illustrated in Figure 1.

⁷ See http://www.govtalk.gov.uk/policydocs/policydocs_document.asp?docnum=905

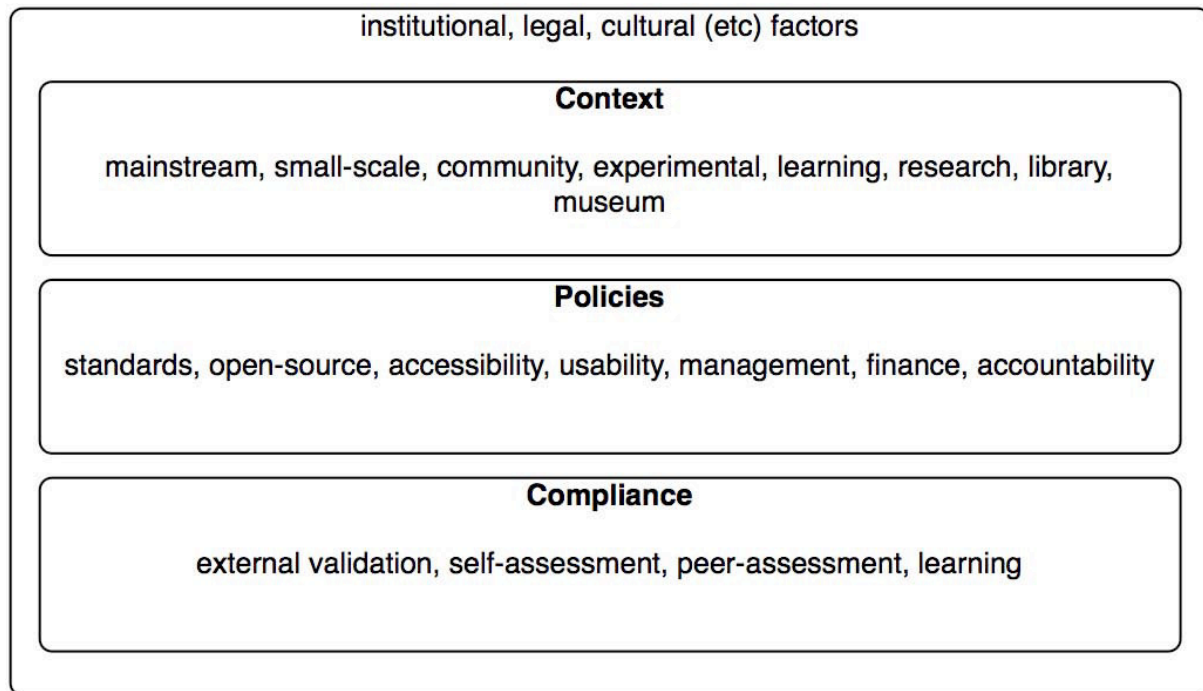


Figure 1: A Layered Approach to Use of Standards and Open Source

This approach uses the following layers:

Contextual Layer: This reflects the context in which the standards or open source software are being used. Large, well-funded organisations may choose to mandate strict use of open standards in order to build large, well-integrated systems which are intended for long term use. For a smaller organisation, perhaps reliant on volunteer effort with uncertain long-term viability, a simpler approach may be more appropriate, perhaps making use of proprietary solutions;

Policy Layer: This provides an annotated description (or catalogue) of relevant policies in a range of areas, including open standards, open source, accessibility and accountability. The areas will include descriptions of standards, the ownership, maturity, risk assessment, etc. It summarises the strengths and weaknesses of the standards;

Compliance Layer: This describes mechanisms to ensure that development work complies with the requirements defined within the particular context. For large, public funded programmes there could be a formal monitoring process carried out by external auditors. In other contexts, projects may be expected to carry out their own self-assessment, or take part in peer-assessment with related projects. In such cases, the findings could be simply used internally within the project, or, alternatively, significant deviations from best practices could be required to be reported to the funding body.

It should be noted that, although it is possible to deploy this three-layered approach within a funding programme or community, there will be a need to recognise external factors, over which there may be no direct control. This may include legal factors, wider organisational factors (for example there are differences between higher and further education, museums, libraries and archives), cultural factors, and available funding and resources etc.

It is also important to note that the contextual approach is not intended to provide an excuse to continue to make use of proprietary solutions which may fail to provide the required interoperability. Rather the approach seeks to ensure that a pragmatic approach is taken and that lessons can be learnt from the experiences gained. In order to ensure that the experiences are shared across the development community (and more widely) it will be important to ensure that systematic procedures are in place to ensure that the experiences are properly recorded and that such experiences are widely disseminated.

A requirement that funded projects should document their decisions on the selection of standards, open source licenses, and open source software, and provide reports based on their experiences in their use will help to ensure that such information is recorded in a systematic way, providing this information in an open and easily accessed

fashion will help ensure that such information can be widely disseminated. The use of a Wiki, with RSS to allow the content to be syndicated and news of changes to the information, can help to support this.

After the selection and deployment of standards there will be a need to ensure that the standards are being used in an appropriate fashion. One means of ensuring that this happens is the use of a quality assurance framework. A similar approach may also be suitable, with minor modifications, for the selection of open source software, and open source licenses for development outputs.

9 Supporting a Contextual Approach

The provision and implementation of a model which provides a pragmatic approach to the selection and use of standards will not guarantee that appropriate decisions are made and that the selected standards are deployed in the most appropriate fashion. There also needs to be a support infrastructure in place which ensures that technical managers, implementers, designers and others involved in research and development activities are able to make technical decisions which are appropriate for the intended purpose.

A support model which is being developed is illustrated in Figure 2.

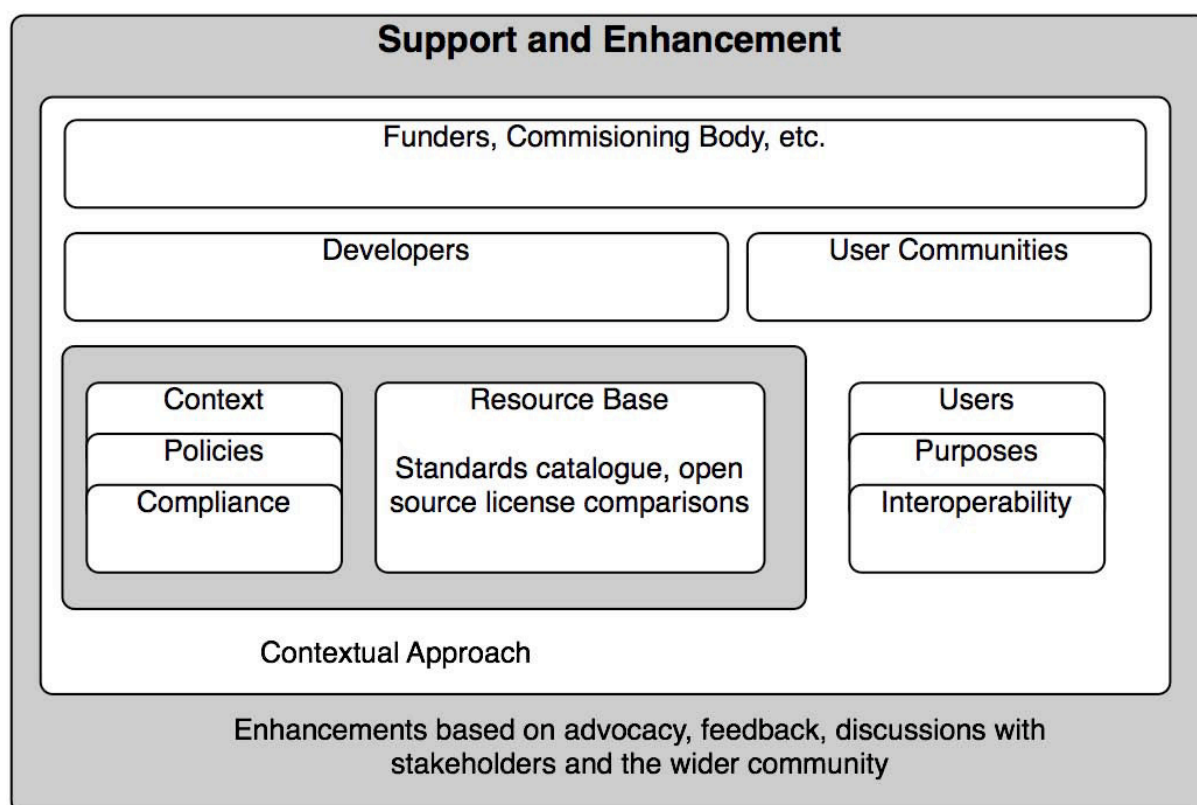


Figure 2: Support Model For Use of Standards

This support model is based on the following features:

The contextual model: This is described elsewhere in this paper. It should be noted that the contextual model primarily intended for use by the development community. The end user community need not be aware of the contextual model that was used as part of the development process;

User engagement: Engagement with the user community will be essential to ensure the sustainability of the approach – it needs to be remembered that the development approach is not an end in itself, but a means for satisfying the needs of the user community.

There are several user communities involved in development activities. The *development community* will typically focus on areas related to the standards, development approach and related areas. The *user community*, in contrast, will often be disinterested in such issues, concerned primarily with use of a service which functions effectively. Although developers should be aware of the needs to address end user needs, it may be difficult to

achieve this goal. It should therefore be a requirement of the *funding body* or organisation which has sponsored development work to ensure that mechanisms are put in place which will ensure that the approaches taken in development will ensure that the needs of the user community are satisfied. In the e-learning space, JISC-CETIS provides a range of Special Interest Groups (SIGs) that have a focus within a particular domain or context, where there is an effort on the part of the organisation to bring together developers and users to promote a better contextual awareness of the role of open standards.

Mechanisms for ensuring the development work is successful in meeting user needs may include:

Advocacy: There will be a need for the development community to promote the advantages of the preferred approaches to development. This could include promoting the advantages of use of open standards. Such advocacy needs to be tailored for the intended target audience, with other developers and end users requiring different approaches;

Feedback: A wide range of feedback will be required. For example, developers will need to provide detailed feedback on the contents of the resource base, funding agencies on the contextual model and implementation experiences, and end users on the end user service;

Engagement: A passive feedback mechanism is unlikely to provide useful feedback. A more effective approach would be to provide more engaging mechanisms that act not only as a one-way transfer of information, but provide richer two-way discussions;

Refinement: The feedback and engagement processes should help to refine those areas in which deficiencies have been identified. This could include over-simplistic or over-complex approaches to the development model.

10 Towards a Contextual Approach to Open Access?

So far in this paper we have looked in some detail at the experiences of advisory services in the adoption and use of open standards and open source software, and how this has led to the development of a contextual approach and support services to assist developers, agencies and users. How applicable is this work to the promotion of open access?

As with open source and open standards, open access is again clearly a “good thing” in principle, that in practice requires an understanding of the context of use, the policy framework within which the organisation operates, and an understanding of the measures that can be used to assess whether open access – or, perhaps, more accurately, the benefits intended to be realised using open access – have actually been achieved in practice. For example, the “green” and “gold” open access options (Harnad, 2004) could be treated in a similar way to the various approaches to open source licensing, and to choices of open standards. The contextual model would offer a resource base providing detailed information on each approach, a connection to the policy context (e.g. mandates), access to communities where experience has already been gathered on use, and a set of measures for conformance, such as community peer review and availability of outcomes for public scrutiny.

A support strategy for open access may use similar mechanisms to those for open source and open access, including advocacy, feedback, and refinement of the resource base in light of user experience and the active engagement and support of a community of use.

In the areas of open standards and open source we introduced the idea of transparency in the decision making process as part of the strategy for a pragmatic approach to adoption. In the case of open access, this would mean organisations and projects publicly documenting their decision on which open access strategy to adopt, or whether not to adopt an open access approach.

As noted earlier, it is also important to note that the contextual approach is not intended to provide an excuse to continue to not support open access. Rather the approach seeks to ensure that a pragmatic approach is taken and that lessons can be learnt from the experiences gained. For example, where existing open access strategies do not meet the requirements of particular contexts, and how new or hybrid strategies can be identified that better suit those contexts.

11 Conclusions

This paper has argued that what is needed is a more contextual approach to the open standards. It could be argued that what we need is not a list of open standards or open source licenses, or open access approaches but an *process for adopting open approaches* which is based on a desire to exploit the potential benefits of open standards, open source and open access, tempered by a degree of flexibility to ensure that the importance of satisfying end users needs and requirements is not lost and that over-complex solutions are avoided. This process could adopt the contextual approach documented in this paper and watch patterns of usage.

References

- [1] ABERNATHY, W. J.; UTTERBACK, J. M. (1978), Patterns of Industrial Restructuring. *Technology Review*, 80 (7), 1-9.
- [2] BASKIN, E.; KRECHMER, K.; SHERIF, M.H., (1998). The Six Dimensions Of Standards: Contribution Towards A Theory Of Standardization. In Louis Lefebvre, A., Mason, R., and Khalil, T. (1998, eds.), *Management of Technology, Sustainable Development and Eco-Efficiency*, Elsevier Press, Amsterdam, p. 53.
- [3] BRAY, T., (2006). Don't Invent XML Languages. *Ongoing(weblog)*. Retrieved from <http://www.tbray.org/ongoing/When/200x/2006/01/08/No-New-XML-Languages>
- [4] Debian Project, (2004). *Debian Free Software Guidelines, v1.1*. Retrieved from http://www.debian.org/social_contract#guidelines
- [5] eLib (1996), *eLib Standards Guidelines*. Retrieved from <http://www.ukoln.ac.uk/services/elib/papers/other/standards/>
- [6] FELDSTEIN, M., (2006). The Blackboard Patent Claims in Plain English. *e-Literate(weblog)*. Retrieved from http://mfeldstein.com/the_blackboard_patent_claims_in_plain_english/
- [7] FIELDING, R. T., (2000). *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, UC Irvine, 2000
- [8] FOGEL, K., (2005). *Producing Open Source Software*. Retrieved from <http://producingoss.com/>
- [9] Free Software Foundation, (2005). *The Free Software Definition*. Retrieved from <http://www.fsf.org/licensing/essays/free-sw.html>
- [10] GEIST, M.,(2006). Patent battle over teaching tools. *BBC News, August 14, 2006*. Retrieved from <http://news.bbc.co.uk/2/hi/technology/4790485.stm>
- [11] GLIMSTEDT, H., (2000). Politics of Open Standards, Modular Innovation, and the Geography of Strategic Patenting in GSM and UMTS Technologies. *Division of Innovation, Lund Institute of Technology*. Retrieved from <http://www.innovation.lth.se/files/Glimstedt%20April%2014th.pdf>
- [12] GLIMSTEDT, H., (2001). Competitive Dynamics Of Technological Standardization: The Case Of Third Generation Cellular Communications. *Industry and Innovation*, 8(1), 49-78. *Routledge*.
- [13] HARNAD, S.; BRODY, T.; VALLIERES, F.; CARR, L.; HITCHCOCK, S.; GINGRAS, Y; OPPENHEIM, C.; STAMERJOHANN, H.; HILF, E., (2004). "The green and the gold roads to Open Access," *Nature Web Focus*, <http://www.nature.com/nature/focus/accessdebate/21.html>
- [14] JISC, (2001), *Standards and Guidelines to Build a National Resource*. Retrieved from http://www.jisc.ac.uk/fundingopportunities/projman_standards.aspx
- [15] JISC, (2005), *Policy on Open source software for JISC projects and services*. Retrieved from http://www.jisc.ac.uk/fundingopportunities/open_source_policy.aspx
- [16] Kelly, B., Guy, M., and James, H., (2003). Developing A Quality Culture For Digital Library Programmes. *Informatica* 27(3) Oct. 2003.
- [17] KELLY, B.; RUSSELL, R.; JOHNSTON, P.; DUNNING, A.; HOLLINS, P.; PHIPPS, L. (2005). *A Standards Framework For Digital Library Programmes*. ichim05 Conference Proceedings. Retrieved from <http://www.ukoln.ac.uk/web-focus/papers/ichim05/>
- [18] KRECHMER, K. (2005). Open Standards Requirements. *The International Journal of IT Standards and Standardization Research*, 4(1), January - June 2006.

- [19] ZUR MUEHLEN, M.; NICKERSON, J.V.; SWENSON, K.D. (2005). Developing Web Services Choreography Standards - The Case of REST vs. SOAP. *Decision Support Systems 40 (2005) 1*, pp. 9-29.
- [20] NOF (2001), *NOF-digitise Technical Standards And Guidelines*. Retrieved from http://www.mla.gov.uk/resources/assets/T/technicalstandardsv5_pdf_7959.pdf
- [21] OSI, (2006). *The Open Source Definition*. Retrieved from <http://www.opensource.org/docs/osd>
- [22] OSS-Watch, (2005). *What is open source software?* Retrieved from <http://www.oss-watch.ac.uk/resources/opensourcesoftware.xml>
- [23] OSS-Watch, (2006). *OSS Watch 2006 Survey*. Retrieved from <http://www.oss-watch.ac.uk/studies/survey2006/execsummary.xml>
- [24] OSS-Watch, (2006b). *Bodington released under Apache License v2.0*. Retrieved from <http://www.oss-watch.ac.uk/resources/bodington-open.xml>
- [25] PFANNES, P. (2002). *Strategic Levers in Standardization Processes in the Mobile Communication Industry*. Thesis, Center for Digital Technology and Management, Munich
- [26] PILGRIM, M. (2005). *The myth of RSS compatibility*. DiveIntoMark (weblog). Retrieved from <http://diveintomark.org/archives/2004/02/04/incompatible-rss>
- [27] RAYMOND, E., (1997). *The Cathedral and the Bazaar*. Retrieved from <http://www.catb.org/~esr/writings/cathedral-bazaar/>
- [28] ROSENBLATT, B., (2005). Opposition Mounts to OMA DRM Patent Licensing Scheme. *DRM Watch, April 4, 2005*. Retrieved from <http://www.drmwatch.com/standards/article.php/3495026>
- [29] STEINBERG, D.H., (2005). Bosworth's Web of Data. *O'Reilly Network, 2005*. Retrieved from <http://www.onlamp.com/pub/a/onlamp/2005/04/22/bosworth.html>

Peer-to-Peer Networks as a Distribution and Publishing Model

Jorn De Boever

Centre for Usability Research, Department of Communication Science, K.U. Leuven
Parkstraat 45 (b 3605), 3000 Leuven, Belgium
e-mail: jorn.deboever@soc.kuleuven.be

Abstract

Content publishing and distribution often occurs in a costly and inefficient manner via client/server networks. Client/server models exhibit negative network externalities in that each additional user causes additional costs by increasingly congesting the system through the consumption of scarce resources. In an era of increasing demand for and size of content, the traditional client/server model produces evidence of its restrictions in terms of cost efficiency and scalability. Content providers – such as publishers, the media industry and users – are exploring new distribution or publishing models that might address the flaws of client/server models. An increasing amount of user generated content, open access and open content initiatives offer content for free, in spite of the fact that the distribution and storing of this content is not free of charge. This reasoning explains the importance of examining innovative distribution models that possibly provide answers to the shortcomings of client/server systems. In some cases, peer-to-peer systems might provide solutions for the flaws of client/server models in that they are characterized by cost efficiency and scalability. The facts that users spend more time online, have an increasing amount of resources (e.g. bandwidth, CPU cycles, content, and storage capacity) at their disposal, store and consume more content and bandwidth, is the basis of the viability of peer-to-peer systems. Peer-to-peer is still associated with illegal copyright infringing activities, although there are several companies exploring new ways for legal and secure content distribution through peer-to-peer networks. In this paper, we try to offer a broad analysis of the opportunities and challenges of several peer-to-peer applications and architectures. We further elaborate criteria in order to understand when the implementation of a peer-to-peer system might be appropriate. These criteria go further than merely technical criteria in that they include social criteria as well, which are as important as the technical ones. If peer-to-peer systems turn out to be a success for content publishing, it may lead to new business models that change the way content is distributed.

Keywords: peer-to-peer; e-publishing; content distribution; classification

1 Introduction

Mass content distribution through the internet often occurs in an inefficient and costly manner. This problem, caused by the limited scalability and high costs of the client/server model, leads the internet to be still a medium of mainly texts and images.

The internet has been marked by a vivid evolution of commercialization during the last decades in that it has become a medium for the masses and it involves more than just information. During the nineties, the internet consisted mainly of client/server models which are uncomplicated methods to manage and control the distribution of content. Throughout the past years, several evolutions have emerged that enticed consumers into wanting more than purely text and images. Several aspects – such as the widespread penetration of broadband internet, higher reliability of connections, the evolution of compression technology, more storage capacity, more CPU power and a large amount of content residing on the personal computers of end-users – changed the way users consume the internet. Internet users are spending more time online and exchange more information and files. The combination of these aspects resulted in an increasing demand for multimedia content that contains e.g. audio and video. In other words besides text and images, people nowadays consume larger, more bandwidth consuming content such as audio and video as well. This shift towards larger content makes it difficult for publishers to gain profit via a client/server model. Several measurement studies of peer-to-peer networks provide evidence for the large and increasing amount of large files such as video that is being shared [1, 2]. Furthermore, users have become more active in a sense that they dispose of an increasing amount of digital tools to create and publish content themselves in a relatively easy way. We observe that users have more content stored at their hard disks that is not accessible to other users that might be interested in this content.

Although compression technologies already offer some possibilities, we still observe multimedia content on the internet to be limited and if available it often turns out to be of poor quality. The widespread penetration of the internet causes content providers to explore new distribution platforms that provide solutions for the disadvantages of the current models. Publishers, the media industry and end users are exploring systems and platforms to publish and distribute online services and content. On the one hand, publishers and media companies attach great importance to examining new innovative models to distribute their digital content in a scalable and cost efficient way. They consider this as a necessity because in the current client/server model, every new consumer implies additional costs. On the other hand, users are generating more content themselves and want accessible systems to publish their content. Hosting user generated content at little or no costs for the users often involves high expenses for the hosts. It is therefore necessary to examine new models to distribute user generated content at little costs.

Peer-to-peer is often associated with illegal file sharing because of the popularity of networks such as Napster, Gnutella, KaZaA and BitTorrent. Although these networks contain(ed) a significant amount of illegal activities, they have demonstrated the opportunities of this disruptive technology. Today, many existing file sharing companies are examining new ways for the legal distribution of content. Furthermore, new companies – like Kontiki, Qtrax and RawFlow – were established that exploit peer-to-peer characteristics for secure content distribution. The question remains whether these peer-to-peer systems will be a viable solution for publishers and consumers. This is why it is important that we provide a better understanding of the characteristics, threats and prospects of peer-to-peer. We argue that peer-to-peer systems possess the capability of turning the internet into a valuable multimedia channel that will provide content providers and users with a rich arsenal of content.

In this paper, we will first explore different types of peer-to-peer applications to provide an overview of the capabilities of these models and we will show that different types of applications are merging. Subsequently, different peer-to-peer architectures will be analyzed in order to provide a classification based on the degree of (de)centralization of the topology and the presence or absence of structured resource location. The appropriateness of peer-to-peer as a publishing model will be described in the final section. This includes an exploration of the question what criteria must be met for a peer-to-peer system to be a suitable application.

2 Characteristics and Challenges of Peer-to-Peer

It is important to understand the pros and cons of systems in order to be able to evaluate them. It is therefore necessary to provide an outline of the characteristics and threats of peer-to-peer systems so that one can comprehend the possibilities and advantages peer-to-peer offer in comparison with other models. There is still no generally acknowledged unambiguous definition of the concept peer-to-peer which causes a discussion about what can(not) be accepted as peer-to-peer. Several authors have tried to formulate their own definitions of peer-to-peer [e.g. 3-5]. In spite of these definitions, we still believe that the following definition is the most accurate and comprehensive: “*The term ‘peer-to-peer’ refers to a class of systems and applications that employ distributed resources to perform a function in a decentralized manner. The resources encompass computing power, data (storage and content), network bandwidth, and presence (computers, humans and other resources)*” [6]. The following principles lay the foundation of peer-to-peer: (1) resources are being shared within peer-to-peer systems, (2) the systems are partially or fully decentralized and (3) the systems are self organizing depending on the extent of (de)centralization [7]. Peer-to-peer systems make use of the unutilized resources of the peers for instance on the level of storage capacity, bandwidth, CPU cycles and content.

Peer-to-peer systems have often been described as the counterpart of client/server networks [8, 9]. In client/server systems, centralized servers manage and control the network, provide services and resources whereas the clients consume these resources. Several client/server networks can hardly meet the demand for resources because of an increasing number of users, higher bandwidth traffic and the arrival of a variety of applications. The major drawbacks of client/server systems in comparison with peer-to-peer is that the client/server models suffer from inefficient allocation of resources and limited scalability which can result in bottlenecks and eventually in single points of failure. Furthermore, additional users stand for additional costs as they consume more bandwidth of the system.

Nodes in peer-to-peer networks do not only act as clients, but they exhibit server functions as well. This is why nodes or peers have been described as servents (SERVER + cliENTS). As said, client/server networks are not scalable and are susceptible to bottlenecks and single points of failure whereas peer-to-peer networks are characterized by: scalability, decentralization, transient connectivity, cost efficiency, fault tolerance, self organization, sharing of resources and autonomy [10, 11]. Other components that often prove to be important in peer-to-peer networks are security, anonymity, resilience and efficiency of resource location [9]. In theory, peer-

to-peer systems exhibit positive network externalities in a way that additional users add value to peer-to-peer networks by introducing extra resources in the system. In this way, users preserve the system and influence the functioning, performance and control of the network by making their resources available. Therefore, it is critical for a peer-to-peer system to be able to cope with the transient presence of nodes, network/computer failures and that the network is able to self organize itself in the absence of centralized coordinating components.

An important challenge for peer-to-peer networks is security [10]. Peer-to-peer systems add risks to the network by distributing control to unknown nodes or peers and it therefore requires new security treatments. Further, the unstable and transient connectivity of nodes has consequences for the availability of resources [12, 13]. The resources of nodes are no longer available to the community if they are offline. Moreover, most users are free riders in that they consume resources while not providing anything to the system [14, 15]. The fact that most nodes are mostly free riders and are only online for a limited period of time makes resource availability a critical factor in the viability of peer-to-peer systems. A reliable peer-to-peer system therefore must be able to detect and recover from failures, guarantee content availability and avoid single points of failure. Subsequently, scalability stands for both opportunities and threats of peer-to-peer [16, 17]. Scalability reveals itself in the load in terms of bandwidth and storage capacity, the number of nodes that can be reached, the number of hops to reach nodes, the amount of resources that can be consumed, etc. without interfering the system's performance. Several file sharing systems contain millions of nodes that send terabytes of data across the network. Although peer-to-peer systems are theoretically inherently scalable, it often turns out to be a major challenge in real terms as we will further explore in the section on architectures.

3 Applications

In this section, we will provide an outline of different existing and new peer-to-peer applications. In this way, we gain insight in the fact that peer-to-peer is more than just file sharing. During the last few years, applications are becoming more integrated so that it becomes harder to draw a line between different types of applications [18].

3.1 Communication: Instant Messaging and Telephony

The first category of applications encloses communication systems such as Instant Messaging (IM) and telephony. These applications furnish the infrastructure mainly for real time or synchronous communication among users [5, 17]. Communication systems try to avoid as much central control as possible in order to reduce costs and to improve fault tolerance. These systems often merge into integrated applications that provide collaborative tools and file sharing on top of communication.

IM is a type of application that can utilize peer-to-peer aspects for their services, which of course does not mean that all IM systems exploit peer-to-peer characteristics. Some IM applications function within a client/server model, whereas other systems – e.g. ICQ, AOL instant messaging and Yahoo Messenger – make use of centralized peer-to-peer systems. The topology of these systems consists usually of a centralized directory server that contains information such as which nodes are online and who might communicate with whom. The communication then takes place directly between peers without intervention of the server.

Voice over IP (VoIP) is another application domain of peer-to-peer systems and has become more widely known particularly because of Skype [19, 20]. Skype is a peer-to-peer VoIP application that provides telephony, IM and audio conferencing via a peer-to-peer system.

3.2 Grid Computing

It has been widely debated whether grid computing can be accepted as peer-to-peer. In either way, grid computing and peer-to-peer networks are both distributed systems that are build to share resources [5, 21]. Grid computing is the coordinated use of resources – computers, processor capacity, sensors, software, storage capacity and data – which is being shared within a dynamic and continuously changing group of individuals, institutions and resources [17, 22]. In contrast to peer-to-peer systems, grids stress the standardized, secure and coordinated sharing of resources with a better guarantee of Quality of Service (QoS). The philosophy behind grids, which is largely the same for peer-to-peer systems, is that we can generate an enormous capacity by coupling several computers and their peripheral equipment in a network. The comparison has often been made with electric power: it should be as easy to get resources from the internet as it is simple to draw electricity from a wall socket [17]. Peer-to-peer and grids might evolve into a convergence in which the benefits of grid computing (interoperability, security, QoS, and standardized infrastructures) and of peer-to-peer (fault tolerance,

scalability and self organization) will be combined. SETI@home (Search for ExtraTerrestrial Intelligence) is probably the most referred project in this area [23]. The processing of the radio signals has been distributed to the personal computers of users to save costs. In SETI@home, a central server receives the data, split it into small units, distributes these units to the clients and coordinates further transactions. The clients process the data using their unused capacity and send the results back to the server. Other similar projects are e.g. FightAIDS@home, Distributed.net, Entropia, Genome@home and Folding@home.

3.3 Collaborative Tools

A third application domain, in which peer-to-peer has already proven its usefulness and value, consists of tools for users to collaborate on certain tasks within groups [18, 24]. This type of software pursues the collaboration of users, even if some of these users find themselves to be outside the corporate LAN. An example of a peer-to-peer groupware tool is Groove Virtual Office, which has attributes such as file sharing, document management, chat, agenda and discussion groups. Security issues such as integrity, authentication and authorization are more critical in a confidential business environment so as to malicious users don't have the possibility to access, read or change the information [25]. It is obvious as well for peer-to-peer groupware applications to have opportunities for e-learning purposes. Peer-to-peer groupware integrates several elements like IM and file sharing.

3.4 File Sharing and Content Distribution

Peer-to-peer content distribution is the most well-known application area of peer-to-peer systems and it contains file sharing systems (e.g. Napster, Gnutella, eDonkey), distributed storage applications (e.g. Freenet) and content delivery networks (e.g. Kontiki). These applications offer companies and users the possibility to publish, store and exchange files and other content [5, 6]. The hype of peer-to-peer file sharing started in 1999 with the arrival of Napster [3]. Napster demonstrated the opportunities of peer-to-peer file sharing which resulted in the development of new systems such as Gnutella, Freenet, KaZaA and BitTorrent. Androutsellis-Theotokis and Spinellis [5] make a distinction between file exchange systems and content publishing/storage systems. On the one hand, file exchange systems are little sophisticated file sharing applications such as the former Napster that only contains some basic functionality and mostly doesn't address issues such as resource availability and security. It is mostly this type of applications that appears in the news because of copyright infringements. On the other hand content publishing and storage applications are more elaborated systems to publish, distribute and store content. These applications focus more on aspects such as security, availability and authorization.

Peer-to-peer streaming is a specific type of content distribution and it probably represents the most successful 'legal' peer-to-peer application. The traditional streaming technologies, such as unicasting and multicasting, are characterized by the fact that additional consumers of the streaming imply more costs. High costs, bottlenecks, single points of failure, lack of scalability and poor quality of most streaming technologies causes e.g. internet television to be still in its infancy. Peer-to-peer streaming, that has some similarities with multicasting, might offer some solutions for these problems by providing cost efficiency, scalability and quality of content [26, 27]. In peer-to-peer streaming applications, clients act as servers as they send units of the stream to other clients in the network. Most commercial peer-to-peer streaming applications integrate centralized components in their architecture to control and secure the content distribution. Examples of peer-to-peer video and or audio streaming are: Rawflow, Octoshape, Coopnet, Splitstream, Peerstreaming and Abacast.

3.5 Wireless and Ubiquitous Peer-to-Peer

Peer-to-peer systems offer some opportunities for wireless systems, such as MANETs (Mobile Ad hoc NETWORKS), and varies from cellular networks to wireless LANs [28]. Wireless communication networks can be considered to be peer-to-peer if the signals are being transferred directly between the appliances. In comparison with personal computers, the capacity of wireless equipment – for instance storage capacity, content and bandwidth – are increasing as well which offers new opportunities for peer-to-peer systems to be applied on mobile phones, Smartphones, PDA and laptops. Wireless peer-to-peer applications evoke other issues than 'traditional' peer-to-peer systems such as the power of batteries and the location of apparatuses when users are moving. The mobility of users combined with a transient connectivity of nodes make that self organization is an even bigger challenge for wireless peer-to-peer systems. Other challenges are emanating from the following characteristics: (1) wireless resources such as storage capacity, bandwidth and processing power are still limited, (2) the performance and capacities are fluctuating and (3) the availability of resources is barely guaranteed without a centralized component. It is therefore necessary for mobile peer-to-peer systems to develop

applications with an efficient search and location infrastructure and routing model as to avoid zigzag movements.

Furthermore peer-to-peer systems exhibit characteristics – e.g. self organization, sharing of resources, collaborating apparatuses – that seem to be similar to some aspects of ubiquitous computing [29]. Similar to peer-to-peer systems, ubiquitous computing architectures must cope with autonomous communicating systems that are marked by transient connectivity. The parallel features of peer-to-peer and ubiquitous computing make that it doesn't seem illogical to integrate these systems.

4 Classifying Peer-to-Peer Architectures

Given that peer-to-peer systems have several different features, we endorse the fact that there might be different ways to classify peer-to-peer architectures. We argue that most peer-to-peer architectures distinguish themselves from each other based on the extent of (de)centralization and on the presence of structure in object location and routing. Based on this we distinguish the following combinations: centralized unstructured, pure unstructured, hybrid unstructured and pure structured systems.

4.1 Degree of Decentralization

Systems might be considered as peer-to-peer when at least some elements are decentralized. The degree to which centralized and decentralized components are applied in the network can vary between systems. In other words, in contrast with what has been suggested in some definitions of peer-to-peer, not all peer-to-peer networks are completely decentralized. We make a distinction between centralized, pure decentralized and hybrid peer-to-peer topologies (Figure 1) [5, 30, 31].

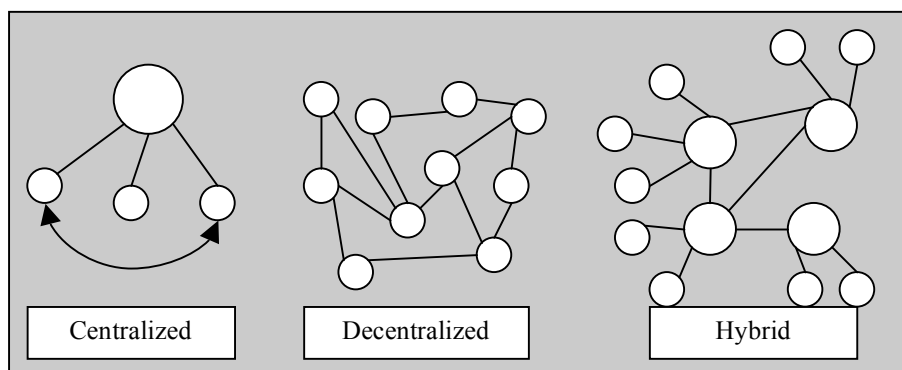


Figure 1: Degree of Decentralization

Centralized peer-to-peer architectures, such as the former Napster, contain a central server that executes vital functions for the system. This central server is mostly used as a directory server that stores an overview of the available nodes and resources in the network. In this way, the central directory server makes it possible for peers or nodes to find, locate and share resources with other peers. Peers eventually exchange data directly between each other without the intermediation of the server which makes it a simple but quite efficient architecture. This type of architecture exhibits the following drawbacks. The whole system stops functioning if the central servers cannot be reached for whatever reason. In other words, the major disadvantages of these systems are the risks of bottlenecks and single points of failure which imposes limited scalability. The advantage of using central directory servers is that if the sought data is available, the search algorithm can mostly guarantee the retrieval of the data.

Pure decentralized architectures consist of nodes that perform functions without the intervention of centralized components. These types of architectures have theoretically unbounded scalability and a high level of fault tolerance. In addition, these systems are autonomous and self organizing in a sense that the peers are responsible for the functioning and viability of the network. In practice, a great deal of these systems has limited scalability because self organization causes a lot of traffic to keep the network running. Another characteristic is that several of these systems have low levels of QoS in the domain of resource location because sometimes only a

limited proportion of the network can be reached. Examples of pure decentralized peer-to-peer networks are Gnutella 0.4, Freenet and Chord.

Hybrid systems are often hierarchical networks that adopt elements of both centralized and pure decentralized architectures in which they combine the advantages (e.g. efficient resource location, scalability) while avoiding the drawbacks (e.g. bottlenecks, limited QoS) of these systems. In hybrid peer-to-peer systems, some peers have more capacities than others and therefore these peers are granted with more responsibilities. These nodes, that perform more functions in the network, are named super nodes or ultranodes. These super nodes are in fact regular nodes that also serve as a kind of directory server, with the difference that these super nodes are dynamic and can suddenly disappear from the network. In this way, nodes with better capacities have more responsibilities in the organization and functioning of the system whereas nodes with fewer capacities are less loaded. This reduces e.g. the possibility of delaying resource location because of links with dial-up connections. This leads to a better performance of the system because of less traffic and better search functionalities. The risks of bottlenecks and single points of failure are limited because of the use of several super nodes in stead of one central directory server. KaZaA and Gnutella 0.6 are some examples of suchlike architectures. Further, we argue that BitTorrent can also be regarded as a hybrid system.

4.2 Degree of Structure

Whether a system is structured depends on how nodes and data are positioned in the network [5, 8, 31].

Unstructured. A system is unstructured when nodes and data are positioned without certain rules and in an ad hoc manner in the network. The location of data is not connected with the topology of the network which results in cumbersome and little efficient search methodologies – such as the ‘query flooding model’ (cf. Gnutella) – that hamper scalability. On the one hand, most unstructured systems are characterized by much consumption of bandwidth in the matter of traffic of messages. Unstructured networks cannot guarantee that data, if available, can be found because the system is often not capable of reaching the whole network. On the other hand, these systems are mostly quite resilient. Another advantage is that these systems – e.g. Napster, Gnutella, KaZaA – mostly support keyword-based search.

Structured. In this type of networks, nodes and data are being placed in a structured way in the network as to be able to efficiently locate data which increases the possible scalability. The nodes, data or other resources are connected to specific locations. Distributed routing tables make it possible to efficiently, i.e. in a smaller number of hops, acquire search results. Structured systems are, in comparison with unstructured systems, more scalable, more reliable and fault tolerant. On the other hand, these systems have the disadvantage that it is more difficult to support keyword-based search because one needs to know the key to be able to locate the associated data. Another shortcoming is that these systems laboriously handle the transient connectivity of nodes whereby the system needs to reconfigure the structure constantly. Examples of structured systems are Chord, CAN, and Tapestry. Freenet is often called a ‘loosely structured’ network because it is not rigidly structured in that the location of the data is not totally specified.

4.3 Centralized Unstructured Systems

These peer-to-peer networks (e.g. Napster and Publius) have a centralized topology and display several client/server characteristics [16]. This type of peer-to-peer networks contains a central server that functions as a directory server [5, 32]. But, this directory server has fewer tasks than servers in client/server networks. In this way, a server in peer-to-peer networks is less loaded than servers in client/server networks. How does this system function? When peers log in to the system, they announce their presence and give some information (e.g. IP address, bandwidth of the connection, number and metadata of files that are being shared, etc.) to the directory server. In this way, there is one server that keeps an index of all available resources in the network. If a peer is searching for information, it sends a query to the server asking for available peers who share the requested information. The server subsequently searches his database and returns the result to the peer who initiated the query. Based on these results, the peer can decide to make a direct connection with a peer from the search results to download the requested data. In a nutshell, the search process is centralized, via the directory server, and the eventual exchange of data or other resources takes place in a peer-to-peer manner. The data is not stored on the server but on the hard disks of the peers.

The most important advantages of this type of architectures are that it is easy to implement and that the server is less loaded in terms of bandwidth and storage capacity. This is because the directory server doesn’t vouch for the sending, distributing or storage of the data. Although this type of architectures use an unstructured search

infrastructure, they do provide good performance to be able to find the data if it is available [9, 11]. Another advantage of a centralized system is that it gives more opportunities to manage and control the network for security. Finally, these systems support keyword-based search which is important to users.

The disadvantages of the system mainly stem from the possible bottleneck at the server which is also hazardous for a single point of failure [11]. This has implications for the scalability of such systems because each additional user induces extra load in terms of traffic and storage capacity.

4.4 Pure Decentralized Unstructured Systems

The most striking feature of pure decentralized unstructured systems – e.g. Gnutella 0.4 – is that there is no centralized component which means that all nodes are directly connected to each other. Nodes function as clients, servers, routers and cache [5, 11]. They act as servers, not only for the storage and transfer of data, but also to search for data. Nodes can be involved as routers to help send messages through the network. Peers have an index of their own data and not of others' data. Therefore, to find the demanded data, it is important to be able to reach as many peers in the network as possible.

The advantage of this category of peer-to-peer architectures is that there is no single point of failure and that it is fault tolerant. The failure of one or even several of the nodes has little impact on the performance of the network. It is essential for these systems to be autonomous and self organizing in order to be able to cope with the transient connectivity of nodes. In the absence of a central infrastructure, the major challenge is to elaborate an efficient search method that is capable of achieving satisfying search results in the presence of transient nodes [9, 11]. Even if sought data is available in the network, unstructured peer-to-peer systems cannot offer guarantees that it would be able to find it. We will explain this with the example of the query flooding model. In the query flooding model, a node broadcasts a query to all his neighbors, his neighbors in their turn broadcast the query further to their own neighbors and so on. This process runs until a limited number of hops is reached according to the TTL (Time-To-Live). This TTL is essential to prevent messages from saturating the network by endlessly flooding it. But, this causes as well that the whole network is often impossible to reach which means that scalability is limited. Scarce content in a large file sharing network for example might be difficult to find because it is too many hops away. Of course several researchers have developed or adapted search methods to address these flaws such as: Random Walkers, Adaptive Probabilistic Search, Breadth First Search, etc. [for a more profound overview see: 33].

4.5 Hybrid Unstructured Systems

Hybrid unstructured peer-to-peer networks – such as KaZaA [34], eDonkey and Gnutella 0.6 [32] – have been developed with the objective of combining the advantages (e.g. better search results and fault tolerance) and circumventing the drawbacks (e.g. scalability and bottlenecks) of centralized and pure decentralized peer-to-peer systems [5, 16, 17, 35]. On the one hand, Napster had a centralized topology and they had to pull the plug on the server which ended the functioning of the whole network. This demonstrated the danger for bottlenecks and single points of failure in centralized topologies. On the other hand, Gnutella 0.4 as a pure decentralized system had to contend with an overload of messages because of the query flooding model. Hybrid peer-to-peer systems try to cope with these problems by introducing hierarchy in the system via the use of super nodes. Super nodes are peers with more capacity, such as bandwidth or storage capacity, than the average peer and therefore they are chosen to perform more functionality in the system. Super nodes mostly have the following tasks:

- Keep record of a directory list with information of a part of the peers and their data;
- Keep record of a directory list with information of some other super nodes;
- Search through the directory list in case a peer sends him a query;
- Redirect queries to other super nodes to be able to have better search results.

In other words, the hybrid architecture includes a combination of a centralized and a decentralized topology and therefore can be regarded as the convergence of these systems. There is not one central server, but there are different servers (super nodes) that all have responsibilities for a part of the node population. The super nodes are interrelated in a decentralized manner, whereas the normal nodes are related to their super node in a centralized way. The super nodes function as directory servers for a part of the peer-to-peer population.

The use of hierarchy in the system increases the chances for better, more efficient and faster search results because these systems utilize the available resources more intelligently [11]. The division of labor is more balanced in hybrid systems because nodes with more capacities get more responsibilities so as to nodes with less capacity don't get overloaded. Slow connections are avoided in this way, which results in an overall better

performance of the system. Data and nodes are inserted in the network in an unstructured way, so that resource location also occurs in an unstructured way.

4.6 Pure Decentralized Structured Systems

Pure structured peer-to-peer systems – e.g. Chord, CAN, Freenet (loosely structured), Kademlia, Pastry – are self organizing networks without centralized components to store and retrieve data. If the content or other resources are available, unstructured networks offer guarantees that it will be able to find it within a limited number of hops [9, 11, 36]. These systems are structured because the resources and nodes are mapped into an address space in the network so as to be able to efficiently retrieve them. The indexing of this address space is distributed among the nodes in the system which makes every node responsible for a part of the indexing. These systems utilize Distributed Hash Tables (DHT) to structure the network: “*Distributed Hash Tables provide a global view of data distributed among many nodes, independent of the actual location*” [36]. DHTs manage the data in the system and it contains a routing system. Data can be efficiently retrieved because the DHT provides the system with a routing scheme to easily find the node that hosts the sought data. A unique key is created and assigned to every data to serve as an ID. The keys are mostly generated using hash functions such as SHA-1 and MD5. Hash functions are operations executed on data with a unique key as a result that has a fixed length regardless of the size of the data. The peers or nodes are responsible for a part of these keys in the address space. Peers are assigned with keys that most closely approach their own ID. It is important to know the unique key of the data to be able to retrieve it. During a search, the query is continuously redirected from node to node and it is getting closer to its destination or key in every hop. Every node in the system has a routing table of several other nodes in the network. A node that receives a query but doesn’t have the sought information locally, routes the query to a node that, according to his routing table, is numerically closer to the destination.

There are two options to store and find data in this type of structured systems [36, 37]. In the first option, the nodes store only the unique keys of the data which serve as pointers to the actual data that is being stored somewhere else. The node that inserted the data is responsible for making the data available. In the second option, the nodes do not only store the keys, but the data as well. The second option implies that nodes store content that other nodes initiated. In this way, even if the node that initiated the data goes offline, the data remains available to other users because another node stores it.

It is a challenge for pure structured systems to maintain and update the routing tables in the transient presence of nodes. The updating process of the nodes’ routing tables causes a load on the network. The resilience and structure of the network might be harmed if a large amount of the population would suddenly (dis)appear with a decreasing performance as a result. Another disadvantage is that these systems don’t support keyword-based search because the search method requires the exact key [9, 11]. Further, a problem of load balancing might occur because nodes might be responsible for: (1) a big part of the address space, (2) a data rich address space and (3) very popular content [36]. Structured peer-to-peer systems exhibit the advantage that they have high levels of scalability and that they have an efficient search method that offers high guarantees for search results.

5 Appropriateness of Peer-to-Peer for e-Publishing and Openness

5.1 Peer-to-Peer Criteria

In this section, we try to elaborate some criteria to decide whether a peer-to-peer solution might be appropriate. We would first like to remark that these criteria are not meant to be formulas for success. One of the most important questions that has been posed is: when and for what types of content are peer-to-peer applications appropriate? It is difficult to formulate an unambiguous answer to the first part of the question. The second part of the question, which is for what types of content peer-to-peer is appropriate, is simpler to answer. In my opinion, peer-to-peer is content independent as it is distribution model and not a content model. Roussopoulos, Baker and Rosenthal [38] tried to answer the question when peer-to-peer might be appropriate. They formulated several criteria to determine whether the implementation of peer-to-peer aspects is the appropriate method for the distribution of a certain kind of content:

- Cost savings: peer-to-peer solutions make it possible for companies and other organizations with limited budget to distribute their content to the masses;
- Relevance of resources: the content must be important to the consumers so that they are more willing to participate in helping to distribute the content;

- Trust: the indispensable need to have peers cooperating is hard to achieve when peers distrust each other;
- Rate of change and criticality: A peer-to-peer application probably will not succeed if there is a high rate of change (peers entering and leaving the system) in an untrustworthy environment. Peer-to-peer has more chances for success when the rate of change is low and the criticality of the information is low.

These researchers formulated the following conclusion: “(...) *the characteristics that motivate a P2P solution are limited budget, high relevance of the resource, high trust between nodes, a low rate of system change, and a low criticality of the solution*” [38]. Although we do agree with the conclusions of these authors, we argue that the practice is more complex and that the appropriateness of peer-to-peer depends on technical (e.g. architecture) and social aspects. The applicability of peer-to-peer systems is independent of content type. We will first discuss the technical aspects to subsequently expand the social aspects.

From a technical perspective, peer-to-peer systems provide solutions for mass content distribution in that they are characterized by cost reduction, scalability and performance. Publishers and content providers might implement a peer-to-peer solution for cost saving objectives. Every additional consumer in a client/server model produces extra costs, whereas this effect is more limited in peer-to-peer networks because additional users mean extra resources for the system. Peer-to-peer systems are especially important for scalability solutions. The connection of the server in a client/server model becomes silted up little by little in an environment of an increasing user population which causes a bottleneck. Peer-to-peer systems can prevent the occurrence of bottlenecks by utilizing the available resources in the peer community. Excellent developed peer-to-peer applications give evidence of great performance in the presence of mass populations. From this, we can conclude that the three main reasons for choosing a peer-to-peer solution are cost of ownership, scalability and performance. Furthermore the architecture of peer-to-peer systems has certain aspects that make some topologies more suitable for specific applications. We explain this with some examples without having the aim to be exhaustive. The larger the consumer mass, the more a decentralized architecture is appropriate in order to avoid bottlenecks. Commercial content providers mostly want to preserve control over the distribution of content so that they are ensured of payment and they are capable of avoiding copyright infringement. Therefore, most commercial and legal peer-to-peer systems have a centralized topology. A centralized topology exposes itself to limited risks of bottlenecks and it is therefore suitable for a limited user population.

From a social perspective, a characteristic feature of peer-to-peer networks is that the performance of the system is not only dependent on technical functionality, but also on user behavior. A critical mass of active participating online peers and content availability are critical to the viability and success of peer-to-peer networks. The number of online peers and the number and quality of content these peers share, determine the value other peers can derive from the network. These last two sentences contain several features that determine whether a peer-to-peer solution might be appropriate:

- Critical mass: it is not necessary for all peers to provide the system with their resources, but it is necessary to have a sufficient amount of peers that contribute to the system so that the content remains available;
- Online: peers have to stay online after they have downloaded content so that others can download this and other content from them;
- Quantitative availability: it is important to have a sufficient amount of content available;
- Qualitative availability: it is not sufficient to have large amounts of content available, but it is critical as well to have resources available other peers are interested in.

From this analysis, we argue that content is suitable to be distributed via peer-to-peer networks if: (1) the distribution of the content is very resource consuming, (2) it is being consumed by a mass, (3) the consuming mass is online enough, and (4) there are enough users willing to cooperate in distributing the content. Publishers, consumers and media companies might consider using peer-to-peer networks if their content meets these criteria.

Critical readers may now wonder whether peer-to-peer is only appropriate for mass content distribution. The answer to that question is no, but there are some important conditions for peer-to-peer to be successful in the presence of small user communities. Peer-to-peer solutions might succeed if the population of peers is often online simultaneously. It is favorable to have a community of peers with strong ties and with similar interests in the sense that they consume the same kind of content. The importance to stimulate users to cooperate via incentives increases in small populations of peers in order to ensure content availability. A centralized peer-to-peer architecture might be an appropriate system for small user communities because the risks for bottlenecks at

the server are more limited. A few examples of content that is mostly consumed by small communities are user generated content and content that is consumed within organizations (e.g. corporate communication).

5.2 Openness

Openness is a new buzz word that supports philosophies of open access [39, 40], open content and open source in which information, knowledge and content is universally available as a public good and is often for free. Whereas this content is mostly free to users, it is often expensive to the organizations hosting this content; e.g. the Budapest Open Access Initiative recognizes this assumption: “*While the peer-reviewed journal literature should be accessible online without cost to readers, it is not costless to produce*” [39]. Comtella (<http://bistrica.usask.ca/madmuc/comtella.htm>), LionShare (<http://lionshare.its.psu.edu/>) and Edutella (<http://edutella.jxta.org/>) are three peer-to-peer systems that support open sharing of educational content such as papers, articles and other educational and research tools (videos, images, presentations, demos, etc.). All these projects have been initiated with the objective of making free and open educational and research material more accessible to all interested persons. Why might peer-to-peer be appropriate for open initiatives? In our opinion, peer-to-peer might be applicable to ‘open’ environments if it meets several of the above mentioned criteria. The storage and distribution of content in open initiatives are often resource and consequently money consuming. There is a large simultaneous online population that possesses many unused resources. These two fundamental aspects, that are at least necessary for the implementation of a peer-to-peer system, are characteristic for several open initiatives. Whether there will be a critical mass of cooperating peers is hard to predict because it depends on the complex interaction of several factors. The main reason for using a peer-to-peer system is cost reduction on the level of storage capacity and bandwidth. If peer-to-peer systems turn out to be successful applications in open environments, it will result in decreasing storage and bandwidth expenses. Decreasing resource expenses reduce the barriers for open initiatives which in his turn can lead to more accessible content.

The internet originally displayed peer-to-peer characteristics in that every computer was mutually connected to other computers in the network and most computers acted as clients as well as servers [17, 41]. In those days, the internet was mainly used for research and military purposes. If the actual open movements will succeed in using peer-to-peer systems for research and educational purposes, then it might be regarded as the *renaissance* of the internet.

6 Discussion and Conclusions

Our analysis suggests that peer-to-peer systems in some cases might provide solutions for the flaws of client/server systems. Client/server models suffer from limited scalability, bottlenecks, cost inefficiency and single points of failure. These characteristics of client/server models set limits for the amount and largeness of available content on the internet. In this paper, we have demonstrated that in some cases, peer-to-peer systems might provide solutions for the drawbacks of client/server systems in that they have already proven their abilities in terms of scalability and cost efficiency. Further, we have shown that peer-to-peer comprises more than file sharing, such as communication, collaboration, and grid computing. The importance of characteristics and (dis)advantages of peer-to-peer systems varies from architecture to architecture depending on the degree of (de)centralization and whether it is structured or not. This can be represented as a pendulum between: (1) risks of bottlenecks, possible single point of failure, more control (centralized) and (2) scalability, fault tolerance, self organization (decentralized). Whether a peer-to-peer system is structured or not determines the efficiency of node and resource location, at which structured systems are more efficient.

One of the major questions of several content providers is when a peer-to-peer solution might be appropriate. Therefore, we tried to elaborate criteria to decide whether a peer-to-peer solution might be suitable. We have to remark that these criteria are not meant to be rules for success as it does not imply that users will adopt and use the system. These criteria imply that peer-to-peer is not always a good solution and that client/server systems will sustain. Besides more technical criteria such as scalability, we paid attention to some social criteria as well. Peer-to-peer systems are not only dependent on technical criteria, but also on social aspects for it are the users that make their resources available and cooperate or free ride in distributing content. More research on social aspects is needed because there is little information on user behavior in peer-to-peer systems. Social research is necessary because users have never had such a powerful impact on a system as the end users influence the performance of peer-to-peer networks by (not) providing their resources to the community. Whether peer-to-peer solutions might be appropriate for open initiatives depends on whether the system meets the aforementioned criteria. It seems likely that peer-to-peer is suitable in some open access and open content systems because scalability, cost efficiency and a large simultaneously online user population are all criteria that are often met in these applications.

The results of this broad analysis provide a better understanding of the capabilities and application domains of peer-to-peer systems. The internet today is being marked by an increasing amount of content and an increasing size of this content which causes more loads on the distribution, storage and consequently costs. That is why publishers and other media companies are trying to find solutions for scalability and cost problems and therefore need to explore innovative platforms such as peer-to-peer systems.

What will the future bring for peer-to-peer? There are several essential issues that need to be remedied in order for peer-to-peer to be able to succeed. There is still a lot of work to be done to address problems of standardization, security, Digital Rights Management and asymmetric connections with unbalanced upload/download ratios. The years 2006 and 2007 might become the turning point for peer-to-peer networks because this is the period that new 'legal' peer-to-peer services have entered the market [42]. Currently, mainly the opportunities of peer-to-peer for video, film and television are being explored by different companies (e.g. Joost, RawFlow, Kontiki, BitTorrent, In2Movies, etc.) and in different workshops [e.g. 43, 44]. Peer-to-peer television is one of the examples that meet the formulated criteria. But it is not all about video. Peer-to-peer is a distribution system in that it is content independent. It is remarkable that almost every 'legal' commercial peer-to-peer system implements centralized components in their architecture. This is probably to ensure control, security and QoS. In this way, these commercial peer-to-peer platforms combine the strengths of peer-to-peer systems with those of client/server models. A lot of non-technical questions remain unanswered, e.g. are users willing to cooperate in a network when they have to pay the consumed content and for what kind of content are they willing to pay. To learn more about the possibilities of peer-to-peer networks, it is essential that the research community explores new applications, environments, content to experiment with peer-to-peer. Peer-to-peer networks have the capacities for a scalable, accessible and cost efficient model for the distribution of content. If peer-to-peer systems turn out to be a success for content publishing, it may lead to new business models that change the way content is distributed. It is hard to predict whether peer-to-peer will become a success for legal purposes. Only the future will tell how peer-to-peer will evolve.

Acknowledgements

FLEET (an interdisciplinary research project on FLEmish E-publishing Trends) is an IWT SBO project, with research partners IBBT SMIT, Cemeso, LSTS, MOFI, ICRI, CUO, MICT, ECDC and TNO.

References

- [1] OECD. Peer to Peer Networks in OECD Countries. 2004.
<http://www.oecd.org/dataoecd/55/57/32927686.pdf>.
- [2] KOLWEY, M.; LECHNER, U. Towards P2P Information Systems. *The Fifth International Workshop on Innovative Internet Community Systems: IICS 2005, Paris, France, 2005*.
- [3] SHIRKY, C. Listening to Napster. In A. ORAM (Eds.), *Peer-to-Peer: Harnessing the Power of Disruptive Technologies*, Sebastopol, CA: O'Reilly & Associates, Inc, 2001, pp. 21-37.
- [4] SCHOLLMIEIER, R. A Definition of Peer-to-Peer Networking for the Classification of Peer-to-Peer Architectures and Applications [poster]. *The First International Conference on Peer-to-Peer Computing, Linköping, Sweden, 2001*.
- [5] ANDROUTSELLIS-THEOTOKIS, S.; SPINELLIS, D. A Survey of Peer-to-Peer Content Distribution Technologies. *ACM Computing Surveys*, 2004, vol. 36, no. 4, pp. 335-371.
- [6] MILOJICIC, D. S.; KALOGERAKI, V.; LUKOSE, R.; NAGARAJA, K.; PRUYNE, J.; RICHARD, B.; ROLLINS, S.; XU, Z. Peer-to-Peer Computing [Technical Report]. 2002.
<http://www.hpl.hp.com/techreports/2002/HPL-2002-57R1.pdf>.
- [7] ABERER, K.; HAUSWIRTH, M. An Overview on Peer-to-Peer Information Systems. *Workshop on Distributed Data and Structures, Paris, France. 2002*.
- [8] STEINMETZ, R.; WEHRLE, K. What Is This "Peer-to-Peer" About? In R. STEINMETZ & K. WEHRLE (Eds.). *Peer-to-Peer Systems and Applications*, Berlin Heidelberg: Springer-Verlag, 2005, pp. 9-16.
- [9] WANG, C.; LI, B. Peer-to-Peer Overlay Networks: A Survey. 2003.
<http://comp.uark.edu/~cgwang/Papers/TR-P2P.pdf>.

- [10] SCHODER, D.; FISCHBACH, K.; SCHMITT, C. Core Concepts in Peer-to-Peer Networking. In R. SUBRAMANIAN; GOODMAN, B.D. (Eds.), *Peer-to-Peer Computing: the Evolution of a Disruptive Technology*, London: Idea Group Publishing, 2005, pp. 1-27.
- [11] SCHMIDT, C.; PARASHAR, M. Peer-to-Peer Information Storage and Discovery Systems. In R. SUBRAMANIAN; GOODMAN, B.D. (Eds.), *Peer-to-Peer Computing: the Evolution of a Disruptive Technology*, London: Idea Group Publishing, 2005, pp. 79-112.
- [12] BHAGWAN, R.; SAVAGE, S.; VOELKER, G.M. Understanding Availability. *The Second International Workshop on Peer-to-Peer Systems, Berkeley, CA, USA*, 2003.
- [13] CHU, J.; LABONTE, K.; LEVINE, B.N. Availability and Locality Measurements of Peer-to-Peer File Systems. *Conference on Scalability and Traffic Control in IP Networks, Boston, USA*, 2002.
- [14] ADAR, E.; HUBERMAN, B.A. Free Riding on Gnutella. *First Monday*, 2000, vol. 5, no. 10, http://www.firstmonday.dk/issues/issue5_10/adar/.
- [15] HANDURUKANDE, S.B.; KERMARREC, A.-M.; LE FESSANT, F.; MASSOULIÉ, L.; PATARIN, S. Peer Sharing Behaviour in the eDonkey Network, and Implications for the Design of Server-less File Sharing Systems. *EuroSys 2006, Leuven, Belgium*, 2006.
- [16] DING, C.H.; NUTANONG, S.; BUYYA, R. Peer-to-Peer Networks for Content Sharing. In R. SUBRAMANIAN & B.D. GOODMAN (Eds.), *Peer-to-Peer Computing: the Evolution of a Disruptive Technology*, London: Idea Group Publishing, 2005, pp. 28-65.
- [17] TAYLOR, I. J. *From P2P to Web Services and Grids: Peers in a Client/Server World*. London: Springer, 2004.
- [18] SCHODER, D., FISCHBACH, K.; SCHMITT, C. Application Areas. In R. STEINMETZ & K. WEHRLE (Eds.). *Peer-to-Peer Systems and Applications*, Berlin Heidelberg: Springer-Verlag, 2005, pp. 25-32.
- [19] BASET, S.A.; SCHULZRINNE, H. An Analysis of the Skype Peer-to-Peer Internet Telephony Protocol – Technical Report. 2004. http://www.rootsecure.net/content/downloads/pdf/skype_protocol.pdf.
- [20] JENNINGS, C.; BRYAN, D.A. P2P For Communications: Beyond File Sharing. *Business Communications Review*, 2006, vol. 36, no. 2, pp. 36-40.
- [21] TALIA, D.; TRUNFIO, P. Toward a Synergy between P2P and Grids. *IEEE Internet Computing*, 2003, vol. 7, no. 4, pp. 96, 94-95.
- [22] FOSTER, I.; IAMNITCHI, A. On Death, Taxes, and the Convergence of Peer-to-Peer and Grid Computing. *The Second International Workshop on Peer-to-Peer Systems, Berkeley, CA, USA*, 2003.
- [23] ANDERSON, D. SETI@home. In A. ORAM (Eds.), *Peer-to-Peer: Harnessing the Power of Disruptive Technologies*, Sebastopol, CA: O'Reilly & Associates, Inc, 2001, pp. 67-76.
- [24] MAUTHE, A.; HUTCHISON, D. Peer-to-Peer Computing: Systems, Concepts and Characteristics. *Praxis in der Informationsverarbeitung und Kommunikation*, 2003, vol. 26, no. 2.
- [25] ASVATHANARAYANAN, S. Potential Security Issues in a Peer-to-Peer Network from a Database Perspective. In R. SUBRAMANIAN & B.D. GOODMAN (Eds.), *Peer-to-Peer Computing: the Evolution of a Disruptive Technology*, London: Idea Group Publishing, 2005, pp. 131-144.
- [26] LIU, Z.; YU, H.; KUNDUR, D.; MERABTI, M. On Peer-to-Peer Multimedia Content Access and Distribution. *The International Conference on Multimedia and Expo, Toronto, Canada*, 2006.
- [27] STOLARZ, D. Peer-to-Peer Streaming Media Delivery. *The First International Conference on Peer-to-Peer Computing, Linköping, Sweden*, 2001.
- [28] KELLERER, W.; SCHOLLMEIER, R.; WEHRLE, K. Peer-to-Peer in Mobile Environments. In R. STEINMETZ & K. WEHRLE (Eds.). *Peer-to-Peer Systems and Applications*, Berlin Heidelberg: Springer-Verlag, 2005, pp. 401-417.
- [29] KANGASHARJU, J. Peer-to-Peer and Ubiquitous Computing. In R. STEINMETZ & K. WEHRLE (Eds.). *Peer-to-Peer Systems and Applications*, Berlin Heidelberg: Springer-Verlag, 2005, pp. 457-469.
- [30] BACKX, P.; WAUTERS, T.; DHOEDT, B.; DEMEESTER, P. A comparison of peer-to-peer architectures. *Eurescom Summit 2002, Heidelberg, Germany*, 2002.
- [31] POUREBRAHIMI, B.; BERTELS, K.; VASSILIADIS, S. A Survey of Peer-to-Peer Networks. *The 16th Annual Workshop on Circuits, Systems and Signal Processing, Veldhoven, the Netherlands*, 2005.

- [32] EBERSPÄCHER, J.; SCHOLLMEIER, R. First and Second Generation of Peer-to-Peer Systems. In R. STEINMETZ; WEHRLE, K. (Eds.). *Peer-to-Peer Systems and Applications*, Berlin Heidelberg: Springer-Verlag, 2005, pp. 35-56.
- [33] TSOUMAKOS, D.; ROUSSOPOULOS, N. A Comparison of Peer-to-Peer Search Methods. *The International Workshop on the Web and Databases, San Diego, Florida, USA*, 2003.
- [34] LIANG, J.; KUMAR, R.; ROSS, K.W. Understanding KaZaA. 2004. <http://cis.poly.edu/~ross/papers/UnderstandingKaZaA.pdf>
- [35] LEIBOWITZ, N.; RIPEANU, M.; WIERZBICKI, A. Deconstructing the Kazaa Network. *The Third IEEE Workshop on Internet Applications, San José, CA, USA*, 2003.
- [36] WEHRLE, K.; GÖTZ, S.; RIECHE, S. Distributed Hash Tables. In R. STEINMETZ & K. WEHRLE (Eds.). *Peer-to-Peer Systems and Applications*, Berlin Heidelberg: Springer-Verlag, 2005, pp. 79-93.
- [37] BALAKRISHNAN, H.; KAASHOEK, M.F.; KARGER, D.; MORRIS, R.; STOICA, I. Looking up Data in P2P Systems. *Communications of the ACM*, 2003, vol. 46, no. 2, pp. 43-48.
- [38] ROUSSOPOULOS, M.; BAKER, M.; ROSENTHAL, D.S.H. 2 P2P or Not 2 P2P. *The Third International Workshop on Peer-to-Peer Systems, San Diego, USA*, 2004.
- [39] BOAI. Budapest Open Access Initiative. 2002. <http://www.soros.org/openaccess/read.shtml>.
- [40] JOHNSON, R.K. Open Access: Unlocking the Value of Scientific Research. *Journal of Library Administration*, 2004, vol. 42, no. 2, pp. 107-124.
- [41] MINAR, N.; HEDLUND, M. A Network of Peers: Peer-to-Peer Models Through the History of the Internet. In A. ORAM (Eds.), *Peer-to-Peer: Harnessing the Power of Disruptive Technologies*, Sebastopol, CA: O'Reilly & Associates, Inc, 2001, pp. 3-21.
- [42] WINSBURY, R. 2006 – The year that P2P comes in from the cold? Mass video broadcasting over the Internet. *Intermedia*, 2006, vol. 34, no. 2, pp. 12-17.
- [43] ARNOLDUS, M. Creative Commons Nederland: Workshop on technical and legal aspects of peer-to-peer television. 2006. <http://creativecommons.nl/nieuws/wp-content/uploads/2006/04/Verslag%20P2P-TV%20Workshop.pdf>.
- [44] KOZAMERNIK, F. EBU Seminar Report: From P2P to Broadcasting. 2006. http://www.ebu.ch/en/technical/trev/trev_306-p2p.pdf.

A Lifeboat Doesn't Do You any Good if it's not There when You Need it: Open Access and its Place in the New Electronic Publishing Paradigm

Ian M. Johnson

Aberdeen Business School, The Robert Gordon University, Garthdee Road
Aberdeen AB10 7QE, Great Britain
e-mail: i.m.johnson@rgu.ac.uk

Abstract

This paper draws on the results of recent research into digital publishing in Latin America sponsored by the European Commission's ALFA programme. It outlines the growth in publishing in the region. It aims to stimulate reflection on the impact of a system in which most of the publishing is supported by institutions rather than commercial companies, and considers authors' aspirations for their work to achieve recognition, attitudes towards peer review and other aspects of journal quality, the indexing and availability of full text journals, and the sustainability of institutionally supported publishing. Examples are drawn from publishing in the field of librarianship and information sciences on which the original research project was focused.

Keywords: Latin America; electronic journals; quality control; findability; sustainability

1 Introduction

"Our actions must embody these new 'realities' because even when people realize that they are on the Titanic and the iceberg is right ahead, we still need to see the lifeboat in order to jump ship"^[1]

Research into the diffusion of innovation notes the significance of communication channels in transferring awareness and understanding of innovations.^[2] A key part of that process has been the scholarly peer-reviewed journal. We are currently in a state of transition in global scientific communication as the new Information and Communication Technologies are becoming more generally available and more powerful. There is no denying the ability of the Internet to disseminate information rapidly, and it is rapidly being accepted that online access to the full text of scholarly papers should become the norm. However, access to 'free' information on the Web has stimulated a perception that the traditional scholarly journals might be replaced by new services compiled, edited, produced, marketed and distributed without the intermediary services of a publisher, a concept that has been promoted by the emergence of pre-print repositories and of electronic journals produced by individuals. The potential of these new models has proved particularly attractive to researchers and their funders, who had become concerned about the so-called 'scholarly communication crisis', and to librarians who have become concerned about the 'serials crisis.'

In the transition between paper and electronic publishing, discontent about the way in which information is handled is rising, and new experimental models are bound to capture attention. Like most phenomena where one order has to be replaced by a new order, a certain amount of disorder or chaos is inevitable in the transition phase. The 'noise' from the chaos has inevitably reached the ears of governments, and the debate has moved into the political domain. In these circumstances, it becomes a debate in which the awkward questions must be asked and answered clearly, or the solution that emerges may be one that will have to be revisited in more critical circumstances. It also requires us to take such parallels as exist and to examine them to assess what might be learned that is relevant.

Latin America provides an interesting paradigm through which to examine Open Access publishing, because the majority of journals published within the region are published by universities or with financial support from national research councils, other public institutions or professional associations. Commercial journal publishing has been inhibited not only by the relatively weak economies in the region, by the poor infrastructure of the book trade^{[3],[4]}, and by the lack of formal training in publishing. Nonetheless, developments in electronic publishing are taking place in Latin America, and it provides some notable examples of Open Access provision. It may thus

offer some realities as a contribution to a consideration of issues in the more industrialised, wealthier countries that could otherwise easily be dismissed as false hypothesising. The paper begins with a review of the growth of scholarly and electronic publishing in Latin America, and then focuses on the key issues in the debate about Open Access: quality, visibility, findability, and sustainability.

2 Methodology

This paper draws on the results of recent research into journal publishing in Latin America, undertaken with support from the European Commission's ALFA programme ('*America Latina - Formacion Academica*'). The aim of the ALFA Programme is to support collaboration between European and Latin American Universities. In common with most of the research and development programmes that the Commission sponsors, it has to be based on a multinational partnership. In the case of ALFA, the requirement is that there should be at least 3 Universities from Latin America and 3 from the member states of the European Union. In this project, the Robert Gordon University's partners were Queen Margaret University College, Edinburgh; Universidad Nacional Autonoma de Mexico (CUIB); Universidad Nacional del Sur, Bahía Blanca, Argentina; Universidad Federal do Parana, Curitiba, Brazil; Hogskolan i Boras, Boras, Sweden; and Universidad Carlos III, Madrid, Spain. The aim of the project was to identify professional journals published in the region with a view to ensuring their wider availability through digitisation and thus contribute to professional education and development. Many of the projects supported by the European Commission's Research and Development programmes have short names that are intended to capture the underlying idea. This project was no exception. REVISTAS – 'journals' in Spanish - became an acronym for **REd Virtual Sobre Todas las AmericaS**, which translated into English as something meaningful: 'a virtual network across the Americas.'

REVISTAS, focused on the feasibility of digitising journals as an aid to professional development in the field of Librarianship and Information Sciences, but journal publishing in this field is probably representative of many, if not most, disciplines in the region. The paper aims to discuss how traditional patterns of scholarly communication in that region are being or may be impacted by the shift to electronic media and the emergence of alternative approaches to publishing in a way that draws parallels between the Latin American experience and that in other countries where electronic publishing has become more widespread.

As well as reviewing much of the literature on the topic, the project team compiled a list of serial titles based on a number of indexes, journal articles and selected library catalogues. A comprehensive search would need to cover both the print and electronic catalogues of every institution that has taught librarianship and information studies, as well as every National Library in the region, and it must be acknowledged that more titles probably remain to be discovered by individuals more familiar with LIS publishing in their own countries. This is almost implicit in the wide disparity between the numbers of journals reported for each country. A final web search was carried out in early March 2007 using the metasearch engine 'Dogpile' to check for online versions of the list of titles that had been gathered to date^[5].

3 The Growth in Scholarly Publishing in Latin America

There is no reliable evidence for the number of scholarly serials published in Latin America, but there is clear evidence of growth in the number of publications appearing in the languages spoken in the IberoAmerican communities. Whilst data from the ISSN International Centre^[6] shows growth in the number of records for English language serials was c.19% between 2001 and 2006, it also demonstrates much faster growth in records for serials published in Portuguese, Catalan, and Spanish.

Language	2001	2002	2003	2004	2005	2006	Increase
Portuguese	13,244	13,277	13,294	13,310	21,324	21,361	61%
Spanish	37,064	39,782	41,859	43,850	48,222	51,112	38%
Catalan	1,034	1,163	1,248	1,340	1,479	1,555	50%

Table 1: Number of ISSN records per language (Source: ISSN International Centre).

Only 11 of the countries in which these languages are used have national ISSN Centres. The data from them suggests that there are at least 24,816 records for serials published in Latin America.

National Centre	2001	2002	2003	2004	2005	2006	Increase
Argentina	7108	7391	7722	7954	10,040	11,006	55%
Brazil	10001	10001	10001	10000	18572	18573	86%
Chile	1510	1559	1758	1813	2065	2244	47%
Colombia	1754	1754	1742	1743	1743	1798	3%
Costa Rica	146	146	146	146	146	146	0
Ecuador	159	159	159	159	159	159	0
Mexico	3432	3432	3431	3431	3431	3431	-
Portugal	3924	3924	3924	3924	3924	3924	0
Spain	18876	21309	22576	24382	26303	27851	48%
Uruguay	1526	1771	2019	2091	2225	2315	52%
Venezuela	1704	1704	1704	1704	1704	1704	0

Table 2: Number of records in national ISSN Centres (Source: ISSN International Centre)

The apparent absence of growth in the number of recorded serials in some countries, rapid increases in other countries, and variations in the number of ISSN records for countries with similar populations suggests that there is probably significant under-recording in the ISSN system. Moreover, Latin American journals have not always registered an ISSN^{[7],[8]}, and it may be speculated that part of the growth in records may be attributed to belated registration. For example, the growth in records for serials emanating from Brazil (8572) is more than those in Portuguese (8117). There may also be some discrepancy between the data held at the national and international ISSN centres because of the difference between these linguistic and geographic analyses. Moreover, some of the serials published in Latin America are published in English. For example, 14% of the 239 Open Access journals indexed by ISI are English language journals that originate in Latin America.^[9] Moreover, the number of titles apparently published in Portuguese and not recorded by the national centres in Portugal and Brazil is 1,136, which may or may not represent the output of the other countries from where publications in Portuguese may emanate (notably Angola, Goa, Mozambique, and Macau).

Not all the serials recorded by the ISSN Centre could be considered to be scholarly journals. More relevant data may be drawn from Latindex, the main directory of journals that is compiled within the region and principally intended as an aid for university libraries, which lists 15,578 titles, including 2,468 online journals.^[10] However, there may also be some under-recording in Latindex. The main index to library science articles about Latin America that is compiled within the region^[11] has recently been demonstrated to have indexed fewer than half the serials in the field that are now known to have been published.^[12]

4 The Transition to Electronic Publishing

According to the ISSN International Centre, some 50,000 serials are available internationally in computerised formats, compared with 1.2 million in print. Although some under-recording may again be suspected, this reflects significant change since the first experiments with electronic publishing commenced in 1992.

As a contribution to resolving the problems of scientific communication in the region, the participants in a Conference convened by the International Council of Scientific Unions in Guadalajara in 1997 argued that the mechanisms for the promotion and distribution of scientific publications must be improved and suggested "...the establishment of a Latin American scientific electronic periodicals collection."^[13]

Their thinking may have been influenced in part by the proximity to the establishment of *SciELO (Scientific Electronic Library Online)*^[14] in Brazil in 1997. Its Open Access service has since spread to several other countries in the region. In addition, many other journals have established an online presence. For example, of the 220 journals in the field of Library and Information Sciences that are known to have been or are currently being published in the region, the full texts of 48 have now been made available online, but only 2 have met the criteria for inclusion in SciELO. Others are moving in the direction of online publication: 8 more journals publish an Electronic Table of Contents and Abstracts online, and 13 publish their Table of Contents.^[15]

To provide access to these open access journals, a number of aggregator services have been established. The Brazilian Nuclear Information Centre maintains *LivRe*, a portal to more than 2,500 journals, not all of them in Spanish or Portuguese.^[16] A more selective service is provided by *RedALyC, Red de Revistas Cientificas de*

América Latina, el Caribe, España y Portugal, which is hosted by the Universidad Autónoma del Estado de México, and provides access to some 300 peer-reviewed journals in Spanish, Portuguese and English.^[17]

Several commercial database publishers also make a selection of journals in Spanish and Portuguese available. Grupo Océano, a Spanish company, has developed 6 databases covering different fields of knowledge.^[18] EBSCOHost has developed 3 databases^[19], whilst Thomson Learning promotes *Informe*.^[20] Dialnet also includes some Spanish language content.^[21] The most recent entrant to the field is ProQuest, which has developed a new collection of full-text scholarly journals *Publicaciones y Revistas Sociales y Humanísticas (Prisma)*.^[22] There appears to be some overlap between these services, and some even include titles that are freely available through SciELO.

5 Visibility

The growth in the number of serials may be explained by the growth of the local economies and consequently in national Higher Education and research systems. Latin American scholars are no different from those anywhere else in the world in the desire for their work to achieve recognition and make some impact in their field. In common with Higher Education institutions all over the world, Latin American Universities are experiencing the need to manage their educational, research and associated assets more effectively and transparently than in the past. They recognise that making their research and scholarly outputs more readily available will contribute to growth in the recognition of both scholars and institutions, and support the development of new and more fertile relationships between academic staff and departments both nationally and internationally, as well as stimulating economic and social development. Making them available could also facilitate much needed changes in teaching and learning, facilitating the development of a pedagogical environment that is information-rich and fosters the student-centred approaches to learning which are the key to success in the Twenty-First Century ‘Information Society.’

There is generally an expectation that - unless research is related to state security or defence, or is commercially confidential - the results will be published, i.e. that a report will be written for the sponsor, and that it will be summarised in whole or in parts in papers in scholarly journals, and perhaps in magazines intended for practitioners or the lay reader. Part of the assumption that these papers will be published is made possible by an understanding that “scientists in the public sector are largely motivated by intellectual curiosity, peer recognition and the promotion of the public interest rather than by private economic gain.”^[23]

The expectation that the results of research will be published as a journal paper is reinforced by the reward system in the academic world – a reward system that is supported by governments. Research Councils in some Latin American countries have given career incentives and financial rewards to academics who publish in journals of high recognition and visibility. Paradoxically, it is often countries that provide support for the publication of indigenous journals that also focus their reward system for researchers on the publication of their work in international journals.^{[24],[25]} Researchers in Latin America naturally want their papers to be published in international journals to improve access to their work and increase its global impact. These tendencies are enhanced in countries in which national research assessment and funding practices favour submission to international journals over submission to national journals. Elite Latin American researchers in all disciplines have therefore sought to maximise the potential impact of their research by submitting their manuscripts to well-established European and North American journals. For example, a study of the productivity of Mexican PhD holders trained abroad found that the majority had selected international journals indexed by ISI as their publishing outlets.^[26]

Since the evaluation of research work can be influenced to some extent by the visibility and reputation of the journal in which the work is published, the choice of highly visible, prestigious journals as publication outlets has become crucial, especially for scientists.^[27] If journals published in Latin America are to raise their attraction power for researchers in the region to select them as outlets for their research papers, they will clearly have to demonstrate that they are of comparable quality. This implies that quality control procedures will be in place, that other researchers will easily be able to find and access those journals, and that they will become sufficiently well established to become well known.

6 Quality control

The publisher of any journal is responsible for decisions about the level of quality control that is exercised by determining whether papers should be submitted to independent peer review. This will normally involve selecting (and often remunerating) an editor whose standing at least matches the perceived or expected standing of the journal, as well perhaps as some degree of oversight over the selection and activities of the members of any editorial advisory board. The editor makes a significant contribution to the standing of a journal by selecting experts from the editorial advisory board and/or others who can confirm that papers meet an acceptable standard in terms of their academic content. Paradoxically, although themselves largely drawn from the academic community, Latin American scholarly publishers and editors have not consistently addressed the crucial issues of quality control that affect the impact of the contributing authors' research. In the absence of any imperative to improve sales and distribution, peer review mechanisms in Latin American journals have been lax.^{[28],[29]} One consequence has been that the journals often duplicate coverage of subjects, or reprint papers from elsewhere, whilst possibly leaving significant gaps in the coverage of sub-disciplines.^[30]

In general, most scholarly publishing in Latin America has been handled by academics.^[31] One commentator observed that scholarly publishing in the region seemed to be operated by highly committed and altruistic academics trained to do research, but not necessarily to run publishing houses.^[32] These academics develop publishing skills on the job or in some countries through targeted professional development schemes. Latin American scholarly journals, supported by or through public institutions, depend on the annual budgetary allocation to enable them to sustain regular publication. They have been affected by regular financial crises in the region^[33], and have not always succeeded in maintaining a regular publication schedule. These institutional journals are frequently not sold through subscription mechanisms but exchanged for journals from other universities or associations. They rarely reach a wide international audience. Library collections often contain incomplete files of a journal.

To overcome the problems that are endemic in journal production in the region, inclusion in SciELO requires adherence to rigorous guidelines requiring peer evaluation and regular production "thus establishing challenges for the enhancement of the scientific output in the participating countries".^[34] However, the consequence of this policy of selectivity is that SciELO currently makes only c.248 journals available online in full text, a small proportion of the total published in Latin America.

Quality control also includes the technical preparation of the journal. Journal publishers incur significant costs in getting the product to the reader. Editorial offices have to be maintained for logging new papers, tracking their progress, and generally communicating with the authors, referees, printers, etc. A small but increasing number of publishers in Latin America are now using Open Journal Systems^[35], for managing journal production, particularly in Brazil where it has been translated into Portuguese (SEER - Sistema Eletrônico de Editoração de Revistas^[36]). Despite the fact that most papers are now submitted in electronic format, there is almost always a certain amount of effort necessary to check citations for accuracy, to create links to CrossRef, and to complete the proof reading and copy-editing. Some publishers and editors of open access journals appear to be attempting to transfer some or all of this responsibility to their authors. Whether that will be acceptable to authors and practicable remains to be seen.

The adoption of online format for some journals had not overcome the problem of irregular publication and consistent access. Some journals had not been published for several years; others had been only short-life experimental projects. Some other, single issues of papers or journals appear to have been converted into Permanent Document Format (pdf) solely at the initiative of their author or editor and made available through a repository or an aggregator such as RedALyC. In some cases, the URL had changed without a link to the new URL being created, or the links to the text of articles were broken. There is already evidence that some online journals are not attracting sufficient papers to maintain a regular publication cycle.

7 Findability

Making the journals or papers available online is of little value unless there are good indexing and abstracting services to guide the potential users to papers that are relevant to their interests. The international visibility of scholarly periodical publishing in Latin America has been the object of a number of studies.^{[37],[38],[39]} Whilst the regions major news magazines and newspapers are indexed in several subscription-based online sources^{[40],[41],[42]}, only a small proportion of scholarly periodicals from developing countries is indexed and abstracted by the major scientific secondary databases.^{[43],[44]}

A central archive of indexing data and a cross-site searching facility for SciELO is based in the original office at BIREME – *Biblioteca Regional de Medicina* in São Paulo. Recently, the SciELO index to its Brazilian journals has recently been uploaded into OCLC's WorldCat.^[45] SciELO Chile will be uploaded shortly, and other SciELO partners are expected to follow. This provides an alternative access point for potential users of the journals included in SciELO, and will arguably raise their visibility and use, at least amongst OCLC's member libraries. OCLC has also recently added to its database the indexes (*Clase* and *Periódica*) that have been compiled by the Dirección General de Bibliotecas at the Universidad Nacional Autónoma de México (UNAM-DGB) for the last 28 years, covering 400 of the region's journals in the arts, humanities, social and pure sciences.^[46] A similar number of journals, possibly the same collection, are now included in SCOPUS. However, the full texts of few of these journals are available online.

Research papers made available internationally through electronic publishing appear to have a higher national and global impact than achieved through publication in an indigenous printed journal. However, it is also important that the indexing service is widely known, and this is by no means the case. An interesting example is provided by the most substantial index to journal articles on Librarianship and Information Sciences from or about Latin America that is compiled within the region, itself the sole survivor of attempts made to establish such a service in several countries. INFOBILA was initiated in 1986 by the Universidad Nacional Autónoma de México Centro Universitario de Investigaciones Bibliotecológicas.^[47] It is based on collaboration with a network that covers 13 countries in addition to Mexico: Argentine, Brazil, Chile, Columbia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Panama, Peru, Uruguay and Venezuela. It has been freely available online since 1997 and has recently been redeveloped with the capacity to include or link to the full-texts of journals. However, it indexes only a handful of the online journals produced in the region. It is also disappointing that it seems possible that INFOBILA may be almost completely unheard of in the countries in which the indexed journals originate. An impromptu survey of the c.350 participants of a conference on digital libraries in Argentina in October 2005 suggested that only about 10 people there were familiar with INFOBILA. There is some evidence to suggest that, as a result of their availability through SciELO, a number of English language journals originating in Latin America and indexed by ISI are attracting more attention and more citations by other researchers than previously. A study of the 5 journals published in English in Brazil that have been indexed by ISI for at least 5 years, and available in full-text on SciELO^[48] for at least 2 years revealed that their impact factor had more than doubled since their inclusion in SciELO.^[49] Interestingly, Thomson ISI has recently agreed to begin including journals in Spanish in its Citation indexes from January 2006, possibly under pressure from its considerable Spanish customer base (as well, perhaps, as incipient competition from new indexing services such as Google Scholar^[50] and SCOPUS^[51]). The impact of this on author preferences for publishing outlets for their research remains to be seen.

Having good and well-known indexes goes only part of the way towards making the full text available. The difficulty in tracing the printed journals has been exacerbated by a relatively large production of new titles with small readerships and short life cycles.^[52] Commercial publishers seeking to digitise some of the region's journals have already experienced difficulty in finding complete sets to digitise, and the searches conducted for the REVISTAS project confirmed the haphazard distribution of copies of many of the printed journals. However, few of the online journals appear to have taken the steps necessary to publicise their existence, or to ensure that their contents are discovered by registering with a variety of aggregators and search engines. In many cases there was no evidence of registration of ISSNs. Coverage by the IberoAmerican e-journal aggregators was poor. *Livre* included only 29 of the 90 librarianship and information sciences journals published in Spanish and Portuguese (including those published in Europe), whilst *RedALyC* included only 8. The principal European aggregator of journals in Spanish and Portuguese (and other open access journals) is *REI*, *Recursos Electronicos de Informacion*, a service maintained by the Universidad de la Rioja in Spain.^[53] *REI* is maintained on behalf of REBIUN, the Spanish University libraries consortium and is not limited to peer reviewed journals, but still included only 15 of the 90 titles. Moreover, the aggregators and indexes are not necessarily well known. *RedALyC* was not known to the REVISTAS partners from the region, nor to senior LIS professionals based in the same city as its host institution. Bypassing the aggregators may overcome their deficiencies. SciELO is now beginning to use CrossRef^[54] to create links to and from the full text of papers in the journals that it hosts, but there are few signs that this practice has yet been more widely adopted.

8 Sustainability

Much of the fabric of online publishing in Latin America remains supported by institutions. Most of the journals are published by universities. SciELO is supported by FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo in collaboration with BIREME - Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde, and CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico. Anecdotal comments

suggest that its principal funders may be questioning why a local agency should be funding what has become an international service. RedALyC is alleged to depend to some extent on international aid from the Spanish government.

The technology lends itself to creating electronic collections that offer a range of features that add value, and that will be increasingly expected as standard features of e-journal services. Features such as cross-file searching, browsing, saved search histories, Table of Contents alerts by email, and citation linking must be introduced into the region's electronic journals, presenting challenges in terms of the availability of both skills and finance. That must raise concerns about the future sustainability of these journals. There is already evidence that some online journals had not been published for several years; others had been only short-life experimental projects.

The insecure financial base must raise concerns about long-term preservation of those electronic journals that do appear, particularly as many of the region's National Libraries do not have a preservation policy that extends to electronic media produced in their country – or the resources to implement one.

9 Discussion and Conclusions

In Latin America, the rewards and recognition for researchers and other academics are closely linked to the perceived quality of their published outputs. The evidence indicates quite clearly that the absence of any commercial imperative to raise quality to improve sales and distribution has had a damaging impact on quality controls. The publishers and editors of the vast majority of the electronic journals that have emerged to date seem largely to be continuing their previous neglect of quality control.

Visibility is clearly an issue for Latin American researchers. They want their publications to be highly visible. The evidence tends to indicate past failure in efforts made in Latin America to raise awareness of the contents of journals and to ensure adequate distribution of copies to meet potential demand. Making the full text of journals freely available online alone has not yet proved any more effective.

“Findability precedes usability”.^[55] The evidence is that the region's printed journals have not been well served by international or indigenous indexing services. Although some efforts are now being made to improve the arrangements for indexing, these only serve to highlight the limited availability of full-text sources.

The final issue to emerge from this study was concerns about the sustainability of publications and related services that depend on institutional support. The evidence tends to indicate that, to date, personal or institutional circumstances have contributed to the short life of many Latin American journals and newsletters. Simply switching to electronic media has not yet entirely resolved these issues.

The problems that have been discussed may be peculiar to Latin America. The examples drawn from this review of the region's literature of Librarianship and Information Sciences may not be exactly paralleled in every discipline, and further research to test the findings from this study on a wider scale is needed, and needed soon. However, the realities of scholarly communication in Latin America should prompt a pause for reflection by anyone interested in securing the future of scholarly communication at a time when the existing system is undergoing significant changes.

The aim of the paper was to use these realities to provide a fresh perspective on some of the global implications of the shift to electronic publishing, particularly to inform the debate about open access publishing, and to point to issues that still appear to need further consideration before significant changes in the system of academic communication are put in place. The situation that now exists may, in some respects, be resembled to the position of the ‘Titanic’ approaching the iceberg. The scale of the problem that confronts us is enormous. Just like the bulk of the iceberg, the vast majority of research papers are out of sight, not hidden below the surface but because they have not yet been written. The arguments about the most appropriate course to steer are complex. Faced with what is perceived as a major threat to scholarly communication, there are members of the research community and the library community who seem to want to abandon ship immediately without any clear idea of whether that is the safe course of action. The experience of institutionally supported publishing in Latin America and the faltering emergence of electronic journals there suggests that open access publishing could prove to be as much use as a safeguard for scholarly communication as a trap door on a lifeboat.

Acknowledgement

The author gratefully acknowledges the contribution made by the partners in REVISTAS and particularly Dr. Virginia Cano to the preparation of this paper.

Notes and References

- [1] REINSBOROUGH, P. De-colonizing the revolutionary imagination: values crisis, the politics of reality and why there's going to be a common sense revolution in this generation. *The Journal of Aesthetics and Protest*, 1 (2), August 2003. [online]: http://www.journalofaestheticsandprotest.org/1/de_colonizing/8.html [Accessed 8 April, 2005]
- [2] ROGERS, E.M. *Diffusion of innovation*. 4th ed. 1995. New York, U.S.A.: Free Press.
- [3] JOHNSON, P.T. A brief overview of the book trade in Spanish speaking Latin America *in Seminar on the Acquisition of Latin American Library Materials (19, 1974, Austin, Texas). Final report and working papers*. 1976. Amherst, Mass.: SALALM Secretariat pp. 55-59
- [4] BABINI, D.; SMART, P. Using digital libraries to provide online access to social science journals in Latin America. *Learned Publishing*, 19 (2), 2006, 107-113
- [5] 'Dogpile' [online]: <http://www.dogpile.com/> [Accessed 1 March 2007]
- [6] ISSN International Centre [online] <http://www.issn.org/> [Accessed 15 April 2007]
- [7] CANO, V. Bibliographic control and international visibility of Latin American periodical publications. *in: Indicators for developing countries, edited by R. Arvanitis and J. Gaillard*. 1992. Paris: ORSTOM. pp. 511-526.
- [8] CANO, V. Characteristics of the publishing infrastructure of peripheral countries: A comparison of periodical publications from Latin America with periodicals from the US and the UK. *Scientometrics*, 34 (1), 1995, 121-138.
- [9] MCVEIGH, M.E. *Open Access journals in ISI databases: analysis of impact factors and citation patterns*. 2004. [online]: <http://www.thomsonscientific.com/media/presentrep/essayspdf/openaccesscitations2.pdf> [Accessed 15 January 2006]
- [10] Latindex -Sistema Regional de Información en Línea para Revistas Científicas de América Latina, el Caribe, España y Portugal. [online]: <http://www.latindex.unam.mx> [Accessed 15 April 2007]
- [11] INFOBILA [online]: <http://cuib.laborales.unam.mx> [Accessed 15 April 2007]
- [12] JOHNSON, I.M.; CANO, V. Electronic publishing in Librarianship and Information Sciences in Latin America – a step towards development? Forthcoming.
- [13] CETTO, A.M.; ALONSO, O., *editors. Revistas científicas en América Latina – Scientific Journals in Latin America*. 1999. Paris: International Council of Scientific Unions; Mexico: UNAM, CONACYT, and Fondo de Cultura Económica. pp. 461-466
- [14] SciELO [online] - <http://www.scielo.org/> [Accessed 7 August 2004]
- [15] JOHNSON, I.M. REVISTAS – online LIS journals in Latin America. Focus on international library and information work, 2007. (forthcoming)
- [16] LivRe [online]: <http://livre.cnen.gov.br/> [Accessed 1 March 2007]
- [17] RedALyC [online]: <http://www.redalyc.com/mx> [Accessed 3 January 2006]
- [18] Grupo Océano [online]: <http://www.oceano.com/oceano/oceano.html> [Accessed 5 December 2005]
- [19] EBSCOHost [online]: <http://www.epnet.com/> [Accessed 15 December 2005]
- [20] Informe [online]: <http://www.gale.com/pdf/facts/inform.pdf> [Accessed 5 December 2005]
- [21] Dialnet [online]: <http://www.dialnet.com.mx> [Accessed 12 April 2005]
- [22] Prisma [online]: <http://www.il.proquest.com/division/pr/05/20050408.shtml> [Accessed 5 December 2005]
- [23] UHLIR, P.F. Re-intermediation of the Republic of Sciences: moving from intellectual property to intellectual commons. *Information Service and Use*, 23 (2/3), 2003, 63-66

- [24] VESSURI, H. Estrategia de valoración de las revistas científicas Latinoamericanas [Strategy for evaluation of Latin American scientific journals]. in: *Publicaciones científicas en América Latina*; edited by A. Cetto and K. Hillerud. 1995. Mexico: Fondo de Cultura Económica. pp. 200-210
- [25] BONILLA, M., and PEREZ ARAGON, M. Revistas Mexicanas de Investigación Científica y Tecnológica. *Interciencia* 24(2), 1999, 102-106.
- [26] LICEA de ARENAS, J.; SANTILLÁN-RIVERO, E.; ARENAS, M.; VALLES, J. Desempeño de becarios Mexicanos en la producción de conocimiento científico ¿de la bibliometría a la política científica? *Information Research*, 8(2), 2003. paper no. 147 [online]: <http://InformationR.net/ir/8-2/paper147.html> [Accessed 15 January, 2006]
- [27] RAVETZ, J.R. *Scientific Knowledge and its Social Problems*. 1971. Oxford: Clarendon Press
- [28] CANO, 1992, *ibid*.
- [29] MENEGHINI, R. Brazilian production in Biochemistry: the question of international vs domestic. *Scientometrics*, 23 (1), 1992, 21-30.
- [30] DIAZ, I.G.; AGUILAR, G.S. Las revistas científicas: su problemática en América Latina y El Caribe. [The problems of scientific journals in Latin America and the Caribbean.] *in: Revistas científicas en América Latina - Scientific Journals in Latin America*; edited by A.M. Cetto and O. Alonso. 1999. Paris: International Council of Scientific Unions; Mexico: UNAM, CONACYT, and Fondo de Cultura Económica, p. 231
- [31] CANO, V. International visibility of periodicals from Ireland, India and Latin America. *Knowledge and Policy*, 6 (3-4), 1992-1993, 55-78
- [32] GOMEZ, Y.J. A proposito de un Ejercicio de Evaluación de Seriadas Científicas. [A proposal for an evaluation of scientific journals.] Paper presented at the *Second International Workshop on Scientific Publishing in Latin America*, Guadalajara, Mexico November 27-30, 1997. [Unpublished]
- [33] BABINI; SMART, 2006, *ibid*.
- [34] GREENRIDGE, E. An overview of the PAHO Virtual Health Library. *in: Models of Cooperation in U.S., Latin American and Caribbean Libraries: the first IFLA/SEFLIN international summit on library cooperation in the Americas*; edited by B.E. Massis. 2003. Munich, Germany: K. G. Saur. pp.45-51
- [35] Open Journal System [online]: <http://pkp.sfu.ca/ojs/> [Accessed 15 April 2007]
- [36] SEER - Sistema Eletrónico de Editoração de Revistas [online]: <http://www.ibict.br/secao.php?cat=SEER> [Accessed 15 April 2007]
- [37] CANO, 1992, *ibid*.
- [38] KRASUKOPF, M.; VERA, M.I. Las revistas científicas de América Latina acreditadas en el ISI [Latin American journals indexed by ISI]. *in: Publicaciones científicas en América Latina*, edited by A. Cetto and K. Hillerud. 1995. Mexico: Fondo de Cultura Económica. pp. 168-175.
- [39] CETTO, A.M.; HILLERUD, K., editors. *Scientific publications in Latin America*. 1995. Mexico: ICSU, UNAM, and Fondo de Cultura Económica. p. 305
- [40] Info-Latinoamérica [online] - <http://www.nisc.com/factsheets/qila.asp> [Accessed November 2003]
- [41] Latin American Newsletters [online] - <http://www.latinnews.com> [Accessed November 2003]
- [42] Prensa Latina [online] - <http://www.prensa-latina.cu/English/> [Accessed November 2003]
- [43] CANO, 1992-3, *ibid*
- [44] GONCALVES DA SILVA, L.; SILVA FERNANDES, R. La cobertura de las revistas Latinoamericanas por los Servicios de Indización: el caso de las revistas brasileras. Paper presented at the *Second International Workshop on Scientific Publishing in Latin America*, Guadalajara, Mexico, November 27-30, 1997. [Unpublished]
- [45] OCLC WorldCat [online]: <http://www.oclc.org/worldcat/default.htm> [Accessed 3 January 2006]
- [46] *Clase and Periódica* [online]: <http://www.dgbiblio.unam.mx/> [Accessed 7 May 2006]
- [47] INFOBILA is available, free of charge, directly through the UNAM-CUIB web site [online]: - <http://cuib.laborales.unam.mx> [Accessed 7 August, 2004]
- [49] ALONSO, W.J., and FERNANDEZ-JURICIC, E. Regional network raises profile of local journals. *Nature*, 415, 2002, 471-472

- [50] Google Scholar [online] – <http://scholar.google.com/> [Accessed 7 May 2006]
- [51] SCOPUS [online] – <http://www.scopus.com/scopus/home.url> [Accessed 7 May 2006]
- [52] GUIMARAES, J.P. Opportunities and common goals for research in the Americas. *in: Science and technology in the Americas, perspectives on Pan American collaboration, edited by J. Stann.* 1993. Washington, D.C., U.S.A.: American Association for the Advancement of Science. pp. 65-72.
- [53] REI, *Recursos Electrónicos de Información* [online]: <http://aps.unirioja.es/biblio/recursos?sub=1> [Accessed 3 June 2006]
- [54] CrossRef [online]: <http://www.crossref.org/> [Accessed 3 January 2006]
- [55] Findability.org [online]: http://www.findability.org/archives/cat_findability.php [Accessed 12 January 2007]

Expectation and Reality in Digital Publishing: Some Australian Perspectives

Bill Martin, Hepu Deng, Xuemei Tian

School of Business Information Technology, RMIT University, Melbourne, Victoria, 3000, Australia
e-mail: {bill.martin; hepu.deng; xuemei.tian}@rmit.edu.au

Abstract

This paper presents a brief summary of the findings of a Web-based survey of the views of Australian publishers, on the potential impact of digital technologies, followed by three case studies conducted between January and April 2007. The survey results indicate that the most influential technologies currently in use in publishing are the Internet and the World Wide Web, with little or any interest being shown in for example, semantic technologies. There is however, widespread realization of the importance of providing enhanced customer value through digital content and delivery channels, with consequent implications for changes to value chains and the emergence of new and transitional business models, which however, are likely to complement rather than replace existing business models. The case studies drawn from a set of eight selected to include a range of value propositions and business models suggest that in Australia publishers are optimistic about the prospects of digitisation but are nonetheless cautious in its uptake and application.

Keywords: digital publishing; business model; value chain; case study; Australia

1 Introduction

Traditionally the publishing industry has played a key role in the dissemination of knowledge and for centuries was a forerunner of what today would be described as a *knowledge-based business* [1]. Until the closing decades of the last century, publishing and associated printing activities were based upon old technologies, with clear implications for business processes and relationships among the main stakeholders in what was basically a linear progression from the creator of content to its publication in print form [1, 6, 7].

The advent of digital technology has potentially limitless implications for publishing both in hard copy and electronic formats [2, 6, 7]. Combined with advances in electronic commerce [8] it offers the prospect of new value propositions and business models for those who are able to take advantage of developments in digital technology. Digital publishing incorporates several characteristics including an infrastructure that gives multiple options with digital content available in various formats and viewing modes according to customer requirements and basic editing processing and updating of information on the server, leading to reductions in processing time and the fast, efficient transmission of content, with subsequent economic benefits [5]. This said, even the latest digital tools and applications are at best enabling mechanisms whose adoption must relate to the overall business strategy and purpose [4, 5].

This paper presents the initial findings from an Australian government-funded research project looking at the implications of digital technologies for the publishing industry in Australia, with particular emphasis upon current and emerging stakeholders, competition, and value propositions and business models, current and potential. The findings (which are still to some extent interim in nature) have emerged from a variety of research activities including literature review, focus groups, a national online survey of publishers and the conduct of case studies. The project adheres to the generally accepted view of publishing as a set of content industries comprising sectors for book, journal, newspaper, directory, magazine, music, maps and multi-media publishing [1, 2]. However, its major focus is on book publishing in Australia. This paper concentrates largely on three case studies conducted during the research.

2 Methodology

Following the conduct of an extensive literature review (including analysis of secondary documentation such as Annual Reports) and of three focus groups, the major research methods employed were those of survey and case

study. After several unsuccessful attempts to obtain access to relevant membership listings, the researchers made use of a commercial listing service. They provided a list of 65 publishing companies throughout Australia. In the event this turned out to be an exercise of somewhat limited value in that the great majority of addresses obtained were those of newspaper and magazine publishers (particularly publishers of rural newspapers), most of whom had no interest in participating in the project. However all those responding were in fact book publishers and their responses, limited in number as they were, tended to support the major assumptions underlying the survey. The case study was operated on the basis of a set of protocols designed in order both to facilitate consistency in the handling of responses to key issues raised in the survey and to ensure the presence of a certain amount of rigour in the conduct of the case study exercises.

3 Analysis of Survey Results

On a more positive note, the conduct of a survey had always been regarded as being part of a triangulation process involving the literature review and focus groups and the conduct of case studies. The data analysis resulted in identification of the general extent of progress made towards planning and implementing digital initiatives, and more specifically, those factors that influenced this process and issues with regard to industry trends, stakeholders and competition, propositions and business models. Of the 65 surveys mailed, and subsequently re-mailed to publishers, only 14 were returned completed. Although a response rate of almost 22 percent from a Web-based industry survey would appear to compare well with the reported norm for such exercises of 4% to 6% [3], the researchers make no claims for significance. The findings will be presented in a forthcoming paper and are summarised here as follows:

- 70% of respondents reported increased growth in revenues from existing products/services and nearly 60% from new products/services.
- The main benefits anticipated from digital technologies are in the areas of new niche markets, repackaging and repurposing of existing content, consumer-generated content and the enhancement of value chains.
- The most profound effects expected from digital publishing are in the areas of specialist business/professional/academic publishing, government and web-based publishing.
- The critical success factors for digital business models were identified as technical robustness, consumer acceptance and financial logic.
- Subscription-based and content creation business models were the most highly regarded, frequently in the context of niche markets.
- Key organisational changes anticipated included:
 - ❖ Introduction of digital media divisions.
 - ❖ Introduction of an integrated platform for all editorial operations, print and digital.
 - ❖ Changes in human resource practices to suit a digital environment.
 - ❖ Organisation-wide promotion of cultural change to suit a digital environment.
 - ❖ Introduction of new strategies for the digital market.

4 Background to the Case Study Element

About half of the candidates for case research emerged from the online survey exercise and the rest were obtained later by direct approach. The three cases reported here are drawn from a group of eight that will be completed as part of the research project. These are all exploratory and descriptive in nature, rendering them suitable for the kind of interpretive research undertaken in the project. The case studies gave researchers the opportunity to meet face-to-face with senior members of publishing companies and discuss the results of the survey analysis with them. The case study instrument was designed to enable respondents to take ownership of the process and respond within the boundaries of meta-level questions [8].

Interviewees were asked a combination of open and closed questions and were free to add anything else they thought important. The case studies operated on the basis of a standard set of protocols relating to research design, operating procedures and data analysis techniques [3]. This was to guard against bias and ambiguity and to ensure as far as possible that a logical chain of evidence could be seen to operate from the initial research questions to the ultimate conclusions [9-11]. This, for example, led to the use of *How* and *Why* type questions for exploring operational links over time and *What* type questions for exploring new phenomena such as digital developments. The protocol specified detailed procedures in relation to data collection during the interview process. Every interview was recorded and transcribed, with the transcriptions being read and independently analysed by two of the three-person research team, one of whom had not participated in the particular interview session under analysis. In addition, the teams of two interviewers both kept separate field notes, which again were later subject to mutual and then third party scrutiny. Finally with regard to data analysis, the strategy was designed to link findings and interpretations not to generalisable outcomes but to contexts beyond the immediate, to extrapolation to other situations and environments [11].

5 The Case Studies

In the three cases reported here, the firms are identified only by the use of numerals. They comprise respectively a university press (Company A), an educational publisher (Company B) and an electronic publisher with close connections to a conference operator (Company C). The major focus will be on their business models, which for present purposes are perceived as a description of the roles and relationships among a firm's consumers, customers, allies and suppliers that identifies the major flows of product, information and money and the major benefits to participants [12]

The business models of the three firms were identified following cross comparison of each company across a range of constructs identified as important to successful business models. These are:

- Customer base
- Value proposition
- Value chains
- Core competencies
- Products and services
- Partners
- Use of and Attitudes to Technology
- Risks and opportunities
- Business models

5.1 Customer Base

There is a considerable similarity in the makeup of the customer bases of these three firms, serving as they do a largely academic or educational market. However, one area of difference is that Company A as well as Company B is engaged in the provision of publishing services to conference organizers. Specifically, the customer base of these three companies can be described as follows:

- For Company A, the customer base has remained remarkably stable for the last 16 years, with the main difference being with regard to expansion into overseas markets. Most of their customers are libraries (notably academic, state and corporate) and small publishers, with a minority of direct sales to end users over the Web.
- For Company B, the customer base is comprised of teachers and pupils in the primary and secondary sectors.

- For Company C, the customers are mainly academics either seeking to publish their own papers or access those of others, on either a subscription or a per item basis. There is also a small but growing segment of custom in the library market and substantial revenue from the provision of publishing services to the associated conference business.

5.2 Value Proposition

All three firms offer customers a range of value propositions including:

- Companies A and C offer the benefits of a full electronic publishing service including provision of software, metadata, file conversion, content management and quality.
- Company B, while offering a digital dimension in the form of PDF formats and Website *question and answer* facilities, has as its major value proposition the ability to delivery content in the form of hard copy textbooks.
- Companies A and C offer the benefits of online access to and delivery of aggregated and indexed content based on a common technology platform and sophisticated search technologies.
- Company A offers provision of an additional marketing, sales and promotion channel to its customers.
- Company A provides archival services.

5.3 Value Chains

The value chains of the three firms are all familiar in scope although that for Company B is much the more traditional: author to publisher to printer to distributor/bookseller to reader [1]. While in essence the same, the value chains for companies A and C are much more geared to a digital environment with the major stages entailing:

- Stage 1: Acquiring content from authors or owners (via licensing or payment).
- Stage 2: Obtaining and converting digital files involving PDF and XML formats, creation of metadata and databases, editing and quality assurance.
- Stage 3: Printing (frequently outsourced) with content held in digital repositories.
- Stage 4: Sales, marketing, promotion through representatives, print media and virtual and physical book shops.
- Stage 5: Archiving content

All three companies were confident of maintaining their place in what they expect to be changing value chains in the near future. They were not concerned about possible disintermediation as a result of technology, but all agreed that booksellers have reason to be concerned.

5.4 Core Competencies

All three companies identified as core competencies the provision of high quality content (in either print or digital formats), the ability to organize content including editorial competencies and the ability to negotiate licensing and royalty arrangements, and the provision of networks of business partners and services including production, distribution, marketing and selling. Those competencies emerging as specific to individual companies included:

- Meta data creation, file conversion and content management (Companies A and C).
- Competencies in current and emerging classroom content delivery methods (Company B).
- Competencies in curriculum development and assessment (Company B).

- Technology competencies (Companies A and C).
- Conference management competencies (Company C).

5.5 Products and Services

With respect to the products and services offered, Company B is clearly different from Companies A and C. This is because the main source of revenue for Company B is through the sale of hard copy textbooks, with a modest trade in e-Books in PDF format and the delivery of classroom content via digital whiteboards.

In contrast Company A earns 96% of its revenue from digital products and services including:

- Bibliographic databases which also form the basis of the search infrastructure.
- Online databases giving access to fully indexed full text journal articles by using a single search interface.
- E-Press: a cover-to-cover aggregation of journals, monographs, conference papers, reports, occasional series and other *grey* literature published in Australia and hitherto not widely available online.

Company A has a minor trade in hard copy books (some 4% of output) and offers a full e-publishing service to a growing client list.

Company C also sells consultancy services (both publishing and technological) , as well as hosting conferences, the revenues from which subsidizes publishing activities including:

- Access to digital content in the form of monographs, single papers and electronic journals.
- Access to journal contents via an archive of titles and abstracts.

5.6 Partners

All three firms have common partnership arrangements in the form of links to authors, printers, marketers, distributors and booksellers. Company B has a particular relationship with schools and Company C with its associated conference business. Of the three, Company A has the most diverse range of partners which in addition to those mentioned in the foregoing include the National Library of Australia, the Copyright Agency, a range of government departments, such as the Attorney General's Department, various research centres in fields as diverse as family studies, criminology, agriculture and languages and a range of small publishing operations seeking to go digital. All these partners in a variety of ways add value to the products and services of the case companies.

5.7 Use of and Attitudes to Technology

There were clear differences in the attitudes of the three case firms towards the take-up and development of technology. Both Company A and Company C had from the outset relied upon the use of technology to gain market share and a competitive edge. They had sought to market a technology-intensive value proposition. Company B was much more pragmatic, linking developments in technology infrastructure and applications much more closely to market demand. Although a multinational company with ample financial and other resources, Company B did not maintain an active research and development program as such, preferring to monitor general developments and if necessary respond appropriately.

Companies A and C were easily the most enthusiastic of the six case companies interviewed prior to the writing of this paper. They both strongly endorsed the potential for *many-to-many* forms of communication including contributions from end users and distributed content and cognition. They were particularly interested in the potential of the Semantic Web and Web 2, not least given their respective histories of involvement with and expertise in metadata creation, file conversion and content management. They are heavy users of XML for the management of often relatively small print runs and the transition from source to print and web outputs using open standards and a high degree of automation. To this extent they see themselves as already beginning to

engage with the notion of the semantic web, but realize that there is a long way to go before this comes to fruition.

This commitment to technological development at both Companies A and C is simply a reflection of their continued appreciation of the value of technology to the sustainability of their businesses. Hence while both Companies A and B outsource aspects of meta data creation and file conversion (in the case of Company C to Mumbai) this has been done more for technical and quality reasons than simply to cut costs. For Company C the Mumbai operation is critical to its global data harvesting activities, which in turn are central to the marketing of conferences and the recruitment of authors and journal editors.

Both Companies A and C have invested heavily in proprietary content management and workflow systems. Key files and databases at Company A are based on Terratext Foundation software developed within the company's parent institution and for which Company A has a permanent licence. Company C, following extensive work with almost 20 standards, has developed a core publishing and workflow management system (CG Publisher) which it claims is the first fully online publishing environment in the world. It manages publishing proposals, version control for drafts and editions, and contracts and automatically places completed texts (print and electronic) into an easily managed self-publishing site, as well as into personal sites for each of the authors. Company C believes that it is largely owing to the existence of its core management systems that it has been able to grow its business ten-fold during the last three to five years

Although nowhere as engaged in the development and application of technology as the other two firms, Company B is by no means oblivious to its importance. In addition to a small-scale involvement in the production of e-Books, Company C delivers content under licence to classrooms using its own range of electronic whiteboards. To date the uptake of this technology has been constrained both by a shortage of relevant content and by school budgets.

5.8 Risks and Opportunities

Company A sees very little on the horizon as regards potential risks, and in particular nothing in the way of threats from new entrants or from developments in technology. In terms of good governance they are focusing on keeping costs down, for instance in relation to royalty and licensing fees and looking at opportunities for improving their delivery infrastructure in order to reduce the unit costs of production. There is little sign of any potential problems from for instance channel or supplier conflict. The company is very comfortable with ongoing developments in Open Access publishing, which it regards as being highly domain and content-specific and where the future may lie in the publication of material that is not saleable on a commercial basis. Company A is currently participating in a local repository experiment, for which they are providing input on software and content management. However they see this more as a goodwill gesture than as a commercial venture. So far as technology is concerned, they have been early adopters of digital opportunities and they would see further opportunities in the digital publishing space owing to their strengths in metadata creation and management and in indexing and searching. They are also intending to pursue new markets comprised of library consortia and large libraries in Asia, the United Kingdom and North America, to repackage and reformat existing materials for corporate and enterprise markets, and to develop new products both with regard to aggregated services and content.

Despite its relatively low key presence in digital markets, the future whether in terms of technological or related change holds few fears for Company B. Hence, although much has been made of potential disintermediation in the value chain for publishing consequent upon the empowerment of authors or on competition from new players in the market, Company B is confident that whereas booksellers may be adversely affected, changes in publisher-author relationships are just as likely to be in its favour. There could be a risk of channel conflict were they to move to any substantial form of direct-to-customer sales, or indeed to any wholesale attempt to deliver content through their Website (hence conflicting with the traditional book selling model). On the other hand, threats from the wholesale digitization of texts, say by Google or Microsoft, are seen as more a matter for old material than for new. Their customers on the other hand, want new and dynamic content. Curricula are constantly changing and publishers have possibly unique expertise, not only in updating content but also in scoping and sequencing it in relation to course changes and more generally in the organization of content. Their view is that if the Internet has taught us anything it is that *more is not necessarily* more when it comes to, timely, relevant and high quality content of credible provenance. The major threats posed to educational publishers in Australia for the foreseeable future are not those of digitization, but rather of government policies, not just as regards the funding of education but also in relation to support for the creation of content. In Australia, governments at both state and Federal level, appear to see the future of digital content production as lying

outside the mainstream publishing industry. For example, a national effort to produce such content through a body called *the Learning Federation*, has largely succeeded only in producing sets of learning objects (animated content intended to illustrate the use of say mathematical or scientific concepts for use by teachers) which according to surveys conducted by the Copyright Agency are used by very few schools. Company B is of the opinion that they could profit from the opportunity to develop and supply customized digital content. Indeed, a diversification in content creation, along with realistic funding for school hardware and software, could result in a transformation of the firm's value proposition to the point where additional revenue streams would accrue not just from digital content but also from the provision of hardware and software.

Ironically one potential area of risk for Company C stems from its inherent technological strengths. The fact that their technology is so sophisticated means that it is costly both to implement and to amend for different purposes. Moreover, there is a very high sunk cost in a relatively small pool of technical staff, with the accompanying risk of knowledge loss and damage to the business through the departure of key people. So far as competition is concerned, Company C is extremely comfortable, given that it owns the conference business which underpins the supply of content to its publishing arm. They perceive potential opportunities through the development of semantic technologies and given their existing expertise in connecting and processing digital documents they believe that they have every reason to be positive about the future.

5.9 Business Models

Authors such as Timmers [13] and Weill and Vitale [12] have categorized business models by type, arguing that for any organization, the business model can be constructed from any two or three atomic models drawn from this categorization. An analysis of the business processes of the three case study companies revealed that all of them contained at least two of the following atomic e-business models:

- Direct-To-Customer: This involves a small but growing B2C model operating as pay-per-view with customers paying either by monthly account or by credit card.
- Content provider: providing content (information, digital products and services) via intermediaries.
- Intermediary (Aggregator): bringing together buyers and sellers by concentrating information.
- Shared infrastructure bringing together a range of players (some of them competitors).

The models shown below will employ a schematic developed by Weill and Vitale [12] wherein the following components and relationships will be depicted as follows:

- Participants: Represented as:
 - ❖ Squares (firms of interest).
 - ❖ Left- and –right-facing pentagons (customers and suppliers).
 - ❖ Split squares (partners – organizations whose products or services help to enhance demand for those of the firm of interest).
- Relationships: Where solid lines between participants indicating a primary relationship and dotted lines an electronic relationship between the parties.
- Flows: Where arrows represent the major flows between participants and can either be money (\$), a product or service, digital or physical (0) or information (i).

Business model for Company A

The company sees itself as having a hybrid business model that involves publishing and aggregating largely on a business-to-business basis. It began basically in 1989 as a cost-recovery model, but since 1997 it has operated as a commercially sustainable (but not-for-profit) publisher and aggregator. It contains elements of all the four atomic models listed above. Figure 1 shows an overview of its business model.

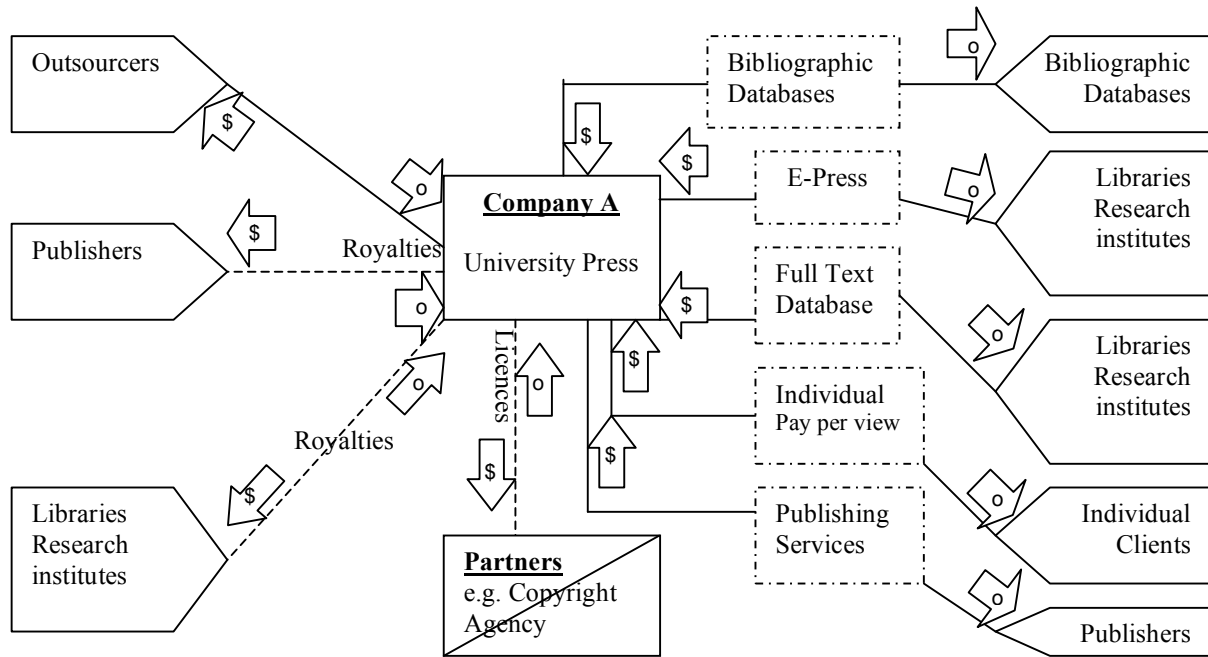


Figure 1: The business model for Company A

Business model for Company B

The business model for Company B is based largely on the traditional market for textbook sales, but again it contains elements of at least three of the atomic models listed above. Figure 2 shows the business model for Company B.

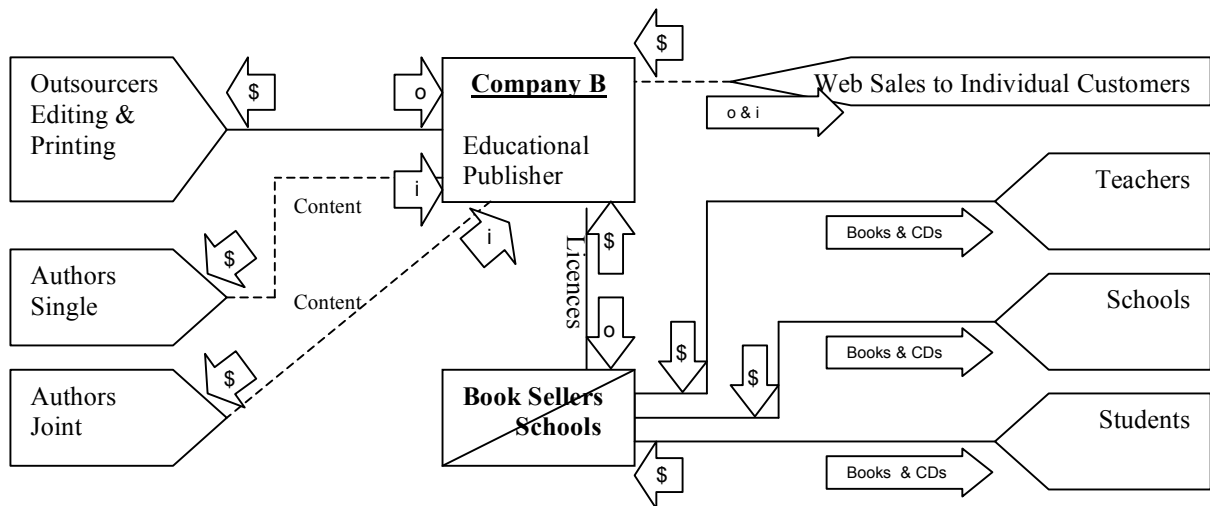


Figure 2: The business model for Company B

Business model for Company C

The business model for Company C is largely that of a full service provider, with elements of direct to customer and content provider models included. Essentially Company C sells publishing services to conference attendees including peer reviewed publication of single papers and sales from an online book store. Figure 3 presents the current business model for Company.

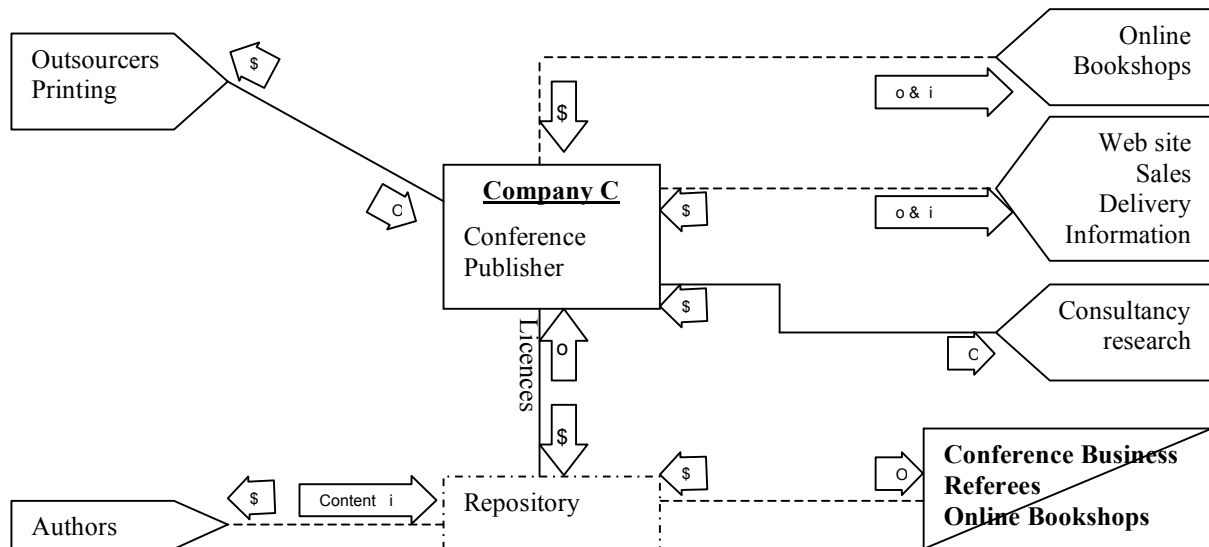


Figure 3: The business model for Company C

6 Conclusions

What has been reported here are findings from three of what will ultimately be eight case studies seeking to identify current and future business models for book publishing in Australia. The case study protocols, the structure of the interviews and the nature of the questions posed were all determined by feedback from focus groups and a national online survey. On the basis of what has been learned from the six cases conducted until now, the researchers perhaps over-estimated the likely impact of technology on the thoughts and deeds of publishers, while underestimating the continued popularity of the printed book. To some extent this is not so apparent in the context of the three cases reported in this paper. Companies A and C are major users of leading edge technologies and see the future very much in terms of the exploitation of technology for business sustainability. Company B a highly successful and profitable multinational publisher of educational texts, remains much more focused on traditional perceptions of value and on channels for its delivery, while maintaining a careful watch on market developments. For Company C this already entails the ability to respond to what for it is a minority demand for digital content, and evidence from the other three cases not covered in this paper suggests that publishers are *hedging their bets* to the extent that many of them have a growing presence in markets for digital products and services. This is certainly the case for example, with the industry partner for our research project (CCH Australia), which while regarding itself as a traditional publisher operating in niche professional markets, nonetheless generates up to one-third of its revenue from digital sources. The overall conclusion, therefore, is that publishers are *making haste slowly* in response to the potential inherent in digital technologies, whose potentially disruptive presence is more than balanced by a range of organizational, commercial and market factors.

References

- [1] COPE, B; KALANTZIS, M (2002), Managing Knowledge: Communication, Learning and Organization Change. In Cope and Freeman eds, *Developing Knowledge Workers in the Printing and Publishing Industries*, Common Ground Publishing, Melbourne.
- [2] COVEY, DT (2003), Copyright Permission: Turning to Dust or Digital. *International Journal of the Book*, Volume 1, Common Ground Publishing Melbourne.
- [3] DUBE, L; PARE, G (2003), Rigor in Information Systems Positivist Case Research: Current Practices, Trends and Recommendations. *MIS Quarterly*, 27, 4, 597-635.
- [4] DAVIS, M; WALTER, M (2003), Next-Wave Publishing Technology: Revolution in Process and Content. *Seybold Publications*, 3, 15, 3-15.
- [5] KLEPER, M (2001), *The State of Digital Publishing*. Prentice Hall PTR.

- [6] JANSEN, B (2003), The Future of the Book: Format and Technology. *International Journal of the Book*, 1, Common Ground Publishing Melbourne.
- [7] MASON D; COPE, B (2001), Australian Book Production in Transition. In Cope and Mason eds, *Creator to Consumer in a Digital Age: Australian Book Production in Transition*, Common Ground Publishing, Melbourne.
- [8] MOLLA, S; HEEKS, R; BALCELLS, I. (2006), Adding Clicks to Bricks: A Case Study of e-Commerce Adoption by a Small Catalan Retailer. *European Journal of Information Systems*, 15, 424-38.
- [9] OLIVER, S; KANDADI, K (2006), How to Develop Knowledge Culture in Organizations: A Multiple Case Study of Large Distributed Organizations. *Journal of Knowledge Management*, 10, 4, 6-24.
- [10] ROWLANDS, I; NICOLAS, D (2005), New Journal Publishing Models: An International Survey of Senior Researchers. *A CIBER report for the Publishers Association and the International Association of STM Publishers*, London, CIBER <http://www.ucl.ac.uk/ciber-pa-report.pdf>
- [11] WALSHAM, G (1995), Interpretive Case Studies in IS Research: Nature and Method. *European Journal of Information Systems*, 4, 74-81.
- [12] WEILL, P; VITALE, M (2001), *Place to Space: Migrating to E-Business Models*, Boston, Harvard Business School Press.
- [13] TIMMERS, P (1998), *Business Models for Electronic Markets*. European Commission, Brussels, <http://webarchive.org/web/20030612192921/http://lists.commerce.net/archives/ecowg/199901/msg00010.html>.

Libraries as Publishers of Open Access Digital Documents: Polish Experiences

Marek Nahotko

Institute of Information and Library Science, Jagiellonian University
ul. Gronostajowa 7, 30-387 Kraków, Poland
e-mail: nahotko@inib.uj.edu.pl

Abstract

This article presents the experience of Polish libraries in the field of electronic publishing. There have been described some solutions applied for creating digital libraries and institutional repositories. Nowadays, Polish libraries seem to be passing from the stage of digitalization of their own collections (usually of historic value) to publishing new digital-born documents in their own institutional or multi-institutional repositories. This activity should be (and is) developed in co-operation with university press companies.

Keywords: digital library; open access; electronic publishing; Poland

1 Introduction

In Poland, a large number of societies dealing with research communication (authors, librarians and users – readers of scientific publications) are satisfied neither with the operation of today's research communication system nor information exchange. Besides, negative effects can be noticed not only in research communication, but also in other information processes and communities, what is visualized, for example, in a constant decrease in the level of books reading at public libraries. The reasons for the dissatisfaction are, among other things, a sharp increase in the prices of publications, copyright-related issues, problems pertinent to intellectual property as well as a still longer and longer time interval between arriving at research results and their publishing. Polish libraries are active participants in the discussions held about the items said, and they seek to extenuate the problems, which come up and to submit some proposals of practical solutions aimed both at reforming the system of publishing and relative processes. Their actions are chiefly focused upon the items connected with access to the resources. The first goal is heading off the 'crisis of journals'; the problem consists in fighting against prohibitive prices as they make it considerably harder for one to find the desired publication. The other is the limitation of effects brought about by 'access crisis', which means not only limitations in permanent access to documents already published, e.g. by preventing one from gaining access to scientific electronic journals when the pertinent license has expired, but also impediments to having access to older publications, still esteemed by users.

2 Methodology

In this article have been utilized the data collected during research into Internet websites, dedicated to projects related to electronic publishing, whose initiators are mainly libraries operating on *dLibra* software. Then, there has been also presented a case study of a digital library which, due to a relatively long period of operation, its experience and achievements, is a good example for illustrating the trends, which nowadays dominate in Poland in the field of e-publishing performed at libraries. While analyzing the problems under presentation, there has also been made use of interviews held with librarians – authors of new forms of publishing; the Internet bulletin board was also used for this purpose. The analysis has covered as well the relation of publishers of Polish scientific journals towards the idea of electronic availability of their publications. The journals placed on the list of the so-called score journals by the Polish Ministry of Research and Sciences have been marked out for the analysis said. An author publishing in such a journal is awarded proper score, highly appreciated when his research achievements are subject to evaluation. According to the Ministry, those are top level journals in their fields; therefore, their publishers should take care of spreading the contents under publication also electronically. Hence, it might be supposed that the position of journals not included on this list must be yet worse.

3 Results

Recently, some initiatives related to electronic publishing, and first of all, to making digital libraries, have appeared in Poland. This phenomenon is typical for the library sector at the beginning of the 21st century. Not only do libraries collect traditionally their resources and render them available to the public, but they also take over some new tasks, for instance, electronic publishing of documents. On the turn of the 20th and 21st centuries, in many forums (for example, at numerous conferences), Polish librarians debate the issue of involving libraries in electronic publishing. Nowadays, another stage which consists in the implementation of practical solutions, has commenced. Initially, attention was focused on the digitalization of libraries' own collections, mainly for their protection and archiving. Later on, there appeared also some projects aimed at electronic publishing of newly born documents, usually born-digital. Such documents are published in digital repositories through the mediation of librarians who administer them.

Position of libraries

In Poland, libraries started their actions related to building digital libraries with the digitalization of their own collections at the beginning of the 21st century. The survey carried out in 2003 showed that the digitalization was performed in 25 libraries, and in 14 of them were special purpose-built laboratories [1]. According to the latest information collected (end of 2006), 115 libraries deal with digitalization – 51 of them are at university level schools, including almost all university libraries (16) and technical university libraries (11); instead, 48 of them are public libraries.

Initially, most initiatives were short-term tasks, whose purpose was mainly taking immediate actions and performing services as ordered by users. In consequence, they turned out to operate with certain irregularities and were affected by various, often subjective factors (equipment, staff, finances). Some libraries established special divisions within their organizational structures; others preferred services performed by external companies.

As digitalization was developing, it was necessary to make certain decisions on the selection and choice of materials for digitalization. Those problems were often solved by library managers and specialists, employed mainly at the divisions dealing with special collections, the collection acquisition and circulation as well as with its protection and preservation. In most libraries, digitalization programs cover first of all old prints, manuscripts, incunabula and 19th century journals and magazines. The main purpose of the actions mentioned was not only protection of valuable resources, but also meeting the new needs of library users.

Reasons for digitalization of library collections:

- The necessity of protecting the collections possessed against destruction and making them accessible to a large number of readers; those collections are of high value from the viewpoint of cultural heritage;
- The need of rendering university press books, textbooks and other learning materials more and more accessible, as well as of making them adapted to use in distant learning (e-learning);
- Readiness for involvement in the promotion of university level schools through popularization and spreading of research and culture potential as well as intellectual production of university staff (series, journals);
- The necessity of becoming active participants in the national strategies eEurope, ePoland and in the strategy UE i2010, as well as the need of participating in the initiative of digitalization of collections of the top Polish libraries.

Lastly, the digitalization process of the resources held in Polish libraries has been quickening its pace. The libraries plan to maintain this pace in the nearest future. Among other things, this process is favored by a reduction in prices of IT hardware and services.

Anyway, the quality of digitalizing operations still needs improving. A part of electronic versions have been compiled from poor quality materials, e.g. old microfilms, which will cause them to be useless soon and the digitalization process will have to be repeated. Another unsettling information is the lack not only of any uniform standards for recording and archiving digital documents, but also of certificates to determine the durability of digital records media (some collections are recorded on CDs). Therefore, neither the future nor the accessibility of such resources is certain.

In order to coordinate the works and to secure a more close co-operation among those libraries which render digital publications accessible, there has been established a consortium called Digital Library Alians [2]. Its aim

is to develop and to intensify the actions related to acquisition, presenting and popularization of digital resources connected with both the cultural heritage in different Polish regions and research resources produced at respective Polish university level schools. In consequence, the co-operation and the funds raised due to joint efforts should lead to a development of regional networks of digital libraries which are supposed to constitute a stable structure, and to an unification of standards and an optimization of the solutions adopted.

Polish libraries – both academic and public – participate in electronic publishing in two ways, by:

- Making digital libraries - which contain documents digitalized (scanned) from originals, collected and stored in traditional (printed) library resources. Originals are often of value, and access to them is hindered. Those are usually historic documents to which the copyright is no longer applicable;
- Creating repositories of digital documents (articles, PhD, MA theses, reports etc.), whose authors are researchers employed at institutions provided with a repository. Those are documents for which the copyright provisions are applicable.

In both cases said, libraries become publishers and editors of electronic documents; in most cases those documents are digitalized copies of traditional publications. So far, in neither case we can say about any traditional roles played in both fields by libraries. Therefore, because of new roles involved, librarians face many new problems to be solved in the matter of new technology and organization application.

There can be distinguished two organizational models of libraries and/or digital repositories in Poland:

- institutional (academic libraries, Polish National Library), including 7 libraries working on dLibra software;
- regional models, focused mainly on major university level schools, sometimes also on regional public libraries; they consist of 2-23 institutions, chiefly libraries, but also of museums and archives.

dLibra Software

Most initiatives presented so far are based on Polish software called dLibra, compiled at the Poznan Supercomputing and Networking Center (<http://dlibra.psnc.pl/>). Nowadays, this software is applied in a few dozen libraries (Fig. 1). This software serves for professional making of collections of digital objects. It allows to collect and to render digital objects available in Internet in various formats (eg. txt, html, pdf, djvu). Each object may consist of any number of files and is described with metadata (MARC, Dublin Core etc.). Each implementation of dLibra software includes the three main elements working in the client-server configuration (Cf Fig. 2):

- The server of the digital library/repository, responsible for the performance of all library functions, usually operating on the dedicated hardware, not accessible directly to end users;
- An application of the editor and administrator (client) which allows them to enter digital objects, their descriptions and execution of other similar functions;
- End user's application (client), based upon Web interface, and allowing one to have customized access to the objects within a collection.

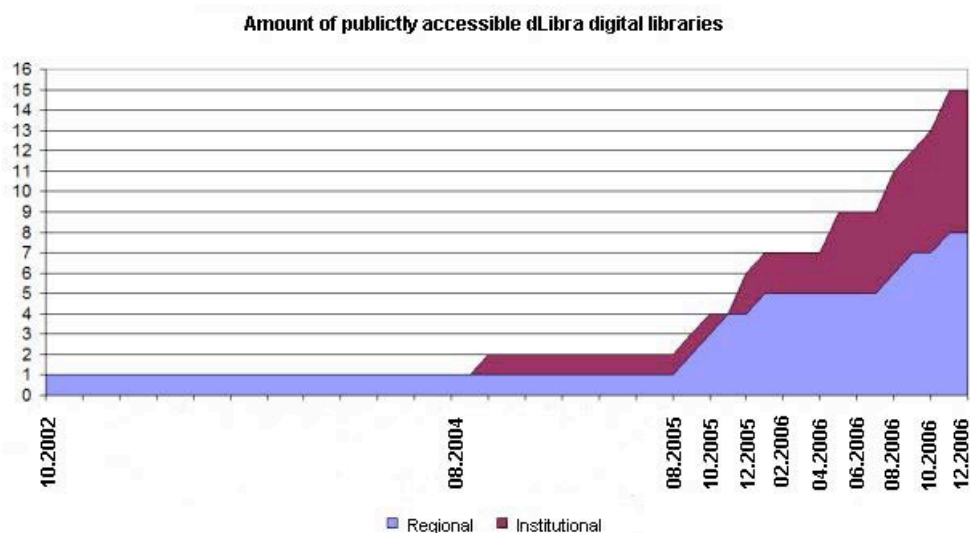


Figure 1: Progress of dLibra digital libraries in Poland

dLibra allows us to implement the majority of international standards, presently under application worldwide, for example RSS, XML, RDF, MARC, Dublin Core or OAI-PMH. It can be upgraded by independent programmers under condition of free access to the newly prepared software.

Publications are placed to the system by their authors directly or with librarian's intermediation. Any author of a publication can modify texts previously compiled, which leads to their new editions. Those editions consist of files which also may have various versions. Editions, in turn, can be published or not; they can be also made accessible for a certain time until the fixed date.

It is also possible to make group publications serving, in turn, for combining single publications which have some common features e.g. successive journal issues or series. Within a group publication may operate other groups, too. Each group is provided with its own description. Publications may be grouped into collections. Each publication may belong to more than one collection. In the case of assigning a group publication to a collection, all publications within a group are automatically assigned to this collection. Collections may be divided into sub-collections, which leads to a tree structure. Collections are provided with their own descriptions, also copied to a sub-collection with a possible modification.

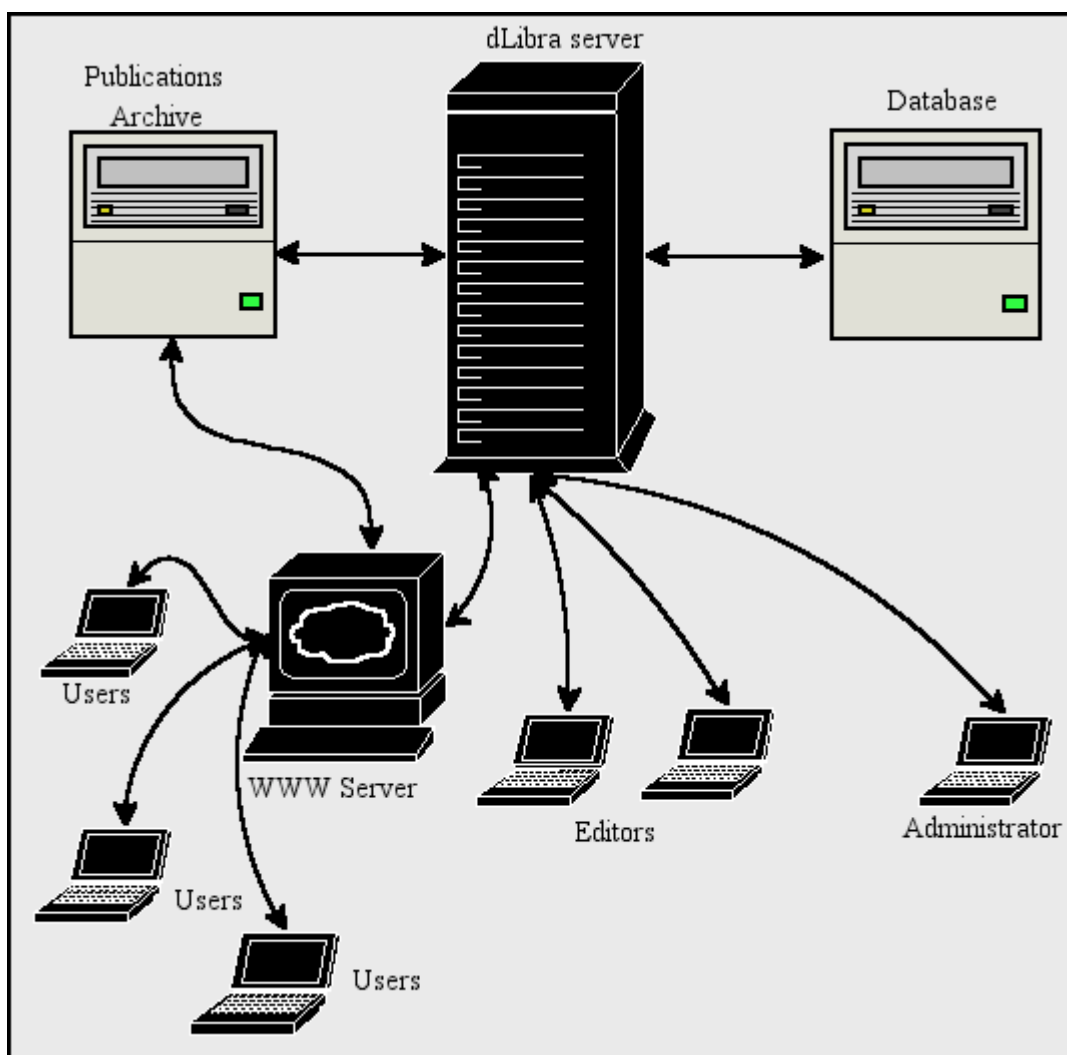


Figure 2: dLibra client-server architecture

Publications collected by dLibra software are indexed by popular search tools, like Google; hence, those are not resources of any hidden (invisible) Web. The architecture designers paid much attention to indexing of descriptions of digital objects by search engines, which has led to quite a good effectiveness. Another functionality, very important for the user, is a possibility of searching the contents of all dLibra resources from the level of each system implementation (Cf Fig. 3, option: 'Search remote libraries'). In consequence, irrespective of the library the user has chosen at the beginning of his/her search, he/she can search all digital libraries consisting one network, with one search tool [3].

Case Study – Kujawsko-Pomorska Digital Library

One of the oldest and largest digital libraries in Poland is the Kujawsko-Pomorska Digital Library (KPDŁ). In order to set it up, in 2003 there was established a consortium of libraries led by the Nicolaus Copernicus University in Torun (north-western part of Poland) and its library, with participation of other two regional high schools. It is also planned to co-operate with local public libraries. Each cooperating institution places their own digital resources on a joint platform, and administers them in the scope of compilation, updating and access rules. KPDŁ is a part of the Project of building open information society ePoland, which in turn constitutes a part of eEurope. EU's financial share was 75%.

The resources of KPDŁ digital objects consist of three collections:

- Research and teaching collection aimed at improving the quality of teaching by securing access to digital copies of textbooks, monographies and research articles;

- Cultural heritage collection which is to include the most valuable rare books, manuscripts, books published in the 19th and 20th century, to archive records, music notes, emigration, cartography and iconography collections;
- Regional records which will include publications, articles and occasional materials on the history of the Region of Kujawy and Pomerania [4].

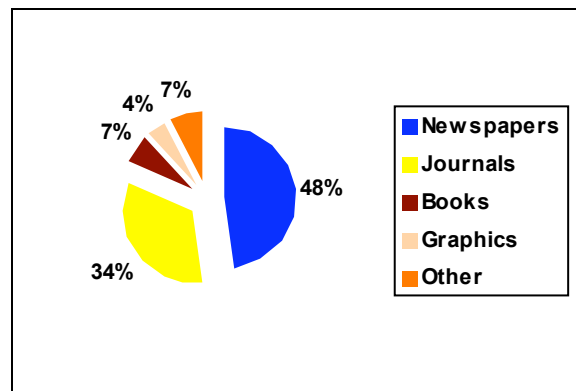


Figure 3: Kinds of objects in KPD L

Collections are divided into smaller groups as needed. Materials are assigned for digitalization by three libraries, which bear joint responsibility for the KPD L resources, namely two academic and one medical (Cf Fig. 3). Access is first of all given to teaching materials in the field of medicine (university notebooks, journals and monographies published before 1945, and a self-published journal ‘Biological and Medical Sciences’). Other branches of science are represented by regional historic journals, subject to digitalization in co-operation with two chief regional public libraries. Instead, the Nicolaus Copernicus University Library offers access to the sources on the history of the region, emigration collections, engravings and Vilnius records [5], so much essential for the history and tradition of this University. The Library also digitalizes teaching materials (e.g. set books for philologists), not subject to copyright restrictions. So as to avoid repetitions, the lists with materials to be digitalized are agreed upon both electronically and at monthly meetings of the editing staff.

The co-operating university level schools publish as well their own, contemporary materials and research papers. The authors of such works sign proper licence agreements in which they may reserve the range of access to their work: no limitations in the whole Internet, at their own university only or access to the users of consortium libraries only. On the same basis, they give their consent to the KPD L for electronic publishing of their texts. The authors hand over their works free of charge.

It was not difficult to select the software for the digital library under project, since libraries operating on dLibra software had been already in existence (Fig. 4). It was acknowledged that such a platform is provided with the fundamental functions, indispensable for any digital library: cataloguing and giving access to text and graphic files, searching of documents through any words taken from the description or contents of the document, collections management, navigation within a publication or limitation of access to a selected group of users. A significant feature of dLibra was the compatibility allowing one to work with the library system Horizon, used in the libraries of the region.

Before a publication can appear in the KPD L, it must go through certain stages (in the brackets are those who are responsible for their performance):

- Section and assignment of documents for digitalization (selecting librarians) based on rules as agreed upon;
- Compilation of objects ready for digitalization as the so-called list (selectors);
- Queuing of the documents assigned to digitalization and queue control (editor);
- Technical works on a document and its handover to the digitalization lab (selectors);
- TIFF format scanning and archiving (technicians);
- Processing of OCR scanned files into DjVu format (CT staff, technicians);
- Compilation of a bibliographic description for a local catalogue in Horizon system (MARC 21 format), conversion into dLibra (Dublin Core format) (catalogers);
- Publication in the digital platform (editor);
- Control of metadata in dLibra and Horizon systems, amelioration of the resources and possible corrections (main cataloger) [6].

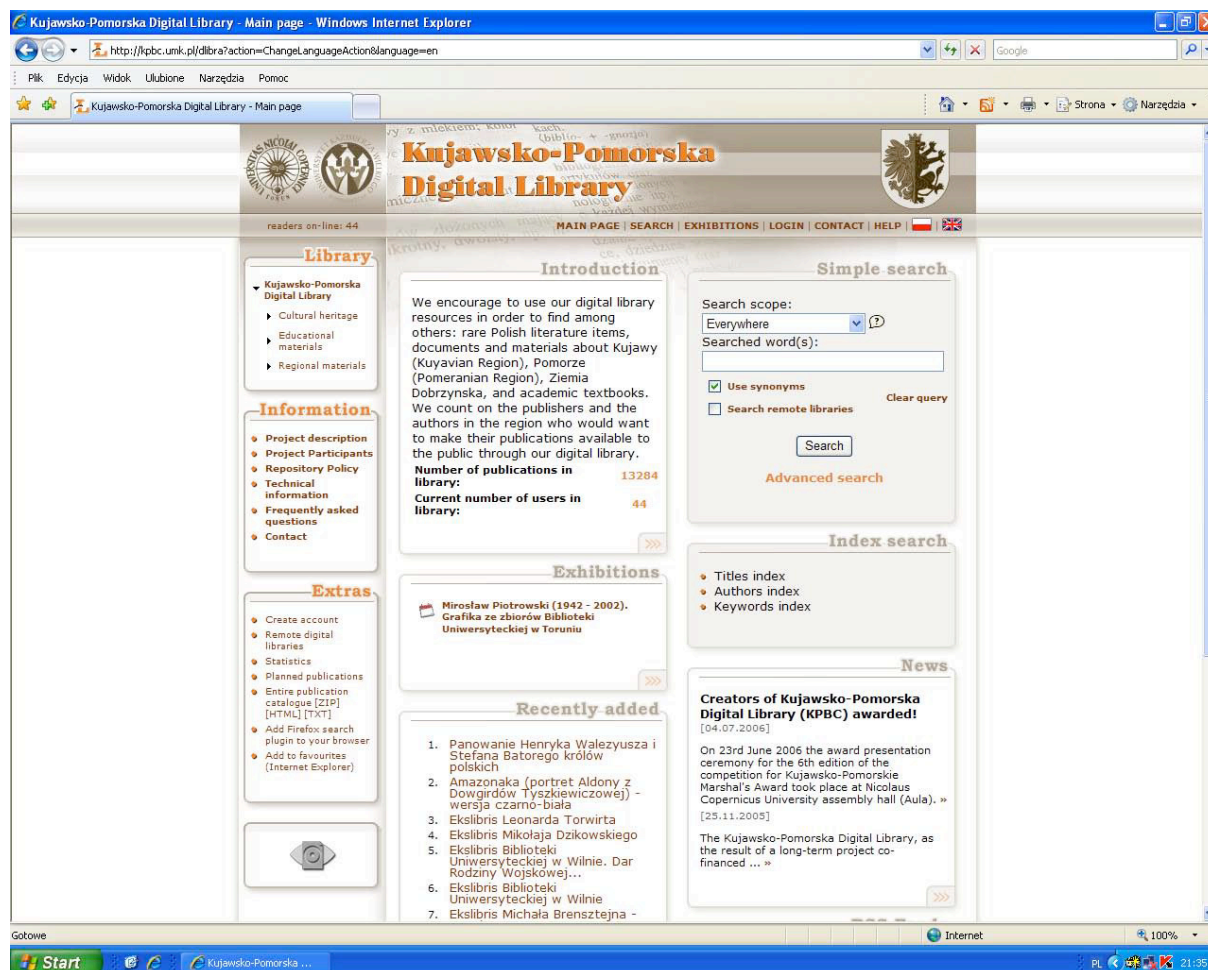


Figure 4: Main page of KPD

Procedures compiled for respective stages provide the following actions:

1. Assignment of documents:
 - a. determination what should be digitalized and how to do it,
 - b. compilation of lists with no more than 15 items, with detailed data for processing (scanning quality, color, and others),
 - c. uploading of lists on a joint disk accessible in Intranet.
 Responsible: librarians responsible for respective collections.

2. Queuing of documents (lists):
 - a. queuing of lists, priority assignment, setting the sequence of scanning,
 - b. keeping the lab informed about any queue and that the performance of a task is possible,
 - c. hand-over of lists to the digitalization lab,
 - d. constant control of the digitalization process and of the compliance with the procedures.
 Responsible: KPD Editor

3. Scanning and archiving:
 - a. a staff member orders documents for scanning by contacting the person who is signed under the list,
 - b. the staff member signs the list for a librarian who supplies the materials, and the former considers the list as a lending form,
 - c. scanning (in the lab) of documents as queued on the list; any remarks about the scanning result are addressed at the editor,
 - d. processing of the scanning result with the programs enclosed to the scanner (picture correction, framing etc.),

- e. entering the file under a standard name (shelf number is a file name or folder name for many files); assignation of catalogue numbers to the box with the carrier of the archival digital version of the document; physical description of the archival digital version (file format, carrier type and recording date, resolution, color detail level etc.),
 - f. transmission of final files for further processing,
 - g. delivery of materials to the division of formal working against receipt, like in b/.
- Responsible: Digitalization lab staff.

4. Processing of the files created in the scanning process:

- a. obtaining of file formats (DjVu, HTML, PDF and others) as planned in the process of assignment and preparation; OCR for some objects,
 - b. handover of files to the editor for further operations.
- Responsible: Digitalization KLab Staff, CT specialists.

5. Publication on the digital platform:

- a. combining an object with a description,
 - b. uploading of files either to one or more dLibra collections as indicated,
 - c. supplementing of a description in dLibra with the archival version metadata, if any.
- Responsible: KPDL Editor.

6. Compilation of a bibliographic description:

- a. cataloguing in Horizon,
 - b. placing final descriptions in the target dLibra collection,
 - c. constant amelioration and quality control of descriptions in dLibra,
 - d. making corrections to dLibra, review of indices,
 - e. handover of books to the bookstacks or to the reading room as specified on the list.
- Responsible: librarians responsible for respective special collections.

7. Control of metadata and transmission of a description to NUKAT [7]:

- a. entering of the new data related to a digital object to the existing records in NUKAT database,
 - b. compilation of new records with a re-routing to a KPDL object,
 - c. combining of existing records in NUKAT database.
- Responsible: chief cataloger.

According to the list said, the respective tasks are carried out by a team consisting of various members who have different levels of skills and qualifications:

1. Project coordinator – administration, finances, cooperation with partners, promotion, content and quality supervision, negotiations on copyright with authors.
2. Coordinator's deputy – supervision over the CT part of the project, hardware, software, contracts with suppliers, tenders, standards.
3. Administrator – project files, finances, reporting, personal matters, correspondence and others.
4. Editor – edition of digital library objects, idea of resources and its administration, coordination of works of the team which compiles and enters documents.
5. CT specialist – software, supervision over dLibra software, engineering solutions, statistics.
6. Chief selection specialist – selection of documents for digitalization, selection of materials from special collections, work coordination.
7. Chief cataloguer – bibliographic description, metadata, amelioration of entire resources, standards.
8. Technicians – digitalization, supervision over the lab, scanning standards, objects archiving.

Situation of the university press publishers

Nowadays, in the process of research paper publishing, the role of the author of a publication, viz. the compilation of a text becomes the easiest one. But problems will start soon after. First, one must get some funds (grant) for publication. When finally apportioned, this money turns out to be halved. That is why the publisher usually refuses any royalty for the author, considers the entire project as an unprofitable task, which has no positive effect upon the development of any enterprise. All advertising and marketing activities spell only more expenses; that is why from the publisher's viewpoint the best solution would be withdrawing the item already published from distribution at all. Other activities carried out by researchers in the course of the publishing procedure, like preparation of reviews are also performed free of charge, which yet worsens the unfavorable situation mentioned.

In addition, access to the information is hindered by improperly arranged book trade. For being commercial entities, bookshops do not deal with academic books as usual [8]. Their activities are targeted at mass consumers who, for example, purchase such items, like Harry Potter; instead, the sale of single copies of scientific texts goes beyond the boundaries of commercial risk.

A large number of university press publishers try to send their items by themselves, often through Internet bookshops; however, not all of them resort to such a solution. Then, there will also appear other problems, for instance, mail-order sale of a low circulation book whose publishing has been refunded, and in consequence, the publisher has already got their profit; such a situation may be seen as a contingency, not profit. University press publishers divide their items into those they have got to publish at cost due to their role played in academia, and those on which they can gain some profits. Of course, professional marketing refers to the latter.

In turn, this means that in the process of publishing of academic items it is necessary to find a new solution in which libraries, especially academic ones, may and should actively participate. Academic libraries begin to take over the functions pertinent to university press publishers. Practically, those publishers are by definition non-profitable entities, always in deficit. This will lead to establishing a kind of electronic library publishing houses, a part of which will become electronic repositories of publications supposed to be used free of charge within a reciprocally advantageous cooperation held with other institutional and regional repositories. Due to such a cooperation, it will be possible to economize on publishing and by giving access to a large number of non-commercial (yet valuable) low circulation publications. Eventually, a large number of university press publishing houses will become superfluous. Their today's number arises from the fact that each university level school/college, however small it may be, has an ambition to have their own publishing house. In consequence, there are many microscopic publishing companies, and in many cases their professionalism and potential are exiguous.

Moreover, the process of publishing academic texts will be accelerated; as of today, it takes years to bring them out as publications. Such a situation results in part from the top-down order coming from the government agencies in the matter of having the quality of research and academic publications evaluated. The aftermath of such an evaluation is the list of scientific journals as published by the Polish Ministry of Sciences; any author who has published in them is assigned a score enhancing the evaluation of his research achievements. Therefore, the editors of the magazines from this list have their hands full for a few year time or more.

4 Discussion

As a result of the actions taken by all persons and institutions involved in contemporary research communication, one can notice a change in the roles assigned in the process of making electronic publications. After a short time, indispensable for preparing such changes, we can expect serious modifications to the operation of publishing companies and assignments performed by their staff; nevertheless, such modifications will also refer to librarians and book dealers. In a more or less conscious way, representatives of those professions get ready for the changes and modifications to come soon. Such changes will also refer to scientific community. Authors become editors and publishers; instead, publishers deal with the aggregation of contents and contribute with their own value added. Librarians turn into digital librarians, which is related to their participation in electronic publishing.

According to the actions enumerated and reported in the case study, the establishment of a repository administered by the traditional library or consortium of libraries entails changes to the organization because, those take over new functions. A part of tasks performed in the digital library demands only that librarians should change their way of working and their habits (e.g. transit from MARC 21 cataloguing to metadata, like Dublin Core). Other actions entail completely different skills, so far typical rather for publishing companies than libraries, e.g. compilation and edition of digitalized contents.

Differences between traditional and electronic publishing can be described by dividing each research communication process into four stages at which concrete functions are performed:

- Description of the idea and conceptions arising from conducted research;
- Certification of values of the described ideas and research results;
- Distribution of ideas and results by making them accessible to prospective readers interested in them;
- Archiving of results so that they might be utilized successively.

It is evident that such functions are performed by each system of research communication, either traditional or contemporary, based upon new digital technologies.

Function	Process	Performed by:	Financed by
Registration	Delivery of a digitalized text	Librarians	Repository
Certification	Review	Researcher – reviewer	Publisher of the printed original
Circulation	Open repositories	Librarians	University level schools, local government
Archivization	Permanent access	Librarians	University level schools, local government

Table 1: Model of Polish repositories

In compliance with the model said, one may state that Polish initiatives are endowed with a certain, separate set of features if compared with similar initiatives developed in other countries. In the Polish model, authors do not provide repositories with their latest publications; instead, old items (often 100 and more years old, to which the copyright provisions are inapplicable), are supplied by the librarians who have scanned them. Since Polish repositories (mainly regional) are often established by public libraries, they are financed from local government budgets.

Eventually, Polish repositories are dominated by archival resources which consolidate the role of the library in the field of archiving and museum functions, but such repositories do not play the main role usually attributed to them, since they make no contribution to the acceleration of research communication. Such a situation may result from the absence of agreements in the field of author's rights and copyright. It is still unclear which solutions will be adopted. Especially, two of them are under consideration: solutions as applied in Wikipedia, and Creative Commons. As of today, respective institutions make their own agreement/ contract models to be used while receiving texts from their authors to be published in the repository.

A solution of today's difficulties related to research communication in Poland seems to be a cooperation between libraries and university publishers so as to make an Open Access publishing system. The base for such a cooperation should be a modification to today's Polish model of academic/research publishing. Scientific institutions publish their own items (mainly journals and series) to be exchanged for those from other institutions (also from abroad). Many scientific centres convert their published items into the electronic form. Those journals are usually of non-commercial nature – they are financed from different sources – directly by government agencies and research institutions.

The transition from the traditional library via digital library to the digital publisher is a part of processes leading to a development of digital research and science. It should facilitate innovativeness, by making new ways of production and popularization of research results. Due to new IT methods and technologies, it is possible to make research communication more streamlined at all its stages – from making texts, via evaluation of their quality, administration, circulation and archivization.

For assessment of Polish academic publishers, of interest could be also some data regarding scientific journals published by small university press publishers and research societies. Parallely be stressed that practically all such journals are brought out by small publishing companies, since there is no large, commercial publishing house which deals with this business. Among the journals placed on the polish government list (available on website WWW: http://www.nauka.gov.pl/mein/_gAllery/13/66/13662.pdf), one may find top score items; if you publish in them, it will contribute greatly to the evaluation of your academic achievements. About 75% of the journals have their own websites. As far as journals with websites are concerned, in 48% of cases, one may gain

access to full texts (Cf Fig. 5). Instead, 52% of them place on such websites contents and/or abstracts (sometimes very general data). It means that in Poland the process of making the contents of scientific journals available to the public in electronic mode has already started, but in this field we are still behind the level of 80-90% of journals available online, as in Western Europe and USA.

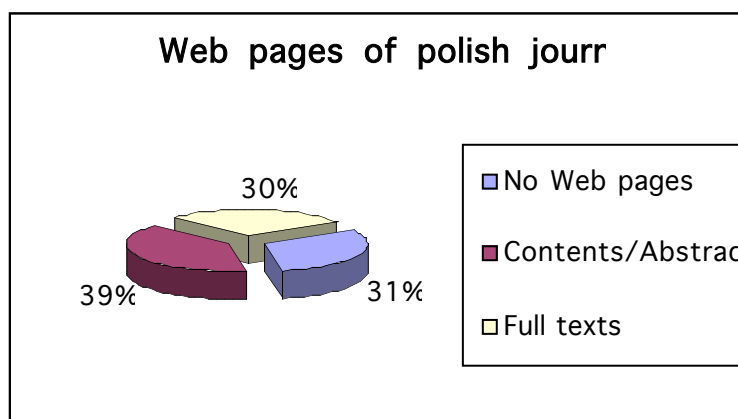


Figure 5: Accessibility of Polish journals on the Web

A large part of those publications is published in English. As far as the magazines from the government list, available in Internet, are concerned, 75% of them are available in English. They are made accessible in wide networks, which can thus secure a possibility of their popularization (wide distribution), also due to the absence of a language barrier. In particular, this refers to journals dealing with medicine, sciences and mathematics.

Quality control often consists in selecting materials by the editors; therefore, there is often no typical reviewing. However, this is not any problem in small circles of specialists in narrow fields of science, since those specialists know each other. In such circles, it is easier to perform quality control according to the reputation of respective researchers and research institutions (faculties, institutes).

5 Conclusions

In Poland, there have appeared some new initiatives related to electronic publications. One of them is the initiative of Interdisciplinary Center for Mathematical and Computational Modelling (ICM) [9] in Warsaw, and the Library of the Warsaw University, which proposes the establishment of a national repository of research texts named DIR. The repository could include as well the collections from the digital libraries already existed. The core of the project is to be the Virtual Library of Science, already in operation (<http://vls.icm.edu.pl>). The cooperating institutions might provide DIR with their own electronic documents (scans or versions made as electronic) with the metadata added. The model of cooperation assumes the storage digital objects by various ways:

- i. In DIR, only;
- ii. In local repository (digital library), only;
- iii. In both places at the same time.

In ICM-DIR, materials would be given a final retouch, and their presentation in Internet could be secured. In consequence, there would be created a central, but scattered collection of Polish science accessible via one searching interface.

A model of open access repository of digital objects in conjunction with new publishing processes as performed in libraries proves to be more and more successful in Poland. The process of delivering the value added as a result of publication is not contradictory to the values of the open access idea. Experience shows they may complement and support each other, particularly when researchers try to expand and develop their research due to the application of new forms and possibilities of electronic publishing. The creation of digital libraries and repositories in Poland is an interesting example of the process of integration of digital libraries, repositories and publishing activities as is carried out by librarians in their traditional libraries. Those processes are under way now; therefore, not all their effects are known yet.

In such a way there appears a change in the publishing paradigm, like in the period following the invention of printing, when texts previously available only in manuscripts went to print on a large scale. Materials previously available in print are nowadays digitalized so as to be included in the worldwide resources of digital objects. At the same time, new born-digital objects are being collected, and their availability will be higher because the original (author) version of all publications is digital and has been ready to use for a long time.

Notes and References

- [1] KOWALSKA, M. (2006). Digitalizacja zbiorów w bibliotekach polskich – próba oceny doświadczeń krajowych. *Biuletyn EBIB*, 11(81), <http://www.ebib.info/2006/81/a.php?kowalska>
- [2] ROŻNIAKOWSKA, M.; MARGAS, M. (2006). „eBiPol” – Biblioteka Cyfrowa Politechniki Łódzkiej na tle innych inicjatyw bibliotek cyfrowych w kraju od strony technicznej, formalnej i projektowej. *Biuletyn EBIB*, 4(74), http://www.ebib.info/2006/74/rozniakowska_margas.php
- [3] KALOTA, T. (2006). Marzenie o polskim systemie rozproszonych bibliotek cyfrowych. *Biuletyn EBIB*, 4(74), <http://www.ebib.info/2006/74/kalota.php>
- [4] CZYŻAK, D. (2005). Kujawsko-Pomorska Biblioteka Cyfrowa – stan zaawansowania realizacji projektu ZPORR. *Biuletyn EBIB*, 9(70), <http://ebib.oss.wroc.pl/2005/70/czyzak.php>
- [5] Nicolaus Copernicus University in Torun continues traditions of polish university in Vilnius, existed since 1579. After 1945 it was removed to todays polish territory, as well as a lot of polish citizens.
- [6] BEDNAREK-MICHALSKA, B. (2006). Kujawsko-Pomorska Biblioteka Cyfrowa – pragmatyka tworzenia biblioteki cyfrowej. *Biuletyn EBIB*, 7(77), <http://www.ebib.info/2006/77/michalska.php>
- [7] NUKAT – National Universal Central Catalogue of scientific libraries in Poland (<http://www.nukat.edu.pl>).
- [8] WOJCIECHOWSKI, J. (2006). Dostęp półotwarty. *Forum Akademickie* no. 11, pp. 24-26.
- [9] ICM is well known in Poland because of central subscription of abroad scientific journals for polish libraries.

Use of Open Access Electronic Journals by Chinese Scholars, and an Initiative to Facilitate Access to Chinese Journals

Ruoxi Li¹; Fytton Rowland²; Zichuan Xiong²; Junping Zhao⁴

¹Chongqing Normal University, Shapingba District, Chongqing Municipality, 400047, Republic of China
e-mail: cc86@163.com

^{2,3}Department of Information Science, Loughborough University, Leicestershire LE11 3TA, United Kingdom
e-mail: ²J.F.Rowland@lboro.ac.uk; ³shawnzec@gmail.com

⁴Tsinghua University, Beijing, People's Republic of China
zhaojunping@tsinghua.edu.cn

Abstract

Surveys were carried out with two groups of Chinese scholars – one group working in China, and a second group working in the UK. The objective was to investigate usage of Chinese-language scholarly journals and the potential for them to use an Open Access business model. The results were compared with those published by the CIBER group at university College London, whose sample of scholars was international in scope. The overseas Chinese group made very little use of journals published in China, and one of the reasons for this was the difficulty of accessing the electronic versions of these journals from the West. We therefore proposed the construction of an English-language website to provide access in the first instance to the full texts of journals published by the members of the Society of China University Journals in the Natural Sciences (SCUJNS), and we created a pilot version of this website.

Keywords: open access; Chinese journals; overseas Chinese scholars

1 Introduction

This paper reports work undertaken while R.L. and J.Z. were Visiting Scholars at Loughborough University in 2006. In their home universities in China they are the editors of scholarly journals published by these universities, and as such they are active members of the Society of China University Journals in the Natural Sciences (SCUJNS), a collective organization representing such university-published journals in China. Many Chinese journals are published directly by universities in this way, but a lot of them are little-known in the West. Not all of them have English-language abstracts or metadata.

Recently there has been an upsurge of interest in the West in the academic journals published in China, in recognition of the amount of research being conducted in that country and reported only in its own literature. As a result a number of papers have appeared in western journals describing the scholarly publishing scene in China [1-3]. Chinese publishers in their turn have shown interest in making their journals accessible to westerners through the English language, with one of the first such initiatives coming from J.Z.'s home university, Tsinghua [4]. Most Chinese journals are not Open Access at present but the subscription fees, tailored as they are to both Chinese cost levels and Chinese affordability, seem very moderate to Westerners.

As part of a programme of research investigating possible futures for Chinese-language scholarly journals published by universities in China, surveys were carried out of two groups of Chinese academics, one group based in China and the other group made up of expatriate Chinese scholars now working in the United Kingdom. The purpose of the surveys was to ascertain the knowledge of, and attitude to, Open Access (OA) journals among these groups. Differences between the home-based and overseas Chinese scholars, and between these groups and the general international group of scholars studied by the CIBER group at University College London [5] were also of interest.

Conversations with the overseas Chinese scholars showed that they made very little use of the Chinese-language literature published in China, either as authors or as readers, even if they had used it when previously resident in China. It appeared that this lack of use was in part a consequence of the Chinese journals being difficult to access from the West, even though many of them are available in electronic versions. We therefore carried out the pilot phase of a proposed operational website that would provide straightforward access from the West in the English language to Chinese electronic journals.

2 Surveys of Chinese Scholars

Method

Chinese scholars who had published papers in seven university journals published in Beijing, Xi'an and Chongqing were sent a questionnaire. About 3000 e-mail invitations to participate were sent to authors in China, and about 1000 paper questionnaires were also distributed. Over 500 responses were received, but exclusion of incomplete questionnaires from the survey reduced the final number analysed to 376, a response rate of 9.4%. In the UK, members of the academic or research staff at Loughborough, Nottingham and Sheffield Universities who had Chinese family names and personal names were approached individually and asked to take part. The majority of them were born in the People's Republic of China, all could read Chinese, and most were now permanently resident in the West; 50 responses from these overseas Chinese scholars were received. The results from the China-based group and the overseas group were compared with each other, and both were also compared with those from the international group of authors surveyed by Rowlands *et al.* [5], the CIBER group, whose questionnaire we used. We are grateful to Dr Ian Rowlands for permission to use their questionnaire, and for helpful discussions.

Results

Of the group resident in China, computer scientists (24.5%) and engineers (28.5%) predominated, but many other disciplines were also represented. Their average age was 31.76 years. In all, 75% of them worked in Universities, and fewer than 1% in business or government. Engineers, mathematicians and computer scientists also dominated the UK-based group.

More than one-third of both groups said they knew 'nothing at all' about Open Access (OA), though more of the China-based group (29%) than of the UK-based group (16%) claimed to know at least 'quite a lot' about OA. Using a chi-squared test, the differences between the China-resident group and the overseas Chinese group were significant at the $p < 0.05$ level, and those between the China-resident group and CIBER's international group were significant at the $p < 0.005$ level. The difference between the overseas Chinese and the CIBER respondents was not significant, however, possibly reflecting the more international orientation of the UK-based group compared with those who remain in China. The younger authors were more ignorant of OA, in contrast to the CIBER group's results, which found older scholars less knowledgeable, and this difference was significant at the $p < 0.005$ level.

About three-quarters of the UK-based Chinese group associated the term 'Open Access' very strongly with 'free to access', a similar proportion to CIBER's international group, whereas only 45% of the China-based scholars thought that this was the defining characteristic of OA. Of CIBER's respondents, 47% did not associate the term 'OA' with 'author pays', but only 23% of our China group and 24% of our overseas group did not associate 'OA' with author pays. This difference was significant at the $p < 0.001$ level. Fewer of the China-based authors than of CIBER's sample had ever published in an OA journal: 15.7% of our China sample claimed to have done so versus 25.7% of the international group surveyed by the CIBER team, the difference being significant at the $p < 0.001$ level.

Fewer of them have self-archived their papers or put them on to an institutional repository: 17.5% of our China-based scholars and 14% of our UK-based ones had, versus 32% of CIBER's group. The difference between our two groups on this issue was not significant, but the differences between each of them and the CIBER group were significant ($p < 0.001$ for the China group and $p < 0.0001$ for the overseas group). The important advantages of self-archiving were seen by our respondents to be wider communication of results (38% said this was 'very important'), speed of dissemination (46% 'very important'), and increased impact (41% 'very important').

Responses to questions measuring attitudes of the respondents towards a possible OA-oriented future scholarly-communication system showed that the Chinese scholars were generally more positive in attitude towards OA journals than were CIBER's international sample, with younger Chinese respondents more optimistic about the likely effects of OA than older ones. However, our respondents differed from CIBER's on a number of points. Only 20.3% of our China group, but 78% of CIBER's international group, thought that printed scholarly journals would disappear altogether. This perhaps reflects the lesser progress towards electronic publication that has been made in China so far. Perhaps connected is another difference: 27% of our China respondents but 55% of CIBER's thought that rejection rates would fall. (High rejection rates, in some disciplines at least, can reflect unaffordable printing costs, and purely electronic journals do not suffer from cost constraints in the same way as printed ones.) It may be, though, that the apparently large differences between the international CIBER group and

out respondents may in part be explained by the much lower average age of our group, given the observed lower level of knowledge of OA among the younger age groups.

When asked whether scholarly publishing in China should become wholly OA, fewer than 30% of scholars in China agreed but almost 60% of the UK-based Chinese respondents agreed, and this difference was significant at the $p < 0.001$ level. The reason for this difference is not clear, given that most of our UK-based group had started their research careers in China, but it may be that those still based in China are aware of the potential financial difficulty in maintaining an OA publishing operation, while those who have moved to the West are aware of the high subscription prices of western journals and regard the cost levels in China as sufficiently modest to make OA a feasible business model. Over 70% of both groups thought that at least a partial conversion to OA should occur in China. It was notable that those who publish frequently were less likely to favour an all-OA future (18% agreeing) than those who publish less (30% agreeing), a significant result at the $p < 0.05$ level; however, there was no relationship here with the respondents' age. It may be that frequently publishing authors are more senior in the research profession, and as such are generally more aware of the cost structures of scholarly publication. Unsurprisingly, those who claimed to know nothing about OA were likely to make a neutral response towards a possible all-OA future, neither agreeing nor disagreeing with the proposal.

Further questions investigated financial issues. We first asked how the respondents' research had been funded. Fewer of the Chinese authors than of CIBER's international group had external grant or contract funding for their work: 40% of the international group said that the research underpinning all of their articles was funded, whereas under 30% of the China authors could say that, while only 16% of international authors said that none of their work was funded whereas over 20% of the Chinese group said this. Again, though, this may in part reflect the different age profiles of the two groups as well as their nationality.

Even so, 85% of the China-based group had paid page charges to a Chinese-language journal and 35% of the UK-based Chinese group had done so. This hints at the possibility that in some cases they had paid page charges out of their own pockets. In contrast, only 6% of the overseas group had ever paid page charges to an English-language journal, versus 38% of CIBER's international group who had paid for publication in a western-language journal at some time. These results seemed to indicate willingness on the part of Chinese authors to contemplate paying for publication in journals in their native language, whereas the major international journals published in English were perhaps perceived to be commercial successes and not in need of this financial support. Those authors who claimed to know a lot about OA were more likely to have paid page charges (92%) than those who knew nothing about OA (78%), a result significant at the $p < 0.025$ level, and in agreement with the CIBER group's findings.

One important point was the amount of these payments; the median amount paid per article by Chinese authors was about 600 yuan, equivalent to about 40 pounds sterling, 60€ or US\$75, a much smaller figure than is charged by western OA journals currently. The median amount paid by the overseas Chinese authors was lower, but it may simply be that their payments occurred longer ago, before they left for the West. Those who had published more frequently tended to have paid larger amounts than those who published less, perhaps explained by the more prolific authors having more research funding. As might be expected, there was a loose relationship between the amount people had paid in the past for page charges and the amount they said they might be willing to pay to OA journals in the future; the median amount they were willing to pay was well under 500 yuan. However, there was a big difference between the China-based and the UK-based group here; only 15% of the China-based group were totally unwilling to pay anything for publication in Chinese journals, whereas 70% of the UK-based group were, a result significant at the $p < 0.0001$ level. It is difficult to account for this large difference; perhaps those who have moved to the West have become used to the more commercially-based scholarly publishing system of North America and Western Europe, and do not see why they should pay for publication. Nor is it easy to square these views with the fairly positive attitude of these respondents towards OA. Willingness to pay for publication in English-language (or other non-Chinese) journals was marginally lower for the China-based group but marginally higher for the UK-based group.

When asked who should pay for the costs of publishing scholarly journals, 65% of respondents said that the scholar's department or faculty, or the research funder, should cover all or most of these costs. Those in biomedicine were the most likely to take this view, and those in engineering the least likely, though it was the view of over half the respondents in all disciplines. Over 80% said that neither authors nor readers should have to pay out of their personal pockets. Over 30% thought that all or most of the costs should be covered by central government – perhaps not a surprising view in China – but around 25% felt that commercial sponsors should make a big contribution.

Interviews

Many of the UK-based Chinese scholars in the survey were also interviewed in the Chinese language by R.L., mostly face-to-face but in some cases by telephone. The full results of these interviews are not published here, but one important finding from them was that, having moved to the UK, they had ceased to make much use of the Chinese-language literature either as authors or as readers. They had adopted the publishing and reading habits of their Western-born colleagues, and concentrated on the major journals published in Western Europe or North America. One reason given by them for this was the fact that they regarded the mainstream Western journals as better than the Chinese ones, a result which accords with observations made in countries as varied as New Zealand and Malaysia [6, 7]. But it is also true that they said that they could not access the Chinese-published journals from their UK universities, since their university libraries did not subscribe to the Chinese journals or to the Chinese aggregation services that link to them, such as Wan Fang [8], China Academic Journals [9] and Chinese Scientific Journals Full-Text Database [10].

Discussion

Chinese science is known to be of high quality, with many scholars born and educated in China now working in major universities in the West. Their early work, and that of others who have stayed in China, is largely reported in the Chinese-language literature published within China. This literature is little-used by scholars in other countries, largely because of the language barrier, but evidence from this survey seems to suggest that it is difficult to gain access to these journals from the West, even when they are in principle accessible in electronic form. Their subscription prices are low by Western standards, but Western universities do not in general subscribe to them so they are inaccessible to research workers in the West, whether Chinese-speaking or not.

Overall, the survey seemed to suggest that knowledge of the OA principle was incomplete among these scholars, even though almost one-third of the China-based group claimed to know quite a lot about OA.

This low level of understanding was reflected in some contradictory results: for example, the overseas Chinese group was more favourable to the idea of an all-OA future for Chinese journals than the China-based group, but less willing than the China-based group to pay publication charges! While few of them supported OA for English-language journals, the respondents were generally sympathetic to the idea that journals based in China might turn to OA, provided that individuals did not have to pay publication charges out of their own pocket. This perhaps reflected a degree of realism about the financial prospects for scholarly journals published in languages other than English. Given the relatively low costs of publishing in China, reflected in the fairly low page charges that some authors had paid in the past, it might indeed be possible for Chinese journals to be published electronically at the expense of research funders and authors' institutions. Chinese publishers – many of the universities, such as the member institutions of SCUJNS – do not publish their journals on a commercial basis or seek to make large surpluses from them, but they do have costs, modest though they may be, to cover.

3 Planned Website for Access to Chinese Electronic Journals from the West

Introduction

Interviews with the overseas Chinese scholars in the survey showed that they made very little use of the Chinese-language literature published in China, either as authors or as readers, even if they had used it when previously resident in China. It appeared that this lack of use was in part a consequence of the Chinese journals being difficult to access from the West, even though many of them are available in electronic versions. There are secondary databases based in China [8-10], but these are available on a subscription basis and few Western university libraries subscribe to them. Thus the full texts of the journals are inaccessible from outside China, even though many of them are in principle available free of charge. In addition to encouraging use of this literature by overseas Chinese scholars, it is desirable to make it accessible to others in the West who cannot read Chinese. As editors of some of these journals, R.L. and J.Z. were also aware that they are little used by non-Chinese speakers, owing to a general lack of English-language web pages to access them, even though individual papers often have short English abstracts. They would prefer that their journals were better-known, and better-used in the West and seek to provide tools to access them better.

Proposal

This project therefore constituted the pilot phase of a proposed operational website that would provide straightforward access in the English language to Chinese electronic journals, especially those in membership of the Society of China University Journals in the Natural Sciences (SCUJNS), the organisation that might operate the website in the longer term. R.L. and J.Z. are both active members of SCUJNS in their capacity as editors of university journals at Chongqing Normal University and Tsinghua University respectively. It was hoped that on their return to China in late 2006 they would be able to obtain funding for the development of the pilot website

into an operational service. If such a service is provided, then it would be expected that visibility and impact of Chinese research work would be greatly improved outside China.

The full texts of papers in Chinese are held on the servers of their publishing organisations (mostly universities) in China. At present about one-third of the 700+ journals published by SCUJNS members are available in electronic form, but this proportion is expected to increase rapidly. The concept is that an English-language website will be created that will provide ready access to these journals, and this website in turn could be linked into sites such as the ALPSP Learned Journal Collection [11] that host many journals from not-for-profit organisations in the west.

The project was named EJUNIC (**E**lectronic **J**ournals of **U**niversities in **C**hina). The main aim of the pilot EJUNIC website design was to create a web interface for publishers to register their journals, and to facilitate overseas readers' access to these academic resources under an Open Access mechanism. The detailed objectives include:

- To create English and Chinese language versions of the website.
- To design a registration system for publishers to mount links to their journals.
- To design a login system for publishers to keep their journal updated.
- To display all the included journals in specific pages that provide title, link, introduction, and contact information.
- To provide a browsing function with an A-Z index of the journals included
- To provide both advanced and simple search functions that allow users to search journals by title, author, keywords, and abstract.
- To establish a harvesting program that automatically collects available articles from the included journals
- To provide long (informative) English-language abstracts of the papers in the journals, and English metadata

Technical aspects of the pilot phase were implemented by Z.X., who was a postgraduate student of Electronic Publishing at the Department of Information Science at Loughborough University at the time. Functions that will be provided in the full implementation are:

Register function – allows publishers to mount links to their journals included in EJUNIC.

Browser function – allows readers to browse all the included journals in our database. An A-Z index is provided.

Search function – allows readers to search a particular term (title of journal or ISSN) to locate the needed material. In an operational version, we also expect to implement further function that will provide readers a powerful text level search engine to locate items within the included journals by more choices of search terms, such as title of article, author, abstract, etc. The function is expected to adopt a harvesting program based on an Open Access standard.

The registration process was fully implemented in the pilot phase, and falls into several stages:

ISSN verification: EJUNIC assigns the journal's ISSN as a unique username. A database was designed to hold ISSNs and the titles of their corresponding journals. Publishers will be asked to provide their ISSN to make sure they are suggesting a valid journal.

Submitting basic information: Once the ISSN is verified, the publisher then moves to the stage of basic information input. The required information in this stage includes a valid URL, contact person, e-mail address, and telephone number. A database (basic_info) holds this information.

Journal verification: EJUNIC is a website based on an Open Access (OA) protocol. It requires that all the journals included are free to access. Therefore, a verification procedure is carried out in order to confirm that the journal is OA.. If the journal is proved to be OA, a password will be sent to its publisher, which ends the whole registration process. Otherwise, we will send an email to inform the publisher of possible reasons of failure.

Technical details

An English-language home page was designed, and links to a number of journals, mostly those for which J.Z. is responsible at Tsinghua University, were implemented. PHP technology was used to bridge the website and a database created by MySQL. PHP is a human-readable language which is easy to write, edit, understand and expand. It is currently recognised as one of the most popular languages that used in network programming. On the other hand, MySQL is widely used in small sized databases, as it provides good flexibility in terms of database management. The designing environment was simulated by an application called APM Express 5.0

(APMEX). APMEX is a software package that associates PHP, MySQL and a database management tool, PhpSQLAdmin. It significantly eases the complex process of PHP and MySQL configuration. The coding process was completed by Macromedia Dreamweaver 8. Cascading Style Sheet technology was adapted to improve the appearance of the website.

Future activities

A further activity proposed for the operational phase is the provision of informative English-language abstracts of the papers in the participating journals. We recognise that this will entail negotiation with the various publishers, and locating people in China with good English-language skills to provide the abstracts.

Initially the journals included will be those published by SCUJNS members, but in a later phase it is hoped that other Chinese journals published by not-for-profit organisations in China will also be brought into the ambit of EJUNIC.

4 Conclusions

Earlier work, such as that of the CIBER group [5], has shown that despite the large amount of debate that takes place today about Open Access, scholars in general are still not well-informed about the OA concept. This work shows that Chinese scholars are, if anything, even less well-informed than their Western counterparts, and even scholars from China who have moved to the West permanently to work are significantly less likely than CIBER's respondents to have published in an OA journal. Despite their relative ignorance of this topic, both the Chinese groups of respondents seemed moderately favourable towards OA for Chinese-language journals, and this perhaps reflects a realism about the modest commercial prospects for these titles compared with English-language journals published by major for-profit or not-for-profit organisations in Western Europe or North America. Certainly, those resident in China seemed willing to pay author charges, and many had done so, even in some cases out of their personal pockets. The advantages of OA that they detected were similar to those mentioned by other groups – wider communication of their work and consequent higher visibility and impact for it. They also mentioned faster publication, which might be seen as an advantage of electronic publication per se, rather than OA. As the general cost level of publishing is lower in China, and many journals are already published by universities directly, it may be easier to progress to an OA publishing model ('the Gold Route to OA') in China than in Western countries. In contrast, fewer of our respondents – in both the China-based group and the UK-based group – than of CIBER's group had posted copies of their articles on institutional repositories, and it may be that the 'Green Route to OA' has made less progress in China than in the West.

Although much has been done to make Chinese research better known in the West [1, 2, 4, 12], and indeed major Western information providers, such as Swets with their 'Gateway to China' service [13], and NetLibrary working in partnership with a Taiwan company [14], are now providing information services, it is clear that scholars working in the West are largely not using the Chinese literature, or publishing in it, even when they themselves are originally from China and can read the Chinese language. It seems that the fact that these services are commercial and charge subscription fees, even where the original journals may be free to access electronically, leads to their being available in the West only to the largest and best-resourced institutions. We therefore proposed that a website be provided through SCUJNS to provide direct, easy and free access to university-published Chinese journals from outside China. This would be an English-language website but would link to the full texts in Chinese held on publishers' own servers, and these would be enhanced by long, informative English-language abstracts. Both the website with its metadata, and the full texts, would be available free of charge. We produced a prototype of this website which was prepared quite quickly using readily available open-source software, and which functioned satisfactorily. It is hoped that it might be developed into a full operational version, with other journals from Chinese publishers other than universities being added to it in a later phase, and that it might be linked to the ALPSP Learned Journals Collection [11] to further enhance its visibility.

References

- [1] STANLEY, A.; YAN S (2007) China Opening Up: Chinese University Journals and Research – Today and Tomorrow, *Learned Publishing* **20**(1), 43-50.
- [2] JIA, X. (2006) The Past, Present and Future of Scientific and Technical Journals of China, *Learned Publishing*, **19**(2), 133-141.

- [3] WANG, S.; WELDON, P. R. (2006) Chinese Academic Journals: Quality, Issues and Solutions, *Learned Publishing*, **19**(2), 97-106.
- [4] ZHANG, L.; YAO, Y.; ZHANG, F. ; DU, WENTAO (2006) The First Comprehensive Chinese University Journal Published in English, - the Tsing Hua Journal, *Learned Publishing* **19**(3), 204-208.
- [5] ROWLANDS, I.; NICHOLAS, D.; HUNTINGDON, P. (2004) Researchers' Attitudes towards New Journal Publishing Models, *Learned Publishing*, **17**(4), 261-274.
- [6] ROWLAND, F. (2005) Scholarly Publishing in New Zealand, *Learned Publishing*, **18**(4), 300-310.
- [7] ZAKARIA, J.; ROWLAND, F. (2006) What are the Prospects for Online Scholarly Publishing in Malaysia? The Cultural Constraint? In Proceedings of the ELPUB 2006 Conference, Bansko, Bulgaria, pp. 229-236
- [8] Wan Fang Data, English version at <http://www.wanfangdata.com> (accessed 12 March 2006)
- [9] China Academic Journals, English version at <http://www.thtf.com.cn/www/web/en/index.asp> (accessed 12 March 2006)
- [10] Chinese Scientific Journals Full-Text Database, English version at <http://dx3.cqvip.com/en/index.htm> (accessed 12 March 2006)
- [11] ALPSP Learned Journal Collection, in partnership with Swets, <http://www.alpsp-collection.org/> (accessed 22 December 2006)
- [12] WANG, J. (2006) Major Chinese Full-Text Electronic Information Resources for Researchers and Scholars, *Serials Review*, **32**(3), 164-171.
- [13] Swets Gateway to China, <http://www.swets.com/web/show/id=84103/langid=42> (accessed 11 April 2007)
- [14] NetLibrary, Chinese Language e-Resources, <http://library.netlibrary.com/ChineseLanguage.aspx> (accessed 11 April 2007)

Managing Expectations for Open Access in Greece: Perceptions from the Publishers and Academic Libraries

Banou G. Christina¹; Kostagiolas A. Petros²

¹ Department of Archive & Library Science, Ionian University
Plateia Eleftherias, Corfu 49100, Greece
e-mail: cbanou@ionio.gr

² Department of Archive & Library Science, Ionian University
Plateia Eleftherias, Corfu 49100, Greece
e-mail: pkostagiolas@ionio.gr

Abstract

In Greece, there seems to be a growing level of awareness regarding open access among scholars, faculty staff and information professionals. Indeed, consensus regarding the necessity of open access initiatives in Greece is gradually established. The present of open access in other European settings may however be revealing the expected, though distinct, future of open access in Greece. This work focuses upon some current aspects for open access and attempts to investigate them for the Greek setting. The investigation includes five (5) important aspects of open access, i.e. a) ETDs management from the academic libraries, b) university repositories development, c) regulation of digital and/or printed scientific material quality requirements, d) cooperation and competition between libraries and academic publishers, e) understanding the role of scientific work dissemination in developing future professionals and scholars. The paper initially provides an outline for the Greek publishing industry, focusing on STM publishers and on the way they take advantage of the changes mainly in editorial and marketing terms, in a hybrid technological era. The Greek publishing industry may be representative of other national small publishing markets. Further, an empirical research is providing in order to illuminate open access from two different points of view: that of STM publishers and that of academic libraries' directors in Greece. The empirical investigation took place in February and March of 2007 and is based on seventeen experts' perceptions. The methods employed are outlined and include the development of the questionnaire for semi-structured interviews. Finally, the unexpected agreement from both publishers and academic libraries' directors regarding open access development is discussed and some specific for Greece conclusions are drawn.

Keywords: open access; publishing industry; academic libraries management; scholarly communication; Greece

1 Introduction: Setting the Scene

In Greece, there seems to be a growing level of awareness regarding open access among scholars, faculty staff and information professionals. Indeed, consensus regarding the necessity of open access initiatives in Greece is gradually established. Academic libraries, and for that matter university authorities in Greece, realize nowadays that cannot purchase access to all the scientific information their researchers expect, although some association agreements and research programs, involving publishers, the National Documentation Centre and other institutions, have assisted [1]. Publishers in Greece have been considering the development of distinct pricing models for making available books, monographs and scientific articles, so they cause further pressure on institutional budgets. Overall, the current scholarly communication model, that the academia employs, seems to currently disconfirm expectations of scholars and of the Greek scholar community as a whole. A novel information and research strategy for academic libraries is required involving scholar publications which are digital, online, free of charge, and free of most copyright and licensing, compatible with printed edition.

The present of open access in other European settings may be revealing the expected, though distinct, future in Greece. Open access for Greece may constitute a greater challenge due to the language barriers, which may form two (2) distinct types of scientific publication: the ones written in Greek and the ones that are not. The expectations of the scholar community relate to the Electronic Thesis and Dissertations (ETDs) management, the development of university repositories and, finally, the regulation of the digital versus printed material quality requirements. The academic libraries in Greece ought to enhance their role within the current scholarly communication setting. On the other hand, it may safely be assumed that the scholarly community in Greece has nothing or little to gain from any publishing pricing model. Scholars mostly wish to publish their work in high

impact journals (or as monographs) either open access or not, realizing gradually that openness has a positive effect on impact factors based on citations and/or other traditional and frequently used measures of research impact.

This work focuses upon some current aspects for open access and attempts to investigate them for the Greek setting. The investigation includes A. an outline of the Greek publishing industry, focusing on STM publishers and on the way they take advantage of the changing environment mainly in editorial and marketing terms, in a hybrid era. In that framework, novel publishing strategies and policies are developed. From that point of view, the Greek publishing industry may be representative of other national small publishing markets. B. the expectations of the directors of the academic libraries about open access. It is interesting and fascinating to illuminate open access from two different points of view: that of STM publishers and that of academic libraries' directors. C. an empirical investigation based on seventeen experts, that took place in February and March 2007 through a semi-structured questionnaire, in order to portray the specific aspects in Greece. The methods employed are outlined. D. the unexpected agreement from both publishers and academic libraries' directors regarding open access development is discussed and some specific for Greece conclusions are drawn.

1.1 The Present Scenario in the Greek Publishing Industry

The publishing industry in Greece may be characterized from the absence of conglomerates and of large foreign publishing houses. Furthermore, it is rather traditional (family owned and managed enterprises) in comparison to international markets. Specifically, the Greek Scientific-Technical-Medical (STM) publishing production [3] represents about one third (35,1%) of the annual book production. There is a steady increase, during the last five years, in the annual production of new scientific titles, that may express a turning point of the Greek STM publishing industry. In 2004, 2692 new scientific titles were published (out of 7.888 new titles of the total annual book production), while, in 1999, 2410 were the new ones [4]. It is significant that small and medium-sized STM, on the one hand, and general publishing houses, that also produce scientific publications on the other, manage to develop the profile of the Greek publishing industry; at the same time, they play a central role in scholarly communication in Greece, collaborating with the academic community.

In regard to the Greek publishing industry, focusing on the STM publishing, a number of specific features can be synopsized as follows [5]. The Greek publishing market has a rather small audience of about 14 million people, due to the Greek language, which is unique among the European languages. Hence, that market has not yet been in the focus of international conglomerates or large publishing groups; on the other hand, the Greek publishing industry is deeply influenced by them in certain terms, such as in patterns of promotion and of management practices. One of the main features of the Greek publishing industry is that almost all the Greek publishing houses are companies, owned and run by members of a family, who continue and try to innovate, respecting the tradition. Concerning the STM publishers, it is characteristic that many of the publishing houses' names consist of the surname of the founder, who, in some cases, still runs the company: Sakkoulas, Papatotiriou, Siokis, Paschalidis, Ziti, etc.

The remarkable increase, during the last fifteen years, in the total annual production of new titles (from less than 3,000 in the beginning of the 90ies to 7,888 titles in 2004) reveals the prosperity and the turning point of the Greek publishing industry [6]. More specifically, concerning the STM publications, there is a steady increase, as it was referred above. Large publishing houses in Greece produce more than 80 titles per year, medium produce 10-80 titles, while the small ones publish less than ten (10) titles annually [7]. Only seventeen are large publishers; five of them are STM. The majority of STM publishers are medium. Generally, Greek STM publishers are competitive to academic presses and to organizations with publishing activities such as scientific institutions, museums, chambers and others.

Concerning the scientific publications, the publishing houses can be categorized as follows: a. strictly STM publishers, b. general publishers, which include in their catalogues scientific texts, c. organizations and institutions that publish or order and encourage publications. The well known and specialised in scholar work publishing houses are the market leaders. They can, through their policy, influence the structure of the book market in Greece. It is significant that the majority of the new titles published annually are works of Greek academics and scientists. Out of 2692 new titles published in 2004, only 770 were translations, something that demonstrates that the Greek STM market develops an interest and a taste in the national scientific production, for which there is need to be promoted. The academic community determines to a great extent, by its special needs and expectations and through collaborations, the STM production. Furthermore, the academic community is an important knowledge producer and co-operates with the publishing houses, not only by its works, but also by editing and being responsible of series. With its high expectations and with a very good judgement, this

particular audience is usually a force for change and for innovation. STM publishers are intended to a specialized and, therefore, specific, rather homogenous and steady reading audience. This target group is easily accessible, through economic ways of promotion. On the other hand, general publishing houses promote and advertise, sometimes even the scientific titles, in such a way so as to attract the majority of readers.

In the last five years, there are “new” needs in the traditional and rapidly changing Greek publishing industry. The profile of the Greek publisher was until recently formed in terms of a family company for a small market, something that gradually is changing; although family enterprises, the STM publishing houses are conscious of the competition, of the need for innovation and of the new role that they are called to play. New policies and strategies, competitive values, and new information technologies demonstrate the need of special studies and life long learning in the STM Greek publishing industry. In an era, in which information is the main product, the challenges for the STM publishers are great, and they should be innovative always bearing in mind that “publishing companies are content-acquiring and risk-taking organizations oriented towards the production of a particular kind of cultural commodity” [8].

1.2 Selected Research Issues for Open Access in Greece: Competition and Co-Operation

This work provides some empirical results which are based on the perceptions of the Greek STM publishers and directors of academic libraries for five (5) important aspects of open access, i.e. a) ETDs management from the academic libraries, b) university repositories development, c) regulation of digital and/or printed scientific material quality requirements, d) cooperation and competition between libraries and academic publishers, e) understanding the role of scientific work dissemination in developing future professionals and scholars. However, the Greek publishing market and academia encompass some unique features, including the language and other exclusive features. These may further establish an additional set of research objectives that are referred here as the “language” and the “digital product” issues.

The STM publishers in Greece are focusing on the conventional printed material production, and hence they do not consider openness as a “real” threat until now. Furthermore, academic libraries have recently invested a significant amount of money and effort in personnel development and information technology, have initiated additional university policies and/or mandates regarding the Electronic Thesis and Dissertations (ETDs) management, the development of university repositories and finally the regulation of the digital versus printed material quality requirements. The academic libraries in Greece are “better” organizations than Greek publishing enterprises in “gathering” scholarly work and have better access to digital distribution channels [9], as they have taken advantage of the new digital technologies [10]. The amount of accessible digital information is increasing due to the advent of information technologies, the Internet and other international networks. However, the diversity in the content of the material and in the languages text are written in, is also significantly increasing. The Greek language forms a barrier that has to be crossed. Large international publishing enterprises are intensive on digital STM publications produced mostly in other languages than the Greek language, while national publishers are resolving quality issues for Greek scientific work and thus they are developing highly prestigious conventional STM products based on the scientific work.

A realization that arose with this research is that “open access” is redefining information “transaction” for the Greek scientific and technical STM market. However, if “transactions” are costless, the most important issue is that the rights of the various parties involved should be well defined and the results of legal actions easy to forecast [11]. It would therefore seem desirable to further focus on the stakeholder’s views, taking their perceptions into account when making economic decisions and/or investigating market characteristics. In that respect, “openness” is less of a substitute product within the competitive forces in the publishing industry in Greece, and it may be rather be treated as a factor that characterizes the nature of competition (and co-operation) within this particular industry. Greek publishers may co-operate and compete with the academic libraries, in regard to the language of the text and/or the nature (printed and digital) of the produced STM material.

2 The Empirical Research Conducted

The objective of the empirical research is to identify and then investigate the perceptions of Greek STM publishers and the perceptions of the directors of academic libraries, through five (5) important aspects of open access: a) ETDs management from the academic libraries, b) university repositories development, c) regulation of digital and/or printed scientific material quality requirements, d) cooperation and competition between libraries and academic publishers, e) understanding the role of scientific work dissemination in developing future

professionals and scholars. In the following paragraphs the methods employed as well as the results of the empirical study are presented.

2.1 Methods Employed

The empirical research was based on semi-structured interviews that were directed to large publishers and directors of central academic libraries in Greece. Therefore, the interviewees were organization representatives (experts) selected on the following criteria: a) have an in depth experience managing STM material, b) academic libraries with central administration as well as publishers with more than 30 titles annual production were selected. The academic publishing houses are not included in the research, mainly, due to their small annual production (less than 30 titles). On the other hand, within the group of STM experts a number of participants represent publishing organizations as the Technical Chamber of Greece.

A questionnaire was designed for the survey and a pre-test was conducted. The questionnaire includes both closed and open-ended questions. For the closed-ended questions a five-point Likert scale was used, ranging from 1="strongly agree" up to 5="strongly disagree", in order to determine the extent of agreement and/or disagreement of the participants to specific statements regarding open access in Greece. The questioner included seventeen (17) questions as follows:

- Research Questions 1 to 12: aimed at investigating characteristics of the experts participated in the survey (organization title, address, telephone, email, full name, position in the organization, education, years of employment) and the organization they represent (number of employees, nature of services and/or products –conventional, digital, hybrid–, presence of electronic thesis and dissertation system, presence of repository);
- Research Questions 13 and 14: aimed at assessing the degree of agreement of the participants to the following statements, "the information provided to the scientific community in Greece is sufficient" & "clear need for the open access development in Greece";
- Research Question 15 for "ETDs management from the academic libraries": aimed at assessing the degree of agreement of the participants to the following statements, "support other activities", "produce economic value", "support scientific work", "support co-operation" & "is a reason for competition between academic libraries and publishers";
- Research Question 16 for "university repository development": aimed at assessing the degree of agreement of the participants to the following statements, "support other activities", "produce economic value", "support scientific work", "support co-operation" & "is a reason for competition between academic libraries and publishers";
- Research Question 17 for "law and regulation of digital and/or printed scientific material": aimed at assessing the degree of agreement of the participants to the following statements, "sufficient for ETDs", "sufficient for repositories", "sufficient for copyright issues", "support co-operation" & "is a reason for competition between academic libraries and publishers".

The research questions 13 through 17 were accompanied with open questions, so that the participants could state in free narrative form their additional comments. The survey was not aiming in providing results that could be generalized, although, in the lines of Behrakis [12], the information recorded and the qualitative analysis provided, produce indicative results based on expert's opinion. Furthermore, it was found that the distance in the scale between 1="strongly agree" and 2="agree" was small, as well as, that the distance between 4="disagree" and 5="strongly disagree". Hence, the initial form of the scale was reduced from five to three [agreement (+), rather (=), disagreement (-)]. The percentages were computed and graphs were produced, while for the open-ended questions content analysis was employed in order to determine frequency of statements of interest [13].

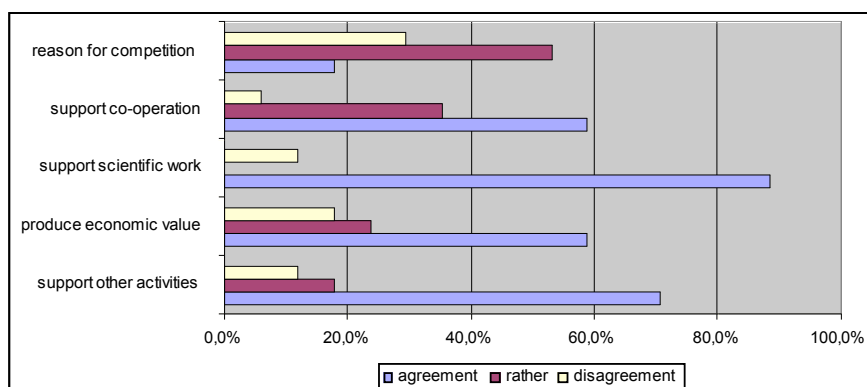


Figure 1: Experts points of view regarding the "ETD's management" aspect of Open Access in Greece

2.2 Results of the Empirical Research

The empirical research was conducted through structured interviews based on a specially designed semi-structured questionnaire, within February and March of 2007 (a pilot study took place in January 2007). Among the twenty two (22) academic libraries and the twenty (20) large Greek STM publishers conducted, representative of ten (10) academic libraries and seven (7) STM publishers agreed to participate in this survey (i.e. the 45.5% of the libraries conducted and 35.0% of the STM publishers). The overall profile of the group of experts was considered sufficient for our purposes in consideration of the following, a) the group of experts is representative including participants from the academia and the STM publishing industry, b) the group of experts consist of highly skilled and educated information professionals (the participants are all university graduates mainly from Library Science and/or Information Science departments, six of them hold a postgraduate diploma, while four of them hold a Ph.D.), c) sufficient working experience (fifteen out of the seventeen were employed in the organizations they represented for more than five years), and d) specific organizational features (six organizations employed more than twenty five professionals, while thirteen organizations provide services of both conventional and digital form). The results for each of the close-ended questions are provided bellow (Table 1 and Figures 1 to 3), while the analysis of the open-ended questions follows.

Research Question	Issues addressed in the survey for Open Access in Greece	Agreement	Rather	Disagreement
13	"the information provided to the scientific community in Greece is sufficient"	53.0%	23.5%	23.5%
14	"there is a clear need for the open access in Greece"	88.2%	0.0%	11.8%
15	"ETDs management from the academic libraries"			
	"support other activities"	70.6%	17.6%	11.8%
	"produce economic value"	58.8%	23.5%	17.6%
	"support scientific work"	88.2%	0.0%	11.8%
	"support co-operation"	58.8%	35.3%	5.9%
	"is a reason for competition between academic libraries and publishers"	17.6%	52.9%	29.4%
16	"university repository development"			
	"support other activities"	64.7%	11.8%	23.5%
	"produce economic value"	52.9%	23.5%	23.5%
	"support scientific work"	82.4%	0.0%	17.6%
	"support co-operation"	64.7%	11.8%	23.5%
	"is a reason for competition between academic libraries and publishers"	17.6%	47.1%	35.3%
17	"law and regulation of digital and/or printed scientific material"*			
	"sufficient for ETDs"	17.6%	11.8%	41.2%
	"sufficient for repositories"	11.8%	11.8%	41.2%
	"sufficient for copyright issues"	17.6%	11.8%	35.3%
	"support co-operation"	11.8%	17.6%	35.3%
	"is a reason for competition between academic libraries and publishers"	11.8%	11.8%	41.2%

Table 1: Results of the survey for aspect in open access in Greece (*6 of the responders could produce a reliable judgment for the statements)

In Table 1, the overall results of the survey conducted are exhibited. In the first column of Table 1, the statements under investigation are provided; while the columns that follow provide the percentages reflecting the perceptions of the participants (STM publishers and directors of the academic libraries). More than half of the experts (53%) state that the information provided to the scientific community in Greece is adequate (23.5% “rather” adequate and another 23.5% “disagree”), whereas the majority of the participants (88.0%) agree that open access initiatives are indeed required (Table 1, research question 14). The majority of the experts agree that the ETDs management from the academic libraries may “support other activities” and “support scientific work”, while they do not think that would be a reason for competition increase between publishers and academic libraries (Figure 1).

In Table 1 (research question 16 “university repository development”) and in Figure 2, the results indicate a significant agreement among the participants in the survey, stating that in Greece the development of university repositories may “support co-operation” (64.7%), “support the scientific work” (82.4%) and “support other activities” (64.7%) within the university communities. Once again the experts did not indicate that this is a reason for increasing the competition between libraries and Greek publishers. In Table 1 (research question 17) and in Figure 3, the experts point of view is presented, regarding “law and regulation of digital and/or printed scientific material” aspect of open Access in Greece. For the research question 17, six of the participants did not express any opinion within the survey, stating that they need further information. However, the majority of the participants in the survey stated that “law and regulations” in the present form in Greece do not support “scientific work” and “co-operation”, and they are not sufficient for the development of “university repositories”, and “ETD’s management” within the Greek universities.

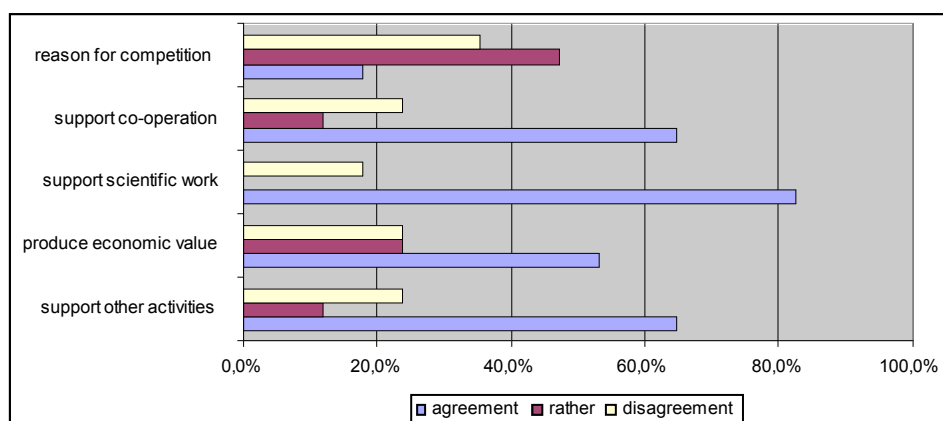


Figure 2: Experts points of view regarding the “repository development” aspect of Open Access in Greece

The experts expressed their views, for the distinct aspects of open access examined in the survey, in a free narrative manner through a set of five (5) open-ended questions attended each research issue (from 13 to 17). The participants frequently commented on the following: **a.** improvements have been achieved in Greece in terms of quality and quantity of the scientific information services over the last 6 to 8 years, **b.** open access in a cost – benefit analysis framework seems to prevail, reducing management cost (although significant investments ought to be made for the management of openness) and gradually reduce the “need” for costly agreements with international publishers, **c.** open access development may support improvements in Greek scholar production in both the Greek and other European languages, through better information provision and “free of subscription charges” high quality scientific communication, **d.** apprehension and support within a centrally regulated legal and investment framework for open access in Greece is required, while Greek scholars should support openness within the university communities. The participants representing academic libraries in the survey stated that open access may be used as a vehicle for further improvements (e.g. in grey bibliography management, technical reports distribution, maintenance cost reduction etc.) and of course professional development. Furthermore, the library directors stated that education and empowerment of the library staff may be a key factor for making openness a reality and that the aspects of open access studied here, may support co-operative initiatives within the academic community. It is worth mentioning that both Greek academic library directors and publishers which participated in this survey, support open access development, and although sceptical, mainly the publishers, they state that with proper regulation, e.g. embargo on the time of scholar material provision, public and private sectors can find common grounds for co-operation.

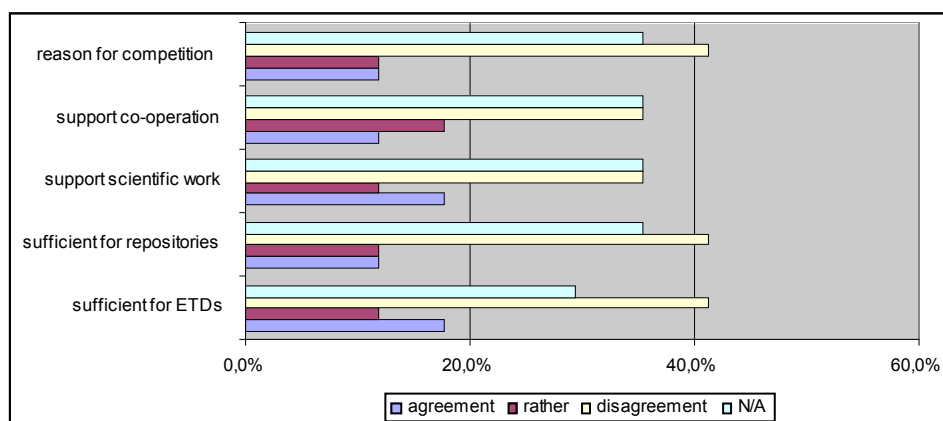


Figure 3: Experts points of view regarding the “law and regulation” aspect of Open Access in Greece

3 Discussion

The directors of the academic libraries pointed out that an important factor for ETDs management and repository development in the universities are the education and the empowerment of the library staff. Furthermore, they thought that centralized university initiatives may provide a way forward regarding the academic quality requirements for either printed or electronic scientific publications. Repositories, in particular may provide articles, pre-prints, post-prints, dissertations, PhDs, monographs or chapters from monographs, proceedings, rare material, to the scientific community, while ETD management undoubtedly enhance the quality of the scholar work production.

Some academic libraries in Greece have developed or they are developing ETD's and/or repositories: In the University of Macedonia in Thessalonica a repository has been developed and a mandate is regulating thesis and dissertations standards for digital submission. Similar efforts are undergoing in the University of Athens, for example at the Faculty of Law, where an ETDs management initiative is under development for the postgraduate courses of Civil Law. Similarly, in most Greek universities (e.g. University of Piraeus, Ionian University etc.) such initiatives are under consideration. Furthermore, a number of Greek universities have an academic press (e.g. University of Crete, University of Macedonia, etc.), that it is closely related to the academic community and the library. The academic libraries directors finally expressed their belief that the unrestricted reading, downloading, copying, sharing, storing linking and accessing scientific work, which open access incorporate, will lead to more efficient scientists and professionals in the future. However, a number of economical, political, legal and technical aspects ought to be addressed as soon as possible for satisfying user needs [14]. Universities ought to provide to all members access to the information; especially, to the information that is produced inside the university campus.

The international publishers usually resist to Open Access due to a. economic investments, b. issues of co-competition (cooperation and competition) with academic libraries c. legal and copyright issues. The STM publishers in Greece publish for a homogenous small market. They are engaged in printed scientific books and conference proceedings, and they collaborate with scholars and universities. Thus, the challenge for STM publishers in Greece is to realize and take advantage of open access rapid development, shaping a novel marketing strategy for the printed and/or the digital publications. It has been indicated that the author would prefer to communicate directly with the reader-user, without the intervention of the publisher. The scholarly community in Greece must become conscious that open access journals have high impact factors and thus, to gain their trust [15]. On the other hand, it has been pointed out, regarding publisher – author relation, that diachronically the publisher has been the one that takes the risk [16]. Innovation has been proved to be highly valued within the publishing industry [17].

4 Conclusions

A significant reason for the frustration surrounding openness in Greece is not so much the concept itself nor the economic issues involved, but the way openness is presented by the decision makers in scholarly communication. Open Access in Greece is possible, authorized, and beneficial for all those involved. The results drawn from this work may enlighten the main trends and the current issues for Open Access in Greece. It is clear that, if a rather unexpected co-operation -fruitful for open access development- among STM publishers, the academic libraries and the scholarly community in Greece is established, this should be on the ground of well defined roles and legal arrangements. A very interesting question to be further investigated in the future arises: Is Openness a threat for small publishing markets, such as the Greek STM market? Publishers in Greece “do not perceive openness as a threat” but on the contrary, if the roles are clearly defined and their investments secured, they might build upon openness innovative novel strategies and co-operative policies with the academic libraries in Greece.

Like in every innovation, such as the concept of open access, dilemmas and/or threats arise, because we don't really know what benefits we are getting from that. Hence, we should establish the research needed to find out the technological, political consequences of open access. Furthermore, empirical results can shed light on whether publishers and academic library directors may find in Greece a common ground for improving the quality of scholarly communication in Greece and in Europe.

Acknowledgements

The authors would like to express their gratitude to all those participated in either the pilot study or the survey presented in this work, i.e. Alivanoglou L. (director of Kleidarithmos publications), Anthi-Kalofolia E. (Ionian University, director of central library), Brindesi H. (Efgenideio foundation, library), Dimakopoulos E. (TEE library), Eleftheroudaki S. (Eleftheroudakis chain-bookstore and publications, publisher), Fragkou A. (University of Macedonia, director of central library), Georgakis K. (A.T.E.I. of Patras, director of central library), Kalambaliki M. (National Technical University of Athens, director of central library), Kamariotakis E. (TEE library, technical publications), Katsirikou A. (University of Piraeus, director of documentation centre), Kokkinos D. (National Technical University of Athens, central library), Moniou D. (Kyriakides publications), Papageorgiou V. (Metaixmio publications, publisher), Pesmatzoglou E. (University library of Attikon Hospital, director of library), Synelli K. (University of Patras, director of central library), Toraki K. (TEE, director), Vardakosta I. (Harokopio University, director of central library), Zachos G. (University of Ioannina, director of central library).

Notes and References

- [1] KOROBILI-XANTINIDOU, S.; MORELELI-CACOURIS, M.; TILIKIDOU, I. “Concepts, reality and suggestions about Greek library management education”, *New Library World*, 104 (1189), 2003, pp. 203-217
- [2] KOSTAGIOLAS, P. .A “Information services for supporting quality management in Healthcare”, *Journal on Information Technology in Healthcare*, 4 (3), 2006, pp. 137-146.
- [3] CLARK, G. *Inside Book Publishing*, third edition, London and New York: Routledge, 2006, pp. 42-56.
- [4] National Book Centre of Greece. *The Book Production in Greece. 2004*, Athens: National Book Centre of Greece, May 2006, pp. 2-3.
- [5] BANOU, C. (2005/2006), “Money and Taste: New roles for the Greek publishers in a changing era. A case-study of small publishing markets”, *The International Journal of the Book*, vol. 3, number 2, pp. 39-46.
- [6] *ibid.*
- [7] National Book Centre of Greece. *The Book Production in Greece. 2004*, Athens: National Book Centre of Greece, May 2006, p. 39.
- [8] THOMPSON, J. B. *Books in the Digital Age. The Transformation of Academic Publishing in Britain and the United States*, Cambridge: Polity Press, 2005, p. 15.

- [9] LYTRAS, M.; SICILIA, M.; DAVIES, J.; KASHYAP, V. “Digital libraries in the knowledge era. Knowledge management and semantic web technologies”, *Library Management*, vol. 26 (4/5), 2005, pp. 170-175.
- [10] DOBREVA, M. “IT applications of the medieval Slavonic written cultural heritage”, *Proceedings. 1st International Conference on Typography and Visual Communication. History, Theory, Education*, Thessaloniki: University of Macedonia Press, 2004, pp. 161-170.
- [11] DEAKIN, S.; MICHIE, J. “The Theory and Practice of Contracting”, in *Contracts, Co-operation, and Competition. Studies in Economics, Management and Law*, edited by Simon Deakin and Jonathan Michie, Oxford University Press: Oxford, 1997, pp. 1-39.
- [12] BEHRAKIS, T. *Multidimensional Data Analysis: methods and practices*. Athens: Livanis, 1999 [in Greek].
- [13] HARWOOD, T.G.; GARRY, T. “An Overview of Content Analysis. *The Marketing Review*. Vol.3, 2003, pp. 479-498.
- [14] XIN L., “Library as incubating space for innovations: practices, trends and skill sets”, *Library Management*, 27 (6/7), 2006, p. 37-378.
- [15] ANTELMAN, K. Do Open-Access Articles Have a Greater Research Impact?”, *College and Research Libraries*, 65.5 (Sep. 2005), p. 372-282.
- [16] SCHIFFRIN, A. *The Business of Books. How International Conglomerates Took Over Publishing and Changed the Way we Read*, London - New York: Verso, 2001.
- [17] STEVENSON, I. “The liveliest of corpses”: trends and challenges for the future in the book publishing industry, *Aslib Proceedings*, 52 (4), April 2000, pp. 133-137.

Towards a Semantic Turn in Rich-Media Analysis

Tobias Bürger; Georg Güntner

Salzburg Research Forschungsgesellschaft m.b.H.
A-5020 Salzburg, Jakob-Haringer-Strasse 5/III, Austria
e-mail: {tobias.buerger, georg.guentner}@salzburgresearch.at

Abstract

Typical application scenarios in the area of rich-media management, such as the continuous digitisation of the media production processes, the search and retrieval tasks in a growing amount of information stored in professional and semi-professional audio-visual archives, as well as the availability of easy-to-use hard- and software tools for the production of rich-media material in the consumer area, lead to an increasing demand for a meaning-based management of digital audio-visual assets. This “semantic turn” in rich-media analysis requires a semantic enrichment of content along the digital content life cycle and value chain: The semantic enrichment of content can be achieved manually (which is expensive) or automatically (which is error-prone). In particular, automatic semantic enrichment must be aware of the gap between meaning that is directly retrievable from the content and meaning that can be inferred within a given interpretative context. Each solution has its benefits and drawbacks. Our paper discusses the relevance of semantic analysis of rich-media in certain application scenarios, compares two possible approaches, a semi-automatic and an automatic approach, and presents a case study for an automatic solution. Following the observations of the case study, we come up with recommendations for the improvement of the semantic enrichment by an manual annotation step.

Keywords: semantic web; multimedia content management; semantic indexing

1 Introduction

As a motivation for the application of semantic technologies in the area of rich-media analysis we want to highlight the following application scenarios: Firstly, the continued digitisation of the media production process at professional content providers and content distributors (e.g. broadcasters, telecommunication companies) not only leads to an exponential growth of highly unstructured digital material, but also to an increased demand for a reliable classification of audio-visual material along all stages of the digital content value chain. Secondly, in the consumer area and the semi-professional area (e.g. corporate media archives, or small and medium sized audio-visual archives) the easy-to-use production tools lead to an unmanageable amount of audio-visual material, that rather later than sooner has to be managed and indexed in some way, whereas the meaning of the digital essences often is locked in the raw content. Thirdly, even if basic metadata is available to describe the content and its meaning, user-centred applications are increasingly demanding the utilisation of the benefits of the true semantic search approach, i.e. inference and reasoning, narrowing down and widening the search by using some kind of formal knowledge representation.

To exemplify the above scenarios, imagine a user who wants to find out recordings of performances of sacred music by Wolfgang Amadeus Mozart in and around the city of Salzburg during the Salzburg Festival 2005. This query is full of hidden semantics (see also figure 1 for a schematic presentation of this query and the associated knowledge model): Location based semantics (what does “in and around the city of Salzburg” mean?), time based semantics (when was “Salzburg Festival 2005?”), factual semantics: e.g. which works are considered to be “sacred works”; which musical forms are known to be “sacred works” in general (e.g. a choral, a mass); which particular works by Wolfgang Amadeus Mozart are sacred works?

The importance of the knowledge related with the query is, that nothing of it has to be encoded in the media essences or their description: all this knowledge can be modelled, described and used without any particular relation with the digital essences.

Our research group is currently investigating different approaches and methods for the combination of media content with semantic annotations and for the usage of pre-existing knowledge (i.e. the “context”) for inferring further knowledge about the content automatically. We have faced the question of annotation from low-level content analysis recently, in the national research project Smart Content Factory (SCF) and we are going to

address the issue of merging new content with existing knowledge in the IST project LIVE (see section 2 and 3). In another project, Smart Content Factory, we tried to automatically derive the semantics of TV news clips in order to make them browse- and searchable. To do this, we combined information extracted from raw multimedia content with domain knowledge about multimedia data.

In LIVE which deals with broadcasting of media events by integrating different videos streams with background information about these media events, we investigate how domain knowledge and background information can be efficiently combined to deduce further knowledge from broadcast live video-streams.

As we have experienced in the Smart Content Factory [1], semantic descriptions of content can enhance fast and easy navigation through audio-visual repositories. Semantics - i.e. the interpretation of the content - is important to make content machine-processable and to enable the definition of tasks in workflow-environments for knowledge workers in the content industries. Some of the recent research projects in the area of semantic (or symbolic) video annotation try to derive the semantics from the low level features of the audiovisual material or from other available basic metadata, e.g. by audio-classification or classification of camera movement. Some of the projects aim at highly automated indexing using the results of automatic speech recognition however error-prone they may be. Most of these approaches are - as also pointed out in [2] - not capable to derive the semantics of multimedia content because in many cases the results of the analysis cannot be related to the media context [3]. For humans the construction of meaning is an act of interpretation that has much more to do with pre-existing knowledge (the “context”) than with the recognition of low-level-features of the content. This situation is commonly referred to as the “semantic gap” [4].

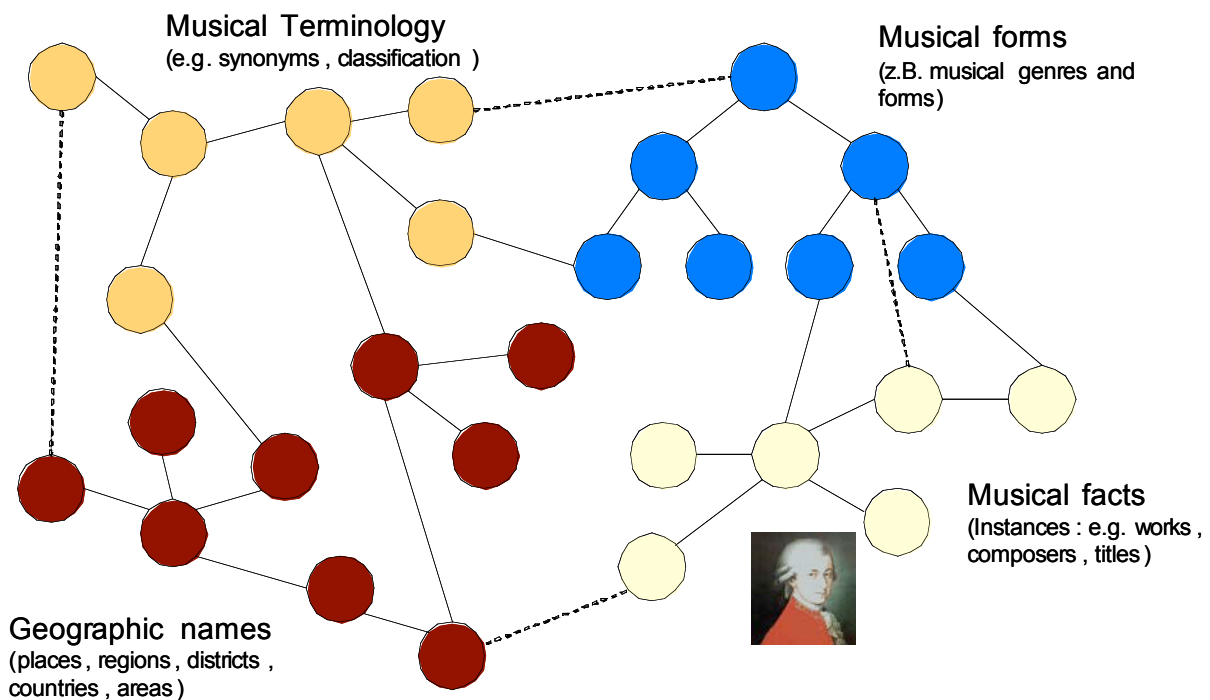


Figure 1: The knowledge-base for the semantic query about Mozart’s sacred works

Two solution paths have emerged for this problem: The first one is to provide rich annotations created by humans as training data for the system to learn features of videos for future automatic content-based analysis. The second approach does not rely on training, but purely on analysis of the raw multimedia content. The training approach is not well suited for scenarios in which a great amount of content has to be annotated before any training and automation can be done or in which the application domain is very broad. The second approach usually only works well in settings where the relevant concepts can easily be recognized. However, most content based services demand richer semantics. As pointed out in section 4, popular examples on the Web show that there are currently many service-based platforms that make use of their users' knowledge to understand the meaning of multimedia content.

In our paper we concentrate on different approaches to close this semantic gap and provide insight into two solution paths, one automatic and one semi-automatic and demonstrate a case study on a prototypical solution for the “semantic augmentation” in the area of audio-visual archives.

2 Methodology: Automatic Vs. Semi-Automatic Semantic Rich-Media Analysis

In general, metadata generation systems can be classified in manual-, semi-automatic- and automatic annotation tools: The aim of automatic and semi-automatic tools for the analysis of rich-media content is to extract as much useful information from the raw media file as possible. Manual annotation tools aim to provide support for users to add metadata by hand.

Currently many systems try to expose the semantics of multimedia data by adding metadata to it. However, most of them do not derive these annotations just from the low-level features detected in the raw media data, but instead for example either analyze the different modalities of a video, analyse the usage context of the media or rely on human annotation/interpretation to derive higher-level semantic features from multimedia data. In this section we want to introduce different approaches that we applied in two research projects making use of semantic technologies for rich-media analysis.

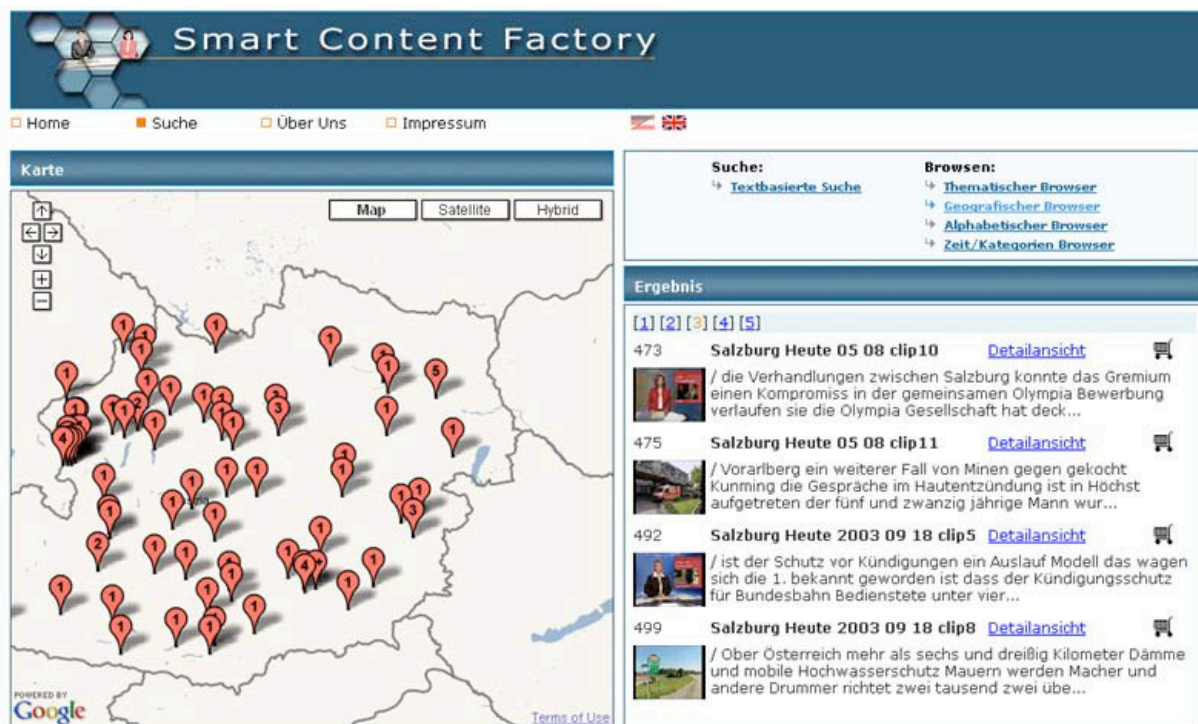


Figure 2: Location based navigation in the Smart Content Factory

Firstly, a project called “Smart Content Factory” (2003-2006) was driven by the idea to develop a system infrastructure for the knowledge-based search, retrieval and navigation in audio-visual archives of news clips. The approach was highly relying on an automatic feature extraction, mainly the speech-to-text transcription during the first phase. In a late phase this automatically extracted features were supported by additional reliable sources available in the digital production process at the Austrian Broadcasting Corporation (ORF). Section 3.1 describes the results and major findings of the selected approach with respect to the semantic indexing. Secondly, an ongoing integrated project called “LIVE staging of media events” (started in 2006) is driven by an approach to combine the methods of automatic and semi-automatic detection, extraction and annotation of content with a knowledge-base under the control of a semantic based media framework. Moreover the framework propagates knowledge and contextual information to a recommender system which thus to some degree becomes aware of the meaning of the media. Section 3.2 describes the current state of this approach.

3 Results: Towards A Semi-Automatic Reliable Semantic Analysis Framework

This section describes the results of two research projects with respect to their selected approach for the semantic analysis.

3.1 Smart Content Factory – An Automatic Approach to Semantic Rich-media Analysis

In a research project called Smart Content Factory [5], we developed a prototype of a system infrastructure for the automatic knowledge-based refinement of audio-visual content repositories based on state-of-the-art digital asset management systems. In the project an automatic approach was used which was based on a two step indexing pipeline:

In the first step a primary index is created by methods provided by state-of-the-art media analysing tools (i.e. Virage Video Logger™ and the Smart Encoding™ process). Video clips are passed to the Factory in MPEG-1 format. A polling mechanism informs the indexing and contextualisation components about newly available video clips and triggers the indexing process. The primary indexing results in the creation of key frames, the automatic detection of scenes, the transformation of speech to text, the recognition of speakers, etc.

Subsequently a semantic indexing process is started which is based upon the results of the previous content based indexing and relies primarily on the speech to text transformation (i.e. “audiologging”, for which different solutions have been tested in the course of the project). The dependency on the results of the audiologging is a weak point in the concept of the Smart Content Factory in so far as the subsequent semantic indexing builds on the results of a per se error-prone automatic extraction process. The semantic indexing is using the Lucene indexing framework [6] and the ontologies and thesauri forming the knowledge base of the Smart Content Factory which are accessible via “pluggable” RDF knowledge components described in the previous section. By applying and using knowledge models during the semantic indexing we create a set of “smart indices”, allowing search, retrieval and reasoning along various dimensions of the information space.

The screenshot displays the Smart Content Factory web application. At the top, there is a navigation menu with links for Home, Suche, Über Uns, and Impressum. Below the menu is a search bar with the text 'Begriff: Salzburg' and an 'anwenden' button. The main content area is divided into two columns. The left column shows a hierarchical tree view of categories, including 'IPTC Beschlagwortungs-System', 'Kriege & Konflikte (1)', 'Sport (6)', 'Soziales (8)', 'Wissenschaft (3)', 'Politik (9)', 'Parteien (2)', 'Salzburg Heute 10 02 clip1 (512)', 'Salzburg Heute 10 02 clip1 (583)', 'Verteidigungspolitik (2)', 'Wahlen (2)', 'EU (1)', 'Regierung (1)', 'Salzburg Heute 2003 10 30 clip7 (395)', 'Freizeit & Lebensstil (5)', 'Vermischtes (3)', 'Kinder (1)', 'Salzburg Heute 09 18 clip3 (484)', 'Veranstaltungen (1)', 'Leute (1)', 'Arbeit (1)', 'Gesundheit (2)', 'Umwelt (6)', 'Bildung (2)', and 'Schulen (1)'. The right column displays search options: 'Suche: Textbasierte Suche', 'Browsen: Thematischer Browser', 'Geografischer Browser', 'Alphabetischer Browser', and 'Zeit / Kategorien Browser'. Below this is a 'Clip Kurzinformation' section for the selected clip 'Salzburg Heute 09 18 clip3'. It shows the title, relevance (38%), a thumbnail image, ID (484), location (Salzburg), duration (PT134120S), categories (Lärm Soziales Kinder), and a snippet of the content: '/ geraten versuchen Salzburg hatte Reiter / gegen den Kindergarten in Salzburg / der indische Hilfe in Salzburg eigen heute /] das lag NDR Ehepaar ein Salzburg eigen...'. At the bottom right, there is a contact link: 'Kontakt - Infos: georg.quentner@newmedialab.at'.

Figure 3: Category based navigation in the Smart Content Factory

One of the key issues of our semantic indexing framework of the Smart Content Factory was the use of an extensible set of formal knowledge models (accessible via the Jena RDF framework [7]):

(1) The first knowledge model ('locations') contains a thesaurus of geographic names. The thesaurus extends the properties of gazetteers by modelling the hierarchical relations between the geographic locations (e.g. 'village' is-part-of 'political district'). The geoname thesaurus is based on data structured according to the ADL Feature Type Thesaurus [8] and is represented in RDF, the gazetteer is structured according to the ADL Gazetteer Content Standard [9]. This model allows specialisation and generalisation of search queries by location concepts. The recognition of location names is further supported by an engine for the recognition of named entities. In the Smart Content Factory we used LingPipe [10] to resolve ambiguities:

Scanning texts for occurrences of common known Austrian place names like 'Wien' (Vienna) or 'Salzburg' is rather easy. But it is a lot harder to semantically distinguish an appearance of the name of the Austrian village 'Haus' from the German word for house (the building). Due to the lack of a bigger training set, the location name recognition's precision/recall is not very high, but it serves as a starting point for distinguishing false positives from true positives. A simple tf/idf-based ranking is used to determine the 'most significant' location which a video is related to. Tf/Idf means 'term frequency-inverted document frequency' [11].

(2) To identify thematic categories we used the IPTC thesaurus (International Press and Telecommunications Council, [12]) which defines a hierarchical structure of thematic news categories (e.g. sports, policy, economy). For our purpose the IPTC thesaurus is represented in RDF according to SKOS Core 1.0, an RDF schema defined by W3C for the description of thesauri and similar types of knowledge models [13]. Similar to the location name identification process, this process identifies terms from a controlled vocabulary provided by the IPTC thesaurus.

(3) A web-based synonym service for (German) words was integrated into the semantic indexing process. The service was created and is maintained by the University of Leipzig ('Deutscher Wortschatz' [14]). The Web service interface is based on the SOAP protocol. By means of this service the index is enhanced with synonyms of non-stop words. All models are either integrated via a Web service interface or stored in a database and are accessed via the Jena RDF framework [7], which also provides a powerful inferencing mechanism (e.g. traversing of hierarchical relations).

In the Smart Content Factory, one of the objectives for the introduction of the knowledge-based index was the enhancement of search and retrieval and navigation support [1].

The main benefits of this approach were the little need for human intervention in the process as the annotations were totally generated automatically. Another benefit was that the approach was extensible as other thesauri/knowledge models could be easily plugged in to recognise for example events or dates. One drawback however was the amount of false positives locations that were recognised, which was mainly due to the bad results of the text to speech engine. Another drawback was the amount of time needed to index a video, which doesn't allow real-time indexing of the video.

Figure 2 shows the application of the location based semantics for the map based search and navigation in the audio-visual archive. Figure 3 exemplifies the category based navigation, using the IPTC thematic thesaurus which is represented in RDF as described above. Both navigation paradigms were highly appreciated by the test user group during a user evaluation in 2006, whereas other forms, e.g. the hyperbolic tree navigation paradigm were ranked low in the users' interest profile.

3.2 LIVE – Real-Time Semi-Automatic Annotation of Videos with the Intelligent Media Framework

The integrated project "LIVE Staging of Media Events" (LIVE; FP6-27312, [15]) aims at the creation of novel intelligent content production methods and tools for interactive digital broadcasters to stage live media events in the area of sports, such as the 2008 Olympic Games. In the terminology of the project, "staging live media events" is a notion for the creation of a non-linear multi-stream video show in real-time, which changes due to the interests of the consumer (end user). From a technical viewpoint, this requires a transformation of raw audiovisual content into "Intelligent Media Assets". LIVE will develop a knowledge kit and a toolkit for an intelligent live content production process including dynamic human annotation and automated real-time annotation. As a consequence novel iTV video formats for live events will evolve.

In the LIVE project we applied the lessons learnt from the automatic approach of the Smart Content Factory to overcome the weaknesses of automatic metadata extraction and extended the system architecture to meet the requirements of real-time semantic indexing. In the first phase of the project we started to design an Intelligent Media Framework that is taking into account the requirements of real-time video indexing to combine several automatic and manual annotation steps. The Intelligent Media Framework thereby integrates the following components of the LIVE production support system:

- (1) The Intelligent Media Asset Information System (IM AIS) providing access to services for the storage of media, knowledge models and metadata relevant for the live staging process and providing services for the creation and management and delivery of intelligent media assets. This will be the central component of the Intelligent Media Framework and will semantically enrich incoming metadata streams;
- (2) The Recommender System, giving content recommendations to the user based on the user's personal profile and on previous user feedback;
- (3) The Metadata Generation System, dealing with the detection, extraction and annotation of knowledge from audiovisual material;
- (4) The Video Conducting System, dealing with the real-time staging of a live event.

The components of the indexing pipeline in LIVE are shown in figure 4: the Automatic Analysis Application, the Human Annotation Tool and the Intelligent Media Framework. The role of the Intelligent media Framework is to accept and handle partial information about particular media items, to add semantic information to the items and to infer and attach contextual knowledge to the items that is probably related to the event that is staged. It furthermore provides knowledge services that offer controlled vocabularies related to the current context of a stream to guarantee the unambiguousness of the terms used.

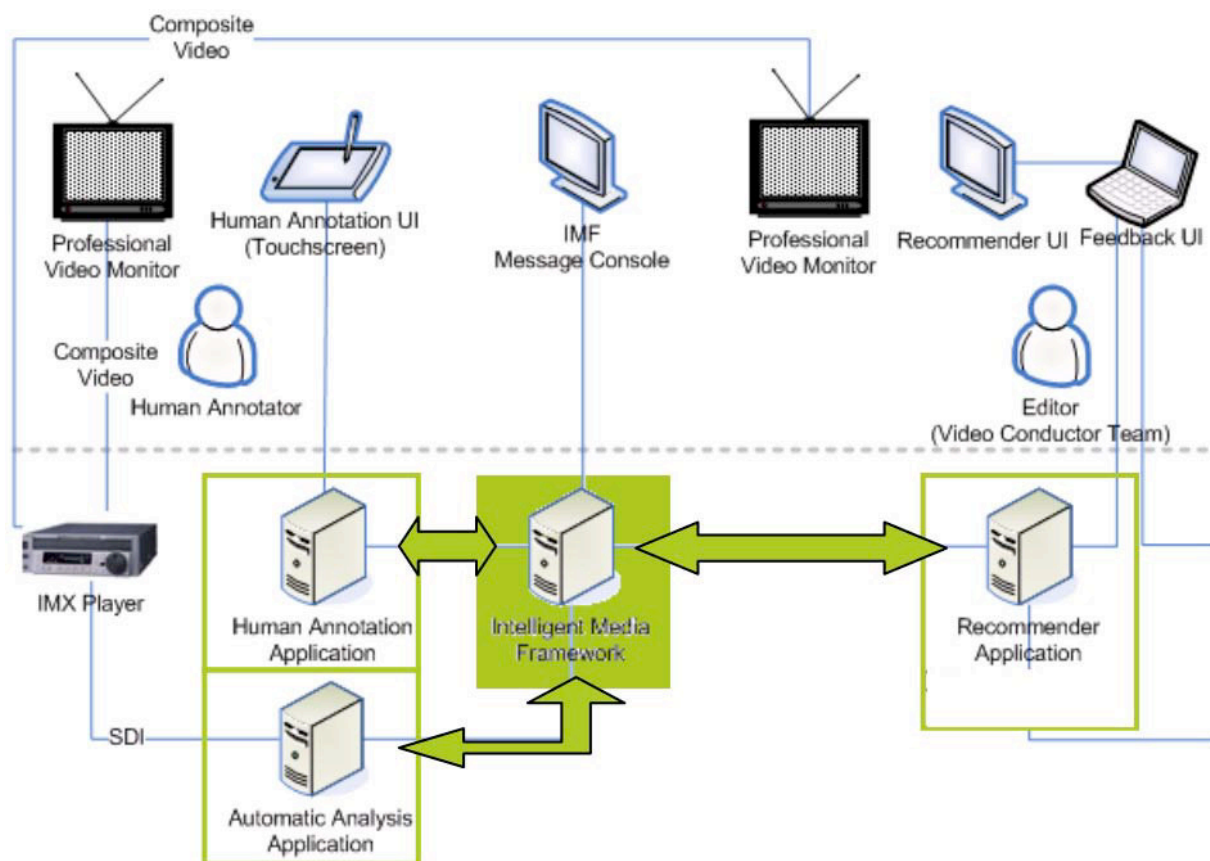


Figure 4: Category based navigation in the Smart Content Factory

The semantic enrichment process in LIVE is twofold: The Automatic Analysis Application detects close-ups, shots, faces, camera-motion, colour schemes, scenes and artists. This information is enriched in (1) a manual step done by the human annotator through the human annotation tool (2) in the Intelligent Media Framework that has knowledge about the context of the analysed media item. In (1) terms from the controlled vocabulary are assigned to the low-level information that was extracted in the basic analysis step. In (2) these terms are used to

attach more semantic information of the current action or event to the media items that is possibly inferred by the current event schedule or other particular information that was detected in the course of this event.

This semantic indexing process is more reliable than the approach from the Smart Content Factory, because it is neither based on error-prone text transcripts nor totally relies on automatic analysis tools. One key enabler of this semantic indexing step is the use of existing information systems at the broadcaster's side that have knowledge about the staged event, the participants and so on. The most important step in this process, however, is the human annotation that is later inferred by the Intelligent Media Framework in a reliable way. This allows us to act in real-time (with a maximum delay of approx. 20ms) and provides high-level metadata helping to bridge the gap between the raw audio-visual essences and their intended meaning.

Figure 4 shows parts of a prototypical demonstration setup shown during a LIVE review meeting in Vienna in March 2007. Both, a human annotation tool and the automatic annotation system, are using a semantic aware middleware, the Intelligent Media Framework, which provides context information and the controlled vocabulary for the annotation process and propagates the detected "meaning" to a recommender system for the professional user (editor, video conductor). The video conductor team decides which live streams and switching possibilities are offered to the consumers.

4 Discussion

In this section we list existing approaches that try to extract knowledge from rich-media items and try to relate these approaches to the LIVE approach. Indexing and metadata generation is a common task to media analysis systems and there are sound algorithms and methods in the area of computer vision, pattern recognition, natural language processing and signal processing that can be applied, most of them applicable for extracting low-level features from the essences. Currently many systems try to expose the semantics of multimedia data by adding annotations to it. However, the common trait of all the following examples is that they do not derive these annotations just from the low-level features detected in the raw media data, but instead for example either analyse the different modalities of a video, analyse the usage context of the media or rely on human annotation/interpretation to derive higher-level semantic features from multimedia data.

Recent research efforts try to combine automatically derived features like speech-to-text transcripts with background knowledge and related information on the Web: an application called Rich News [16] deals with automatic annotation and extraction of semantics from news videos: In a first step, the system extracts text from speech and then it tries to extract the most important topics from that. With these extracted topics, the system starts a Google search to find news stories on the web that cover the same topic(s). Google did exactly the same in a research project [17] to enhance American TV news with background information from the Internet: They extract important topics from the subtitle channel that is broadcasted with every news show and give consumers the possibility to get background information about these news items by displaying links to related web pages. MediaMill [18], a system that was developed at the University of Amsterdam, uses all modalities of a video to derive the semantics of it, which is especially important when the visual content is not reflected in the associated text like close captions or speech-to-text transcripts. Besides these examples from the research community there are also some major trends in industry: Some of the established industry-strength systems come from the classical document management area with its sophisticated full text indexing and information retrieval methods. The industrial players are now extending these methods to audiovisual content (e.g. Convera's RetrievalWare [19] or Autonomy's IDOL Server [20]). Other vendors come from the digital asset management sector and extend their systems to full text indexing and knowledge management methods (e.g. Virage's VS Archive [21]). These systems rely heavily on metadata and on full-text indexation which in turn, is based on speech-to-text extraction, but also make use of statistical methods to classify content according to taxonomies or thesauri. The leading search engines (e.g. Google or Yahoo) are currently extending their search features to audio-visual media, but they are still to a high degree relying on metadata and on text transcripts provided along with the media assets. Besides that, Google and Yahoo both released versions of their search engines to discover contents in videos: Google for example has been indexing news stories from an U.S. broadcaster. There the search index is based on the closed caption provided along with news clips. They also scan the Web for videos and images: In that case - as also described in [22] - additional metadata are generated from adjacent information on the website (e.g. text blocks or the video file name).

However, there are also other services dealing with content which are already very popular among a large user community: Two of the most popular content-based services are Flickr [23] and Last.FM [24], both of which get their users to classify (tag) images or music files within these systems. These systems do not classify the content according to predefined categories, but instead, taxonomies evolve from the users' tags (folksonomies). These

tags can be regarded as user-based knowledge about certain items that on these platforms, is used to filter or recommend content to other users. The main requirements in LIVE are (1) reliability of annotations as fast decisions have to be made based on them and (2) the real-time assignment of these annotations. The mentioned approaches are shortly discussed according to their usefulness in LIVE.

As mentioned above Web sites such as Flickr [23], or also Riya [25] recently began to apply algorithms for automatic extraction of content metadata (e.g. shapes, colour or texture features) and some of them already use high level pattern recognition technology such as face detection (e.g. Riya) However the metadata provided by them is not reliable enough. Other approaches that tend to provide reliable metadata information like MediaMill [18] or the commercial tools like Virage's VS Archive [21] perform the necessary analysis not fast and accurate enough. Manual annotation tools like Vannota [26], Advene [27] or M-Ontomat-Annotizer [28], to name just a few existing annotation tools, are too complex to be utilizable in the LIVE real-time situation. LIVE or the IMF also goes one step beyond the research as it tries to include contextual information in the analysis as much as possible, however at the moment solely in the LIVE domain.

5 Outlook and Conclusion

In the Smart Content Factory we experienced that it on the one hand is possible to automatically extract low-level features from video and augment them through the use of semantic technology but on the other hand this information is sometimes error-prone as it has to be based on incomplete or wrongly extracted features. In LIVE wrong or incomplete information could lead to wrong decisions in the live staging process which to some extent forces us to augment extracted information manually to reach a high degree of reliability. The use of controlled vocabularies and semantically enhanced metadata throughout the whole LIVE system introduces a common language that will lead to fast and reliable decisions during the staging process. The Intelligent Media Framework is responsible for the augmentation of the automatically extracted information that is additionally manually refined. Our next steps in the development of the IMF are the further addition of contextual information to the metadata sets of single media items, such as information related to the event or the athletes.

Acknowledgements

Smart Content Factory is one of the lead projects at Salzburg NewMediaLab (SNML [29]), the Austrian centre of excellence in the area of digital content engineering and new media technologies. Salzburg NewMediaLab is funded by the Austrian Federal Ministry of Economics and Labour and the Province of Salzburg. The project started in 2003 and ended in September 2006. It was coordinated by Salzburg Research. Further partners were ORF (the Austrian Broadcasting Corporation), X-Art ProDivision and Joanneum Research (all Austrian partners). LIVE (Live staging of media events, FP6-27312, [15]) is an integrated, multidisciplinary initiative that will contribute to the IST strategic objective "Semantic based knowledge and Content Systems" and "Exploring and bringing to maturity the intelligent vision". The project is funded by the European Commission within the 6th Framework Programme. The project started in 2006 with a duration of 45 months. It is coordinated by Fraunhofer IAIS. Further partners are Salzburg Research, the University of Ljubljana, the University of Bradford, the University of Applied Sciences in Cologne, the Academy of Media Arts Cologne, ORF, Pixelpark and ATOS Origin.

References

- [1] BÜRGER, T.; GAMS, E.; GÜNTNER, G.: *Smart Content Factory - Assisting Search for Digital Objects by Generic Linking Concepts to Multimedia Content*. In: Proceedings of the Sixteenth ACM Conference on Hypertext and Hypermedia (HT '05), 2005.
- [2] BLOEHDORN, S. et al.: *Semantic Annotation of Images and Videos for Multimedia Analysis*. In: Proceedings of the 2nd European Semantic Web Conference, ESWC 2005, Heraklion, Greece, May 2005.
- [3] BÜRGER, T.; WESTENTHALER, R.: *Mind the gap - requirements for the combination of content and knowledge*. In: Proceedings of the first international conference on Semantics And digital Media Technology (SAMT), December 6-8, 2006, Athens, Greece.
- [4] SMEULDERS, A. W. M. et al.: *Content-Based Image Retrieval at the End of the Early Years* In: IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 22 No. 12, December 2000.
- [5] Smart Content Factory project web-site: <http://scf.salzburgresearch.at/> – Last visited: 10.04.2007

- [6] The Apache Jakarta Project: Lucene. From: <http://jakarta.apache.org/lucene> - Last visited: 10.04.2007
- [7] Jena – A Semantic Web Framework for Java. From: <http://jena.sourceforge.net/> - Last visited: 10.04.2007
- [8] University of California, Santa Barbara: Alexandria Digital Library Feature Type Thesaurus, <http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver070302/> - Last visited: 12.10.2004
- [9] University of California, Santa Barbara: *Guide to the Alexandria Digital Library Gazetteer Content Standard*, <http://www.alexandria.ucsb.edu/gazetteer/ContentStandard/version3.2/GCS3.2-guide.htm> - Last visited: 12.10.2004
- [10] LingPipe: <http://www.alias-i.com> – Last visited: 10.04.2007
- [11] Term frequency-inverted document frequency (tf/idf): <http://en.wikipedia.org/wiki/Tfidf> – Last visited: 10.04.2007
- [12] International Press and Telecommunications Council. From: <http://www.iptc.org/> - Last visited: 10.04.2007
- [13] MILES, A. J. *SKOS Core - Guidelines for Migration* .<http://www.w3.org/2001/sw/Europe/reports/thes/1.0/migrate/> - Last visited: 01.04.2007]
- [14] University of Leipzig, Institute of Computer Sciences. Project “Deutscher Wortschatz“ (German dictionary). From: <http://wortschatz.informatik.uni-leipzig.de/> - Last visited: 28.04.2005
- [15] LIVE – Live staging of media events; project web-site: <http://www.ist-live.org> – Last visited: 10.04.2007
- [16] DOWMAN, M.; TABLIN, V.; URSU, C.; CUNNINGHAM, H.; POPOV, B.: *Semantically enhanced television news through web and video integration* In Proceedings of the Workshop on Multimedia and the Semantic Web at the European Semantic Web Conference (ESWC 2005), 2005.
- [17] HENZINGER, M.; CHANG, B.-W.; MILCH, B.; BRIN, S.: *Query-Free News Search*. In Proc. of the 12th World Wide Web Conference, pp. 1-10, 2003.
- [18] SNOEK, C. G. M. et al.: *MediaMill - Exploring News Video Archives based on Learned Semantics*. In: Proceedings of the ACM Multimedia Conference 2005, 2005
- [19] Convera - <http://www.convera.com/> - Last visited: 10.04.2007
- [20] Autonomy - IDOL Server: <http://www.autonomy.com/content/Products/IDOL/> – Last visited: 10.04.2007
- [21] Virage: <http://www.virage.com> – Last visited: 10.04.2007
- [22] FERGUS, R. ; PERONA, P.; ZISSERMAN, A.: *A Visual Category Filter for Google Images*. In: Proceedings of the 8th European Conference on Computer Vision, Prague, Czech Republic, 2004.
- [23] Flickr: <http://www.flickr.com> – Last visited: 10.04.2007
- [24] Last.FM: <http://www.last.fm> – Last visited: 10.04.2007
- [25] Riya: <http://www.riya.com/> – Last visited: 10.04.2007
- [26] Vannotea: <http://liris.cnrs.fr/advne/> – Last visited: 10.04.2007
- [27] Muvino: <http://vitooki.sourceforge.net/components/muvino/code/index.html> – Last visited: 10.04.2007
- [28] M-Ontomat-Annotizer : <http://www.acemedia.org/aceMedia/results/software/m-ontomat-annotizer.html> – Last visited: 10.04.2007
- [29] Salzburg NewMediaLab (SNML), “Kompetenzzentrum für Neue Medien” (competence centre for new media technologies and digital content engineering): <http://www.newmedialab.at/> – Last visited: 10.04.2007

Towards an Ontology of EIPub/SciX: A Proposal

Sely M S Costa¹; Claudio Gottschalg-Duque²

¹ University of Brasília, Department of Information Science
Campus Universitário Darcy Ribeiro, Brasília, DF, Brazil
e-mail: selmar@unb.br

² University of Brasília, Department of Information Science
Campus Universitário Darcy Ribeiro, Brasília, DF, Brazil
e-mail: klauss@unb.br

Abstract

A proposal is presented for a standard ontology language defined as EIPub/SciX Ontology, based on the content of a web digital library of conference proceedings. This content, i.e., EIPub/SciX documents, aims to provide access to papers presented at the total editions of the International Conference in Electronic Publishing (EIPub). After completing its 10th years in 2006, EIPub/SciX is now a comprehensive repository with over 400 papers. Previous work has been used as a basis to build up the ontology described here. It has been presented at Elpub2004 and it dealt with an Information Retrieval System using Computational Linguistics (SiRILiCo). EIPub/SciX ontology constitutes a lightweight ontology (classes and just some instances) and is the result of two basic procedures. The first one is a syntactic analysis carried out through the Syntactic Parser-VISL. This free tool, based on lingsoft's ENGCG parser, is made available through the Visual Interactive Syntactic Learning, a research and development project at the University of Southern Denmark, Institute of Language and Communication (ISK). The second one, carried out after that, is a semantic analysis (concept extraction) conducted through GeraOnto, an acronym that stands for “generating an ontology”, which extracts the concepts needed in order to build up the ontology. The program has been developed by Gottschalg-Duque, in 2005, in Brazil. The ensuing ontology is then edited via Protégé, a free, open source ontology editor. The motivation to carry out the work reported here came from problems faced during the preparation of a paper to Elpub2006, which aimed to present data about a number of aspects regarding the EIPub/SciX collection. While searching the collection, problems with the lack of standardization of authors and institutions names and the non-existence of any control of keywords had been identified. Such problems seem to be related to an apparent absence of “paper preparation” before entering into the SciX database. Lack of preparation, in turn, has brought about the desire of finding a solution, which is expected to support the work of those interested in searching the collection to retrieve information. EIPub/SciX ontology, therefore, is seen as that helping solution to support EIPub information retrieval.

Keywords: ontology; Elpub conferences; information retrieval

1 Introduction

Electronic publishing constitutes a hot topic of discussion within the academic environment, particularly in the study of scholarly communication. Such interest is due to the opportunities provided by the web and the Internet for a document to be available world wide in electronic format. As a free, democratic environment, the Internet provides a huge amount of information, which, on the other hand, presents a challenge for those who seek relevant hits. The digital content available today actually represents a great chaos to those interested in finding relevant information for research or in scrutinising a document collection for the same purpose.

There have been a variety of approaches to help with this matter, such as those that made possible to develop thesauri, ontologies, taxonomies, topic maps and other resources. They have been developed with the aim of facilitating the intellectual work. Therefore, a well-organised collection should be supported by one of them. In this context, an ontology can be considered one of the richest resources used for the automatic treatment of electronic documents, since it constitutes a set of definitions of a *formal* vocabulary.

In this paper, we present a proposal for a standard Ontology language defined as SciX ontology, based on the content of SciX, a digital library of conference proceedings. SciX can be viewed as a response to the need of organising and making available a collection of papers presented in an annual international conference. It is

actually a web digital library that provides access to papers presented at the International Conference in Electronic Publishing (EIPub). The conference completed its 10th years in 2006, and SciX is now a comprehensive repository with over 400 papers. The collection comprises papers presented at the two traditional tracks of sessions, namely, general and technical, as well as abstracts of keynote speeches, workshops, round tables and special sessions presentations as well as other kinds of contribution.

2 Motivation and Expectations

During EIPub2006, the 10th version of the conference, a short paper was published with observations on a few quantitative data [1]. The analysis of papers from the 10 years conferences showed that SciX content does need a standardisation process of its data in order to improve search and retrieval for research purposes.

Since the work of Paul Otlet and Henri La Fontaine, regarding documentation, retrieval of relevant content of a document is deemed the key factor of success in any information service/product. In this regard, ontologies have the capability of significantly improving retrieval needed in information services. The proposed ontology will certainly help the exploration of SciX content in both quantitative and qualitative approaches, in the extent that ontologies constitute a set of *classes* (for example, author, title, key-words), *individuals* (for example, Leslie Chan, University of Toronto) and *properties* (<http://www.utsc.utoronto.ca>) that allows a sounder work on the data available.

It is interesting to note that a vocabulary ontology expressed in a formal specification, such as the Web Ontology Language (OWL), makes possible machine processing of information (in a very basic level), rather than simple data, adding expediency to web content search and retrieval. Based on this understanding, the authors of this paper have decided to develop an ontology to help the work of researchers or practitioners interested in using SciX data for research. The leading objective is to provide an “information resource”, allowing the generation of a richer domain-specific knowledge, which is a formal specification of a controlled vocabulary.

The work carried out on the EIPub/SciX collection in 2006 has actually allowed the standardisation of authors and institutions names. The output, however, has not been aggregated to that collection until now. Taking into account that the digital collection so far reproduces the information provided by authors themselves, the work on keywords standardization requires a controlled vocabulary, gradually built up while processing SciX collection. Nevertheless, despite more than 10 years, it seems possible to create this control and help future authors to rely on the output for better stating keywords pertaining to their papers.

The ontology makes it possible to define nodes of semantic relationships and make inferences concerning the topics covered by authors. In addition, a number of relationships between concepts are possible to identify, which, in turn, can respond to the need of standardising them.

This paper, therefore, reports the experience of developing a semi-automated process of extracting concepts from an electronic document collection, in order to create an ontology. It is semi-automatic because of the non-automated procedures concerning part of the data extracted from the database. The idea is to develop an optimised, interesting output of a scholarly papers collection, with the aim of making it easier to be handled by researchers. Through its semantic net, EIPub/SciX ontology is intended to provide better conditions for the user to find the information needed more efficiently. These standard metadata, besides other possibilities, can contribute to define a new environment based on EIPub/SciX Ontology concepts.

It is noteworthy to call attention to the fact that authors have always used different ways of informing both their own names and their institutions names wherever, and whenever they publish. Moreover, different authors define the same topic differently. Concerning EIPub, another aspect that deserves attention is related to the conference sessions' title, which do not always represent a 'core topic' to which the content of those session papers converge. This, in turn, makes an accurate content analysis difficult to carry out.

Such ambiguities and inconsistencies are probably related to the way EIPub has been conducted. It is observable that, as the conference progressed, a number of procedures started to be implemented to make it more organised. At least three of these improvements are clearly identified. Firstly, preoccupation with well-formatted/presented papers based on a well-planned template, along with guidelines about what is expected from authors, has helped improve the content entered into SciX. Secondly, requirements of an abstract, keywords and other data, not present in some of the first editions of the conference, seem to have had an impact on such improvement. Thirdly, the definition of the conference sessions' titles, over the last editions (and 1997's!!), well depicts the content of papers presented in those sessions. Nevertheless, for the enhancement of the access and use of this

content, a standardisation is urgently needed. EIPub/SciX ontology, with no doubt, can definitely contribute to that. In spite of still being in an incipient stage as yet, the ontology has the potential of accomplishing the aforementioned purpose.

3 Methodology (Theoretical Approach and Methodological Procedures)

In order to carry out the study, a number of procedures have been performed, in accordance to what has been stated in theories that underlie content analysis and retrieval in a web environment. As it has been asserted, *the use of ontology as a formal explicit specification of a shared conceptualisation, can help to solve the problem of inefficiency, overloaded “fake information”, ambiguity and chaos* [2]. These authors draw attention to the fact that the use of automatic semantic analysers (despite its earlier own approaches with some incipient, but encourage results) makes possible extract the conceptual structure, describe phrases and use semantic relationships between words and concepts to establish connections between them. This structure, which constitutes a meta-level description, is a representation that brings order to the collection of documents, so it can be understood as an ontology, in the sense defined by Gruber [3] and others. That is, in computer science, the term ‘ontology’ expresses explicit formal specifications of terms in a domain and the relationships among them. Notwithstanding the improvement afforded by an ontology, two problems remain, ambiguity and inconsistencies. Names ambiguity have been approached as one of the major problems in retrieving information from a database. As observed by Han et al [4], *“because of name variations, identical names, name misspellings or pseudonyms, two types of ambiguities in research papers and bibliographies can be observed”*. They are authoring multiple name labels and multiple authors sharing the same name label. The same authors point out that *“it may affect the quality of scientific data gathering, can decrease the performance of information retrieval and web search, and even may cause incorrect identification of and credit attribution to authors”*.

The solution, that is, name disambiguation, has been approached in a variety of ways and has always been related to the creation of authority files. Auld, cited by French et al [5], stressed that this sort of strategy has been called ‘authority work’ and have mostly benefited from computational procedures.

It is interesting to emphasise that name ambiguity can be related to a number of entities, such as authors, institutions, journal or conference titles and so forth. Ambiguities in institutions names have been approached by French et al, who looked at techniques to aid in detecting variant forms of strings in bibliographic databases. They highlight Taylor’s approach [6], whose first principle of authority control is concerned with all variants of a name being “brought together under a single form so that once users find that form, they will be confident that they have located everything relating to the name”. This ‘single form’ has been defined as ‘canonical name’ and, in the work of French and his colleagues, consisted of deciding on a set of canonical affiliation strings and then, assigning each affiliation string in the database to one of these canonical strings.

As can be inferred, disambiguation of names is crucial to the work proposed here, as the simple creation of the ontology itself could not solve this sort of problem by itself. It has been partially and preliminary performed by Costa et al [7] in order to carry their analysis out. Nevertheless, it has been a non-automated process in the sense that no computational procedure was used.

The result, however, corresponds to Taylor’s first principle of authority control and is used in this second work upon EIPub/SciX collection. That is why it is still a ‘semi-automatic’ extraction process. Further work will develop an automated procedure for the creation of canonical names of both authors and institutions.

As regard the ontology and the procedures developed in order to create it, the work was based on the previous model developed by Gottschalg-Duque [8], adapted for this application (Figure 1), as it does not yet include the indexing module. As can be observed, the whole process consists of the stages *file conversion, natural language processing and ontology creation and editing*. The implementation of the modules and sub-modules (figure 2 shows a detailed view of the natural language processing module, comprised of two sub-modules, syntactic and semantic) is done by means of three programs, which are Syntactic Parser, GeraOnto and Protegé.

The stages involved in the analyses and in the ontology creation are succinctly described further, and show how each of them is performed, along with the indication of the software used. It is important to highlight that GeraOnto, because of patent problems, does not allow giving any detail.

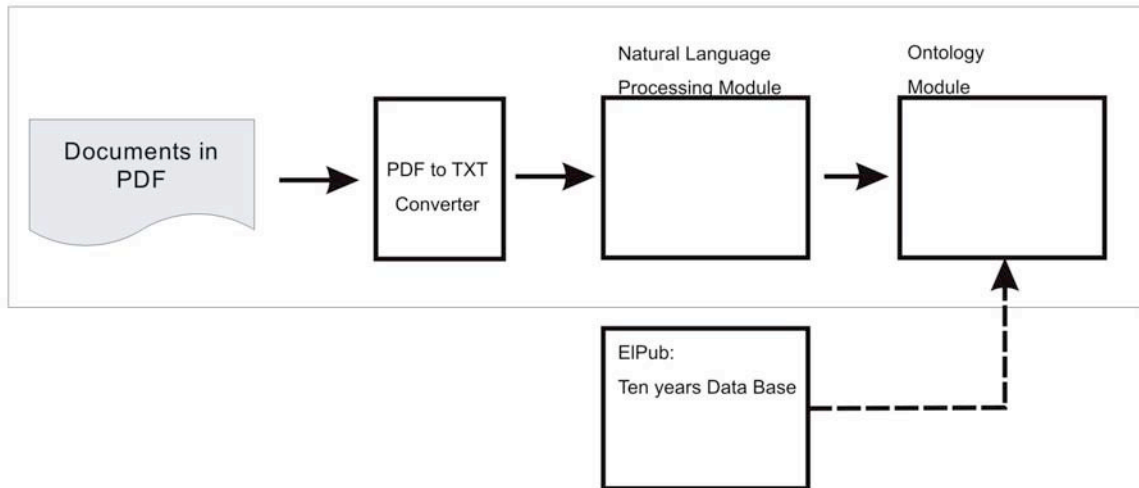


Figure 1: The process of creating the EIPub/SciX Ontology

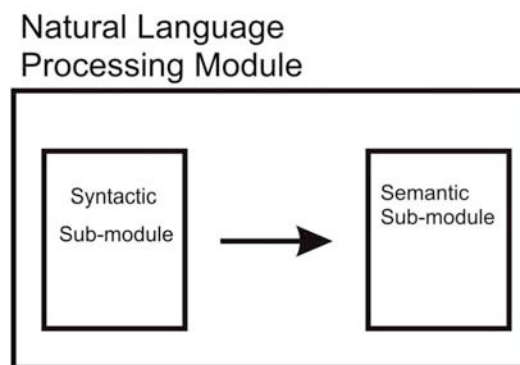


Figure 2: Detail of the two sub-modules of the natural language processing

The ontology construction policy adopted pointed to the definition of what constitutes the relevant concepts that should compose the ontology structure (Figure 3).

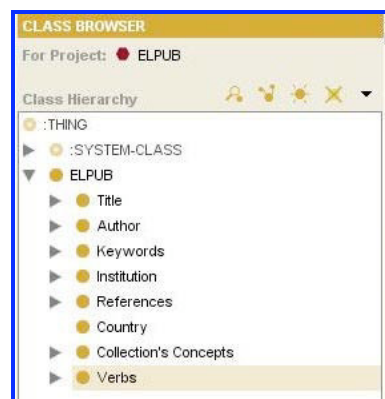


Figure 3: The ontology structure

4 Results

The procedures carried out comprised the following steps and produced results as exemplified in figures 4, 5 and 6:

- Visit SciX site and collect the entire collection of EIPub papers;
- Transfer the collection into a native database;
- Manually extract titles, author's and institution's names, as well as keywords;
- Replace authors and institution names in the native database by the canonical names created by Costa et al [11]. It is interesting to point out that, for institution names, canonical affiliation strings have been created by applying the rule of putting names given by authors in a standard order. That is: university, faculty/school/institute, department and programme/project, whatever appears. For authors names, the rule was to adopt the most complete form of a name;
- Convert all pdf files into txt files;
- Send the texts (from the introduction to just before the references) to a syntactic analyser (**Syntactic Parser - VISL**), which automatically performs the analysis and generates a syntactic tree with all syntactic tags (example in figure 4);

```

SOURCE: live
1. tekst
A1
PARTIAL TREE:The rules could not construct a complete tree
|-D:adj Electronic
|-S:IA-O/C:cl
| |-S:ping publishing
| |-P:v constitutes
| |-Od:pron one
| |-A:g
| | |-H:prp of
| | |-D:g
| | | |-D:art the
| | | |-D:adj hottest
| | | |-H:n topics
| | | |-D:cl
| | | |-P:v discussed
| | | |-A:g
| | | | |-H:prp amongst
| | | | |-D:n researchers
| | | | |-A:g
| | | | | |-H:prp from
| | | | | |-D:g
| | | | | | |-D:art a
| | | | | | |-H:n variety
| | | |-D:g
| | | | |-H:prp of
| | | | |-D:n disciplines

```

Figure 4: Syntactic Parser output

- Send the syntactic tree to *GeraOnto*, which extracts the semantic elements (noun phrases and verbs) of interest for the construction of the ontology. These are concepts that can or cannot be composed of more than one term or concept;
- Insert these concepts into *Protege*, which edits the EIPub/SciX Ontology, using SciX's record identifiers as its slots (examples of the output are shown in figures 5, 6 and 7).

- different techniques
- Natural Language
- words
- phrases
- With high-quality OCR tools
- digital form
- to make
- the Web
- □The
- scholarly skywriting□
- the well-established business model

Figure 5: Concepts automatically extracted from a text

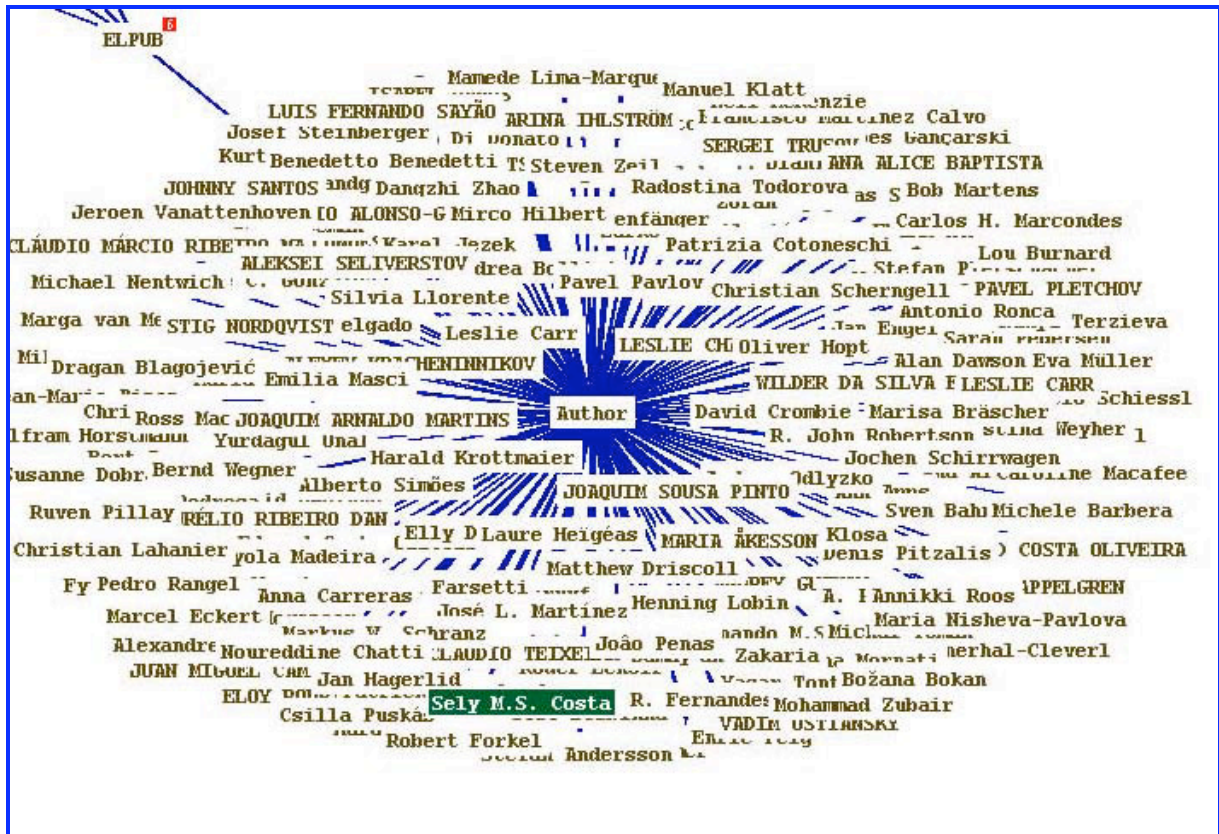


Figure 6: Graphic presentation of the ‘author’ super class and its sub-classes

Notes and References

- [1] COSTA, S. M. S.; BRÄSCHER, M; MADEIRA, F.; SCHIESSL, M. Ten years of ElPub: an analysis of its major trends. In: Martens, B.; Dobрева, M. (Eds.) *Digital spectrum: integrating technology and culture*. Proceedings of the Elpub conference. Bansko : FOI-COMMERCE, 2006. pp. 395-399.
- [2] GOTTSCHALG-DUQUE, C. ; LOBIN, H. Ontology extraction for index generation.. In: COSTA, Sely M. S.; ENGELEN, Jan; MOREIRA, A. C. S. (Eds.) *Building digital bridges: linking culture, commerce and science*. Proceedings of the 8th ICCC International Conference on Electronic Publishing. Brasília, 2004. pp. 111-120.
- [3] GRUBER, T. *What is an ontology*, 1996. Available at: <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html>
- [4] HAN, H.; ZHA, H.; GILES, C. L. Name disambiguation in author citations using a K-way spectral clustering method. In: *International Conference on Digital Libraries archive*. Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries. New York : ACM Press, 2005.
- [5] FRENCH, J. C.; POWELL, A. L.; SCHULMAN, E. Using clustering strategies for creating authority files. *Journal of the American Society for Information Science*, 2000, vol. 51, no. 8, pp. 774-786.
- [6] FRENCH, J. C.; POWELL, A. L.; SCHULMAN, E.; PFALTS, J. L. cite the work of Taylor, published in 1984. Their article is about authority files and has been published in 1997, in the proceedings of the ECDL 1997.
- [7] COSTA, ref. [1]
- [8] GOTTSCHALG-DUQUE, C. *SiRILiCO uma proposta para um Sistema de Recuperação de Informação baseado em Teorias da Lingüística Computacional e Ontologia*. Belo Horizonte, 2005.

On the Evolution of Computer Terminology and the SPOT On-Line Dictionary Project

Jiri Hynek; Premek Brada

Department of Computer Science & Engineering, Faculty of Applied Sciences, University of West Bohemia
Univerzitní 22, 306 14 Pilsen, Czech Republic
e-mail: {jhynek, brada}@kiv.zcu.cz

Abstract

In this paper we discuss the issue of ICT terminology and translations of specific technical terms. We also present SPOT – a new on-line dictionary of computer terminology. SPOT's web platform is adaptable to any language and/or field. We hope that SPOT will become an open platform for discussing controversial computer terms (and their translations into Czech) among professionals. The resulting on-line computer dictionary is freely available to the general public, university teachers, students, editors and professional translators. The dictionary includes some novel features, such as presenting translated terms used in several different contexts – a feature highly appreciated namely by users lacking technical knowledge for deciding which of the dictionary terms being offered should be used.

Keywords: terminology; dictionary; language; lexicography; translation; wiki; information technology

1 Introduction

Ordinary users and experts alike get the feeling that Czech translations of computer documentation (including books, help files, reference guides) are of inferior quality [1,2]. It is likely that the same applies to other languages. Readers often prefer original documents in English, which discriminates against those readers that either have no access to these materials, or lack the required linguistic knowledge.

It is generally believed that what makes a translator's work difficult is the technical terminology. But in reality, the main problem is to grasp the *meaning* of specific terms, the actual thing that is denoted by the term, use of the term in the *context*, occurrence of the term with other terms, and the term's stylistic features.

The situation is complicated by the fact that translations are created (literally made up, invented) by technical staff and software developers who lack linguistic skills and „feeling“ for the natural language; or at the opposite end of the spectrum, by „professional“ translators who lack skills and terminology in the specific ICT (information and communication technology) sub-domain.

Complaints about the poor quality of translations are often backed by the lack of uniformity in addressing specific technical terms. In other words, different authors give the same thing different names. Activities aimed at standardizing computer terminology are very rare; one of the very few is Microsoft Terminology Translations [3] available for 59 different languages at the time of writing. The glossary provided in the form of a CSV file contains more than 12,000 English terms plus the translations of the terms. Unfortunately, translated terms are not provided for all the terms in the list, and some translations seem to be controversial.

In this paper we would like to discuss the issues which are encountered when translating terms arising in a rapidly developing domain like the ICT, including taxonomy of the terms in view of their development status. Secondly we present a project of an on-line dictionary aiming at supporting the work of translators and localization developers. The structure of the paper is as follows. In the following section we discuss the work of translators and the tools they have available. Section 3 describes the taxonomy of terms and some issues with respect to translation. Section 4 presents the SPOT on-line dictionary, the goals of the project, key distinguishing characteristics of the tool, and its current status. The paper is finished with a conclusion.

2 Dictionaries and Resources for Translators

We are living in the age of an information explosion. We would need more than one year to study the amount of information produced worldwide within a single day. It is up to translators to cope with all the changes that take

place in the field of their specialty. Searching for technical terminology is one of the most time-consuming tasks of every professional translator. Fortunately, the majority of support data can be found by means of search engines, such as Google. Online dictionaries are also available, although the contents of their free versions are often limited, or they are too general for specific purposes. We are happy to observe the trend of utilizing “collective wisdom” in special Internet projects (Wikipedia being the best example), with users selflessly sharing their expertise with their fellows. Our project is one of those.

The SPOT online dictionary presented here (see <http://spot.zcu.cz/>) is not the only project of its kind at our University. There is the English-Czech GNU/FDL dictionary available at <http://slovník.zcu.cz/online/> and www.wordbook.cz, which is based on the i-spell database [4] – see Figure 1 below.

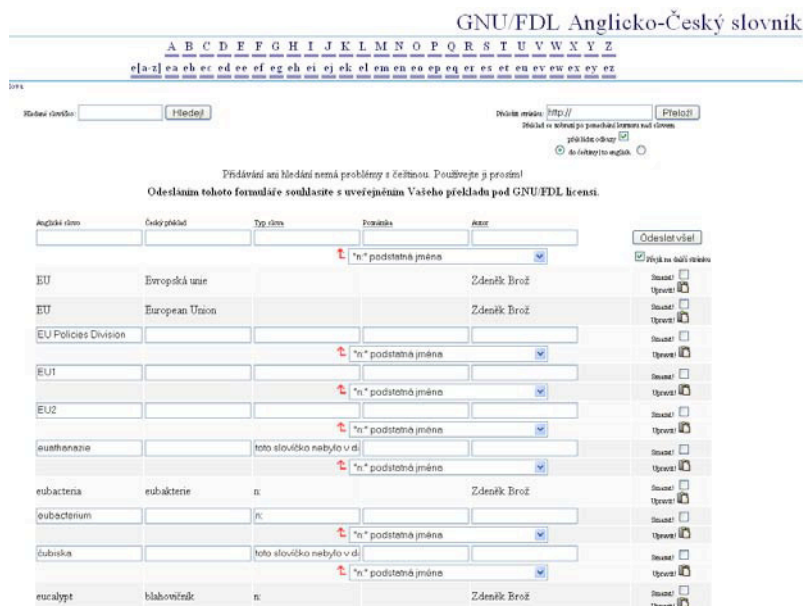


Figure 1: English-Czech dictionary based on the i-spell database

Of course, there are various general dictionaries and encyclopaedias freely available online. These represent an excellent source of linguistic information as well as wisdom, but they are unsuitable for our objective. You can visit the following:

- Merriam-Webster OnLine at: www.webster.com
- Encyclopaedia Britannica online at: www.britannica.com
- Oxford Advanced Learner's Dictionary at: www.oup.com/elt/oald/
- Wikipedia at: <http://wikipedia.org/>
- Citizendium at: <http://en.citizendium.org/>
- Dictionary, Thesaurus and Encyclopedia at: www.reference.com
- EuroWordNet (multilingual database with wordnets) at: www.illc.uva.nl/EuroWordNet/
- Free On-Line Dictionary of Computing at: <http://foldoc.org/>

A comprehensive list of on-line dictionaries for multiple languages suitable for both general and specific purposes can be found at <http://a-z-dictionaries.com/online-dictionary.html>, or www.yourdictionary.com.

2.1 CAT Tools for Professionals

Today's ICT localization projects often involve millions of words of software documentation, help files, warnings, error messages and other texts to be translated. In order to keep the documentation consistent, memory-based computer-aided translation (CAT) tools are a must. They have a substantial impact on both the translation quality and the productivity of all people involved in the localization project. The roll-out of CAT tools dates back to the beginning of the 1990s. Here are a few examples of today's most popular CAT tools (listed alphabetically):

- Deja Vu (www.atril.com)
- IBM Translation Manager (www.ibm.com)
- MetaTaxis (www.metataxis.com)

- SDL Trados (www.trados.com)
- SDLX (www.sdl.com)
- Star Transit (www.star-ag.ch)
- Systran (www.systransoft.com)
- WordFast (www.wordfast.net)

Additional resources for translators and linguists can be accessed via www.multilingual.com, where you can find a collection of more than 1600 links divided into 41 categories (such as Automated Translation, Dictionaries, Internationalization Tools). There are also conferences dedicated to translation and localization industries, such as Localization World (www.localizationworld.com). The localisation community is also supported by The Localisation Research Centre at University College Dublin (www.localisation.ie). Useful links to various translation resources can be found at www.translation.net/links.html. You can download various glossaries at the translators' directory Go Translators (www.go-translators.com). An example of a free memory translation database for multiple languages can be accessed at www.open-tran.eu.

3 English ICT Terms: From Old-timers to Troublemakers

We are witnessing rapid development in the ICT domain, with quickly changing terminology as one of the consequences. And the language of ICT professionals is suffering badly. The impact on other languages is serious – in the case of Czech, for example, the majority of terms are more or less adopted directly from English. Some terms find their Czech equivalent immediately, some are developing, and for some we still have no equivalent. For those that enjoy categorization, the following classes of computing terms can be identified.

3.1 Old-timers

This category includes many terms dating back to the 1960s. Their meaning has settled, as well as their translation into other languages. We can subdivide into:

- Morally obsolete terms (such as *bubble memory*, *punch tape*, *punch card*, and recently, also *floppy disk* and *diskette*);
- Stabilized and commonly used terms (such as *display*, *plotter*, *button*, *printer*, *mail server*, *dialog window*, *RAM*, *search engine*).

3.2 Novas and Supernovas

English is quite a flexible and even playful language with the ability to create and absorb new words (e.g. *text me* for “send me an SMS message”, or abbreviations like *B4*), and new terms therefore appear very quickly, sometimes with an associated hype. Other languages need not be that flexible, which then poses problems in translating these new words, often created ad hoc. Unfortunately, this happens more than often in our mother language.

Compared to the category above, the number of terms labelled as Novas is relatively small. More difficult is the decision taken by a translator or a publication/magazine editor as to how to have these localized. This category includes: *cookie*, *spam*, *phishing*, *blog*, *freeware*, *emoticon*, *code closure*, *social bookmarking* and so on.

3.3 Troublemakers

The category of Troublemakers is relatively large. Thanks to Troublemakers, computer dictionaries get published and sold in large volumes, electronic dictionaries flourish on the Internet, and long debates among academics and language purists are held. It is up to computer users, editors and authors which terms shall prevail and which shall become obsolete. We can, again, subdivide into:

- **Homonymous Troublemakers** – identical computer terms with different meanings, such as *collector* (of a transistor) and (software) *collector* (e.g. data collector, portable collector representing a SW component), or *plug-in* (meaning an amplifier) vs. *plug-in module*, a SW component);
- **Synonymous Troublemakers** – The same (or virtually the same) meaning expressed differently, sometimes erroneously (such as *cross reference* vs. *cross-index* vs. *link*, *local menu* vs. *context menu* vs. *pop-up menu* vs. *shortcut menu*, *pull-down menu* vs. *drop-down menu* (or a *list*) *submenu* vs. *child*

menu, tool palette vs. toolbox, custom vs. personal, scroll bar vs. slide bar, clickable map vs. interactive map).

3.4 Terms on the Move

- Over time, some IT terms **shift and/or extend their meaning**, such as the former *monitor*, meaning today's *screen*, and which now commonly refers to SW utilities; *link*, formerly used in the context of network connections (e.g. *link control protocol*) is now more often used for *web links* or *object linking*;
- **Everyday terms acquire new meaning** in the context of IT: e.g. *signature* is now commonly used in *method signature*, *virus signature*, or *digital signature*; *little endian* and *big endian* (adopted from Gulliver's Travels by Jonathan Swift) now refer to the method of storing multi-byte data; *pool* acquired the meaning of *fund*, i.e. a source of something (*thread pool*, *resource pool*); *key* becomes an "identifier" (*database primary key*); *root* (such as in *tree root*) becomes a type of user, or is used for *root folders*; *builder* has a new meaning, such as in *application builder*, *list builder*, or *expression builder*; the word *seamless* has become a buzzword in the context of application integration; *heap* currently represents a type of memory; *thread* today refers to a sub-process (such as in *multithreading*); *stamp* has acquired a more abstract sense; *field* (such as in *sports field* or *mine field*) got the meaning of *entry/item*; *host* no longer refers just to persons, but also servers (either hardware or software); *wizard* is commonly used in *installation wizard* or *test wizard*; *garbage* now refers to meaningless data or data no longer needed; *docking* is used for toolbars or laptops instead of (space) ships;
- We are witnessing the process of **heavy verbalization**, e.g. *to right-click* (in place of "click the right mouse button"), *to cache* (in place of "save in cache"), *to host* ("to act as a host"), *personalize*, *televise*, *deserialize*, or *visualize*.

3.5 Esoteric Terms

- **Esoteric IT terms** are used by a relatively narrow group of IT specialists; these include, for example: *setter* and *getter* (in object-oriented programming), *undeploy*, *abstract factory*, *uptime*, *design pattern*, *marshalling*, *serialization*, *proxying*, *entity bean*, *refactoring*, *tight coupling*, *box model*, *locale*, and *hot-swap*;
- **Application- or corporation-specific terms**: these are quite special purpose, found in a limited number of applications, or used only within specific corporations or teams. For example: *rolling period*, *purge*, *context root*, *governance*, *updater*, *Apple menu*, and *Start menu*. Please note that these terms may have different meanings depending on the organization/application.

Both situations in which esoteric terms arise, summed by the "narrow user group" characteristic, make their localization complicated, since specialists are happy to use the original (non-translated) terms, while translators lack the technical background knowledge to make up suitable translations.

4 SPOT On-Line Dictionary

Aware of the issues discussed above, the authors started a new project to overcome the difficulties in creating suitable translations in the rapidly changing ICT domain. Our objective is to help the community of translators (plus editors and IT professionals) to either quickly find the correct translation for new, unusual, or tricky terms, or – if no such translation is known – to create one with the assistance of the "collective wisdom" of their colleagues.

Our work is novel in several aspects. Although there are various general-purpose dictionaries available that help in achieving the first objective, we are focusing strictly on the area of ICT, which undergoes frequent changes from the linguistics point of view. The SPOT dictionary is initially starting with a well-established terminology corpus, and its further development will be supported by our extensive experience in localization projects, besides the knowledge of leading IT professionals and editors.

Our approach is also new in that we "settle" the final version of Czech translations within the community of users, under the supervision of ICT specialists. It is anticipated that we will capture the latent interest of the community of translators, readers, editors and companies involved in localization projects (such as translations of computer publications, documentation and help files, translation of game scripts, and office software localization).

While we believe that professional editorial supervision is absolutely essential for acceptable “settling” of translated terms, the role of the mass of users is seen as a key differentiating point to classical approaches in translation and also as a critical success factor. Similarly to several successful Web 2.0 projects like Wikipedia (www.wikipedia.org), the SPOT project thus hopes to bring into fruitful cooperation professionals and users alike.

Finally, we also offer new features facilitating the work of large or distributed localization teams. Controversial terms are always on-line, rather than stored locally with a localization team member. It is essential that everyone uses the latest version of translated terms. SPOT also eliminates the need to redistribute up-to-date versions of dictionaries, which gets time-consuming in the case of large translation projects.

Last but not least, the added value of our dictionary is greatly enhanced by showing translations in various *contexts* based on on-line search of the Internet. The user can see which of the translations offered should be used, as it can be derived from the context information shown (see Figures 3, 4, 5 below).

4.1 The Dual Role of SPOT

As suggested by the previous paragraphs, SPOT will serve two complementary purposes: a reference dictionary of computer terms, and a platform for “settling” these terms.

4.1.1 Reference Dictionary

The basic function of this Internet dictionary is to provide Czech translations for specific ICT terms, either by browsing or by searching (see Figure 2 below). The quality of the translations is guaranteed by the initial English-Czech corpus based on the English-Czech ICT Dictionary written by the first author. Opportunities to provide valuable add-ons on top of the initial corpus are given by the fact that the dictionary is on-line.

brada | Dictionary control | Administration | Logout

SPOT: On-Line Czech-English Dictionary of ICT

SPOT: On-Line Czech-English Dictionary of ICT is a Czech-English database of computer terminology based on know-how of computer experts, editors of professional journals and translators of ICT documentation. [More information...](#)

Search

Help: Please enter the term to search for Czech/English equivalents.

Browse the Dictionary

In alphabetical order
 A (39), B (16), C (60), D (25), E (24), F (26), G (11), H (10), I (28), J (2), K (2), L (13), M (17), N (10), O (19), P (48), Q (2), R (17), S (63), T (15), U (14), V (3), W (10), X (0), Y (0), Z (0)

Browse these areas
 Algoritmy a programování (0), Architektura počítačových systémů (0), Bioinformatika (0), Databázové systémy (0), Dolování a správa znalostí (0), DTP, typografie, pre-press (0), Elektrotechnika, elektronika (0), Hardware (0), Internet a internetové technologie (0), Kryptografie a kryptoanalýza (0), Mobilní zařízení a prvky (0), Modelování a simulace (0), Operační systémy (0), Organizace, normy (0), Počítačová grafika (0), Počítačová bezpečnost (0), Počítačová lingvistika, NLP (0), Programovací jazyky a překladače (0), Přenos dat, komunikace (0), Robotika (0), Rozhraní člověk-počítač (0), Sítě a distribuované systémy (0), Software (0), Softwarové inženýrství (0), Teorie grafů a diskrétní matematika (0), Umělá inteligence (0), Webové aplikace (0),

Browse these projects
 Acrobat (245), UML (229),

Browse by word status
 Neschváleno (1), Schváleno (474),

Help wanted.
 Registered users ucan enter additional words or edit the dictionary contents. [More information...](#)

Copyright © 2006 Pavel Cvrček © Jan Kravál, Site Map | Contact info

Figure 2: SPOT interface

Firstly, editors can assign categorization *tags* to individual terms, in a manner similar to some popular web services such as Flickr for photographs [5] or StumbleUpon for website links [6]. Users will thus be able to confine their search to a specific ICT area, or study other terms related to their area of interest:

- Algorithms and Programming
- Artificial Intelligence

- Communications
- Computer Graphics
- Computer Linguistics, NLP (Natural Language Processing)
- Computer Modelling and Simulations
- Computer Networks and Distributed Systems
- Cryptography, Cryptanalysis
- Data and knowledge mining
- Database Systems
- DTP, Pre-press
- Electrical Engineering and Electronics
- Hardware
- Internet and Web Technologies
- Man-machine Interface
- Mobile Devices
- Operating Systems
- Programming Languages
- Robotics
- Software

Secondly, we can enrich the information about the terms and their translations by displaying their occurrences in various *contexts* based on on-line search of the Internet. We use *Google search API* to obtain on-demand results, configured so that only sites with high relevance to ICT are searched. The probability that the context results will be meaningful is thereby increased. Having spent many years working on large localization projects, we are aware that showing a term's usage in the context is very useful and highly appreciated, especially by translators.

In its current implementation, SPOT can display the following context information (see Figures 3, 4, 5):

- Web sites, no restriction,
- IT webs only,
- Wikis,
- On-line dictionaries,
- Blogs (to be implemented).

A list of authoritative IT web sites, wikipedias, on-line dictionaries and blogs is maintained by the system's administrator.

Disadvantages of the web-based Corpus

Corpus gathered from the web is not always reliable and Google cannot provide information that linguists would like to have. In addition,

- Users have no control over the content,
- Web sites are full of metadata that are unnecessary for corpus building,
- We need to focus on pages in a specific language only,
- Specific jargon is used in web blogs and chat rooms.

In spite of the above, we believe that the advantages of providing the user with context information outweigh the drawbacks. Searching for context information via Google is exactly what most translators groping in the dark would do in the first place.

Context information prepared ahead – the Database Corpus

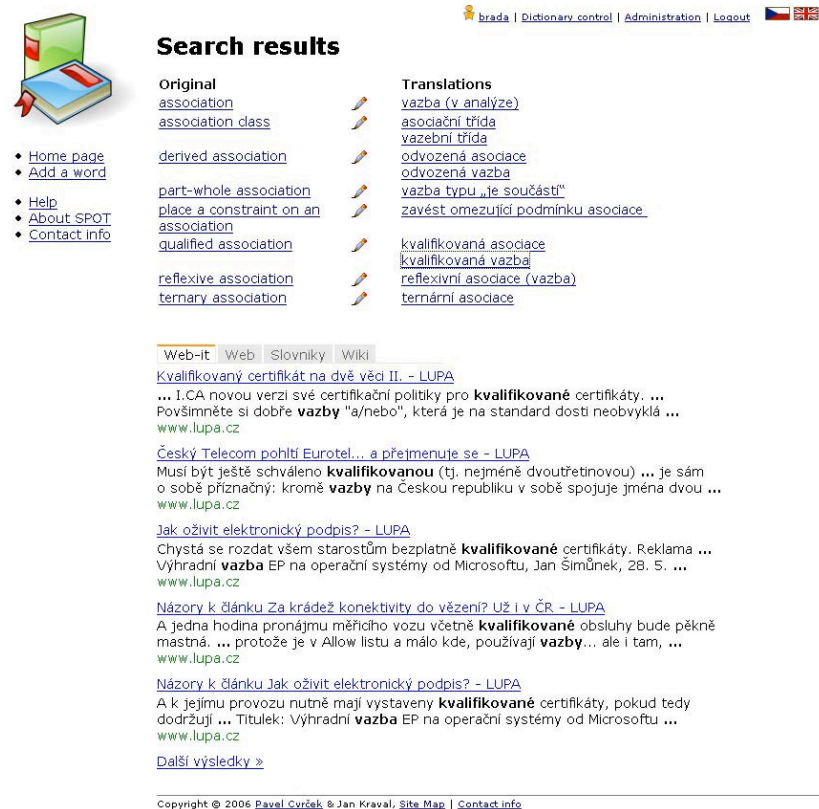
Initially, we were planning to build the corpus manually by collecting context information from the web, selecting the best samples manually and storing the resulting corpus in a text database. We were attracted by the design of the WebBootCaT (see <http://corpora.fi.muni.cz/bootcat/> or <http://sslmit.unibo.it/~baroni/>), which is a tool for building domain-specific corpora to support translators [7].

For more information on web corpora building, see also KWIC Finder (www.kwicfinder.com) ("Web as Corpus"), WebCorp (www.webcorp.org.uk), or Web as Corpus Toolkit (www.drni.de/wac-tk/).

There are two reasons why we rejected the idea of corpus building within a database:

1) SPOT users are on-line and it is unlikely that corpus building on the fly would cause any time delays; context information gathered via Google search API is presented instantly; using a database involves additional overhead plus necessitates database maintenance.

2) ICT terminology evolves very quickly and new terms are created on virtually a daily basis; it would be prohibitive to maintain database corpus up-to-date for this field of speciality.



The screenshot shows the search results for the term 'association' in the SPOT On-Line Dictionary. The interface includes a navigation menu on the left with links for Home page, Add a word, Help, About SPOT, and Contact info. The main content area is titled 'Search results' and is divided into two columns: 'Original' and 'Translations'. The 'Original' column lists various types of associations, and the 'Translations' column provides their Czech equivalents. Below the search results, there are search filters for 'Web-it', 'Web', 'Slovníky', and 'Wiki'. The results are filtered to show only IT-related content, with examples from LUPA and Wikipedia. The footer contains copyright information for 2006 Pavel Cvrček & Jan Kravál.

Search results

Original

- [association](#)
- [association class](#)
- [derived association](#)
- [part-whole association](#)
- [place a constraint on an association](#)
- [qualified association](#)
- [reflexive association](#)
- [ternary association](#)

Translations

- [vazba \(v analýze\)](#)
- [asociační třída](#)
- [vazební třída](#)
- [odvozená asociace](#)
- [odvozená vazba](#)
- [vazba typu „le součástí“](#)
- [zavést omezující podmínku asociace](#)
- [kvalifikovaná asociace](#)
- [kvalifikovaná vazba](#)
- [reflexivní asociace \(vazba\)](#)
- [ternární asociace](#)

Web-it | Web | Slovníky | Wiki

[Kvalifikovaný certifikát na dvě věci II. - LUPA](#)
... I.CA novou verzí své certifikační politiky pro **kvalifikované** certifikáty. ...
Povšimněte si dobře **vazby** "a/nebo", která je na standard dosti neobvyklá ...
[www.lupa.cz](#)

[Český Telecom pohltí Eurotel... a přejmenuje se - LUPA](#)
Musí být ještě schváleno **kvalifikovanou** (tj. nejméně dvoutřetinovou) ... je sám
o sobě příznačný; kromě **vazby** na Českou republiku v sobě spojuje jména dvou ...
[www.lupa.cz](#)

[Jak oživit elektronický podpis? - LUPA](#)
Chystá se rozdat všem starostům bezplatně **kvalifikované** certifikáty. Reklama ...
Výhradní **vazba** EP na operační systémy od Microsoftu, Jan Šimůnek, 28. 5. ...
[www.lupa.cz](#)

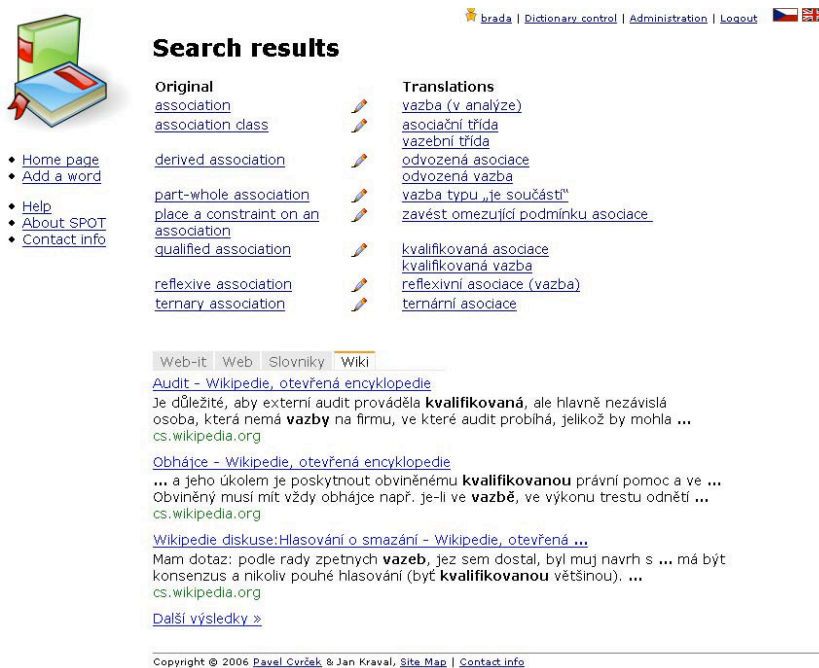
[Názory k článku Za krádež konektivity do vězení? Už i v ČR - LUPA](#)
A jedna hodina pronájmu měřičho vozu včetně **kvalifikované** obsluhy bude pěkně
masťná. ... protože je v Allow listu a málo kde, používají **vazby**... ale i tam, ...
[www.lupa.cz](#)

[Názory k článku Jak oživit elektronický podpis? - LUPA](#)
A k jejímu provozu nutně mají vystaveny **kvalifikované** certifikáty, pokud tedy
dodržují ... Titulek: Výhradní **vazba** EP na operační systémy od Microsoftu ...
[www.lupa.cz](#)

[Další výsledky >](#)

Copyright © 2006 Pavel Cvrček & Jan Kravál, [Site Map](#) | [Contact info](#)

Figure 3: Term translations shown in the context of the web, confined to IT sites only



The screenshot shows the search results for the term 'association' in the SPOT On-Line Dictionary, filtered to show only pre-defined wikis. The interface is similar to Figure 3, but the search filters are set to 'Wiki'. The results are filtered to show only Wikipedia content, with examples from the 'Audit' and 'Obhájce' articles. The footer contains copyright information for 2006 Pavel Cvrček & Jan Kravál.

Search results

Original

- [association](#)
- [association class](#)
- [derived association](#)
- [part-whole association](#)
- [place a constraint on an association](#)
- [qualified association](#)
- [reflexive association](#)
- [ternary association](#)

Translations

- [vazba \(v analýze\)](#)
- [asociační třída](#)
- [vazební třída](#)
- [odvozená asociace](#)
- [odvozená vazba](#)
- [vazba typu „le součástí“](#)
- [zavést omezující podmínku asociace](#)
- [kvalifikovaná asociace](#)
- [kvalifikovaná vazba](#)
- [reflexivní asociace \(vazba\)](#)
- [ternární asociace](#)

Web-it | Web | Slovníky | Wiki

[Audit - Wikipedie, otevřená encyklopedie](#)
Je důležité, aby externí audit prováděla **kvalifikovaná**, ale hlavně nezávislá
osoba, která nemá **vazby** na firmu, ve které audit probíhá, jelikož by mohla ...
[cs.wikipedia.org](#)

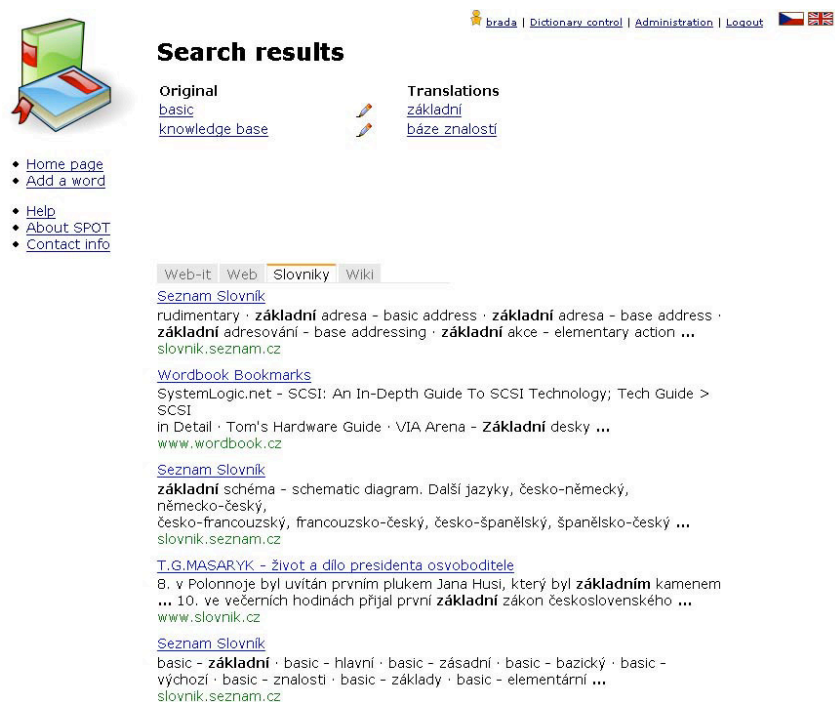
[Obhájce - Wikipedie, otevřená encyklopedie](#)
... a jeho úkolem je poskytnout obviněnému **kvalifikovanou** právní pomoc a ve ...
Obviněný musí mít vždy obhájce např. je-li ve **vazbě**, ve výkonu trestu odnětí ...
[cs.wikipedia.org](#)

[Wikipedie diskuse:Hlasování o smazání - Wikipedie, otevřená ...](#)
Mám dotaz: podle rady zpětných **vazeb**, jez sem dostal, byl muj navrh s ... má být
konsenzus a nikoliv pouhé hlasování (byť **kvalifikovanou** většinou). ...
[cs.wikipedia.org](#)

[Další výsledky >](#)

Copyright © 2006 Pavel Cvrček & Jan Kravál, [Site Map](#) | [Contact info](#)

Figure 4: Term translations shown in the context of pre-defined wikis



The screenshot shows the SPOT dictionary interface. At the top right, there are navigation links: 'brada | Dictionary control | Administration | Logout' and flags for Czech and English. The main heading is 'Search results'. Below it, there are two columns: 'Original' and 'Translations'. Under 'Original', the word 'basic' is listed with its Czech equivalent 'knowledge base'. Under 'Translations', the word 'základní' is listed with its Czech equivalent 'báze znalostí'. On the left side, there are links for 'Home page', 'Add a word', 'Help', 'About SPOT', and 'Contact info'. Below the search results, there are tabs for 'Web-it', 'Web', 'Slovníky', and 'Wiki'. The 'Slovníky' tab is active, showing a list of search results for 'základní' from various sources like 'Seznam Slovník', 'Wordbook Bookmarks', and 'T.G.MASARYK - život a dílo presidenta osvoboditele'.

Figure 5: Term translations shown in the context of other on-line dictionaries defined by the system's administrator

Last but not least, there is a set of *minor features* that can be helpful for SPOT users. Apart from the translation(s) deemed correct, the dictionary can show incorrect or unsuitable translations of a term when such were labelled by the editors. This information can be a valuable guide for translators and learners alike, helping them avoid common mistakes. Also, since registered users are able to vote for the translations, popular and well-accepted Czech equivalents become easily visible.

4.1.2 Platform for Terminology “Settling” by Voting

The main advantage of SPOT is that it can act as a platform for the natural development of quality Czech equivalents to the original English terms. With very little effort it can be adapted for any language.

Internet facilitates quick and efficient communication, and it is the inherent property of “collective wisdom” that we are planning to utilize. SPOT will let its users propose Czech translations of unlisted or “unsettled” terms, vote for these translations and discuss them. Based on our long-term observations, users that are likely (and willing) to suggest language equivalents are those that are truly concerned about their form, such as professional translators, professional engineers authoring technical documentation, or academics.

As a necessary complement, the dictionary also includes features for editors to decide upon the final version of individual translations (see Figure 6 below). The final choice on the correct vs. wrong translations will be taken by a small team of editors chosen from the most renowned professionals in the field of ICT, possessing sufficient linguistic knowledge.



The screenshot displays the SPOT online dictionary interface. At the top, there is a navigation bar with links for 'admin', 'Správa slovníku', 'Administrace aplikace', and 'Odhlásit'. The main content area is titled 'bigwig' and includes a small icon of a book. Below the title, there is a list of metadata: 'Přidal: admin admin', 'Stav: Schváleno', 'Úzus: Ne zvolen', and 'Kategorie: (žádná kategorie není přiřazena)'. A 'Projekt:' field is set to 'UML'. There are buttons for editing and deleting the entry. Below this, the 'Překlady' (Translations) section shows a translation for 'hlavoun (velké zvíře)' with its own metadata: 'Přidal: admin admin', 'Zdroj překladu: Obecná znalost', and 'Stav: Schváleno'. A '+ Přidat překlad' button is present. The 'Komentáře' (Comments) section states 'Dosud nebyl přidán žádný komentář.' and includes a 'Přidat komentář' button. Below this is a form for adding a comment, with a note that HTML tags are not allowed. The form has a 'Předmět:' field and a large text area. A 'Přidat komentář' button is at the bottom of the form. At the very bottom of the page, there is a copyright notice: 'Copyright © 2006 Pavel Cvrček & Jan Kravál, Mapa webu | Kontakt'.

Figure 6: Editing, updating and commenting on term translations

The process of establishing an accepted term translation is as follows:

1. Users add (or import from a CSV file) original terms (in English), assigning them to specific categories;
2. Other users propose additional suitable translations, possibly with references to sources of occurrence, and explanatory comments;
3. Registered users can vote on translations and discuss controversial ones;
4. When the discussion has settled, the editor marks the most suitable translation as “Official”, and the remaining versions as either “Usable” or “Unusable / Non-recommended”.

Since anyone can become a registered user, SPOT supports the idea that translations may become shared work, and consequently a shared responsibility of the “ICT general public”. This is a different approach than the prevailing practice, where a few linguists work on the official localization terms in isolation.

SPOT will also find practical use amongst members of localization teams while creating specialized dictionaries for specific translating projects (see Figure 7). The development of a local corpus proceeds as follows:

1. A special section in SPOT is reserved for the localization team (specific project / customer / product);
2. On-line “settling” of controversial translations (editing, voting) takes place within the team;
3. Terms are propagated instantly into the rest of the dictionary corpus; nonetheless, project-specific terms are designated as such. There is a feature to export this partial corpus to other formats, such as CSV, and to import the terms into the team’s Computer Aided Translation tool – CAT.

Figure 7: Listing terms filtered by a translation project

The advantages this usage of SPOT brings to the localization teams are manifold, rooted mainly in its on-line implementation:

- Controversial terms are always either on-line, or directly in CAT, rather than stored locally with a localization team member;
- Everyone uses the latest (i.e. currently agreed) version of terms;
- No need to redistribute up-to-date versions of dictionaries.

Below you can see an extract from the MARTIF file used for interchanging newly translated terms among team members (export from Termstar CAT tool while working on a localization project):

```
<langSet lang='eng-us'>
<ntig><termGrp>
  <term>Add a Favorite Place...</term>
  <termNote type='termType'>full form</termNote>
  <termNote type='TS_CreateId'>Jiri Hynek</termNote>
  <date type='origination'>20060917T161007Z</date>
  <termNote type='TS_UpdateId'>Jiri Hynek</termNote>
  <date type='modification'>20060917T161007Z</date>
</termGrp></ntig>

</langSet>

<langSet lang='ces-cz'>
<ntig><termGrp>
  <term>Přidat oblíbené místo</term>
  <termNote type='termType'>full form</termNote>
  <termNote type='TS_CreateId'>Jiri Hynek</termNote>
  <date type='origination'>20060917T161007Z</date>
  <termNote type='TS_UpdateId'>Jiri Hynek</termNote>
  <date type='modification'>20060917T161007Z</date>
</termGrp></ntig>
```

4.2 SPOT Implementation – The Current Status

We have currently implemented basic features that include searching, adding new translations, tagging and commenting, dictionary administration by the editorial team, and namely showing translated terms in the context. Under way is the implementation of discussion forums to translations, as well as voting for individual terms suggested by users or editors. Cross-referencing (“see also” links) will be featured in a subsequent release as well. The application is implemented on the Java 5 platform, utilizing the Spring framework and PostgreSQL database system.

The SPOT dictionary can be accessed at <http://spot.zcu.cz>, where interested readers will also find links to project-related pages. For the data model of the current implementation, see Figure 8 below.

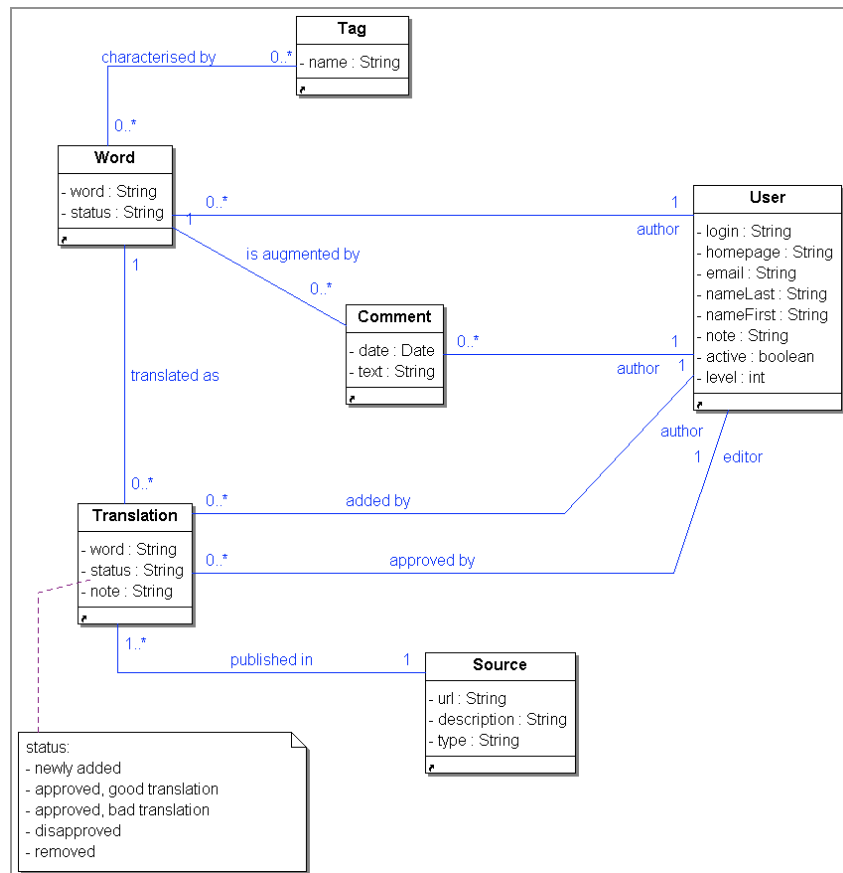


Figure 8: SPOT data model

5 Conclusion

It is not the task of university teachers or pure linguists to propose suitable equivalents of specialized English terms in their native language. Rather than these people, it should be up to the editors of computer magazines, students, and translators of IT documentation. Terminology must adapt to users, not the other way round.

We believe that by implementing the solution proposed herein, we can not only improve the quality of computer documentation translated into Czech, but also the general culture in this area. Indirectly, this may help to increase readers' preferences for translated publications, while curtailing discrimination against those who either do not have the means to obtain the original books, or the language skills to read them.

Hopefully, SPOT will help in at least the partial standardization of terminology being used, and become a useful source of information for persons involved in the technical or academic writing process.

Acknowledgements

This work was partly supported by the Ministry of Education of the Czech Republic under Grant No. 2C06009 within the National Program for Research II.

Notes and References

- [1] VIRIUS M. On the quality of Czech translations of technical literature (in Czech: Nad kvalitou českých překladů odborné literatury). Proceedings of OBJEKTY 2002. Praha 2002. ISBN 80-213-0947-4.
- [2] MONDSCHNEIN P. Officially dishonoured Czech (in Czech: Oficiálně przněná čeština.) <http://games.tiscali.cz/specials/prznenacestina/index.asp>. Published 8th September 2004.
- [3] Microsoft Terminology Translations, available at: www.microsoft.com/globaldev/tools/MILSGlossary.msp
- [4] The i-spell database, available at: www.lasr.cs.ucla.edu/geoff/tars/
- [5] Flickr – Online photo management and sharing application, available at: www.flickr.com/
- [6] Stumbleupon – Personalized recommendations of websites, videos, pictures and more. Available at: www.stumbleupon.com/
- [7] BARONI, K.; POMIKÁLEK, R. 2006. WebBootCaT: instant domain-specific corpora to support human translators. In: Proceedings of EAMT 2006 - 11th Annual Conference of the European Association for Machine Translation. Oslo, Oslo University (Norway), ISBN 82-7368-294-3.

Scientific Heritage in Bulgaria Makes First Digital Steps

Milena Dobreva¹; Nikola Ikonov^{1,2}

¹ Digitisation of Scientific Heritage Dept., Institute of Mathematics and Informatics
bl. 8, Acad. G. Bonchev St., Sofia 1113 Bulgaria

e-mail: dobreva@math.bas.bg

² Phonology and Speech Communications Lab, Institute for Bulgarian Language
Shipchenski Prohod 52, Sofia 1113 Bulgaria

e-mail: nikonov@ibl.bas.bg

Abstract

The paper presents recent initiatives in creation, delivery and management of scientific heritage digital resources in Bulgaria. The local and international tendencies will be sketched. Then the work of the Department for Digitization of Scientific Heritage at the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences and more specifically the joint projects with the State Department of Archives and the Central Library of the Bulgarian Academy of Sciences will be described. Finally, we will present a SWOT analysis of the local situation and suggestions for urgently needed actions.

Keywords: digitization; BulDML; musical periodicals; DIGMAP

1 Introduction

The high expectations about the digital libraries are seen from their listing amongst the three flagship initiatives of the strategy "i2010 – A European Information Society for growth and employment" (the other two areas are caring for people in an ageing society; and the intelligent car). Amongst the basic aims in the joint European efforts are to *avoid duplication*, to *cooperate in networking and standards*, as well as in *developing common and more cost-effective solutions* [1].

Without any doubt, the availability of high quality digital content is in the basis of humanitarian and social research. Most countries from the South-Eastern Europe are still far from moving in line with the European Commission (EC) guidelines in this area. The *EC Recommendation of 24 August 2006 on the digitization and online accessibility of cultural material and digital preservation* [2] emphasizes the importance of setting up of large and sustained digitization facilities, encouraging partnerships between cultural institutions and the private sector, solving the problems around orphan works as well as developing clear quantitative targets for digitization efforts. Furthermore, the *Council Conclusions on the Digitisation and Online Accessibility of Cultural Material and Digital Preservation (2006/C 297/01)* [3] suggest an action plan which can not be followed in countries where there is neither national strategy nor large scale digitisation facilities or recognized competence centres promoting digitization activities.

Recently at the closing of Conference on Scientific Publishing in the European Research Area Access, Dissemination and Preservation in the Digital Age (Brussels, 16 February 2007), Ms. Viviane Reding, EU Commissioner on IS & Media stressed:

*"... if we do not actively pursue the preservation of digital material now, we risk having a **gap in our intellectual record**. If you allow me another historical reference, we do not want to experience the digital equivalent of the destruction of the Alexandria Library. Scientific assets are just too valuable to be put at risk."* [4].

Countries like Bulgaria, which still do not have a national framework for digitization of cultural and scientific heritage, are even in more danger of deeper digital divide. Not only the structured effort to digitize and preserve is missing, but there is yet another danger. The extensive brain drain causes a gap in the community of those who could work on content provision, the experts expected not only to take care of digitisation of cultural and scientific material, but also to place it in the local and wider European context. The experienced researchers

which are still active in their profession do not have to whom to transfer their knowledge because local research career is not attractive to the young generation.

2 The Current Situation in Bulgaria

As it was already mentioned, the concerns communicated on the top EC level are not exactly matching the local Bulgarian situation. However, common and more cost-effective solutions, cooperation in standards and practical work, and care to avoid duplication should be strictly followed in a country with a population of 7, 2 million which hosts over 5 million cultural and scientific heritage objects.

Another important issue of the cultural and scientific heritage institutions nowadays for Bulgaria is the adoption of brand-new IT applications in the sector. Especially new and emerging technologies which are presented in [5] if used at all in Bulgaria appear just in small demonstration projects. Amongst the current problems in Bulgaria, we should mention:

- The absence of a *national strategy*, which leads to lack of co-ordination between separate local initiatives;
- The lack of understanding and practical solutions on importance of such issues as *common quality standards* and *interoperability*;
- The gaps in the local laws and *legislative regulations* related to digitization lead to difficulties for the decision makers in the cultural and scientific heritage sector institutions;
- The need for better *co-operation on regional and European level*, since most of the cultural heritage is one we all share;
- The ambiguity of legal *copyright issues* which leads to serious problems in persuading researchers to share their knowledge in digitization projects affecting the level of presentation of materials, and restricting the depth of presentation. Copyright issues are related to the primary sources on the one hand; on the other hand the issues of legally using the results of research work during digitization are completely unclear.

If we would have to summarize the current situation in Bulgaria, we could draw the following basic conclusions:

- Bulgarian collections are of European importance but they still are not accessible in electronic form;
- Experience exists basically in the pre-digitisation stages of work such as cataloguing, and text encoding, but mass digitisation projects are just about to start in several libraries;
- Digitisation work per se has not been done, thus the country does not match current EC priorities;
- No regular governmental programme (respectively, funding) is available, digitization in Bulgaria strongly depends on external financial support. The Ministry of Education and Science had one stand-alone call for projects on cultural heritage in 2006. Through it several libraries in Bulgaria currently start their own digitisation projects which are not interoperable. Currently, the State Agency for Information and Communication Technology is working on a national strategy for the accelerated development of information society in Bulgaria in 2007-2010 and digitization is included amongst the priority areas as a general topic. This is a positive sign that the field of work becomes officially recognised, but this is not sufficient for the success of the efforts of various institutions;
- Regional cooperation in the field is realistic. SEEDI (South Eastern Europe Digitization Initiative, [6]) is a joint effort to develop awareness about digitization of cultural and scientific heritage in the region along the Lund Principles of the European Union. It is based on the acceptance that researchers and institutions from the region face common problems and share common scientific and cultural heritage, which still cannot be widely accessed in electronic form. The cooperation within SEEDI is bringing together researchers from regional and European centres with similar scientific and practical interest in digitization and by supporting cooperation between them. For that purpose core groups of specialists are created in order to consult, assist, monitor and develop innovative technologies and digitization projects in collaboration with the local cultural and scientific heritage institutions.

Thus, currently there is not only a niche but an urgent need for several interconnected efforts: creating a national framework, boosting wide-scale digitization work, promoting cooperation of local institutions and improving the excellence in the profession.

3 The Experience of the Digitization of Scientific Heritage Department

In 2004, the Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences created the Digitization of Scientific Heritage department which hosts the first and still unique in the country Bulgarian digitization line for scanning books and archival documents. The department has set the ambitious goal not only to do research work but to serve as a competence centre, providing following basic activities:

- assistance to technology and content providers;
- development of state of the art workflows and best practices in various areas of digitisation of scientific and cultural heritage;
- implementation of new technologies related to the digitization of cultural and scientific heritage;
- organization of specialised trainings;
- active participation at the elaboration of a National strategy for digitization of cultural and scientific heritage;
- contribution to the international cooperation within regional and European initiatives;
- representation of the country at international fora;
- methodological guidance and practical implementation within participation in various projects.

4 Current Activities

4.1 Cooperation with the General Department of Archives

The Digitisation of Scientific Heritage department cooperates with the General Department of Archives at the Council of Ministers of Bulgaria (GDA) since 2004. GDA is contributing with defining the priorities for selecting materials for digitization; developing the strategy for preparing descriptions and metadata; preparing specification of the search tools and their future improvement. The Digitization of Scientific Heritage department is contributing with providing its know-how for scanning and optical character recognition, workflow choice, digital image processing and with ensuring the necessary equipment and qualified personnel.

The selection and preparation of documents for digitization (single documents, parts of the archival funds and complete archival funds) is based on the holdings of the Regional Unit “State Archive” – Sofia and includes interesting materials, related to the management of Sofia Municipality, the history of Sofia University, the archives of the former Bulgarian communist party, archival funds of the Monarchy Institute, The Parliament, the Council of Ministers, etc. Selected materials contain valuable manuscripts and printed documents, photographs, sketches, geographical maps and rare books, etc.

The joint work has already brought practical outcomes. In March 2005 both institutions released a multimedia disk “Sofia. Religious spaces”, containing items displayed during an exhibition of the same name, and including scanned documents, digital copies of canonical and dogmatic books, and photographs of paintings, ritual clothes and cult objects.

Recently GDA and the Digitization of Scientific Heritage department started another joint project aimed at the electronic publishing of archive documents related to the Temporary Russian Governance which was established after the liberation of Bulgaria and ruled in the period 1878–1879. This collection of documents is being prepared as a combination of digitised images and full text. It will be organised as a semantic web portal. Scientists and general public will benefit from the availability of full-text transcriptions and tools for semantic search of the archival documents, mostly hand-written sources in Bulgarian and Russian.

Another ongoing effort is aimed at building an electronic archive of documents issued by the Bulgarian Ministry of Education in the 40ies and 50ies of the 20th century [7]. The department provides the methodological guidance in this project. The collection of documents is stored in the Archive of the Ministry of the People’s Education within the State Archival Fund of the General Department of Archives. This is a mixed collection which contains quite diverse documents - official documentation which follows specific templates; letters; notes, certificates; photographs; newspapers, etc. The text documents are printed, typewritten or handwritten. The basic aim of this work is to provide access to different users (specialists in education, historians, and the citizens) to the educational documentation of this historical period. The long-term goal is to build a joint collection of such documents from Bulgaria and Greece.

4.2 Joint Work with the Scientific Archive of the Bulgarian Academy of Sciences

Both institutions cooperate since 2005. The Scientific Archive of the Bulgarian Academy of Sciences stores precious documents, related to the history of the Academy. A pilot project was initiated in December 2006, aimed at the digitization of personal archives of famous Bulgarian scientists. As a kick-off both institutions selected the archive of Marin Drinov, one of the founders of the Academy of Sciences. It contains valuable documents, letters, personal notes and pictures. All of them were digitized and then prepared for electronic publishing. Thus scientific archives will become easily accessible both for the wide public and the researchers.

4.3 Involvement in the Digsaw Project

DIGMAP (Discovering our Past World with Digitised Maps) is a project which is supported through the eCONTENTplus programme [8]. It proposes to develop solutions for geo-referenced digital libraries, especially focused on historical materials and in the promoting of our cultural scientific heritage. The final results of the project will consist in a set of service available in the Internet, and in open-source software solutions that will be able to be reused in other services. The main service will be a specialized digital library, reusing metadata from European national libraries, to provide discovery and access to contents. Also, relevant metadata from third party sources will be reused, as also descriptions and references to any other relevant external resource. Ultimately, DIGMAP will pursue the purpose to become the main international information source and reference service for old maps and related bibliography. DIGMAP will develop solutions for georeferenced digital libraries, especially focused on historical materials and in the promoting of our cultural and scientific heritage. The final results of the project will consist in a set of services available in the Internet, and in reusable open-source software solutions.

The project will make a proof of concept reusing and enriching the contents from the **National Library of Portugal (BNP)**, the **Royal Library of Belgium (KBR/BRB)**, the **National Library of Italy in Florence (BNCF)**, and the **National Library of Estonia (NLE)**. In a second phase, that will be complemented with contents and references from other libraries, archives and information sources, namely from other European national libraries members of TEL – The European Library [9]. DIGMAP might become an effective service integrated with TEL - in this sense the project is fully aligned with the vision “**European Digital Library**” as expressed in the “i2010 digital libraries” initiative of the European Commission.

The technology will be developed by the Department of Information Systems and Computer Engineering of the **Instituto Superior Técnico — Lisbon (IST)**, in cooperation with the Group MERCATOR of the **Polytechnic University of Madrid (UPM)**. The project started in October 2006, and will have the duration of 24 months. The project coordinator is Prof. José Borbinha, of the IST. The technical work will be co-ordinated by the IST, UPM and KBR. The evaluation of the results will be co-ordinated by the NLE. The liaisons with external entities and advising groups will be coordinated by the BNCF. The dissemination will be co-ordinated by the BNP.

The **Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences (IMI)** will provide assistance to the evaluation, liaison with the Southern East Europe, and to the dissemination. In particular, the Digitisation of Scientific Heritage department already discussed what are the available maps and books relevant to the project in the collections of the General department of archives, Central Library of the Bulgarian Academy of Sciences, and National Library ‘Ivan Vazov’ in Plovdiv. Electronic records on these objects are not available yet and the participation in this project would be a chance for exposure of local materials.

4.4 Digitisation of Bulgarian Mathematical Heritage

One of the principal activities of the department is the digitization of mathematical publications of Bulgarian mathematicians, with the aim to build BulDML – Bulgarian digital mathematical library. Following journals and documents are currently being digitised:

1. Physical-Mathematical Journal (in Bulgarian), 1958 – 1991, 1993;
2. Serdica (articles in different languages): 1975, 1995-2000 г. ;
3. Archive on the development of the Union of Bulgarian Mathematicians: Book with records of board meetings, 1905-1936. In addition to the digital images full texts are also entered. Research on obtaining photographs of the mathematicians whose names appear in the records is being done – this will allow us to offer a complete resource on the dawn of Bulgarian mathematics;

4. Full collection of books and publications of Nikola Obreshkov, a famous Bulgarian mathematician.

The works of the Department are not only limited to mathematical publications. Efforts have been made in order to enlarge the spectrum of digitization activities. As a result additional fields were covered by the Department, thus ensuring more versatility and flexibility:

4.5 Digitisation of Historical Musical Periodicals

This work aims to prepare digital copies and descriptions compatible with the Retrospective Index of Musical Periodicals (RIPM) [10]. Currently the following sources has been digitised and described:

- Gusla (1891), printed text, musical fragments, illustrations
- ASO (1934), printed text, illustrations
- Materials from the archive of Stoyan Kenderov (musicologist)

One specific difficulty is the identification of a repository which holds Bulgarian historical musical periodicals. For a variety of reasons, these publications are difficult to find.

4.6 Digitisation of Historic Newspapers

This is an effort started recently which aim is to produce collection of digitised old newspapers: Daga, Lampion, Mir, Dnes (selected issues from the period 1890-1930). This activity is a joint project with the Central Library of the Bulgarian Academy of Sciences. The idea is to offer digital images of newspaper pages as a whole, and access to the texts of the separate articles. Photographs which appear in the newspapers will also be included into a collection of images.

4.7 Electronic Records of Manuscripts

The department prepared descriptions of Old Bulgarian manuscripts in TEI conformant XML (the total number of these descriptions is 806). Manuscripts are not digitised because what to be done with them is a matter of library policies. However, the detailed description is an important preliminary activity.

4.8 Towards The Creation of a National Digitization Network

The “Digitization of Scientific Heritage department” is the initiator of the creation of a nation-wide network of institutions – museums, libraries, archives and research centers, which are intending to start mass digitization, thus avoiding the implementation of scattered non-effective small-scale projects.

There are a number of good examples of joint work with such institutions. We present here briefly only two case studies, which illustrate well the trend of creating synergies in the digitization field.

5 SWOT Analysis

The SWOT analysis is used to evaluate the *Strengths*, *Weaknesses*, *Opportunities*, and *Threats* involved in a process or more specifically in a project or in any other situation of an organization or individual requiring a decision in pursuit of an objective. It involves monitoring and analyzing the environment internal and external to the process in question. In order to provide a precise picture of the current situation we have tried to summarize all relevant conditions to the digitization of scientific and cultural heritage in Bulgaria.

Strengths	Weaknesses
<ul style="list-style-type: none"> • Experience already available. The positive influence is that some institutions already have the feeling what efforts are needed. • Good contacts with colleagues from the region and other EC countries. This is important for being in line with the current practices. • Trainings/specialists meetings done on a regular basis. The circle of specialists from the community of practice grows although this is quite slow process. • Established professional bodies. The existence of departments such as the Digitisation of scientific heritage in IMI is important since it is in contact with many institutions which will play the role of future content providers. 	<ul style="list-style-type: none"> • Lack of established and working national strategies in the field of digitization, online accessibility and preservation. This leaves all decisions on specific actions to the institutions which in fact would play the role of content providers. In most cases their ideas and vision on digitisation are quite simplified. • Strong dependence on external funding. This is in controversy with the need to set up national priorities. External funding is not reflecting the national vision on importance. • Scattered experience. The experience which exists is for small initiatives, not for large projects/programmes.
Opportunities	Threats
<ul style="list-style-type: none"> • Great amount of work to be done, space for creativity. Since digitisation and online accessibility are in the beginning here, this gives space for creative approaches and innovative solutions. • Local specifics may provide interesting cases. For organisations which seek extension of their activities to Bulgaria, the local cultural materials might be very interesting and enrich their vision on the work which they are doing. 	<ul style="list-style-type: none"> • Copyright issues. The unclearness on copyright issues and how they should be approached and solved may create tension for those who do digitisation work. • Various levels of relevant experience The vision of museums, libraries and archives differ. These institutions still do not have designated digitisation units. • Small projects, scattered efforts. The danger is that small project repeat similar efforts and choose solutions where operability is not guaranteed. • Lack of crosswalks. There is no responsible body which would collect data on the standards used in different institutions, respectively there are no crosswalks which could help to build a big shared resource. • Work in conditions where neither governmental nor institutional policies are well established. In most cases this will mean that institutions will reinvent the wheel.

Table 1: SWOT analysis: digitisation in Bulgaria

6 Conclusion

The analysis reveals the basic problems that put obstacles on the way to the mass digitization in Bulgaria. At the same time it contains as well the potential possibilities to improve the situation. On this basis we have formulated following basic tasks, that in our opinion, will boost the digitization activities in the country, and will put the overall process on European level:

- Creating a joint infrastructure for the key cultural and scientific heritage institutions work;
- Establishing a common methodological network for institutions which take care for different types of heritage;

- Finding common standards for encoding and data interchange for the locally-specific features and workflows assuring quality;
- Overcoming the practice of small scale isolated initiatives and promoting a trend to structured complementary activities;
- Introducing areas such as data protection and integrity and digital curation which are currently not used in the cultural heritage sector in Bulgaria;
- Affecting the training and educational gap in the digital preservation and access field, specialists learn from their own pitfalls, not from structured programs;
- Drawing a “map” of existing resources and expertise – this will facilitate the participation in further EU initiatives.

The cultural heritage which we have inherited from the past is quite rich – over 5 million objects in Bulgaria, comparable to its 7,2 million inhabitants. In the present this heritage is underrepresented in the digital space.

To change this, serious efforts are needed in the future. We wrote this paper with the intention to mark the common lines along which the digitisation of cultural and scientific heritage is developing in Bulgaria, and to contribute to future cooperation and exchange of experiences.

We should not forget that what happens in our country is part of the general development worldwide, which currently seems well manifested as follows:

Information technology is now so pervasive and so necessary in our society that we must find ways to effectively manage its costs and its impacts across multiple organizations. The best way to do this is to forge partnerships based on a set of common requirements that individual organizations can refine to meet specific business needs and mission priorities. In terms of implementation, this can take the form of a distributed network where organizations can draw from shared knowledge and leverage a technical infrastructure while operating independently. [11]

What is said relates completely to the digitisation of cultural and scientific heritage. And yet, there is another issue which we should not forget. The production of digital objects nowadays is a gigantic industry:

IDC estimates that the world had 185 exabytes of storage available last year and will have 601 exabytes in 2010. But the amount of stuff generated is expected to jump from 161 exabytes last year to 988 exabytes (closing in on one zettabyte) in 2010.

Chuck Hollis, vice-president of technology alliances at EMC Corp., the data-management company that sponsored the IDC research ... said the new report made him wonder whether enough is being done to save the digital data for posterity.

"Someone has to make a decision about what to store and what not," Hollis said. "How do we preserve our heritage? Who's responsible for keeping all of this stuff around so our kids can look at it, so historians can look at it? It's not clear." [12]

Such questions are raised in the countries which are already ahead in the digitisation work and whose heritage is much better exposed in the electronic space, compared to the Bulgarian one. Yet, we still should find better ways to digitise it and make it visible.

Acknowledgements

We would like to acknowledge the support of the project DIGMAP (Discovering our Past World with Digitised Maps, Programme eContentplus – Project: ECP-2005-CULT-038042)

Notes and References

- [1] The website of i2010 is http://ec.europa.eu/information_society/eeurope/i2010/index_en.htm; see also REDING V. *The role of libraries in the information society*, speech at the CENL Conference, Luxembourg, 29 September 2005. <http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/05/566&format=HTML&aged=1&language=EN&guiLanguage=en>.

- [2] *EC Recommendation of 24 August 2006 on the digitization and online accessibility of cultural material and digital preservation*
http://europa.eu.int/information_society/activities/digital_libraries/doc/recommendation/recommendation/en.pdf
- [3] Council Conclusions on the Digitisation and Online Accessibility of Cultural Material, and Digital Preservation, (2006/C 297/01, Official Journal of the European Union, 7.12.2006, 5 pp.
http://eur-lex.europa.eu/LexUriServ/site/en/oj/2006/c_297/c_29720061207en00010005.pdf
- [4] REDING V., *Scientific Information In The Digital Age: How Accessible Should Publicly Funded Research Be?*, Closing speech, Conference on Scientific Publishing in the European Research Area Access, Dissemination and Preservation in the Digital Age, Brussels, 16 February 2007.
<http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/07/90&format=HTML&aged=0&language=EN&guiLanguage=en>
- [5] ROSS, S; DONNELLY, M.; DOBREVA, M., *Core Technologies for the Cultural and Scientific Heritage Sector* (Technology Watch Report 3), European Commission, ISBN 92-894-5276-5, 2005.
- [6] Website of SEEDI, South-Eastern European Digitisation Initiative,
<http://www.ncd.matf.bg.ac.yu/seedi/>
- [7] DEVRENI-KOUTSOUKI, A., *Electronic Presentation of Bulgarian Educational Archives: an Ontology-Based Approach*. International Journal Information Theories and Knowledge 2007, 8 pp, (to appear).
- [8] DIGMAP project website
<http://digmap.eu/>
- [9] The European Library website
<http://www.theeuropeanlibrary.org/>
- [10] The Répertoire international de la presse musicale (RIPM)
<http://www.ripm.org/>
- [11] PARDO, T. et al., *Building State Government Digital Preservation Partnerships: A Capability Assessment and Planning Toolkit*. Center for Technology in Government, University at Albany, SUNY. 2005, p. 16.
http://www.ctg.albany.edu/publications/guides/digital_preservation_partnerships/digital_preservation_partnerships.pdf
- [12] http://blogs.warwick.ac.uk/hsirhan/entry/how_much_data/

Digitisation and Access to Archival Collections: A Case Study of the Sofia Municipal Government (1878-1879)

Maria Nisheva-Pavlova¹; Pavel Pavlov¹; Nikolay Markov²; Maya Nedeva²

¹ Faculty of Mathematics and Informatics, "St. Kliment Ohridski" University of Sofia
and Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
e-mail: marian@fmi.uni-sofia.bg; pavlovp@fmi.uni-sofia.bg

² General Department of Archives at the Council of Ministers of Republic of Bulgaria
5, Moskovska Str., 1000 Sofia, Bulgaria
e-mail: sofia@archives.government.bg

Abstract

The paper presents in brief a project aimed at the development of a methodology and corresponding software tools intended for building of proper environments giving up means for semantics oriented, web-based access to heterogeneous multilingual archival collections. Some widespread international encoding standards for archival description and for representation of structured electronic versions of various kinds of documents have been used. An analysis is made on the applicability of appropriate Semantic web methods and technologies in order to provide versatile, user-friendly access to archival collections based on the semantics of their contents. Some practical results concerning the digitisation of a collection of archival documents from the period of the organization of the Sofia Municipal Government (1878 – 1879) and the development of a website presenting this collection are described in the paper.

Keywords: digitisation; metadata encoding; ontology; semantic annotation

1 Introduction

Recently Computer Science and information technologies play an important role in numerous successful projects directed to digital preservation of collections of handwritten, typewritten and printed archival documents, photographs etc. which are considered as significant scientific or cultural heritage.

In particular, there is an increasing number of electronic publications of archival collections which are of interest to narrow domain specialists (archivists, historians, linguists etc.) and to the general citizen [1, 2]. However, all these electronic publications give the user access tools oriented to the "standard" archivist's point of view: it is only possible to browse the full archival structure traditional for the particular country, so the search of documents is very difficult and the given search means are too limited.

The paper presents an ongoing project aimed at the development of a methodology and corresponding software tools intended for building of proper environments giving up means for semantics oriented, web-based access to distributed digitised archival collections. Moreover, we suppose that these collections are heterogeneous, i.e. they may include diverse types of materials (official handwritten, typewritten or printed documents, letters, photographs, newspapers, maps etc.) and the texts of the documents within them may be written in different languages. The practical experiments have been performed on a collection of archival documents from the period of the organization of the Sofia Municipal Government (1878 – 1879).

International encoding standards as well as Semantic web methods and technologies have been used. The main difference with other similar projects is in the exploration of the idea that the usage of proper general-purpose and domain-specific ontologies can minimize the resources necessary for the development of tools for adequate, semantics oriented access to heterogeneous (including distributed) multilingual archival collections. More precisely, the project has the following main objectives:

- To define suitable metadata to accompany digitised documents from archival collections in accordance with the international standards, the Bulgarian traditional experience and the needs of the target groups of users;

- To study the various aspects of creation of an appropriate ontology for the mentioned collection (e.g. the scope of the ontology, the corresponding linguistic problems etc.);
- To explore the necessities of the typical users of the discussed archival collection (experts in various domains and general public) in order to give proper kinds of access to this collection. In particular, providing versatile, user-friendly access to the collection based on the semantics of its content;
- To develop a framework (that will be intended for users who are professional archivists) for application of Semantic Web methods and technologies to digitised collections of archival documents.

2 Representation of the Archival Documents

In this paper we present an ongoing effort aimed at creating an electronic version of an archival collection which consists of approximately 980 original handwritten documents from the period of the establishment of civic authorities of Sofia, building the administrative system, the order and law authorities, communal health services and educational system etc. around and after the end of the Russo-Turkish war (1877 – 1878). This is the period when the building of the fundamentals of the Bulgarian state and municipal institutions has been initiated and the basic rules of the contemporary Bulgarian language have yet to be drawn up. Thus the documents within the collection are of great scientific, historical and social value and are of interest to archivists, historians, linguists etc. Because of these reasons we consider it expedient to include in the electronic version of our collection not only digital images of the chosen archival documents but also structured electronic transcriptions of their full texts and proper descriptions of the collection as a whole as well as descriptions of its parts (known as archival units) and all particular documents in it.

2.1 Description of the Structural Parts of the Archival Collection

The discussed descriptions have been prepared in conformity with the traditional practice of Bulgarian archivists. The structure of Bulgarian archives consists of four levels of hierarchy: archival funds, inventory lists, archival units and individual documents. The descriptions at all levels have been structured and accompanied with proper sets of metadata according to the requirements of the EAD encoding scheme [3].

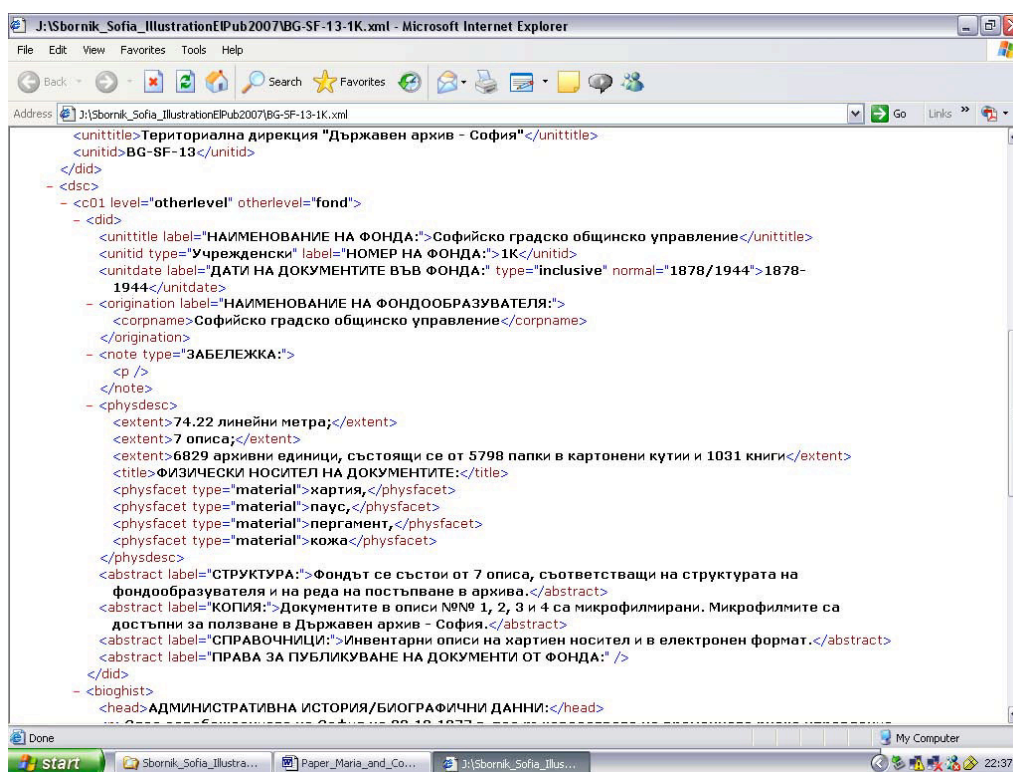


Figure 1: Part of the description of archival fund 1K according to EAD standard

EAD (Encoded Archival Description) is an encoding standard for archival description created and used by archivists to structure and exchange electronic records containing metadata about archival collections. EAD

provides a proper framework for seeing particular archive documents in relation to the whole archive collection. Through the use of multiple levels of description the collection can first be described as a whole and then as smaller parts, which get more specific at each level, until at the lowest level the individual archive documents are described.

For example, the EAD – compliant description of an archival fund contains data about the type of the fund, the dates (starting and final years) of creation of its documents, its logical structure and physical extent, the genre(s) and language(s) of its documents, the substances, technologies and methods of creation of documents and other materials in it as well as some short information about the administrative history of the corresponding corporate body, the history of the fund etc. Fig. 1 shows a part of the description of archival fund 1K (part of which is the discussed collection) according to EAD standard.

2.2 Representation of Electronic Transcriptions of Full Texts of Archival Documents

As it was already mentioned, we maintain two different digital forms of each original archive document: its digital image (in PDF format since this is the most convenient way to have exactly one file containing the image of each particular document independently from the number of the pages of the document) and an electronic transcription of its full text (in XML format). The digital images of the original documents are intended mainly for visualization purposes while the electronic transcriptions of the documents and their EAD encoded descriptions will be used to support various types of search and document retrieval activities. For the representation of the structured electronic transcriptions of the full texts of archival documents we use the TEI standard [4].

The Text Encoding Initiative (TEI) may be considered as an established standard for encoding of structured electronic versions of various kinds of documents. The TEI is a flexible encoding framework for electronic documents, allowing the content of the documents to be presented to users in a variety of ways.

We explored the structure and the contents of the various kinds of documents within the collection (instructions, orders, reports, records of sessions, letters, requests, petitions etc.) and created a generalized model of these documents. A proper set of elements and attributes from the TEI document type definition was adopted to describe this model.

The *type* attribute of <teiHeader> defines the class of the corresponding document: instruction or order (“предписание” in the “official” Bulgarian language typical for that historical period, as shown in fig. 2), record of session, request, petition etc.

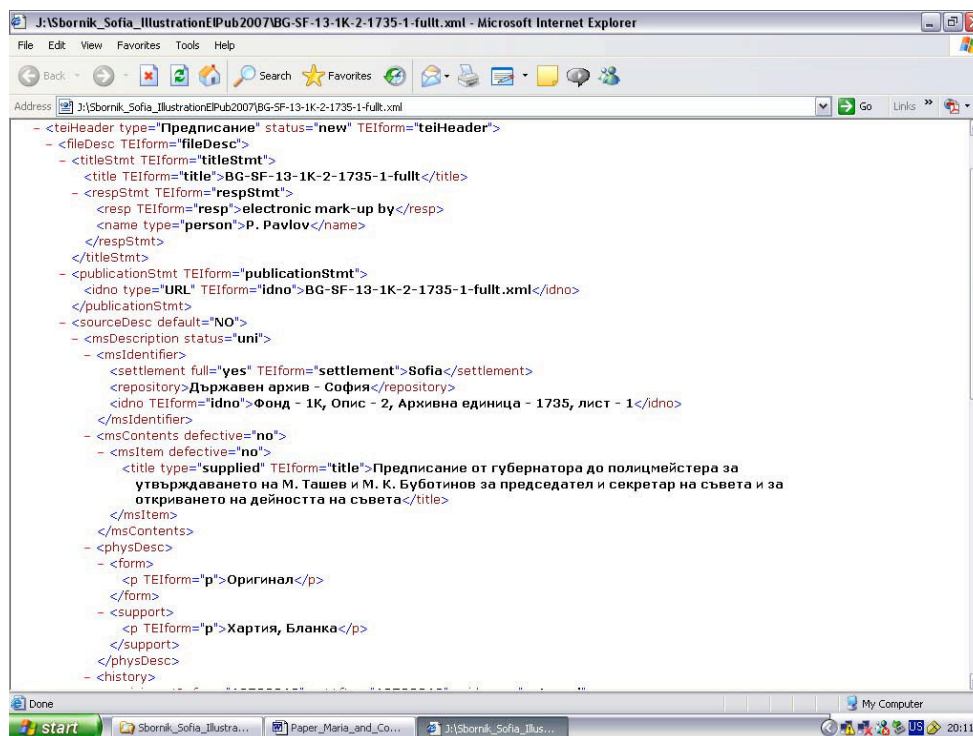


Figure 2: Part of the <teiHeader> element of the electronic transcription of an archival document

The <fileDesc> element (fig. 2) contains a bibliographic description of the computer file which represents the structured electronic transcription of the document as well as some corresponding information about the document in itself. In particular, the nested elements of <fileDesc> provide information concerning the publication place of the electronic version of the document, the name(s) of the person(s) responsible for the creation of the electronic transcript, the original or supplied name of the document, some physical description of the document etc.

The <profileDesc> element provides a description of the non-bibliographic aspects of the document, e.g. the situation in which it was produced and the persons who contributed to its creation (with their names, positions and ranks). Here we should especially mention the nested element <keywords> which is intended to play the role of a semantic annotation of the corresponding document (fig. 3). This semantic annotation has been used for document retrieval purposes as shown in Section 3.

The <body> element contains the main text of the document divided into separate divisions of different types. Each division consists of a set of paragraphs formatted in the same way as the corresponding paragraphs in the original document. A division of type “doc” includes a part of the text of the document in itself. Divisions of type “decision” (fig. 3) contain the resolutions made on the document by the corresponding official (e.g. the mayor or the president of the city council). Two other types of divisions (“execution” and “note”) contain some notes concerning the state of the accomplishment of the decisions or orders formulated in the document, some additional instructions to other persons or officials responsible for the execution of the resolutions etc.

```

<fileDesc TEIform="fileDesc">
  <profileDesc TEIform="profileDesc">
    <listPerson default="NO" TEIform="particDesc">
      <person sex="m">
        <persName TEIform="persName">Алабин</persName>
        <occupation evidence="external">И. д. Губернатора</occupation>
      </person>
    </listPerson>
  </profileDesc>
  <textClass default="NO" TEIform="textClass">
    <keywords TEIform="keywords">
      <list type="simple" TEIform="list">
        <item TEIform="item">общинско самоуправление</item>
      </list>
    </keywords>
  </textClass>
</fileDesc>
</teiHeader>
<text TEIform="text">
  <body TEIform="body">
    <div type="head" sample="complete" part="N" org="uniform" TEIform="div">
      <p TEIform="p">№ 146</p>
    </div>
    <div type="doc" sample="complete" part="N" org="uniform" TEIform="div">
      <p TEIform="p">Утвърдив Председателю Городского Совета М. Ташева и секретарем онаго М. К. Буботинова уведомляю о сем Ваше Высокоблагородие предлагаю с завтрашнего числа открыть действие Софийского Городского Совета.</p>
    </div>
    <div type="decision" sample="complete" part="N" org="uniform" TEIform="div">
      <p TEIform="p">[На гърба резолюция № 5/10.02.1878:] Настоящее предписание Г. Гражданского Губернатора препровождаю в Городской Совет [известить] Г. Ташева и Буботинова для сведения и для открытия с завтрашнего 11 февраля действие в Гор[одском] Совете. Полицимейстеръ в Шт Кап Пауль</p>
    </div>
  </body>
</text>
</TEI.2>

```

Figure 3: TEI – conformant representation of the text of an archival document

There is a minimal overlap between the metadata held in the EAD encoded descriptions and the TEI encoded document transcriptions. Our experience in the implementation of the project indicates that this overlap causes no serious problems.

3 Access to the Collection

The final version of the discussed project will give the user the opportunity to switch between two types of interface to the chosen collection. The first one is based on the principles of the “standard” archivist’s view to an archival collection. The second type of provided on-line access to the collection may be described as the semantics oriented one. Fig. 4 shows a screenshot of the current version of the homepage carrying into effect the indicated types of access to the discussed collection.

The interface to the archival collection oriented to the standard archivist's point of view allows the user to browse the hierarchical structure of the collection as a whole (fig. 5). At the archival fund and inventory list levels the user has an access to the EAD encoded description of the corresponding unit (in XML format) and to a properly visualized form of the same metadata (in PDF format).

The user interface at archival unit level allows one to browse five different forms of each particular document in the corresponding archival unit (fig. 6): the EAD encoded description of the document (in XML format), a proper visualization of this description (in PDF format), the TEI encoded electronic transcription of the full text of the document (in XML format), a proper visualization of the electronic transcription of the document (in PDF format) and a digital image of the original document (again in PDF format). Short historical data accompany this type of interface to the collection.

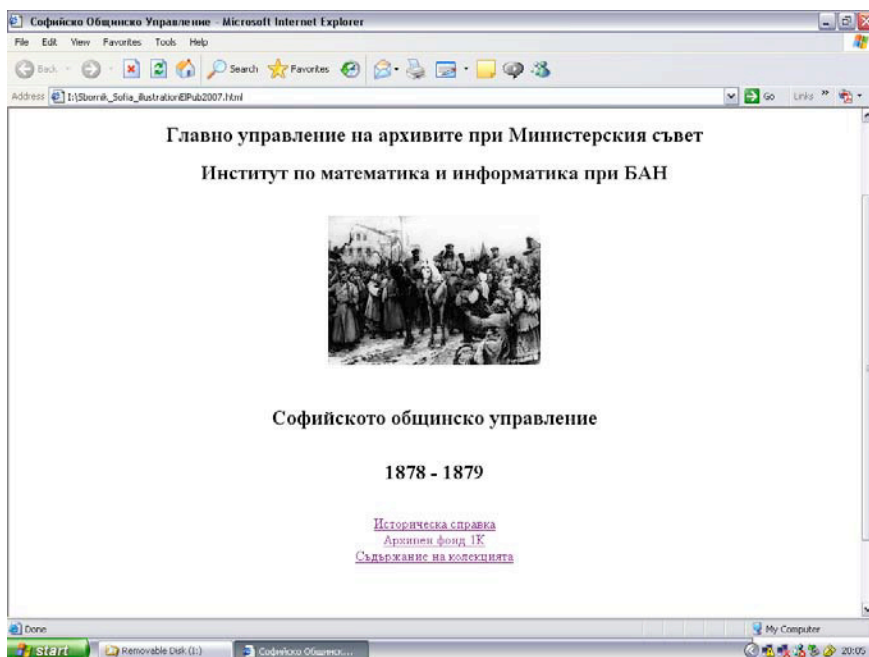


Figure 4: The homepage providing various types of access to the collection

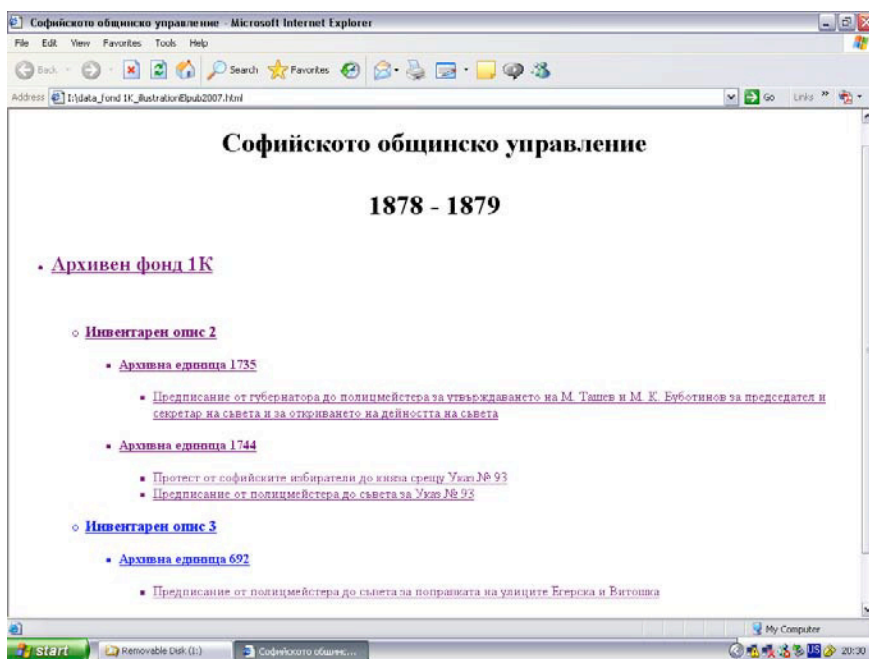


Figure 5: Interface to the collection supporting the standard archivist's view (at archival fund level)

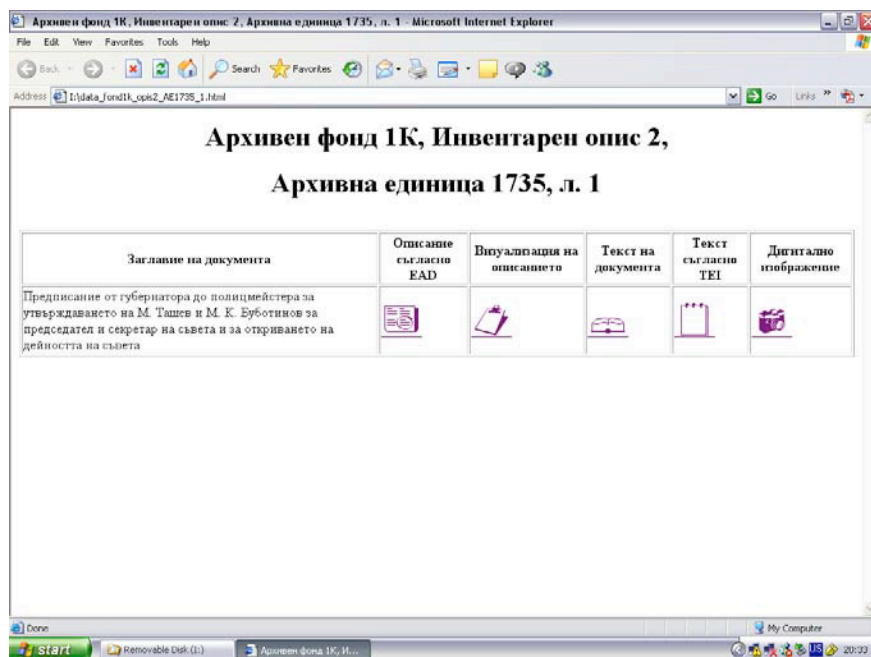


Figure 6: Interface to the collection supporting the standard archivist's view (at archival unit level)

The other type of provided access to the discussed archival collection is based on the use of explicitly represented knowledge describing different aspects of the semantics of the collection as a whole and its structural parts. A set of access tools (often called “finding aids”) realizing various types of document search and retrieval (chronological, oriented to the kinds of documents within the collection, subject oriented etc.) has been under development for the purpose. The search engines of most of these tools use the values of the corresponding elements of the TEI encoded versions of archival documents. In particular, the subject oriented document retrieval is based on the use of the semantic annotation of the documents. The semantic annotation consists of appropriate words and phrases (chosen from a subject ontology) which describe the content of the document.

Recent Artificial Intelligence textbooks define an ontology as “a shared and common understanding of some domain that can be communicated across people and computers” [5-7]. According to [5], “an ontology can be defined as a formal, explicit specification of a shared conceptualization”. Ontologies can therefore be shared and reused among different applications. Moreover, there are at least five serious reasons to create ontologies [8]:

- to share common understanding of the structure of information among people and software agents;
- to enable reuse of domain knowledge;
- to make domain assumptions explicit;
- to separate domain knowledge from the operational knowledge;
- to analyze domain knowledge.

The development of ontologies is still a difficult task, because so far there are no common platforms and verified methods which would prescribe what procedures should be followed in the process of creating an ontology. Nevertheless, there are some reasons to expect that the situation may change in the near future. First, one can find some well-defined principles for design and implementation of ontologies [5]. Second, there is a number of libraries containing already created ontologies (see e.g. [9]) and some of them could be used in the development of new domain-oriented ontologies as examples of good practice. In any case the existence of proper subject ontologies may significantly increase the effectiveness of the implementation of semantics oriented access tools.

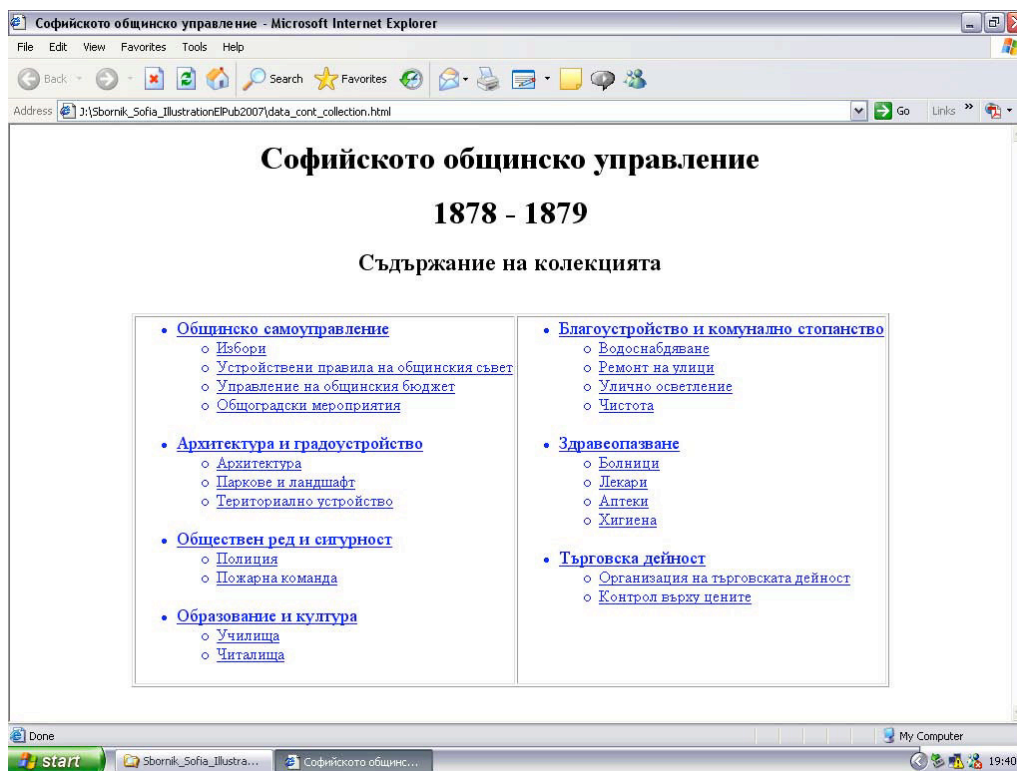


Figure 7: Interface to the collection supporting the subject oriented document retrieval

In the discussed project we use a subject ontology (covering the main types of municipal activities) especially developed for the purpose. This ontology is prepared using Protégé/OWL [2, 10].

Fig. 7 shows a screenshot presenting the web page which supports the subject oriented document retrieval in the current version of our project. The topics viewed on the screen belong to a subset of the concepts at the highest two levels of the mentioned ontology based on the assumption for typical requests for information according to the characteristics of the discussed historical period. Our future plans include some ideas to generate automatically the list of searchable topics using the results of the preliminary examination of the professional needs of the main groups of potential users.

On the other hand, the semantic annotations of the documents within the collection contain proper terms from all levels of the same ontology. When the user chooses a topic from the list shown on fig. 7, the corresponding access tool finds all documents which contain in their semantic annotations terms matching the user query (i.e. identical to the term chosen by the user or semantically related with it).

A tool for search in the full texts of the document transcriptions is provided as well. We intend to use in its implementation some of our former results concerning the development of tools for knowledge based (ontology driven) search in collections of digitised manuscripts [11]. The main idea here is similar to but more sophisticated than the one discussed above. When the user defines his query, the search engine augments it by words and phrases semantically related to these used in the original query (according to a set of available proper ontologies). Then the obtained new query is augmented once more using some synonyms of the main terms and the corresponding terms in Russian or French language (depending on the language in which each particular document is written) from a set of appropriate dictionaries. The final form of the query is processed in a standard way. As a result of the user query processing, the texts of all documents in the collection containing words or phrases semantically related to the one given by the user are properly visualized. The discovered parts of the text matching the concept(s) given as a user query are highlighted.

More complex user queries in the form of conjunctions or disjunctions of “atomic” ones will be processed as well. Some ideas already implemented in our former work [11] will be used for the purpose.

4 Conclusion

In this paper we presented a work in progress directed to the exploration of some open questions concerning the development of proper mechanisms and tools providing adequate web-based access to digitised archival collections. The most valuable expected results of our project could be formulated as follows:

- A methodology for application of international standards, ontological knowledge and Semantic web technologies for the development of software tools providing semantics oriented access to heterogeneous multilingual collections of archival documents;
- A model and a prototype of a website which gives the users an interface supporting various types of access to a chosen archival collection.

The analysis of the current results of the implementation of the project gives us a reason to believe that its final version will be compatible with and even more sophisticated in certain aspects than some popular projects like the BAMCO site [1], the LEADERS project [12] etc. The main advantage of our approach is the proper use of ontological knowledge describing the semantics of the individual documents in the archival collection as well as the semantics of the collection as a whole and the semantics of its structural parts. It allows users with different profiles to study and analyze the documents within the corresponding collection from multiple points of view using a single environment for the purpose.

Acknowledgements

This work has been funded by the EC FP6 Project “Knowledge Transfer for Digitisation of Cultural and Scientific Heritage in Bulgaria” (KT-DigiCULT-BG) coordinated by the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences. The authors are thankful to Dr. Matthew Driscoll from Copenhagen University, Denmark, for the useful advices concerning the TEI encoding of the electronic copies of archival documents.

References

- [1] Brown Archival & Manuscript Collections Online (BAMCO) site. <http://dl.lib.brown.edu/bamco/introcontent.html>, last accessed on April 4, 2007.
- [2] KNUBLAUCH, H.; FERGERSON, R.; NOY, N; MUSEN, M. The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. Third International Semantic Web Conference, Hiroshima, Japan, 2004.
- [3] The Encoded Archival Description (EAD). <http://www.loc.gov/ead/>, last accessed on April 4, 2007.
- [4] The Text Encoding Initiative (TEI). <http://www.tei-c.org/>, last accessed on April 4, 2007.
- [5] Gruber, T. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human-Computer Studies*, Vol. 43, 1995, pp. 907-928.
- [6] GUARINO, N. Formal Ontology, Conceptual Analysis and Knowledge Representation. *International Journal of Human-Computer Studies*, Vol. 43, 1995, pp. 625-640.
- [7] BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. *Scientific American*, May 2001, pp. 35-43. <http://www.w3.org/2001/sw/>, last accessed on April 4, 2007.
- [8] NOY, N.; MCGUINNESS, D. Ontology Development 101: A Guide to Creating Your First Ontology. http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html, last accessed on April 4, 2007.
- [9] Protégé Ontologies Library. <http://protege.cim3.net/cgi-bin/wiki.pl?ProtegeOntologiesLibrary>, last accessed on April 4, 2007.
- [10] OWL Web Ontology Language Overview. W3C Recommendation, 10 February 2004. <http://www.w3.org/TR/owl-features/>, last accessed on April 4, 2007.
- [11] PAVLOV, P.; NISHEVA-PAVLOVA, M. Knowledge-based Search in Collections of Digitized Manuscripts: First Results. *Proceedings of the 10th ICC International Conference on Electronic Publishing (Bansko, 14-16 June 2006), FOI-COMMERCE, Sofia, 2006*, pp. 27-35.
- [12] LEADERS: Linking EAD to Electronically Retrievable Sources. <http://www.ucl.ac.uk/leaders-project/>, last accessed on April 4, 2007.

The Digital Scholar's Workbench

Ian Barnes

Department of Computer Science, The Australian National University, Canberra, ACT Australia
Ian.Barnes@anu.edu.au

Abstract

In this paper I present the reasoning behind the development of a new end-to-end publishing system for academic writers. The story starts with investigating digital preservation of word processing documents. What file formats are suitable for long-term preservation of text? I believe that the answer is a high-quality structured XML format like DocBook XML or TEI. The next question is how do we get word processing documents into that format without incurring a prohibitive cost? Conversion is possible, but it requires human intervention at some point. It would be far too expensive to have archivists editing every document by hand on ingest, so how can we get authors to do the necessary work, particularly as most academics aren't at all interested in digital preservation of their work? The answer I propose is to offer them more than just an archiving solution. Instead of just getting preservation, they get a full end-to-end digital publishing solution, the digital scholar's workbench, tailored to their needs for document interoperability, collaboration and publication in multiple formats... oh, and they get preservation too.

Keywords: word processing; scholarly communication; digital preservation; XML; electronic publishing

1 Introduction

Word processing documents are a problem for digital repositories. They are not suitable for long-term storage, so they need to be converted into an archival format for preservation. This is not just a technical issue, but a people issue also. Most university repositories struggle to get academics to deposit their work: making the process more difficult will only make this problem worse. On the other hand, filling a repository with word processor documents that nobody will be able to read in a few years time, is a waste of time, effort and money. In this paper I will address the following questions:

1. What file formats are suitable for long-term storage of word processed text documents?
2. How can we convert documents into a suitable archival format?
3. How can we get authors to convert and deposit their work?

The answer I will propose is an end-to-end digital publishing system that allows authors to continue writing in a word processor, much as they do now, but provides an attractive range of file format conversion and publishing services: the Digital Scholar's Workbench.

While the vast majority of material generated by universities is text, most research on digital preservation concentrates on images, sound recordings, video and multimedia. You could be forgiven for thinking that this is because text is simple, but unfortunately that's not so. Even relatively short text documents (like this one) have complex structure consisting of sections (parts, chapters, subsections etc) and also of indented structures like lists and blockquotes. A significant part of the meaning is lost if that structure is ignored (for example by saving as plain text). In Section 2 I will briefly review some of the previous work in this area.

Most text documents created today are created in a word processor. In Section 3 I will discuss file formats, and give a tentative answer to Question 1 above. The file formats generated by word processors are generally not sustainable, so we need to consider converting documents to better formats. Many archives have chosen PDF, but this has serious problems. XML is a better answer, but it's not a complete answer as XML is not a file format, but a meta-format, a framework for creating file formats. We have to choose a suitable XML file format for storing documents.

In Section 4 I will discuss possible methods available for converting word processing documents into a suitable XML format, addressing Question 2 above.

In Section 5 I give a description of my own current work in progress, the Digital Scholar's Workbench, a web application designed to solve some of the problems with preservation and interoperability of word processing documents.

2 Previous Work

There is a lot of published research on digital preservation, but not much of it that I found deals in any detail with preservation of *text*. The DiVA people in Uppsala University Library are archiving documents in XML [1]. They use a custom format which is basically DocBook XML for describing the document itself (content and structure), with a wrapper around the outside allowing for collections of related documents and for comprehensive metadata. At ANU we are considering something similar for large or complex documents, using RDF to describe the relationships between the parts. Slats [2] discusses requirements for preservation of text documents, and the relative merits of XML and PDF. Like several other authors with similar publications, she recommends storing documents in XML, but fails to specify *what* XML format to choose. Anderson et al [3] from Stanford recommend ensuring that documents are created in a sustainable format rather than attempting conversion and preservation later, as I will recommend below. This leaves open the question of what to do with existing documents.

3 File Formats

3.1 Preservation Formats Versus Access Formats

A *preservation format* is one suitable for storing a document in an electronic archive for a long period. An *access format* is one suitable for viewing a document or doing something with it. Note that it may well be the case that no-one ever views the document in its preservation format. Instead, the archive provides on-the-fly conversion into one or more access formats when someone asks for it. For example, the strategy I recommend below is to store DocBook XML or TEI, but convert the document to HTML for online viewing or PDF for printing. Some file formats may be suitable for both purposes. XHTML has been suggested, with CSS for display formatting. As XHTML is XML (and particularly if the markup is made rich with use of the `<div>` element to indicate structure), it may be an adequate preservation format, at least for simple documents. As it can be viewed directly in a web browser, it is eminently suitable as an access format.

3.2 Criteria for Sustainability

What features does a good preservation format have? How do we judge? Lesk [4] gives a list of required features for preservation formats (The points in italics are his, the comments that follow are mine.):

1. *Content-level, not presentation-level descriptions*. In other words, structural markup, not formatting.
2. *Ample comment space*. Formats that allow rich metadata satisfy this requirement.
3. *Open availability*. This means that proprietary formats are not acceptable. Remember what happened to GIF images when Unisys claimed that they were owed royalties because they own the file format [5]. What would happen if Adobe decided to do the same with PDF or Microsoft with Word?
4. *Interpretability*. It should be possible for a human to read the data, and also for small errors in storage or transmission to remain localised. This implies a strong preference for plain text rather than binary file formats. A small error in a compressed binary file can render the entire file useless.

Stanescu [6] looks at this topic from a risk management point of view. Slats [2] discusses criteria for choosing file formats, coming to very similar conclusions.

3.3 Word Processing Formats

Microsoft Word

The vast majority of all text documents created today are created in Microsoft Word using its native `.doc` format. It would be great if we could just deposit Microsoft Word documents into repositories and be done with it, but unfortunately that won't do, for a few important reasons:

- Word format is owned by Microsoft corporation.
 - They could choose to change the format at any time, possibly forcing repositories to convert all their documents.

- They could change the licensing at any time.
- Word format is a *binary* format. There is no obvious way to extract the content from a Word document. If the document is corrupted even a little, the content can be lost.
- Word is not just one format but many. Storing documents in Word format would force repositories to support not one but several file formats, or alternatively to engage, every few years, in a process of opening *every* stored document in the latest version of the software, and saving it using the most recent incarnation of the format and fixing any problems. Either way, this is an unacceptable cost.

Microsoft's new Office Open XML `.docx` format [7] is an improvement, but is still unsuitable for archiving. A `.docx` file is a compressed Zip archive of XML files. Compressed files are particularly prone to major loss if corrupted. Also some data is stored as strings that need parsing [8], rather than using XML elements or attributes to separate the different data items. This makes automated processing of these files much more difficult. Microsoft have released the Office Open XML specifications publicly, along with assurances that the format is and will always be free [9]. Despite the mistrust of many in the open source community, who remember the GIF/Unisys controversy [5], this appears to be genuine.

Open Document Format

Open Document Format (ODF) [10] is the native file format of OpenOffice.org Writer [11], the word processor component of the OpenOffice.org open source office suite. Open Document Format is an OASIS and ISO standard and a European Commission recommendation. It is also supported by KOffice and AbiWord.

An ODF file is a Zip archive containing several XML files, plus images and other objects. The Zip archiving and compression tool is freely available on all major platforms, so there should never be a problem getting at the content of an ODF document. Using a Zip archive means that the files are prone to catastrophic loss of content with even minor data corruption.

If we are going to archive word processing documents, I believe that ODF is a better option than Microsoft Word format in any of its variations. Even Office Open XML is still a proprietary format.

One possible preservation strategy would be to convert all word processing documents to ODF for storage. This can be done easily using OpenOffice.org itself as a converter. The conversion could be set up as part of the repository ingest process so that it would be almost totally painless for users. Conversion to ODF preserves all the formatting of most Word documents, with only minor differences. For complex documents that use lots of floating text boxes, these minor differences can make a mess of the appearance of the document. For documents that use embedded active content (chunks from live spreadsheets etc), the embedding will probably fail. For most "normal" documents, even complex ones, the conversion is good.

The main disadvantage of this strategy is that Open Document Format is still a word processing format, not a structured document format. What does this mean, and why is it a problem?

- Word processing formats are at heart about describing the appearance of the document, not its structure. For serious processing it's the structure we want. In 20, 50 or 100 years, most readers will probably not care about the size of the paper, the margins, the fonts used and so on. Even today, if we're going to serve up a document as a web page, those details are irrelevant. Sometimes these details can even be a disadvantage, for example if the document insists on fonts that are unavailable on your computer. On the other hand, the division of the document into sections will always be relevant, useful and important, and must be preserved.
- Word processing formats are flat. That is, the document is a sequence of paragraphs and headings. What we'd really like is a deep structure with sections, subsections and so on, nested inside each other (as in DocBook or TEI). We want this deep structure because it makes structured searches and queries possible, and makes conversion with XSLT much easier. It is possible to do automated conversion from flat to deep structure [12], but at the moment only with documents that conform to a well-designed template. We are working to extend this to less carefully prepared documents, but the process is likely to require human supervision.

The other disadvantage of Open Document Format is that even for simple documents it is extremely complex. Unzipping a one-page document of about 120 words results in a collection of files totalling 300K in size. The formatting information is stored in a complex, indirect way. This makes it relatively difficult to locate the meaningful content and structure and transform it into other formats for viewing or other uses. Instead of leaving documents in this complex format and having a hard job writing converters (XSLT stylesheets) for all possible future uses, it would be better to store documents in a simple, clear, well-structured format that makes converters easier to write.

Other word processing formats

There are several, but none of them has much market share, nor do any of them have any particularly conspicuous advantages. Probably the best strategy with these is to convert them into Word or Open Document Format, then treat them in the same way as the majority of documents. OpenOffice.org will open many file formats, so it can be used as a generic first stage in any process of converting documents into useful formats. Use OpenOffice.org in server mode to open all documents and save them in Open Document Format, then process them into something better.

3.4 PDF

Many repositories have adopted PDF as their main format for text documents, both for storage and for access. PDF has some good points:

- It is easy to create, either using Adobe Acrobat software or using the PDF Export feature available in both Microsoft Word and OpenOffice.org Writer.
- It can be viewed on all platforms using the free Adobe Acrobat Reader software.
- It is extremely effective at preserving the formatting of a document. For some applications (for example in legal contexts) this may be of vital importance.

However, there are some serious problems with using PDF as a storage format [13]:

- The format is proprietary, owned by Adobe. While it is currently open, the company could decide to change this at any time.
- There are some compatibility problems between different versions.
- Documents may rely on system fonts. There is an option in PDF to embed all fonts in the document, but not all software uses this, and some PDF viewing software either cannot locate the correct fonts or doesn't know how to substitute suitable alternatives. Failing to embed all fonts can result in a serious degradation of the on-screen appearance of a document, or in a complete failure to display the content.
- PDF includes extra features like encryption, compression, digital rights management and embedding of objects from other software packages. These all present difficulties for archivists.

PDF is an excellent access format for printing to paper. Any good preservation system should be able to generate PDF renditions of documents for this purpose. PDF is not so good for viewing on screen, as it ties document content to a fixed page size. However, for the reasons given above, it is not a good preservation format.

3.5 RTF

RTF stands for Rich Text Format. It is a Microsoft specification [14], but they have published it, so one could argue that it is an open standard. It is certainly widely interoperable, with most word processors capable of reading and writing RTF. There are problems with using RTF as a preservation format:

- It is still proprietary, with all the risks that entails.
- There seem to be parts of the specification that are not in the publicly available specification document, and which have changed over the years.
- The specification is not complete and precise, leaving many little quirks.

The National Library of Australia has chosen RTF as its main preservation format [15]. I think a well-chosen XML file format has significant advantages over RTF, but it might well be worth retaining RTF as an access format, since it has good interoperability, at least for relatively simple documents.

3.6 XML

XML [16] is widely accepted as a desirable format for document preservation. See for example the assessment of XML on the US Library of Congress digital formats web site [17] and the related conference paper by Arms & Fleischauer [18]. The main reasons are:

- XML is a free, open standard.
- XML uses standard character encodings, including full support for Unicode. This makes it capable of describing almost anything in any language.

- XML is based on plain text. This gives it the best possible chance of being readable far into the future. Even if XML and XSLT are no longer available, the raw document content and markup will still be human-readable. (This will be true even if the *meaning* of the markup has been lost, although formats designed with preservation in mind should try to make the meaning apparent through the choice of element and attribute names).
- XML can easily be transformed into other formats using XSLT [19].

This last point is very important. It means that documents which are stored in XML can be viewed in multiple formats. A minimal solution would generate HTML for on-screen viewing and PDF for printing. However, just saying “XML is the answer” isn’t enough. XML is only really useful when documents conform to a standard DTD or schema. Having an XML-based preservation strategy means choosing one or more (but preferably very few) XML document formats. It also means having a workable method for converting documents into that format.

DocBook XML

DocBook [20] is a rich and mature format that has been in use for about 15 years. It was originally an SGML format designed for marking up computer documentation (particularly the O’Reilly books), but its application is wider, although it still seems a bit awkward and sometimes ill-matched to non-technical writing. DocBook is an OASIS [21] standard.

DocBook is huge, with over 300 elements. This makes it hard to learn, and cumbersome to use directly; few people create DocBook documents by hand. However this is of no concern to the ordinary author if the transformation from word processor formats to DocBook is done automatically. It *may* be a concern for the unlucky person who has to write stylesheets for converting documents to and from DocBook. Fortunately though, Norm Walsh (the guiding force behind DocBook) and others have written a comprehensive set of XSLT stylesheets [22] for converting from DocBook XML into numerous formats including XSL-FO (and hence PDF) and XHTML. This is a huge headstart.

For converting word processor files to DocBook XML, the complexity and number of elements doesn’t matter, since the conversion process will probably target only a small subset of DocBook. This is the approach I have adopted with the Digital Scholar’s Workbench.

TEI

TEI stands for the Text Encoding Initiative [23]. It is designed mostly for the preservation of literary and linguistic texts. Like DocBook, TEI is huge. Furthermore, it’s not exactly *a format*, but a set of guidelines for building more specialised formats. One such is TEI-Lite, which has proved very popular, and is used by several repositories.

TEI may be better-matched than DocBook to some scholarly work, particularly in the humanities. It does have some serious shortcomings however:

- It uses abbreviated element names like <p> for paragraph (where Docbook uses <para>). This is presumably to make it easier to key in by hand, but it is a problem for sustainability since it may make it more difficult to recover the meaning of the markup in the distant future.
- While it has a set of customisable XSLT stylesheets [24], the impression I get is that they are less mature and less comprehensive than the DocBook XSL stylesheets [22].

Whether or not the TEI XSLT stylesheets are up to the job, TEI needs to be considered as a serious candidate for a preservation format for some scholarly writing. Ideally a full solution to the preservation problem would support both DocBook and TEI, allowing authors or curators/archivists to choose the most suitable format for preserving each work (or collection of works).

XHTML+CSS

Since XHTML [25] is both valid XML and can be displayed by web browsers directly (with the formatting controlled by a CSS [26] stylesheet) this has been suggested as a possible archival format. I don’t recommend it, for the following reasons:

- XHTML is essentially a flat format, which means it’s harder to do useful conversion into other formats in the future. It’s possible to use the <div> element creatively to add lots of structure, but if you’re going to do that, you’re much better off using a well-defined structured format like DocBook or TEI.

(Why? Because in those formats the structural elements are rigorously defined, while in XHTML you can use divs however you like, making it hard for processing applications to know what to do. See the section on Custom schemata below.)

- CSS relies on consistent use of the “class” attribute in the XHTML. There is no standard for doing this. Same problem as above.
- CSS is not XML, so parsing it to convert it into some new format in the future is much harder than with XML formats.

XHTML might be a good solution for low-value documents that archivists cannot afford to convert into DocBook or TEI. In these cases a reasonable strategy might be to store the document in Open Document Format and add an automatically generated, perhaps poor quality, XHTML+CSS version for easy viewing and searching. This could either be stored in the repository alongside the ODF version, or could be generated on the fly by a front-end like Cocoon [27].

Custom schemata

One of the biggest traps in the XML world is the idea that you create your own document schemata that perfectly match your particular needs. A university could create document types for lectures, lab exercises, reading lists, research papers, internal memos, minutes of meetings, rules, policies, agendas, monographs and so on. There are serious problems with this approach [28].

The first problem is maintenance. Each format will require converters to turn word processor documents into that format, and XSLT stylesheets for rendering into whatever viewing formats are needed: a reasonable short list would be HTML, PDF and plain text. What happens next is that someone wants to add an element to one of the document types. Every time this happens, you have to modify all the stylesheets for that document type. With multiple formats, there is likely to be demand for conversion between them: converting a stored research paper into a lecture, for example. The number of conversions needed grows fast.

The second problem is loss of interoperability. One of the long-term goals of the whole repository project is that one should be able to retrieve a chunk of something from the repository, and drop it into another document of a different type. The use of custom schemata acts against these goals.

We’d also like people elsewhere to be able to use the documents that we go to so much trouble to preserve in our repository, but we can’t expect them to know all about our special document types. So then we would have to create even more converters for exporting the documents in well-known interchange formats.

4 Converting Documents Into Archival Format

Having decided on a suitable archival format, the second issue is how to convert documents into that format. In a nutshell, this is solved by using OpenOffice.org to convert multiple formats (including Microsoft Word) into Open Document Format, then unzipping the result and applying one or more XSLT transformations to the pieces.

For some parts of the processing, particularly creating deep structure from a flat sequence of headings and paragraphs, XSLT can be quite cumbersome. Direct manipulation (for example using one of the various DOM bindings: Java, Python, Perl, PHP...) is an alternative worth considering.

The only drawback of DocBook (and the same applies to TEI) is that most word processing documents do not contain enough structure information to allow for easy automated conversion. In order to convert word processor documents into DocBook (or TEI), some human effort is required:

- The best scenario is that the document was created using a well-designed word processor template, so that every paragraph has a style name attached to it. These styles can then be used as hooks by an automated conversion process in order to deduce structure. The USQ ICE project [29] is an example of this approach, as is the Digital Scholar’s Workbench (see below). One of my current interests is in the possibility of creating a heuristic “structure guesser” that can use formatting information (indents and justification, space above and below, type size, weight and style etc) to make educated guesses about the structural roles of different paragraphs in documents that were not created using a good consistent set of styles. This is unlikely to ever be a perfect hands-off process however, and will probably always require some human supervision.

- For legacy documents or authors who refuse to use a template, the word processing document will have to be edited by an electronic archivist to get it into a state where it can be converted to DocBook (or TEI). This would require trained staff, and costs time and money.
- For documents that are extremely poorly formatted, or that exist only on paper, another alternative is to send them out to be rekeyed. This is expensive, but for high-value documents or for small projects it may be worth it. A few thousand dollars for typing and marking up a book may compare well with the cost of setting up the infrastructure to do automated conversion, training staff to do the technical editing (cleaning up the markup, making it conform to a template) and so on. One important possibility worth investigating here is of having documents re-keyed in Word using a good template, and then converting to DocBook automatically. A first inquiry about this suggests that it costs roughly three times as much to mark up text in DocBook XML as it does to rekey it in Word. That means we can potentially save two-thirds of the cost if we do the conversion to DocBook with an automated process [30].

5 Usage

The main problem faced by institutional repositories (electronic archives) is no longer technical but social. Very few academics deposit their work. They're not interested, and it's too much trouble. After they finish preparing a piece of work for publication—sometimes a very time-consuming and frustrating process that can take days—they want to move on to the next piece of work. The last thing they want to do is start all over again reformatting and preparing their work for archiving, and then having to type lots of metadata for search and indexing.

I believe that the key to solving this problem is seeing archiving as just another form of publishing. Just like a journal, an archive has its technical requirements in terms of format, metadata and so on. Rather than having to go through this time-consuming and frustrating process by hand, more than once, it would be much better to create a system that can do the whole thing automatically or at least semi-automatically. This is what a good, end-to-end electronic publishing system should be able to do. Once we have a document and its associated metadata in a good, structured XML format, all this should be possible, and more including:

- Sending to a journal
- Submitting to a conference
- Depositing in an archive
- Posting to the department web site
- Posting to a personal blog
- Running off preprints to send to colleagues
- Adding the abstract to the department annual report
- Registering with research productivity measures

A system that offered all these services might be able to attract users from among the academic community. This is the challenge we are trying to meet with the Digital Scholar's Workbench. It's worth taking a few lines to discuss this in software engineering terms. What I'm proposing here looks very like "requirements creep", the process by which a simple software development project gets hopelessly bogged down by a constantly expanding list of requirements. Usually developers are instructed to get a list of requirements early in the process and then lock it down so that the size and complexity of the project can be contained.

In the case of this project, that would have defeated the purpose of the work entirely. A single-purpose piece of software that converts Word documents into DocBook XML already exists. (It is a commercial product called UpCast [31].) Apart from the expense involved in making it available across the university, the problem with this software is that very few ordinary authors will take the trouble to learn how to use it. This is because there is no incentive to do so. Academics don't care about preservation, so almost *any* effort is too much.

Some in the archiving community advocate taking an authoritarian stance and *requiring* academics to archive their work. Apart from any philosophical objections, this approach has another problem, that of quality of metadata attached to documents. If academics are forced to fill in lots of metadata forms, there will be a temptation to save time by entering rubbish. This defeats the whole purpose of archiving. If no-one can find your work, what is the point of preserving it?

Instead of forcing people to do something they don't want to do, the approach I support is to provide a tool that is so useful they will want to use it. Archiving comes for free as part of the package—once someone is using this tool, clicking a button to archive completed work is no trouble at all. Metadata can be scraped from the

document, or at worst only has to be entered once when the document is created. Then it can be used and re-used whenever the document is published or archived or submitted to a journal or conference.

The point here is that what looks like requirements creep is actually a deliberate strategy to create something that will actually be used. By surrounding the archiving functionality with features academic writers will actually want, we make it far more likely that work will be deposited in the archive.

6 The Digital Scholar's Workbench

The Digital Scholar's Workbench [32, 33] is a prototype application that implements some of the ideas in this report. At the moment the Workbench is a web application that converts suitably structured word processing documents into archival quality DocBook XML, and then from there into XHTML for onscreen viewing and into PDF for printing.

In order to work with the current version of the workbench, documents must be written using the USQ ICE template [29]. This template has a single set of all-purpose styles [34] designed so that it is possible to automate the conversion to DocBook XML. To many people, this seems like a major restriction, and a very common response is that people simply won't do that. Experience shows however [35], that authors *will* work with a template if:

- it is sufficiently rich to capture their documents without restricting them too much; and
- they can see the benefits in terms of time saved wrestling with their documents and trying to convert them into other file formats for publication.

This approach is backed up by Liegmann, who states that, "All of us ... have experienced that you need to tolerate and to adhere to a structured framework in order to profit from its advantages." [36] At the time of writing, the Digital Scholar's Workbench has the basic functionality of being able to convert word processing documents written using the USQ ICE template into DocBook XML and from there into XHTML and PDF. Further development will focus on:

- making its support for Word and Writer documents more robust, and supporting more of the features, available in the word processing software;
- improving the links to repository software, so that documents can be deposited, together with associated metadata and any linked images or other resources, with one click;
- adding one-click publishing to blogs, websites and learning management systems;
- reformatting papers ready for submission to journals and conferences, via a plugin mechanism, so that once an export plugin has been created it can be contributed back to the community for use by others;
- support for complex multi-part documents like books and theses;
- articulation with desktop publishing software to produce high-quality typeset PDF output;
- improved metadata entry and storage;
- linking with a version control system, so that authors have access to all previous versions of their documents at all times;
- round-tripping of documents back to Word or Open Document Format to enable seamless collaboration between co-authors using different word processing software;
- platform- and software-independent bibliography management (perhaps outside the limited and difficult-to-use systems built in to Word and Writer);
- adding support for TeX and LaTeX documents;
- support for presentation slides; and
- attempting to remove the requirement that documents be written using only the set of styles in the ICE template.

The prototype workbench will eventually be made available to developers and early adopters as an open source software project, probably through SourceForge [37].

The Digital Scholar's Workbench is built on open-source technology. The current version uses the Apache Cocoon [27] web application framework, which incorporates the Xalan XML parser [38], the Xerces XSLT processor [39] and the FOP XSL-FO processor [40]. It also uses OpenOffice.org in headless mode to transform Word documents into Open Document Format. It relies on the USQ ICE template. (This architecture may change over the coming months.)

7 Example/Colophon

This document was begun in OpenOffice.org Writer on Linux, using the template provided by the conference organisers. When I couldn't format the reference list correctly using the bibliographic support built in to Writer, I moved the document to Microsoft Word on my Mac, and used EndNote [41] for the bibliography formatting. At first this didn't work; something Writer had done to the document meant that Word and EndNote couldn't talk to each other. In order to fix this, I had to start again with a fresh copy of the conference template and cut and paste my content across. Within the time constraints, there appeared to be no practical way to extract my bibliographic data from Writer and import it into EndNote, so I had to enter 43 references by hand. I estimate that I spent at least one entire working day simply wrestling with my word processing and bibliographic software, rather than working on the actual content of the paper.

Imagine now that the Digital Scholar's Workbench existed as described above. None of this would have been a problem. Transferring the paper from Writer to Word would have been accomplished via the DocBook interchange format. I'm not sure how the bibliography support will work—this may well be the hardest part of the work planned, or it may be that a service like Zotero [42] can do everything required—but there would certainly be no need to rekey entries from a bibliographic database. With an appropriate output filter, the workbench could have reformatted my paper to conform to the conference template, saving me the trouble. Alternatively, the conference organisers could have accepted the paper in DocBook XML, giving them the flexibility to typeset the proceedings for publication, and turn them into a web site.

8 Conclusion

This project was originally about digital preservation only. The Australian Partnership for Sustainable Repositories wanted to know how to preserve word processing documents. Answering that question led to the need to convert documents into structured XML for preservation. From there, the question was not just "How?", but also "How will we get people to actually do this?" The prototype Digital Scholar's Workbench is an attempt to answer both questions. The first is a purely technical question, but the second one is a people question, in the realms of the sociology of knowledge production. There is probably a long way to go with this, and the form of the software will change as we get feedback from focus groups of academics. The principle is that there is little chance of getting our users to do what the university archive wants them to do—convert their work into structured XML and deposit it in the archive—without offering them something they want, namely vastly simplified workflows for their everyday writing and publishing tasks.

Acknowledgements

This work was funded by the Australian Commonwealth Department of Education, Science & Training, through the Australian Partnership for Sustainable Repositories (APSR), which is part of the government's Systemic Infrastructure Initiative, part of "Backing Australia's Ability—An Innovative Action Plan for the Future". I thank the Australian National University's Digital Resource Services program, part of the Division of Information, for hosting me while I did this work: Program Leader Peter Raftos, APSR Project Leader Adrian Burton, and my colleagues Chris Blackall, Leo Monus, Scott Yeadon and Margaret Henty. I also thank the ANU's Department of Computer Science for releasing me from teaching while I did this work. I also thank Peter Sefton and his team at the University of Southern Queensland for many useful discussions particularly relating to the USQ Integrated Content Management system (ICE) which has a lot in common with the Digital Scholar's Workbench. I also thank them for creating the ICE word processor templates which I use in the current version of the Digital Scholar's Workbench. This paper is based on, and extends, an earlier report written for and published by APSR [43].

Notes and References

- [1] MÜLLER, E.; KLOSA, U.; HANSSON, P.; ANDERSSON, S.; SIIRA, E. Using XML for long-term preservation: Experiences from the DiVA project. Sixth International Symposium on Electronic Theses and Dissertations. Berlin, 2003. URL: <http://edoc.hu-berlin.de/conferences/etd2003/hansson-peter/HTML/>
- [2] SLATS, J. Practical experiences of the digital preservation testbed: Office formats. File formats for preservation. Vienna, 2004. URL: http://www.erpanet.org/events/2004/vienna/presentations/erpaTrainingVienna_Slats.pdf
- [3] ANDERSON, R.; FROST, H.; HOEBELHEINRICH, N.; JOHNSON, K. The AIHT at Stanford University: Automated preservation assessment of heterogeneous digital collections. D-Lib Magazine 2005;11. URL: <http://dlib.org/dlib/december05/johnson/12johnson.html>
- [4] LESK, M. Preserving digital objects: Recurrent needs and challenges. 2nd NPO Conference on Multimedia Preservation. Brisbane, 1995. URL: <http://www.lesk.com/mlesk/auspres/aus.html>
- [5] GIF. Wikipedia, 2006. URL: <http://en.wikipedia.org/wiki/GIF>
- [6] STANESCU, A. Assessing the durability of formats in a digital preservation environment. D-Lib Magazine 2004;10. URL: <http://dlib.org/dlib/november04/stanescu/11stanescu.html>
- [7] Microsoft Office Open XML formats overview. Microsoft, 2005-6. URL: <http://www.microsoft.com/office/preview/itpro/fileoverview.mspx>
- [8] D'ARCUS, B. Citations in "Open" XML. 2006. URL: <http://netapps.muohio.edu/blogs/darcusb/darcusb/archives/2006/06/08/citations-in-open-xml>
- [9] Ecma international standardization of OpenXML file formats frequently asked questions. Microsoft, 2006. URL: <http://www.microsoft.com/office/preview/itpro/ecmafaq.mspx>
- [10] OpenDocument. Wikipedia, 2006. URL: http://en.wikipedia.org/wiki/Open_document_format
- [11] Writer. OpenOffice.org, 2006. URL: <http://www.openoffice.org/product/writer.html>
- [12] BALL, S. Multi-level non-uniform grouping of very large flat structured documents. AusWeb04, The Tenth Australian World Wide Web Conference, 2004. URL: <http://ausweb.scu.edu.au/aw04/papers/refereed/ball/paper.html>
- [13] ERPA Advisory. ERPANet, 2004. URL: <http://www.erpanet.org/advisory/list.php>
- [14] Rich Text Format (RTF) Specification, Version 1.8. Microsoft, 2004.
- [15] Recovering and converting data from manuscripts collection discs. National Library of Australia, 2002. URL: <http://www.nla.gov.au/preserve/digipres/recovering.html>
- [16] Extensible Markup Language (XML) 1.0 (Third Edition). World-Wide Web Consortium, 2004. URL: <http://www.w3.org/TR/REC-xml/>
- [17] Sustainability of digital formats: XML. Library of Congress, 2006. URL: <http://www.digitalpreservation.gov/formats/fdd/fdd000075/shtml>
- [18] ARMS, C.; FLEISCHHAUER, C. Digital formats: Factors for sustainability, functionality and quality. IS&T Archiving Conference. Washington DC, 2005. URL: http://memory.loc.gov/ammem/techdocs/digform/Formats_IST05_paper.pdf
- [19] XSL Transformations (XSLT) Version 1.0. World-Wide Web Consortium, 1999. URL: <http://www.w3.org/TR/xslt>
- [20] WALSH, N.; MUELLNER, L. DocBook: The definitive guide. O'Reilly, 1999. URL: <http://www.docbook.org/>
- [21] OASIS. OASIS Consortium, 1993-2006. URL: <http://www.oasis-open.org/>
- [22] STAYTON, B. DocBook XSL: The complete guide. Sagehill, 2005. URL: <http://www.sagehill.net/book-description.html>
- [23] The Text Encoding Initiative: Yesterday's information tomorrow. TEI Consortium, 2006. URL: <http://www.tei-c.org/>
- [24] RAHTZ, S. XSL Stylesheets for TEI XML. 2006. URL: <http://www.tei-c.org/Stylesheets/teic/>
- [25] XHTML 1.0 The Extensible HyperText Markup Language (Second Edition). The World-Wide Web Consortium, 2002. URL: <http://www.w3.org/TR/xhtml1/>

- [26] Cascading Style Sheets, level 2 CSS2 Specification. The World-Wide Web Consortium, 1998. URL: <http://www.w3.org/TR/REC-CSS2/>
- [27] The Apache Cocoon Project. Apache, 2006. URL: <http://cocoon.apache.org/>
- [28] BRAY, T. Don't invent XML languages. 2006. URL: <http://www.tbray.org/ongoing/When/200x/2006/01/08/No-New-XML-Languages>
- [29] Integrated Content Environment. University of Southern Queensland, 2006. URL: <http://ice.usq.edu.au/>
- [30] MONUS, L. Personal communication. 2006.
- [31] upCast. Infinity Loop, 2003-2007. URL: <http://www.infinity-loop.de/products/upcast/>
- [32] BARNES, I.; YEADON, S. One-click DSpace ingestion with the Digital Scholar's Workbench. Open Repositories 2006. Sydney, 2006. URL: http://www.apsr.edu.au/Open_Repositories_2006/barnes_yeadon.ppt
- [33] BARNES, I. Integrating the repository with academic workflow. Open Repositories 2006. Sydney, 2006. URL: http://www.apsr.edu.au/Open_Repositories_2006/ian_barnes.pdf
- [34] SEFTON, P. OpenDocument or not, you still need to Use Styles. 2005. URL: http://ptsefton.com/blog/2005/09/13/opendocument_or_not_you_still_need_to_use_styles
- [35] SEFTON, P. Personal communication. 2005.
- [36] LIEGMANN, H. Long-term preservation of electronic theses & dissertations. Sixth International Symposium on Electronic Theses and Dissertations. Berlin, 2003. URL: <http://edoc.hu-berlin.de/conferences/etd2003/liegmann-hans/HTML/liegmann.html>
- [37] SourceForge. Open Source Technology Group, 2001-2006. URL: <http://sourceforge.net/>
- [38] Xalan. Apache, 2005. URL: <http://xml.apache.org/xalan-j/>
- [39] Xerces. Apache, 2005. URL: <http://xerces.apache.org/xerces-j/>
- [40] FOP (Formatting Object Processor). Apache, 2006. URL: <http://xmlgraphics.apache.org/fop/>
- [41] EndNote 9. Thomson ResearchSoft, 2005. URL: <http://www.endnote.com/>
- [42] Zotero. Center for History and New Media at George Mason University, 2006. URL: <http://www.zotero.org/>
- [43] BARNES, I. Preservation of word processing documents. Australian Partnership for Sustainable Repositories, 2006. URL: http://www.apsr.edu.au/publications/preservation_of_word_processing_documents.html

Evaluating Digital Humanities Resources: The LAIRAH Project Checklist and the Internet Shakespeare Editions Project

Claire Warwick; Melissa Terras; Isabel Galina; Paul Huntington; Nikoleta Pappa

School of Library, Archive and Information Studies, University College London
Henry Morley Building, Gower Street, London, WC1E 6BT, United Kingdom
e-mail: c.warwick@ucl.ac.uk; m.terras@ucl.ac.uk, i.russell@ucl.ac.uk, p.huntington@ucl.ac.uk

Abstract

The following paper presents a case study of the way that the research done by the LAIRAH project may be applied in the case of a real digital resource for humanities scholarship. We present an evaluation of the Internet Shakespeare Editions website according to the checklist of recommendations which we produced as a result of our research. The LAIRAH (Log analysis of Internet Resources in the Arts and Humanities) project based at UCL's School of Library Archive and Information Studies, was a fifteen month study to discover what influences the long-term sustainability and use of digital resources in the humanities through the analysis and evaluation of real-time use. Our research objectives were to determine the scale of use and neglect of digital resources in the humanities, and to determine whether resources that are used share any common characteristics. We also aimed to highlight areas of good practice, as well as aspects of project design that might be improved to aid greater use and sustainability. A further aim was to determine whether digital resources that were neglected. In our study we concluded that well-used projects share common features that predispose them to success. The effect of institutional and disciplinary culture in the construction of digital humanities projects was significant. We found that critical mass was vital, as was prestige within a university or the acceptance of digital methods in a subject. The importance of good project staff and the availability of technical support also proved vital. If a project as to be well-used it was also essential that information about it should be disseminated as widely as possible. Even amongst well-used projects, however we found areas that might be improved, these included organised user testing, the provision of and easy access to documentation and the lack of updating and maintenance of many resources. The paper discusses our recommendations, which were presented as a check-list under four headings: content, users, maintenance and dissemination. We show why our findings led us to make such recommendations, and discuss their application to the ISE case study.

Keywords: digital humanities; user studies; good practice resource construction

1 Introduction

The following paper presents a case study of the way that the research done by the LAIRAH project may be applied in the case of a real digital resource for humanities scholarship. In it, we present an evaluation of the Internet Shakespeare Editions website according to the checklist of recommendations which we produced as a result of our research (<http://www.ucl.ac.uk/slais/research/circah/features/>).

The LAIRAH (Log analysis of Internet Resources in the Arts and Humanities) project (<http://www.ucl.ac.uk/slais/research/circah/lairah>) based at UCL's School of Library Archive and Information Studies, was a fifteen month study to discover what influences the long-term sustainability and use of digital resources in the humanities through the analysis and evaluation of real-time use. Our research objectives were to determine the scale of use and neglect of digital resources in the humanities, and to determine whether resources that are used share any common characteristics. We also aimed to highlight areas of good practice, as well as aspects of project design that might be improved to aid greater use and sustainability. A further aim was to determine whether digital resources that were neglected might be re-used. As a result of this research the project created a list of recommendations for features that, if possible, the idea successful digital resource ought to have. We also made recommendations to aid the UK's Arts and Humanities Research Council (AHRC), who funded to project, to develop their strategy for funding usable digital resources for future humanities research.

Numerous studies have been carried out into the information needs and information seeking practices of humanities scholars, over recent years [1-5]. We are not aware, however, of any literature that has used quantitative methods, particularly deep log analysis, (described below) to measure the levels of use of digital humanities resources. Our research also concentrates not on the generality of resources, but on the question of

what *kind* of digital resource is most useful for researchers. Although this has been approached by other projects, evidence has been entirely self-reported. Our research is also the first study which has enabled a comparison of the preferences that users report to quantitative evidence of what they actually use.

2 Methodology

The research was funded by the AHRC as part of its ICT Strategy Scheme. We therefore studied digital resources for humanities research which were based in the UK and non-commercially funded. In the first phase of the project we used deep log analysis of web servers of three humanities portal sites in the UK to determine whether it was possible to assess levels of use of digital resources accessed through these portals. These were the Arts and Humanities Data Service, (AHDS) Humbul Humanities Hub, and Artefact (the last two have now merged to become Intute Arts and Humanities) We discuss this analysis in more detail elsewhere, however, in essence we used the data that web server logs automatically record to determine how the sites were used, in terms of levels of use, which parts of the site were used, where users came from and where they went after leaving the site [6]. Although absolute levels are difficult to gauge our research suggested that roughly a third of the resources remained unused. As a result of our analysis we then chose a sample of projects to be studied in more depth. In the following paper, therefore, the results are mainly based on our qualitative methods. We show how the resulting recommendations may be applied to the analysis of an actual digital humanities research resource, and how our work is adding to the dissemination of good practice in digital resource construction and sustainability.

Our qualitative methods involved the selection of a sample of twenty one well used projects for further study. These were studied to see whether there were any common elements of good practice amongst resources that were well used. A representative from the team which constructed the resource, ideally the PI, was interviewed and any documentation available though the project website studied.

To determine whether neglected projects could be reused, we ran two user workshops where participants were asked to examine a sample of eleven resources which were both used and neglected and to discuss their opinions of them. To identify the neglected resources we used the results of the log analysis and also contacted representatives of the AHDS, who gave us additional information about which resources they felt to be most commonly used, or entirely neglected. We also wished to know whether there was any reason why neglected resources were not used, in addition to possible lack of knowledge about them. We did not wish to create bias by telling participants which were used and which neglected, and asked whether they could determine which resources were neglected, and why they felt that this might be. We were surprised to find that participants were highly critical of resource quality, and tended to identify well-used projects as neglected, rather than the opposite.

As a result of our qualitative research we made a number of recommendations for good practice in digital resource construction. These are presented in detail in the project report [7], however, for ease of use, we also produced a simple checklist, intended for those who either are, or would like to be, the producers of scholarly digital resources for humanities research which may be found at <http://www.ucl.ac.uk/slais/research/circah/lairah/features/>. We hope, however that such recommendations may also be more widely applicable, and relate to other sectors of digital resource publication.

In the following paper, we use the checklist that we created and demonstrate how it may be used to evaluate a real digital resource, the *Internet Shakespeare Editions Project* (ISE) (<http://ise.uvic.ca/index.html>). This also provides a framework for a more detailed discussion of the findings from the qualitative phase of our research. We have used this site because the ISE team approached us after the initial findings of the LAIRAH project had been made public, and asked if we would be willing to use the checklist to evaluate their site. They will be using the results of the evaluation for further development of the ISE, however it also presents us with an ideal opportunity to show how the checklist can be employed in the case of an actual digital research resource. Thus each section begins with the recommendation, we then explain the basis on which we made it, as a result of the findings of our research, and the results of the evaluation are then discussed.

3 Results

3.1 Content

3.1.1 The ideal digital resource should have an unambiguous name that indicates its purpose or content

Our log data initially showed that the names and search terms that are used are significant. For example, resources entitled 'census data' were, not surprisingly, popular. However, a similar resource appeared neglected, perhaps because it was called 'Enumerator returns for county X'. Since keyword searching cannot automatically link synonymous terms, a search on 'census' would not have found the latter resource. Discussion at the workshops also revealed that participants could be confused by misleading titles. Some participants thought that a resource entitled 'The Channel Tunnel Rail Link Archive' would be neglected, since they assumed that it contained digitised records of a railway or engineering company. In fact it is a very well used archive of archaeological documentation for the excavations carried out before the link was built. Conversely a resource called 'The Imperial War Museum Concise Art Collection' was praised, since it was immediately clear to users what it contained, and gave the reassurance of a trusted brand in the museums world, thus participants assumed the contents would be of high quality.

The case study resource is called Internet Shakespeare Editions. This is an excellent description of the site, which also offers numerous additional resources for Shakespeare scholars. However it would be unrealistic to try to describe these in the site's name.

The URL, <http://ise.uvic.ca/index.html>, however does not reflect the name and may be difficult to remember, as compared, for example, to the Old Bailey Online project, who have acquired a domain name which is easy to recall. (<http://www.oldbaileyonline.org/>). However, one important facet of the current URL is that it places the ISE within the domain of a respected university. This type of institutional brand helps users to trust to integrity of the site and the quality of its contents. When a Google search was performed on the keyword 'Shakespeare' ISE came in the 50-60 screen of results, out of a total of 53,000,000 hits. This is a creditable performance, considering the popularity of Shakespeare as a topic. However, it might be improved by encouraging as many users, English departments, and libraries as possible to create links to the ISE page.

3.1.2 The ideal digital resource should concern a subject that is either popular in a wide community or essential for a smaller expert one

The log data demonstrated that certain subjects and themes were particularly popular, for example, warfare, census data, witchcraft, Shakespeare and women's suffrage. We do not know exactly what purpose the resources are being put to, whether high level academic study, family history research or a school history project, for example. However, it is clear that digital resources concerning certain popular subjects are likely to be well used. Nevertheless participants at our workshops stressed the importance of resources which might be vital to research in a relatively small community, whose work would be significantly impoverished without them. It would also be unwise to concentrate research funds on a small number of popular subjects, and neglect less popular areas, since we cannot know which topics, perhaps now neglected, may be widely studied in future.

The ISE website evidently concerns Shakespeare, which is both very popular in the wider community as well as being an important research topic for academics. The website offers two types of navigation, by Academic divisions- done through the metaphor of a building- and by Subject area. This is an interesting way to help different user groups to access the content. As part of the LAIRAH research we found that users felt confused by many sites that were only designed for experts in the subject, and assumed a high level of knowledge about resources and how they should be used. This dual path is thus good practice, since it allows users to access materials in a way that is most useful for them. The functions of the different parts of the site, and the different methods of navigating them are also very clearly described in the About section page, 'How to use this Site' <http://ise.uvic.ca/Foyer/index2.html>. This is a simple and very helpful page title, given that we have found that clear signposting is invaluable, especially for less expert users (discussed below in section 3.2.3)

3.1.3. The ideal digital resource should retain its server logs, and make them available to their funding agency and researchers, subject to confidentiality agreements

During our research we found that it could be relatively difficult to obtain log data, even from large publicly funded portals. Humbul was reluctant to allow us to use their data, even if anonymised, because of worries that individual users might be identifiable, because of their personalisation features. We were able to reassure them that this is not possible, and that any individual machine IP addresses would not be made public in any reports. However, the time taken to do the anonymisation held up our research considerably. Artefact were willing to give us the data, but had not had the technical support to be able to keep it, and what was kept was lacking in detail. Thus we were able only to access a small amount of data.

Many individual projects may be even less likely to realise the importance that their web logs may have as a potential research resource. They may not realise that they should be kept, nor the level of detail of logging that should be made possible, they may also lack technical support to do so. We therefore recommended to the AHRC that if resources were publicly funded they should be asked to keep logs and that as a function of the grant such logs should be made available to researchers for the purposes of monitoring and evaluation. This would avoid the kind of delays we experienced while permissions were negotiated. The ISE server logs have been retained and are made available. We hope to be able to analyse them in detail as part of the next phase of the LAIRAH research, if a funding application is successful.

3.1.4. The ideal digital resource should keep documentation and make it available from the project website, making clear the extent, provenance and selection methods of materials for the resource

The participants at our workshop were concerned that in many cases they could not find enough information about the content of the resource, how it was selected, and its provenance. They also wished to know more about the team that constructed the project, and the expertise of its members, and some of the more technically expert participants would have liked to have found out more about the technical methods and standards applied. They also felt the lack of the kind of information about sources that are found in the print world in citations and bibliography. All of this meta information helps to increase the trust that users have in the quality of digital resources.

However, in our study of even well used projects we found that levels of documentation were extremely variable. Some projects were extremely well documented, these tended to be in subject areas like archaeology and linguistics, where documentation of resources has always been an essential research practice. However, many projects kept little or no organised documentation. It could also be hard to find. Ideally documentation should be easily located from the project website. However, in many cases it was absent, was accessible only through the AHDS, or not at all. In other cases, although some documentation could be found via the website, it was complex, and difficult to locate or incomprehensible to the non-expert reader. One of the most effectively documented resources had been compelled to do so in the terms of their grant from the New Opportunities Fund. We therefore recommended that the AHRC should consider making documentation a deliverable of any funded project.

The ISE site is relatively unusual in that it is extremely well documented and there is ample information available. Most of the information is available from the 'Academic divisions-Foyer', with links from SubjectArea- About ISE'. It is especially important that the 'About' link is available from the top menu, making it as easy as possible for users to access the documentation. The 'About' page is also kept relatively short, is expressed in simple terms and has links to further material. This is ideal and should be retained, since our research has found that users become confused if documentation is too dense, complex or presented on a long page which requires them to scroll. The documentation includes:

- History of the site (<http://ise.uvic.ca/Foyer/ISEoverview.html>)
- Editorial Guidelines
- Details of people on Advisory Boards for different sections [Editorial Board, Advisory board Performance materials, Theatre History and Technical Design and Implementation]
- Details of editors of the plays and poems
- Technical information about the design of the site.
- Information about new site

3.2 Users

3.2.1 The ideal digital resource should have a clear idea of whom the expected users might be; consult them as soon as possible and maintain contact throughout the project via a dedicated email list or website feedback

Very few of the projects that we studied had any contact with their users, nor did they tend to consult them or provide much user interaction on the website, beyond a ‘contact us’ email link. This is a wasted opportunity, since contact with users should help project teams understand the needs of those who will use the site, and adapt it accordingly. This should in turn increase levels of use of the site.

The ISE website states that their aim “is to inspire love of Shakespeare’s work in a world-wide audience”. It would seem that they expect to have a global impact. There is no further explicit indication of their expected users. However, their division of SubjectArea and Academic would seem to suggest that academic users are expected, although the site is welcoming for non-experts, since it gives detailed descriptions of the material it contains and how this might be used. Contact with users is achieved through a discussion section and the provision of contact email links.

3.2.1.1 Discussion [<http://ise.uvic.ca/Annex/discussion.html>]

This page informs users that ‘When complete, this section of the site will provide an informal forum for the discussion of Shakespeare, his works, life, and the performance of his plays.’ The discussion section will be launched in April 2007, and should provide an excellent opportunity for users to interact with each other and the site’s creators. It should also be noted, however, that such forums tend to be labour intensive to keep updated. We have found that users distrust the quality of a site if there is evidence that it is not entirely up to date. Thus, once the discussion section is launched, sufficient resources will need to be allocated to it to ensure that it does not appear outdated.

3.2.1.2 Contact email links

An email contact address is linked to from the following pages:

- About [<http://ise.uvic.ca/Foyer/about.html>]
- Policy on copyright [<http://ise.uvic.ca/Foyer/copyright.html>]
- Guidelines for the acquisition and copyright of performance materials [<http://ise.uvic.ca/Foyer/PerfGuide.html>]
- Shakespeare in performance –FAQ [<http://ise.uvic.ca/Theater/sip/faq.html>]
- About Shakespeare in performance [<http://ise.uvic.ca/Theater/sip/about.html>]

Each contact link is placed within very different contexts and is used for different types of information, thus it may be that the addition of an overall ‘contact us’ link on the top menu would be beneficial for users.

3.2.2 The ideal digital resource should carry out formal user surveys and software and interface tests and integrate the results into the project design

Once again few of even the well used resources in our survey had carried out any kind of formal user testing. One project subsequently regretted this, since it had worked very hard to produce a complex search interface, only to find that in practice it was too complex for the majority of their users.

The ISE carried user testing before the launch of the new design of the site in November 2005. The sample was of about 15 participants, and designed to represent the needs of different user groups. It included students at different levels, English faculty of various ages, skilled computer personnel with no great knowledge of Shakespeare, and several general readers. The results were used to aid the design of the navigation, both to encourage initial entry to the website, and to navigate within the site. ISE made a significant number of modifications as a result of the testing. This testing represents good practice in resource design, and makes ISE relatively unusual in the field of digital humanities resources.

Positive feedback from students and teachers who have used ISE has also been placed the website at <http://ise.uvic.ca/Foyer/corporate.html>. This shows that users have found the contents helpful in their work. The

site has also been given various awards for excellence from internet bodies and those concerned with academic study in general and English literature in particular.

3.2.3 The ideal digital resource should be designed for a wide variety of users, and include information to help the non-expert to understand the resource and use its contents

At the workshop participants thought that many resources appeared only to be designed for subject experts, and therefore deterred the more general user. They argued that the inclusion of simple instructions would be very helpful for the non-expert, and would not affect the experience of the expert user. This proved to be important, since the participants quickly gave up trying to use a website if they were unable to work out how it should be used. Simple, clear signposting should therefore help to increase levels of use of digital resources, since it encourage non-experts to persist with user of the site.

The ISE website seems to be designed for a wide variety of users. A large section dedicated to Shakespeare's life and times (<http://ise.uvic.ca/Library/SLT/intro/introsubj.html>) seems to cater for the novice. Whilst Scholarly Articles on Shakespeare and the Internet (<http://ise.uvic.ca/Annex/Articles/index.html>) appear to be directed at the Shakespeare scholar.

Access to the main site is achieved by clicking on the image of the library which makes up most of the top page. The visual metaphor is appealing in many ways, and undoubtedly attractive, and the ISE team may not wish to spoil it with instructions for use. However, however it may not be evident to all users that they should click this image and our research shows that users are quick to give up on using a resource if they do not find obvious clues about how to use it. In many cases such instructions were not provided by the project team, as the use of the resource may have seemed entirely obvious to its producers. The ISE are planning to address this potential difficulty by creating a more obvious link to their newsletter page- (see discussion below in section 3.4.2)

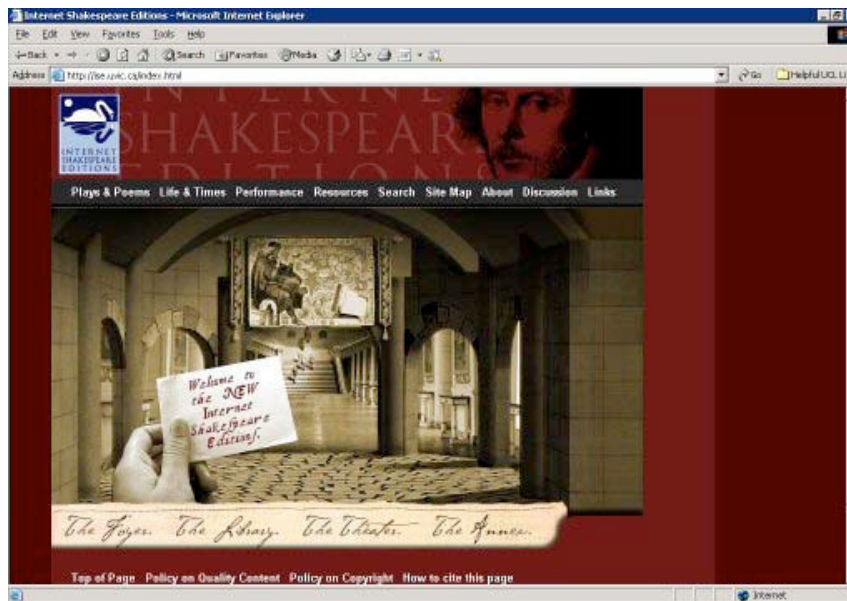


Figure 1: The Top page of the ISE website

Navigation is described in the 'How to use this site' page of the Foyer section. This contains a great deal of very useful information about the contents of the site, who created it, the type of material included, its provenance and extent. This is also easily located by following links from the page. This is vital since this is the kind of information which encourages users to trust the quality of a digital resource. The information provided by ISE should be more than sufficient to reassure users of its high academic standards.

3.3 Management

3.3.1 The ideal digital resource should have access to good technical support, ideally from a centre of excellence in digital humanities

It was not surprising to find that many of the well used projects were associated with centres of excellence in digital humanities. The Humanities Research Institute at the University of Sheffield for example was the base for

several projects whose use was prominent in the log data. This is understandable, since it is difficult for individual humanities researchers to keep up with the latest developments in digital techniques and technical standards. Thus it is vital for researchers to have access at least to a computer support officer, and ideally to such a centre, whose staff will understand not only technical aspects of resource construction, but also the demands of humanities research.

The ISE is based at the Humanities Computing and Media Centre at the University of Victoria in Canada (UVic) and thus has access to a high level of technical expertise and advice. There is also an advisory board in Technical Design and Implementation (<http://ise.uvic.ca/Foyer/techboard.html>). The page includes names and details of the four members. Between them they have both technical and digital humanities expertise. This is an ideal arrangement, since it gives the ISE access to the latest information about technical developments and good practice in digital humanities.

3.3.2 The ideal digital resource should recruit staff who have both subject expertise and knowledge of digital humanities techniques, then train them in other specialist techniques as necessary

The recruitment and training of staff to work on digital humanities resources was a particular challenge for the leaders of the projects whom we interviewed. It could be especially difficult to find staff who were not only technically adept, but also had sufficient knowledge of humanities research that they understood the material itself, and were thus able to mediate between the needs of researchers and technical functionality.

It was most usual for humanities specialists to be recruited, but they then needed to be trained in various computing techniques. This often proved difficult since the amount and quality of training available through universities was often disappointing. PIs also commented that training new researchers often took longer than expected, which could be a significant problem, when the project was operating on time-limited funding.

The ISE has obviously been able to recruit good project staff. The site lists all research assistants (undergraduate and graduate), from 1992 to the present day. Most have subject expertise and/or digital humanities knowledge, plus other areas of expertise depending on their function. The ISE is also very well provided with academic members, editors and advisors. There is one Editorial Board and three Advisory Boards for Performance Material, Theatre History and Technical Design and Implementation. There is also detailed information about the editors of the online editions. This is important, since detailed information about the academic qualifications and technical expertise of the project team helps to reassure users that the material to be found on the site is of the highest academic standards. It is also important that the ISE list the institutional affiliations of board members, since such affiliations appear to act in a similar way to the trusted brand status of commercial sites, such as the BBC for news resources. Once again they help users trust the quality of the resources to be found.

3.3.3 The ideal digital resource should have access to short term funds to allow to retain expert staff between projects

A further problem where staffing was concerned was that in the UK most non-commercial digital humanities projects are made possible by short term grants of public money, usually from the AHRC. This funding is relatively scarce and to it is very difficult for projects to obtain continuous funding, and retain skilled staff. This resulted in wasted resources, since new staff had to be appointed and trained for each new tranche of funding granted, rather than PIs being able to rely on a cadre of expert employees, as is more often the case with scientific funding. We therefore suggested that the AHRC might consider making available small amounts of short term funding, to allow employees to be retained for a few months in the hope of securing further long term funding, as is the practice with some UK science funding councils.

The ISE is relatively fortunate in this regard, in that the Canadian funding system appears to make it easier to access small amounts of grant money on a continuing basis, both from universities themselves, and from public funding. The uninterrupted recruitment of research assistants noted on the site suggests continuous availability of funding. The University of Victoria, The Social Sciences and Humanities Research Council of Canada and the Innovation Development Corporation are listed as 'supporters' (<http://ise.uvic.ca/Foyer/acknowledge.html>). This suggests ongoing access to funding to develop the resource. This is ideal, but rare outside North America, and it is to be hoped that such funding will continue to be available to the ISE.

3.4 Dissemination

3.4.1 The ideal digital resource should have an attractive usable interface, from which all material for the project may be accessed without the need to download further data or software

The need to have an interface that is attractive and usable may seem too obvious for any need for comment. However, workshop participants found that many of the interfaces to the resources in the sample, even well used ones, were problematic. This may be because very few projects had any contact with experts in HCI, or interface design. It is possible to apply for funding for a professionally designed site, but none of the projects in our sample had done so. As a result participants found most unattractive compared with the professional interfaces of commercially produced resources. This is significant, given that it appears that most users make decisions about websites extremely quickly [8], and so an unwelcoming interface is likely to deter many users before they have even accessed the resource's contents.

We also found that several projects, especially databases for historical research, required users to register to use the data, for which they had to be given a password before they were allowed access. Data would then have to be downloaded and used with specialist software. Such registration is sometimes unavoidable, for example for reasons of copyright. However, it was a serious deterrent, for all but the most determined users, and thus should be avoided if at all possible.

The user studies performed by the ISE have helped the team to develop the interface and navigation to help users find their way around the site. Although some users may find the visual metaphor a little confusing, in general the interface is attractive and easy to use. The site is extensive and complex, with a large amount of material, which by nature will mean that the first time user will need to spend some time exploring to find what they need. The different sections are generally well signposted however, and explanations are provided about what kind of content may be found in each section. The pages are clear and well-written, and should be easily comprehensible by web users. This is important, since users of the web tend to skim pages, and take in less content than if they were reading printed material. Pages therefore need to be concise, divide into easily comprehended sections, and be written in a clear and accessible style [9]. All of the data needed may also be accessed without need for further software or to download the data for local use.

3.4.2 The ideal digital resource should maintain and actively update the interface, content and functionality of the resources, and not simply archive it with AHDS

As discussed above, it is important that web sites should be updated regularly. Now that many commercial sites are updated constantly, users have an expectation of currency, and our research has shown that they may therefore distrust the quality of resources that appear not to be actively updated.

Some of the ISE pages contain information about when the site was updated. There is also a detailed description explaining how ISE has been created, updated and what ideas there are for the future. (<http://ise.uvic.ca/Foyer/ISEoverview.html>). In fact there is an active policy of updating the contents of the ISE site. A full-time student updates content in the database of Shakespeare in Performance: this is usually done daily. She also checks the links regularly. A minor update of the Life and Times section (mainly the bibliographies) is planned for April 2007. The texts are updated as they are completed, but this section of the site is still under construction as the ISE team are looking at various ways of displaying annotations and textual variants.

Our study showed that most scholarly websites can learn from the commercial sector when it comes to providing information about how and when their site is updated. Evidently, for a site such as ISE, there is no need continually to update all pages, however practice in providing such information should be consistent throughout the site. Another way to deal with this is to provide a link to news about the site, and provide information about when new content is added. Some projects use the front page for such updates (<http://www.ucl.ac.uk/english-usage/>), however another approach is to link from the front page to a news, or what's new page (for example <http://ahds.ac.uk/>). An RSS feed such as that used by the AHDS is often used commercially to encourage users to revisit sites of interest and could also be used if significant numbers of changes are being made. All of these measures reassure users that updating is happening. The ISE plan to introduce a regular newsletter, the first issue is being prepared for the end of April 2007 and the front page will have a hand that invites the user to link to the newsletter rather than simply providing a welcome page. They expect the newsletter to be a regular feature, issued three or four times a year, and plan to send it to a large readership base of libraries and English

Departments, who it is hoped will create links to the site as a response. They also plan to use RSS feeds to automate the provision of information about page updating.

3.4.2.1 Maintenance

Maintenance is a problematic issue for non-commercial digital humanities resources, since after funding runs out, there may be no resources to make sure that the resource is maintained. Although a recent funding call from the UK's JISC (Joint Information Systems Committee) requires universities to guarantee that they will maintain funded resources for at least ten years [10] this involves a commitment of server space and personnel to do so, a cost which the institution may not feel able to bear in the long term. There can also be the danger that if a member of staff leaves or retires the university may not feel obliged to maintain the resource, and thus, in the worst case, it could be entirely lost. It is also clear that simply leaving data on a server without active maintenance and updating is not satisfactory. One project in our survey was no longer updated or maintained, and the original researcher who created it was aware that not only was the website seriously outdated, but the functionality of the database itself was gradually deteriorating. No-one was paid to maintain the resource, however, or had time to do so voluntarily, and thus only ten years after its construction it was becoming unusable.

The Humanities Computing and Media Centre at UVic, runs the ISE server. The whole system is, however, backed up automatically, off site, by the University Computing and System Services who also perform basic system-level maintenance. This ensures that the data is safely maintained at present, and the ISE are currently negotiating with UVic for continuing support for infrastructure. The ISE is constituted as an independent, non-profit organisation, therefore, it could if necessary exist independently of the university, and be moved to a different server. However it would be ideal if UVic were able to commit themselves to long term maintenance of the data.

In the UK all research that creates digital output, and is funded by the AHRC, must be offered for archiving with the AHDS, which preserves the data, although it does not, of course update it, or maintain a website. However, this option is not available in Canada, and thus individual projects and their host institutions must negotiate such archiving on a piecemeal basis. For example, ISE archive all artefacts for the database at 600dpi TIFF files. All other files are handled by a version control system (Subversion) which keeps track of all changes. This is good practice individually, however without a central service like the AHDS to offer archiving facilities and advice about good practice, many projects may not be aware of the standards to which they should adhere, and thus their maintenance strategy might not be as rigorous. This piecemeal approach is therefore not ideal for digital resources, and at worst potentially poses a very serious threat to their long-term sustainability.

3.4.3 The ideal digital resource should Disseminate information about itself widely, both within its own subject domain and in digital humanities

The strongest possible correlation in our study between a characteristic of a resource and its use was dissemination. All the projects in our study had worked hard to provide information about their work, by giving papers and seminars in both the digital humanities and publishing sectors, and within their own subject areas. This is an important new role for academics, since previously they would have written books, and relied on publishers to market them. In the case of a digital resource, the scholar is now the publisher, and so the responsibility of disseminating information about their work not falls to them.

The ISE has a good level of web visibility. A simple link analysis shows over 5,000 links to any page in <http://ise.uvic.ca>. Publications and papers given as a result of ISE research are also listed in the annex there on a page entitled: 'Scholarly Articles on Shakespeare and the Internet' (<http://ise.uvic.ca/Annex/Articles/index.html>). This shows that the ISE members have been active in disseminating information about their research and the project itself. These include both conferences and journals in English literature and conferences on Humanities computing, although papers do not appear to have been published in any humanities computing journals. The website also cites an example of use of ISE:

*"An idea I discussed at a meeting of the Shakespeare Association of America three years ago, that ambiguous readings and imprecise entrances or exits could be indicated by animation, has been received with a possibly surprising enthusiasm by the editing community in Shakespeare studies:
<http://ise.uvic.ca/Annex/Articles/SAA2002/rich4.html>"*

4 Conclusions

This paper has shown how the findings of our research on the LAIRAH project have been used to construct a check-list of recommendations. We have further shown how such recommendations may be applied in the evaluation of an example digital humanities resource.

The Internet Shakespeare Editions project is an example of excellent practice in the construction of digital humanities resources. It maintains consistently high standards both of content, presentation and technical web design. This evaluation shows that the ISE performs very well when judged according to the recommendations made in the LAIRAH checklist. In many aspects it out performs many of the well used resources in the LAIRAH research sample.

It is an attractive, usable resource, with a wealth of useful content, which is comprehensively documented. It is able to maintain levels of funding, which allow it to recruit able research staff, and it is well supported by a humanities computing centre, by expert editors and well qualified advisory boards. All of these factors help to ensure that users will recognise the content as trustworthy and of good quality. It is also clear that information about the resource is widely disseminated, both in digital humanities and English literature. The resource is actively updated and should continue to be maintained by the University of Victoria. Like all digital resources, especially those where no national archiving system exists, it will inevitably face problems of long term sustainability. However, the team is aware of these, and is taking steps to try to ensure the resource's future.

Acknowledgements

LAIRAH was funded by the AHRC ICT Strategy scheme. We would like to thank all those who took part in workshops, and agreed to be interviewed during the project. We especially thank Michael Best, Roberta Livingstone and the Internet Shakespeare Editions project team for allowing us to use ISE as a case study.

Notes and References

- [1] BARRETT, A. The information seeking habits of graduate student researchers in the humanities. *The Journal of Academic Librarianship*. 2005, vol. 31, no. 4, pp. 324-331.
- [2] TALJA, S.; MAULA, H. Reasons for the use and non-use of electronic journals and databases - A domain analytic study in four scholarly disciplines. *Journal of Documentation*. 2003, vol. 59, no. 6, pp. 673-691.
- [3] HERMAN, E. End-users in academia: meeting the information needs of university researchers in an electronic age Part 2 Innovative information-accessing opportunities and the researcher: user acceptance of IT-based information resources in academia. *Aslib Proceedings*. 2001, vol. 53, no. 10, pp.431-457.
- [4] BRITISH ACADEMY. *E-resources for Research in the Humanities and Social Sciences - A British Academy Policy Review*. 2005. Available from Internet: <http://www.britac.ac.uk/reports/eresources/report/sect3.html#part5>
- [5] DALTON, M. S.; CHARNIGO, L. Historians and their information sources. *College & Research Libraries*. 2004, vol. 65, no. 5, pp. 400-425.
- [6] WARWICK, C.; TERRAS, M.; HUNTINGTON, P.; PAPPAS, N.; GALINA, I. 'If you build it will they come? The LAIRAH survey of digital resources in the arts and humanities. *Literary and Linguistic Computing*. 2007 (forthcoming).
- [7] WARWICK, C.; TERRAS, M.; HUNTINGTON, P.; PAPPAS, N.; GALINA, I, The LAIRAH Project:Log Analysis of Digital Resources in the Arts and Humanities. Final Report to the Arts and Humanities Research Council. Arts and Humanities Research Council. 2007 Forthcoming
- [8] LINDGAARD, G.; DUDEK, C.; FERNANDES, G.; BROWN, J. Attention web designers: you have 50 milliseconds to make a good first impression. *Behaviour & Information Technology*. 2005, vol. 25, pp. 115-126.
- [9] MORKES, J; NIELSEN, J., Concise, SCANNABLE, and Objective:How to Write for the Web. *Useit.com*, 1997. Available from Internet: <http://www.useit.com/papers/webwriting/writing.html>
- [10] JISC. *JISC Circular 03/06: JISC Capital programme*. 2006. Available from Internet: http://www.jisc.ac.uk/fundingopportunities/funding_calls/2006/06/funding_circular03_06.aspx

Feasibility of Open Access Publishing for Journals Funded by the Social Science and Humanities Research Council of Canada

Leslie Chan¹; Frances Groen²; Jean-Claude Guédon³

¹ Department of Social Sciences, University of Toronto Scarborough, Toronto, Ontario, Canada
e-mail: chan@utsc.utoronto.ca

² Trenholme Libraries, McGill University, Montreal, Quebec, Canada
e-mail: francès.groen@mcgill.ca

³ Département de Littérature Comparée, Université de Montréal
email: jean.claude.guedon@UMontreal.ca

Abstract

This paper reports on the results of a feasibility study on open access publishing for humanities and social sciences journals supported by the Social Sciences and Humanities Research Council of Canada's (SSHRC) Aid to Scholarly and Transfer Journals Program. The study is part of a broader effort of the SSHRC to better understand the landscape of Open Access and how best to implement this principle into the current research programs funded by SSHRC. As such, the study was designed to assist SSHRC in making policy and program decisions regarding its Aid to Scholarly Journals Program. In particular, this study focused on the current publishing practices of SSHRC funded journals, with the ultimate goal of understanding the financial implications for these journals if they were to provide open access to the journal content. The more immediate goal of the study was to gain better knowledge of the general level of understanding among journal publishers and editors on the impact of open access and on their scholarly societies' publishing program.

Keywords: social sciences and humanities; open access journal; funding policy; research impact; citation analysis

1 Introduction

Open Access (OA) is the process by which peer-reviewed research publications resulting from public funding are made freely available through the Internet to all potential users. The purpose is to remove the price barrier and other permission barriers that restrict the dissemination and growth of further research. Though a subject of much debate, OA is now widely seen as a means to improve the accessibility and impact of publicly funded research. Evidence demonstrating that openly accessible publications are more highly cited are emerging [1] and new tools and infrastructure for maximizing the usage and innovative applications of research results are being developed, not only for the natural and medical sciences, but also for the humanities and social sciences [2].

The Social Science and Humanities Research Council (SSHRC) of Canada is the largest funding agency of humanities and social science research in Canada [3], and it is also among a growing number of government funding agencies around the world actively addressing the issue of open access [4]. In 2004, SSHRC's Council adopted OA in principle and instructed SSHRC staff to consult broadly with the research community as to the best way to implement this principle into the current research programs funded by SSHRC. SSHRC has chosen to promote open access for journals because the Council understood that open access improves scholarly communication while ensuring that research is disseminated and useful to all citizens, including the public and private sectors [5].

Between 22 August 2005 and 31 October 2005, SSHRC staff conducted a survey across a significant range of actors, including researchers and scholars, scholarly associations, publishers, editors and librarians to elicit comments and views on the subject of open access. A total of 130 submissions were received (researchers and scholars 84; university presses 2; journal editors 26; librarians 12; scholarly associations 5).

The largest and arguably most significant number of responses came from the scholarly community; and within that group 54 of the 84 expressed their support of open access although many had operational concerns. The second largest group, the journal editors, was more divided with 14 supporting in concept open access and 12 opposed. However, all expressed concerns with the financial issues in the transition to open access [6].

While the findings of this consultation are useful and important for further study, the reality remains that the input from scholars, editors and publishers is very limited in quantity. It was evident that further study would be necessary if SSHRC were to move forward on the open access agenda. It was also clear that the preservation of the integrity of the present system of scholarly communication in the humanities and social sciences had to be guaranteed, and that the transition to a new model of scholarly communication must be judiciously implemented. Over time, a valuable system has been developed for the nurturing of humanities and social sciences journals in Canada and this could be replaced by a new system only if it were of greater value.

SSHRC also recognizes that while financial support for research is crucial, the dissemination and uptake of research is equally important. Research left unread or not built upon has no impact and no financial and social return. It is time to re-examine what returns financial support for SSH journals are bringing to scholarship and to the Canadian public.

The world of publishing in general has been radically altered by the introduction of electronic publishing in the last two decades. New modes of production, of access, of ownership of information, and of financing, have been changing scholarly communication in fundamental ways. In this context and in the light of the initial SSHRC investigations, the authors of this paper were invited to conduct a study of the feasibility of open access publishing for journals currently receiving support under the Aid to Research and Transfer Journals Program of SSHRC.

While the immediate goal of the study is to gain a better understanding of current journal publishing practices and general knowledge of OA amongst SSH journal editors, the longer term goal is to provide evidence on which firm and sustainable policy on OA could be developed and implemented by the Council.

2 Materials and Methods

To guide the research process and to keep the scope in check, the following questions were used as guideline:

- To what extent are SSHRC funded journals already available in digital form?
- What are the costs and savings associated with the delivery of these journals in digital form?
- What are the perceived incentives and barriers to moving towards an electronic only version of these journals?
- What are the perceived incentives and barriers to open access publishing of these journals?

2.1 Sources of Data

To answer the questions posted above, we drawn data from a number of sources:

2.1.1. Data from Funding Application

Between 2004-2007, 161 journals received funding of varying amounts from *SSHRC's Aid to Research and Transfer Journals Program*. To gain an understanding of the financial health and support resources of these journals, we first conducted a review of the records of the grant applications, which included the operating budgets of the various journals. A preliminary analysis of the journal contents and titles revealed that a broad range of topics with a number of titles are published in the fields of history, literature, law, economics and education, and a considerable number of titles in a broad range of Canadian area studies. The breadth and variety of the titles in both official languages of Canada led us to conclude that the research should not be based along disciplinary lines, but should be carried out within the broadly defined domains of social sciences and humanities.

2.1.2. Online Questionnaire

A web based questionnaire in both official languages of Canada was developed and invitation to participate was sent by email to the journal editors or key contacts for the journals. Respondents' identities were kept anonymous. The questions were intended to elicit responses from editors regarding the journal's delivery medium, funding support from scholarly associations, the use of commercial aggregators, electronic publishing platform, and support and concerns towards open access. The full list of questions is provided in the appendix.

2.2.3 Citation analysis

To evaluate the citation impact of SSHRC funded journals on scholarship, an analysis of the journals based upon ISI Journal Citation Report was undertaken. While there are well known concerns with the using is ISI JSR to access the impact of scientific literature in general and particularly with humanities and social sciences (see Discussion), the analysis is intended to serve as a snapshot of the overall visibility of SSH journals published in Canada and how these journals compared with journals in their respective fields.

2.2.4 Interviews

To supplement the results from the web questionnaire and to get more in-depth and qualitative information on some of the challenges and opportunities faced by journal publishers, a number of journal editors, publishers and library directors, were selected for interviews, either in person, by phone, or through email. The interviewees were asked to provide their view on the feasibility of open access for the production and distribution of SSHRC funded journals. Their views were integrated into the discussion and recommendations put forth to SSHRC.

3. Results

3.1 Funding

The 161 journals that were successful in the 2004-7 competition received a total of \$6,582,255, with grants ranging from \$2,906 to \$73,370 over the three year period. Of these titles, 29 journals (18% of the titles) received the maximum grant for a total of \$2,127, 730 or 31% of the total funds allocated

In addition to SSHRC support, some journals also receive support from Heritage Canada and from the Government of Quebec, based on publicly available information on the Internet. In Quebec, the Fonds de Recherche sur la société et la culture du Québec (FRSCQ) in their 2004 competition awarded \$2,185,155 to 36 journals, 28 of which also received funding from SSHRC. Heritage Canada reported funding from both Canada Post Corporation and Canadian Heritage as of July 2005 and there are SSHRC titles on this list. However, amounts given to individual SSHRC titles are generally negligible.

3.2 Findings from the Questionnaire

A web-based questionnaire in both official languages was sent to editors of journals supported by SSHRC [7]. The survey was opened to respondents between May 1 and July 31, 2006. It received a rate of response of 42 % (67 out of 161). Of the 67 respondents, 56 were English and 11 were French speakers.

More than 80 % of the respondents reported that articles published in SSHRC funded journals are available electronically. For journals that are online, about half came online between 2002 and 2006. A small number of journals were already on-line in the 1990's beginning with 1993. The use of aggregators was highly preferred as a means of providing electronic access, with 84.4 % of English respondents reporting using a variety of aggregators, including Érudit [8], an electronic platform for journal delivery. For the 9 French responses, 100% reported using Érudit.

Of the commercial aggregators listed in the questionnaire, Proquest was the most heavily used, followed by Ebsco and Érudit. Unfortunately, Blackwell was inadvertently omitted from the list of possible aggregators due to a programming glitch, so the number of journals using Blackwell is not clear. Slightly over 40% of the 54 respondents respond that the most recent issues are available online. The rest have no recent issue available online. 55% of 53 English respondents reported that they do not receive compensation from an aggregator on a pay-per-use basis, while 10 of the 11 French respondents reported no compensation from aggregators.

For journals published by scholarly association, 54% of the 39 English respondents reported that the journal did not receive financial subsidy from the host association, while 7 of 9 French respondents reported the same.

When asked if they are in favour of open access in principle (leaving economic issues aside for the moment), 78% of the 54 English respondents said yes, while 6 of the 10 French respondents reported yes. With regard to the timing for open access, 74% of the 49 English respondents were in favour of the moving wall solution and 14% were for immediate open access. 91% of the 11 French respondents favoured a moving wall solution and only one respondent favoured immediate open access.

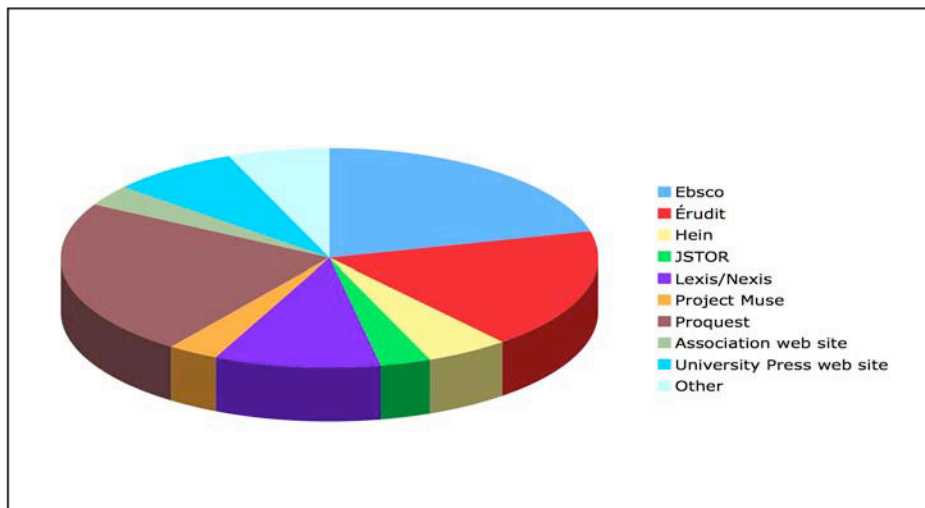


Figure 1: Chart showing the proportion of aggregators used by SSHRC funded journals

When asked if SSHRC should make it mandatory that SSHRC supported journals be available for open access, 84% of the 57 English respondents opposed. Similarly, 82% of the 11 French respondents were not in favour of mandatory open access.

72% of the 50 English respondents were in favour of SSHRC providing funding to support institutional repositories designed to support secure and open access to research publications. Only half of the 10 French respondents were in favour of the same.

71% of 52 English respondents supported the idea that SSHRC should provide financial support for journals to become open access. Of the 11 French respondents, 64% said yes.

82% of the 49 English respondents agreed that SSHRC should provide support for open access journals and consider eligibility criteria appropriate for these titles, while 64% of 11 French respondents agreed. Respondents were also asked to provide suggestions on what these criteria might be, and most agreed that the evaluation criteria for funding support for open access journals should be based on the same quality evaluation criteria used for subscription based journals, with the proviso that the requirement for 200 paid subscribers [9] be removed for OA journals and be replaced by other metrics more suitable for the electronic environment.

3.3 Findings of Citation Analysis

Impact factors have long been an essential criterion to evaluate journals, particularly journals belonging to the same specialties. It is well known that using impact factors is problematic. They must be handled with caution and they can support comparisons between journals only when they belong to closely related fields. Citation cultures can vary considerably from one discipline to another, making comparison of journals across discipline even more problematic [10]. Furthermore, in the case of the humanities, where monographs remain the dominant currency, and where citations are used in extremely complex ways, impact factors have generally not been used [11].

A preliminary survey of the titles supported by SSHRC shows that they are divided between 71 humanities journals (broadly defined) and 90 social science journals. The results given below apply only to the 90 social science journals and they must be treated prudently, but they nevertheless offer some valid insights, especially when they are used to compare journals covering roughly the same fields of study.

With these caveats in mind, we looked at the rankings of SSHRC-supported journals in the Journal Citation Reports (JCR) published by the Institute for Scientific Information (ISI). In particular, we compared the impact factors of listed journals between 1997 and 2005 and compared them to the leaders in their respective field. 1997 corresponds to the earliest year covered by JCR; 2005 is the most recent year available.

Of the 161 titles searched (90 titles are in the social sciences), and of these, only 21 titles (23 %) had an impact factor assigned by ISI in 1997 or 2005. 2 titles with an impact factor in 1997 had lost it by 2005. Conversely, 3

titles had an impact factor in 2005 but not in 1997. This means that only 19 titles had an impact factor in 2005 (or 21% of the social science titles supported by SSHRC).

Journal title	Impact factor (IF) 2005	Rank in assigned subject area	Impact factor (IF) 1997	Rank in assigned subject area	Highest impact title in subject area for 2005	1997 and 2005 IF as % of leader IF
Alberta Journal of Educational Research	NIL		0.028 ¹	100 out of 102	J Learn Sci 2.792	0%
Canadian Geographer	0.491	31 out of 38	0.294	24 out of 31	J Econ Geogr 3.222	15%
Canadian Journal of Administrative Sciences	0.191	69 out of 71	0.057	57 out of 59	Mis Quarterly 4.978	4%
Canadian Journal of Agricultural Economics	NIL		0.129	142 out of 161	Quart J Economics 4.775	0%
Canadian Journal of Behavioural Science	0.345	78 out of 101	0.348	64 out of 108	Ann Rev of Psychol 9.784	3.5%
Canadian Journal of Criminology & Criminal Justice	0.300	19 out of 27	0.213	17 out of 19	Crime Justice 2.588	12%
Canadian Journal of Development Studies	0.300	32 out of 38	NIL		J Rural Studies 2.818	11%
Canadian Jrl of Dietetic Practice & Research	0.237	50 out of 53	NIL		Prog Lipid Res 11.372	2%
Canadian journal of economics	0.635	84 out of 175	0.153	139 out of 161	Quart J Economics 4.775	13%
Canadian Journal of Political Science	0.176	73 out of 84	0.452	23 out of 73	Am Pol Sci Rev 3.233	5.5%
Canadian Journal of Sociology	0.383	63 out of 94	0.333	58 out of 95	Am J Sociology 3.262	12%
Canadian Journal on Aging	0.224	22 out of 24	0.480	12 out of 26	J Gerontol A – Biol 3.500	6.4%
Canadian Modern Language Review	0.304	37 out of 42	0.044	36 out of 40	J Mem Lang 2.815	11%
Canadian	0.648	52 out of	0.426	51 out of 108	Ann Rev of	6.6%

¹ ISI provides impact factors with an unrealistic number of decimals. In calculating the percentages, we have rounded off the results to two decimals.

Psychology		101			Psychol 9.784	
Canadian Public Administration	0.067	25 out of 25	0.193	19 out of 24	J Pub Admin Res Theor – 1.451	4.6%
Canadian Public Policy	0.295	20 out of 25	0.200	18 out of 24	J Pub Admin Res Theor – 1.451	20%
Industrial Relations	1.657	2 out of 16	1.148	3 out of 17	Br J Ind Relat 1.689	98%
International Journal	0.284	38 out of 50	0.358	27 out of 50	Int Security 2.630	10.8%
Isis	0.778	4 out of 29	0.344	15 out of 26	Biol Philos 1.055	74%
Pacific Affairs	0.353	20 out of 33	0.344	15 out of 35	China J 1.174	30%

Table 1: The 21 titles that appear with impact factor in 1997 and/or 2005

It must be immediately noted that no SSHRC-supported humanities journal appears in the analysis. This is not surprising: the *Journal Citation Reports* includes only two categories, science and social science; sampling the social science list of 1747 titles showed that few humanities journals were present. Only a few history and ethics journals were spotted.

For the social science titles, the following results emerged:

- The majority of SSHRC-supported journals simply do not appear in ISI.
- Those that appear, with very few exceptions, hold a very modest rank. Their impact factor compared to the leading publication in their own field is often minuscule.
- Only two titles are ranked in the top ten of their respective fields:

The absence of French-language journals is not surprising given ISI's general bias in favour of English-language publications. In the case of English-language journals, the exclusion from ISI's lists means a very low status: one cannot expect them to be read very much, and, therefore, they cannot be cited very much either. This in turn probably results from a general lack of accessibility: in other words, many Canadian scholarly journals are probably not very widely available (or visible) in foreign libraries, even when they are integrated in aggregators' packages. The relative invisibility of Canadian journals also brings into question the promise that aggregators can significantly enhance the visibility of their journals. Humanities journals, for reasons already mentioned, remain excluded from this particular analysis.

Impact factors of Canadian journals have on the whole increased 31% between 1997 and 2005, but this may be due to a variety of factors, including the growth of the ISI lists across the years. As more journals are scanned by ISI, more citations are collected, which should lead to higher impact factors.

4 Discussion

4.1 Transition to e-publishing

It appears that for SSHRC-funded journals, the transition to e-publishing is well underway. However, a number of the more established journals are still available in paper only. Moreover, when editors speak about e-publishing, they may have in mind a quick-and-dirty conversion of scanned images into pdf format with (perhaps) some ability to carry out full-text searching. Some digitization operations appear to be left in the hands of aggregators who use such fast solutions in providing articles online. The issue of meta-data is rarely addressed in a lucid way. Neglecting these issues threatens long term preservation or interoperability of formats across time. If libraries have access to inferior digital files, they will not be able to participate in the preservation effort and valuable scholarship could be lost forever.

4.2 The Role of Aggregators or Third Party Providers

The survey showed that most on-line access to SSHRC funded journals is provided through a variety of platforms, developed by both profit and non-profit organizations that offer a variety of services and interfaces. In some cases the publisher is also the on-line provider [e.g. Blackwell]; in other cases the vendor is a third party provider who acquires the rights from the journal publisher, usually but not necessarily an association, to provide electronic access to a title via a package of electronic journals that the vendor sells to libraries [e.g. Proquest]. Usually the aggregator will offer for a fee an electronic copy of an individual article in a journal which it controls if the requester does not have access through a subscription.

The qualitative part of the survey also reveals that there is considerable confusion, on both sides of the linguistic divide, regarding electronic publishing and Open Access. This is doubtlessly the result of the ease with which journals may be accessed when an institution subscribes to an electronic package of journals. These journals appear openly accessible only to individuals who are members of a particular university community. In fact, beyond the circle of readers with access privileges, they are “toll gated.”

However it was not clear from the survey whether journals that joined an aggregator resulted in increased usage of the journal, as many reported that they did not receive any additional revenue from aggregators. Nor were aggregators generally open to providing journals with usage statistics, as journals are generally bundled into packages and licensed to libraries in complicated schemes.

4.3 Compensation to the Journals from Aggregators

Our survey also indicates that many of the respondents to the survey appeared to be unsure of the exact nature of their contractual relationship to the publishers and/or aggregators, for example with regard to rights ownership as well as financial compensation. In some cases, it even appears that some contractual agreements between aggregators and journals are not being fulfilled.

The use of aggregator services comes with a cost. The economics of the services are difficult to study since many of these arrangements are confidential and aggregators are reluctant to discuss them. Aggregators are important here they are widely used by SSHRC funded journals. In theory they provide value-added services to end users; and in any economic analysis represent part of the cost of scholarly communication. However, given the low citation ranking of many of the journals who also use aggregators, it is not clear whether joining an aggregator and restricting access to the journal contents represent good return on investment.

Many editors of SSHRC journals who completed the grant application forms were open about these arrangements but it is impossible to say if this is the case in all successful applications. It is reasonable to consider that the SSHRC application form should be revised to allow the journal editor to identify specifically the aggregator used and the cost arrangements that have resulted.

4.4 Support for a Modified form of Open Access was Strong

The result to this question is interesting in that it emanates from a set of individuals that actually wear two hats: on the one hand, editors are also researchers and they know, from that perspective, what is good for them; on the other hand, as editors responsible for the financial well-being of journals that often need careful nurturing, they are concerned about the economic effects of Open Access on their publication. This probably explains the muted agreement in favour of some modified form of Open Access, in particular the request for a moving wall, the purpose of which is to minimize financial risks for the journal due to perceived lost of subscription.

The example of many journals in the *Érudit* collection seems to indicate that most journal editors feel fairly confident about not losing revenue with a two-year moving wall. This looks conservative to the authors of this study. There is also the perception that unlike literature in the sciences, papers in the humanities and social sciences have longer “half-life” and therefore a longer moving wall is necessary. Currently, there is no empirical evidence to support or refute this perception.

4.5 Mandating Open Access Is Clearly Not Endorsed By Editors

Academics do not like being forced into anything and, even though they may favour Open Access, they are intent on preserving their ability to choose freely. Obviously, using the argument of public funding to force Open Access on journals may generate a revolt. The fear is that a forced march toward Open Access could be

destructive given the uncertain financial implications.

A far more compelling case can be made on the basis of the public good that will come to the Canadian people when journals that cover topics such as adoption, mothering, social policy issues, immigration, refugees, the environment, Shakespeare and the theatre, to name just a few, are available readily to all citizens. Adult education and broad learning will advance. Such access can only improve the knowledge and well-being of the Canadian people, but the economic case for such social benefits has yet to be made and remain an important area for future study [12].

4.6 Support for Institutional Repositories

The apparently different attitudes of francophones and anglophones with regard to institutional repositories may be the unexpected consequence of the presence of *Érudit*. Since most francophone journal editors involved with *Érudit* seem to accept two-year moving walls, it may be that they wonder what the uses of institutional repositories are. It must also be remembered that *Érudit* itself incorporates a depository which further confuses the issue. On the English side, the distinction between repositories and OA journals may be a good deal clearer precisely because they are handled in very different locations: the repositories are generally in the hands of librarians while the journals are in the hand of a publisher, a scholarly association, or one (or several) aggregators.

Repositories will remain important to ensure the long-term preservation of the national scholarly heritage and librarians are very much needed in this role. It is one of their traditional functions and publishers are certainly not the best placed to take on this role. Publishers appear and disappear, while libraries remain stable. Even Elsevier has agreed to leave the preservation issue in the hands of the Royal Dutch Library. Many a small publisher of Canadian scholarly journals will disappear before Elsevier does.

4.7 SSHRC Funding for Open Access

The last response confirms hints and trends already noted above. The researcher part of the editor wants Open Access; the editor is willing to go there if there is no risk. Should SSHRC find the ways to finance Open Access, the probability is that most Canadian editors would follow the Brazilian SciELO model [13] and accept Open Access without any hesitation. In fact, they would welcome it as it would certainly enhance the international visibility of their publications. And once Open Access is guaranteed for the electronic version of the journal, the issue of a paying paper version can become an interesting strategy to bring revenue to associations or similar organizations. In any case, what is urgently needed here are some experiments and data gathering to properly assess the economics of OA publishing and the added funding needed.

Concern also exists about SSHRC's potential intent to fund open-access journals and, in particular, the impact of funding open-access titles upon the funding of traditional journals. The fear, it appears, is to see a limited pie divided into a greater number of smaller slices.

Should SSHRC decide to finance open-access journals, maintaining quality was the essential issue from the perspective of editors; on the other hand, editors were silent about relying on the number of subscribers as a criterion of funding. It appears that, in the electronic world, especially with the various packages offered by aggregators to their customers, the evaluation of usage has to be revised and can no longer safely rest on numbers of subscribers.

4.8 Moving Beyond ISI Impact Factor

Results of the citation analysis suggest that authors publishing in the SSHRC-supported journals will not be readily cited given their low visibility, at least according to ISI's JCR. With regard to impact factors, SSHRC-supported journals display characteristics similar to those observed in most journals from the developing nations. They are national journals rather than international journals, in the sense that their visibility abroad is very limited. Like journals from developing countries, SSHRC-supported journals often suffer from a vicious circle: low impact factors induce low submission rates of generally less significant articles that attract little attention and, therefore, few subscriptions. In other words, and, given ISI's claim that they select the best journals in any given field, this survey raises the general issue of perceived quality and most important, usage of SSHRC-supported social science periodicals (including law journals). The survey also raises the question regarding return on SSHRC's investment as journal articles that are not widely read and cited translate into low research uptake and impact. The question that SSHRC must address is whether it makes sense to implicitly encourage

journals to close off access to the content for the sake of a limited number of subscribers, number that are required by SSHRC's funding criteria. Or does it make more sense to trade-off the limited economic return from subscription with a potentially much larger return on readership, which may in turn leads to higher submission, usage and visibility.

Given the fact that most journals supported by SSHRC do not have impact factors, another issue emerges: what alternate evaluation criteria should be applied to these journals, particularly non-subscription based open access journals, applying in the next funding round? Obviously, when titles are available in electronic format on the Internet, new kinds of metrics can be applied, such as hits, downloads, links and, of course, citations. Development and implementation of such new indexes for the evaluation of open access journals is clearly a priority for the scholarly community and for SSHRC. In this regard it is encouraging to see the growing number of studies and projects that aim to provide alternative and better measurement and metrics of usage and research impact, particularly for literature that are openly available [14].

5 Conclusions

The results of the study indicate that many of the journal editors understand that providing Open Access will greatly improve the visibility and citation impact of their journals. However, many editors also worry about the financial conditions under which the transition to OA can be managed. As it stands, the return on research investment, at least as measured in citation counts, is poor for most of the SSHRC subsidized journals. Providing a special source of funding to offset possible losses of subscription revenue could become a strong incentive to move toward Open Access. Given that many of these journals have small subscription revenues, the needed financing, which could take the form of a kind of insurance policy, ought to be quite limited, but the precise amount is difficult to determine at this stage and a separate study would be required.

There is also considerable consensus that SSHRC should support open access journals and encourage journals that wish to experiment with conversion to Open Access to work collectively in a SSHRC-supported experiment designed to better understand the financial implications, author's uptake, and usage of publication before and after becoming Open Access. The experimentation will provide better data to gauge the financial viability of Open Access publishing. These results will be useful in turn to examine whether scaling up the process to a larger number of journals is desirable [mention the new funding program in a footnote?].

Perhaps the most valuable consequence of this study has been the important recognition that there is no magic way to move into electronic publishing and Open Access. Testing, exploring and experimenting while consulting and dialoguing should be the principles under which any kind of action plan should be undertaken.

With regard to electronic publishing, environmental pressures as well as various forms of inducements on the part of aggregators or some publishers have led to a transition carried out in such a wide variety of ways that "chaos" might well be the best term to describe it. In the process, SSHRC is finding itself subsidizing some very profitable commercial aggregators, while denying support to some innovative Open Access journals that are deserving of help except for the fact that they do not have any paying subscribers.

A good reason for this chaotic transition to e-publishing may well have been the consequence of the inability to create orderly experiments so as to identify best practices and enhance the sharing of new know-how. Only in Québec has there been the semblance of an organized move toward electronic publishing [15], but it may have been done in such a centralized manner that it may not fit the ethos of the rest of the country. Nonetheless, it remains a valuable source of experience. Elsewhere, there are dispersed endeavours to produce electronic journals, most of the time on tiny scales [16]. On the non-commercial front, only a very few university presses have developed in-house capacity in this regard. Whether they are willing to share this know-how is far from obvious.

Finally, it seems clear that SSHRC must take on a leadership position in this regard. As other granting councils in the USA and in Europe (particularly the UK, Germany and France) have amply shown, this is to SSHRC's advantage. More fundamentally, if SSHRC does not show some national leadership, no one else will do so, except perhaps in the form of some bid to become the monopolistic device for SSH publishing in the country. Clearly, no one wants this outcome. No one wants one university press, or, even worse, a large commercial press to become the sole provider of e-publishing services to Canadian SSH. At the same time, it is clear that the emerging digital environment is challenging all publications to globalize in an effective manner.

It is natural that journals in similar disciplines be grouped together so that a particular journal platform tends to become well known for its coverage in, for example, economics or law, and a number of journals from a variety of publishers in many countries might be housed in that disciplinary platform. But is the notion of a national platform be useful to researchers used to work on well-focused issues with information coming from all over the planet? In other words, how does one reconcile the idea of a national strategy for scientific and scholarly publishing with the universal characteristics of validated knowledge? These questions lie beyond the scope of the present study, but they are part of the changing landscape of Canadian scholarly communications as it impacts SSHRC-supported journals and they should not be neglected.

Afterword

We are happy to report that in late March 2007, SSHRC announced a new one-year experimental program in support of open access journals [17]. The program adopted several recommendations from our initial report submitted to SSHRC in August 2006 [18]. Amongst the key innovations of the program is the adoption of usage based metrics and cost per article as basis for funding. Of course peer review and the expertise of the editorial board remain as primary quality criteria, but the addition of alternative usage and impact metrics should allow innovative open access to gain the funding support they deserve. We eagerly await the outcomes of this experimentation and we hope this program will generate the much needed economic and usage data for better planning and support of a broader range of open access journals in the humanities and social sciences.

Acknowledgements

We would like to express our thanks to all the journal editors who participated in the online questionnaire. We also like to thank the following individuals who generously gave their time for interviews: Alan Burke, Lynn Copeland, Ann-Marie Corrigan, Carlin Craig, Gunther Eysenbach, Rowley Lorimar, Carol Moore, Brian Owen, and John Willinsky. This study was supported by a Social Science and Humanities Research Council's President's Grant.

Notes and References

- [1] ANTELMAN, K. (2004). Do open-access articles have a greater research impact? *College & Research Libraries*, 65(5): 372-382. (Online). Accessed Jan. 15, 2007, from http://eprints.rclis.org/archive/00002309/01/do_open_access_CRL.pdf. See also the regularly updated bibliography of citation impact studies maintained by Steve Hitchcock: <http://opcit.eprints.org/oacitation-biblio.html> , accessed March 1, 2007
- [2] Following the NSF funded report on Cyberinfrastructure for the natural sciences, the American Council of Learned Societies and the Mellon Foundation also funded a parallel study supporting the development of cyberinfrastructure for the humanities and social sciences: <http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf>
- [3] <http://www.sshrc-crsh.gc.ca/>
- [4] See the The ROARMAP list of the strongest funder and university policies: <http://www.eprints.org/openaccess/policysignup/>
- [5] SSHRC's original position on OA is no longer available on its web site http://www.sshrc.ca/web/about/council_reports/news_e.asp, first accessed in August 2004. But SSHRC's position on OA is even more clearly stated in the context of its recent announcement (March 29, 2007) on a new "Aid to Open Access Research Journals" funding program: http://www.sshrc.ca/web/apply/program_descriptions/open_access_journals_e.asp , accessed April 1, 2007
- [6] The survey result was made available by David Moorman, Senior Policy Advisor at SSHRC, at a meeting on March 9, 2006: http://open.utoronto.ca/index.php?option=com_content&task=view&id=234&Itemid=226
- [7] SurveyMonkey, www.surveymonkey.com, was used to administer the questionnaire
- [8] Érudit is a digital publishing and dissemination platform that originated at Les Presses de l'Université de Montréal in 1998 and has since evolved into a network that support a large variety of journals, mostly from the province of Quebec. <http://www.erudit.org/>

- [9] The 200 existing subscribers was one of the requirements for journals to qualify for SSHRC funding. However, this clearly exclude journals that are already open access but still require financial support. It was also clear that some journals that were affiliated with a scholarly association were using the association's membership to inflate the number of subscribers, thereby qualifying them for the grant.
- [10] ARCHAMBAULT, E.; VIGNOLA-GAGNE, E; COTE, GREGOIRE; LARIVIERE, V; GINGRAS, Y. (2006) Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics* 68(3):329-342
- [11] HICKS, D. (2005). The four literatures of social sciences. In *Handbook of Quantitative Science and Technology Research*. Edited by H. F. Moed, W. Glänzel and U. Schmoch. Page 473-496. Springer Netherlands.
- [12] Studies on the enhanced economic benefits of Open Access on research and development have recently been made, for example: HOUGHTON, J.W., STEELE, C. AND SHEEHAN, P.J. (2006) *Research Communication Costs in Australia, Emerging Opportunities and Benefits*, CSES Working Paper No. 24, Centre for Strategic Economic Studies, Victoria University, Melbourne. Available <http://www.cfses.com/documents/wp24.pdf> ; HOUGHTON, J.W., SHEEHAN, P.J. (2006) *The Economic Impact of Enhanced Access to Research Findings*, CSES Working Paper No. 23, Centre for Strategic Economic Studies, Victoria University, Melbourne. Available <http://www.cfses.com/documents/wp23.pdf> . But the social benefits of OA to scholarly literature have yet to be well studied and documented.
- [13] SciELO stands for Scientific Electronic Library Online www.scielo.br . It is a pioneering project in providing open access to scientific journals published in Brazil, and now from other Latin American countries.
- [14] BRODY, T.; HARNAD, S.; CARR, L. (2006) Earlier Web Usage Statistics as Predictors of Later Citation Impact. *Journal of the American Association for Information Science and Technology (JASIST)* 57(8):1060-1072. Available: <http://eprints.ecs.soton.ac.uk/10713/>
HARNAD, S. (2007) *Open Access Scientometrics and the UK Research Assessment Exercise*. In *Proceedings of 11th Annual Meeting of the International Society for Scientometrics and Informetrics* (in press), Madrid, Spain. Available: <http://eprints.ecs.soton.ac.uk/13804>
BOLLEN, J.; VAN DE SOMPEL, H.; SMITH, J.; LUCE, R. (2005). Toward alternative metrics of journal impact: a comparison of download and citation data. *Information Processing and Management*, 41(6):1419-1440. Available: <http://public.lanl.gov/herbertv/papers/ipm05jb-final.pdf>
BOLLEN, J.; VAN DE SOMPEL, H.(2006). Mapping the structure of science through usage. *Scientometrics*, 69(2):227-258. Available: <http://library.lanl.gov/cgi-bin/getfile?LA-UR-05-7070.pdf>
- [15] We are referring to Érudit, see note number 8.
- [16] Just exactly how many self-started open access journals in the humanities and social sciences produced in Canada is not known and a systematic survey of these journals and their funding and editorial practices would be an important future project.
- [17] Details of the funding program, eligibility criteria, and adjudication process are available on SSHRC's web site. Accessed April 1, 2007
http://www.sshrc.ca/web/apply/program_descriptions/open_access_journals_e.asp
- [18] http://www.sshrc.ca/web/about/publications/journals_report_e.pdf

Appendix I: Questionnaire Sent to Journal Editors or Key Contacts

A. Transitioning to electronic publishing

1. Are the articles in your journal available electronically on the Internet?

Yes

No

If “no”, please skip to section B.

2. When did your journal become available on-line?

3. Is your journal available electronically through an aggregator or a portal?

Yes

No

4. If “yes” to question 3, please specify which aggregator or portal you are using:

Blackwell

Ebsco

- Érudit
- Hein
- JSTOR
- Lexis/Nexis
- Project Muse
- Proquest
- Association web site
- University web site
- University Press web site
- Other (please specify)

5. Are all issues, including the most recent, available on-line?

Yes

No

6. If you answered “no” to question 5, please specify which years have been digitized.

7. Who owns the digital rights to your journal?

B. Questions regarding your publisher

8. If the publisher of your journal is not your scholarly association, is your publisher financing part of the activities related to the publication of your journal? (for example, editorial stipend, peer review process, etc.)?

Yes

No

9. Does your journal receive compensation from an aggregator on a pay-per-use basis?

Yes

No

C. The issue of Open Access

10. The Open Access movement: Putting peer-reviewed scientific and scholarly literature on the internet. Making it available free of charge and free of most copyright and licensing restrictions. Removing the barriers to serious research. “Open Access News”, <http://www.earlham.edu/~peter/fos/fosblog.html>.

11. Are you in favour of Open Access in principle (leaving economic issues aside for the moment)?

Yes

No

12. In order to provide Open Access to your journal, you will have to devise a new business plan for your journal. Which business plan would you favour?

13. All issues immediately available, including the latest (true Open Access)?

A “moving wall” with the latest issues available only through subscriptions, and the earlier issues available in Open Access?

A publishing fee for all accepted articles?

A choice between “b” and “c” offered to authors according to their ability/willingness to use funds from various sources to publish?

14. Should SSHRC provide financial support for journals to become Open Access and non-subscription based?

Yes

No

15. Should SSHRC provide support for Open Access journals and consider eligibility criteria appropriate for these titles

Yes

No

16. If you have suggestions or comments on what these criteria might be, please list them here.

The Research Impact of Open Access Journal Articles

Yaşar Tonta; Yurdagül Ünal; Umut Al

Department of Information Management, Hacettepe University
06532 Beytepe, Ankara, Turkey
e-mail: {tonta, yurdagul, umutal}@hacettepe.edu.tr

Abstract

The availability of scientific and intellectual works freely through scientists' personal web sites, digital university archives or through the electronic print (eprint) archives of major scientific institutions has radically changed the process of scientific communication within the last decade. The "Open Access" (OA) initiative is having a tremendous impact upon the scientific communication process, which is largely based on publishing in scientific periodicals. This exploratory paper investigates the research impact of OA articles across the subject disciplines. The research impact of OA articles as measured by the number of citations varies from discipline to discipline. OA articles in Biology and Economics had the highest research impact. OA articles in hard, urban, and convergent fields such as Physics, Mathematics, and Chemical Engineering did not necessarily get cited most often.

Keywords: open access articles; research impact; scholarly communication; citations analysis

1 Introduction

There are some 24,000 scientific journals publishing 2.5 million articles each year. Scientific journals are expensive. The economic model of publishing is based on subscription and licensing. Price hikes in the publishing sector within the last 30 years are well beyond the inflation rates. This has been primarily due to lack of competition. Some publishers can easily become monopolies, as no two journals can publish the same article in view of copyright restrictions. Moreover, those who use the scientific journals (scientists) and those who pay for this service (usually libraries) are different, which results in what is called the "price inelasticity" in economics and empowers the scientific journal publishers further [1]. As scientific journal prices increase, some libraries cancel some of their subscriptions because they cannot afford the price hikes. Publishers then increase prices further to make up the lost income. Consequently, some more libraries discontinue their subscriptions. In response, to make up the lost income, publishers increase the prices again. This vicious circle is not only the main cause of the so called "serials crisis," but also it affects the scientific communication process. Interestingly, the lack of competition in scientific journal publishing enables some publishers to increase their market shares by increasing prices. When the price of an already expensive journal is further increased, libraries tend to cut off subscriptions to cheaper but prestigious journals in order to keep the more expensive ones [2].

Scientific research and its outcome (e.g., scientific journal articles) get supported primarily by public money. Articles are given by scientists to commercial publishers free of charge and refereed by scientists free of charge. Yet, the same scientists pay dearly, through their libraries, to subscribe to the very same journals despite the fact that their salaries are paid for by public monies and their libraries are supported by public funds. The triple payment of public money to support research projects, to pay for salaries of scientists, and to fund libraries is emphasized by the following comment: "What other business receives the goods that it sells to its customers from those same customers, a quality control mechanism provided by its customers, and a tremendous fee from those same customers?" [2]. Universities and governments have recently begun to scrutinize the scientific communication process. Web access to research articles created new opportunities and showed that alternative or complementary economic models can be experimented with [3, 4].

One of these models is what is called Open Access (OA). OA is defined as "free (...) access to" scientific publications. "A complete version of the work (...) is deposited (and thus published) in at least one online repository (...) maintained by an academic institution, scholarly society, government agency, or other well-established organization that seeks to enable open access, unrestricted distribution, inter operability, and long-term archiving" [5]. OA increases the research impact by making articles available, free of charge, to all those interested. Two parallel and integrated strategies to create a more effective and equitable scientific communication process are suggested: (1) researchers "self-archiving" their articles that are published in

refereed journals in their web sites or institutional repositories and making them available through the Internet; and (2) researchers publishing their articles in OA journals. More than 90% of commercial publishers support self-archiving. There are currently more than 2,500 OA journals published in all subjects.

Several prominent institutions including OECD and UN support OA. Recently, some universities decided to mandate researchers to self-archive their published articles. A bill (Federal Research Public Access Act) mandating OA to publicly-funded scientific publications in the United States is likely to become enacted in the near future. The European Commission (EC) recommends OA to EC-funded research reports [6, p. 87]. Governments allocate billions of dollars of taxpayers' money to research. For instance, the annual budget (28 billion dollars) of the US National Institute of Health alone is higher than the GDP of 142 nations [7]. OA increases the impact of the publicly-funded research and triggers new research projects, thereby increasing the return on investment [8-11].

In this paper we look into the research impact of OA journal articles in sciences, social sciences, and arts and humanities. The term "research impact" in this study is defined as the number of times that each article is cited in the literature. Journal articles representing nine disciplines were selected from the Directory of Open Access Journals (www.doaj.org). Citations to each article were identified through Elsevier's Scopus. The research impact of articles in different disciplines was compared to find out the underlying trends. Findings were discussed in light of why OA is supported in varying degrees in sciences, social sciences, arts and humanities.

2 Literature Review

It has for long been observed that scientific communication processes differ in sciences, social sciences, and arts and humanities. While scientists publish their contributions primarily in journals as articles, social scientists and scholars of arts and humanities prefer monographs as the main outlet of their contributions. Whereas journal articles constitute 90% of all publications in sciences, books and monographs in social sciences constitute 40% of all publications [12]. The intensity of production also differs from discipline to discipline. In chemistry it is not uncommon for a researcher to produce several journal articles in a given year whereas a social scientist would publish a single article perhaps every other year or so. Some social scientists and humanities scholars may not even bother to publish journal articles but concentrate on publishing a few monographs instead throughout their academic careers. "Disciplinary cultures" have an impact on scholarly communication processes and the ways by which researchers in each discipline communicate their findings [13].

The emergence of the Internet and electronic publishing in the early 1990s has profoundly changed the scientific communication patterns. While physicists and computer scientists, for instance, reacted very quickly and began to use electronic publishing as a means of disseminating research results over the Internet, social scientists and arts and humanities scholars were somewhat slow to react. For some researchers the acceptance of electronic publishing in support of scientific communication was "not just a matter of time": field differences have to a large extent determined the acceptance levels [14]. Electronic publishing is seen as a transitory period by some researchers, for example. Some do not trust the electronic media while others see electronic journals inferior compared to printed journals. Copyright concerns discourage some researchers. Reasons are too numerous to discuss in detail here. These cultural issues shape the scholarly communication and explain the degree of use of electronic journals across the fields [15].

Field differences and disciplinary cultures also played an important role in OA movement since mid-1990s. Similar concerns shied away some researchers from self-archiving their contributions through their personal web sites or institutional archives. While almost all articles in sciences (e.g., physics and mathematics) have currently been open access, the percentages are much lower in social sciences, arts and humanities (e.g., 60% in economics, 25%-30% in political science, psychology and sociology, and less than 20% in anthropology and geography) [16, p. 88]. Only 5% of social scientists self-archive their papers.

As mentioned earlier, OA makes scientific papers more visible and increase their research impact [8-11]. OA articles get cited more often by other researchers, thereby bringing their authors more recognition and prestige, and providing them incentives to do more research. The *Proceedings of the National Academy of Sciences* (PNAS) is a prestigious journal with an high impact factor (IF) publishing both OA and non-OA articles. OA articles published side by side with non-OA articles at PNAS were cited more quickly and twice as many times than non-OA articles [17]. This finding is somewhat contradictory with that of an earlier study [18] that analyzed the impact factors and citation patterns of OA journals in ISI databases and found that OA journals usually have lower IFs than non-OA journals in their subject categories. It appears that OA articles help increase the IF of a prestigious journal even further.

Earlier studies tended to measure the research impact of OA journal articles mainly by using the Web of Science (WoS) database of ISI (now Thomson Scientific). WoS at that time did not index that many OA journal titles. The situation has changed in 2004, however. Elsevier's Scopus and Google's Google Scholar (GS) citation databases were introduced almost at the same time in November 2004. These databases track citations that come from refereed journals as well as those from resources available on the Web. The overlapping citations between WoS and Scopus, and WoS and GS are not as high as one would expect (58% and 31%, respectively, for articles in library and information studies) [19]. Scopus covered the library and information studies (LIS) literature more comprehensively and retrieved 26% unique citations that were not retrieved by WoS. The percentage of unique citations retrieved by GS was somewhat lower (21%). In different studies, WoS retrieved higher citation counts for articles that were published in 1985 in the *Journal of American Society for Information Science* and for articles in oncology and condensed matter physics in 1993 than Scopus and GS [20, 21]. This is primarily due to the fact that the WoS database goes back to 1900s while the Scopus database cover citations since 1996. (Information is not available for GS.) Jacso [22] reviewed these three citation databases in more detail and compared them in terms of their major features such as database subject coverage and composition, number of records, and search and retrieval characteristics.

3 Research Questions

As reviewed earlier, the research impact of both OA and non-OA articles has been addressed in the past. There is a considerable difference between scientific disciplines in terms of both the rates of research impact and the acceptance of OA as a means of dissemination of research results. Antelman [11] found that OA articles in mathematics and electrical and electronics engineering have a greater research impact than that in political science and philosophy. In a different study Antelman [16] identified different degrees of acceptance of self-archiving in six social science disciplines (economics, sociology, geography, political science, anthropology, and psychology). Based on Becher and Trowler's [13] and Whitley's [23] studies, she posited that "differences between disciplines can be characterized in terms of the degree of mutual dependence between researchers and the degree of task uncertainty in defining shared problems, goals, and procedures" [16, p. 92]. The interdependency in social science disciplines is low and common issues and objectives are defined ambiguously. Moreover, the rates of self-archiving practice were found lower in divergent social science disciplines that concentrate on rural issues (e.g., anthropology, geography, sociology and psychology) and higher in convergent ones that concentrate on urban issues and have close relationships with other disciplines (e.g., economics) [16, p. 92].

Antelman's interpretation of her findings seems interesting. If such a relationship between self-archiving rates and different scientific disciplines exists, one would think that a similar relationship may also hold true for varying degrees of research impact of OA articles in different fields. This paper aims to explore the conjecture that OA articles in the interdependent, convergent and urban disciplines would have higher research impact than that of independent, divergent and rural disciplines.

What is meant by hard/soft, urban/rural, and convergent/divergent fields is that "Physics represents hard science, which is convergent and urban in its social aspects; history is a soft discipline, relatively convergent and rural; sociology is a soft, divergent, and rural discipline; whereas biology is both mostly rural science, and also a mixture of soft and hard elements" [24 p. 68].

Nine fields under three groups were identified along this continuum of hard/soft, urban/rural, convergent/divergent and interdependent/independent scientific fields. In the first group, physics, mathematics, and chemical engineering represent hard and applied sciences that are convergent and urban in their social aspects. In the second group, economics, biology, and environmental science represent disciplines that have both hard and soft components. Economics is a more urban discipline than both biology and environmental science in this group. In the last group, sociology, psychology and anthropology represent soft, divergent, and rural disciplines. According to Whitley's [23] dimensions, disciplines in the first group have "high degree of mutual dependence and low degree of task uncertainty" while the ones in the last group have the opposite. The disciplines in the last group lie somewhere in between.

This paper addresses the following research questions:

- Does the research impact of OA articles differ across the fields in sciences, social sciences, and arts and humanities?
- If it does, do OA articles in hard, urban and convergent fields receive more citations (hence higher research impact) than those in soft, rural, and divergent ones?

4 Methodology

What follows is a detailed account of the sampling process of articles published in OA journals. The Directory of Open Access Journals (DOAJ, www.doaj.org) lists more than 2,500 OA journal titles. It was used to select OA journals representing nine disciplines (physics, mathematics, chemical engineering, economics, biology, environmental science, sociology, psychology and anthropology). The detail of each journal title (subject, year, language) was recorded (January 2007). Non-English journal titles and titles that did not have enough back issues (since 1999) published were excluded from the sample frame. DOAJ (www.doaj.org) assigns one or more subject headings to each journal title. Journal titles with a single subject heading were preferred.

Journal titles not covered by Elsevier's Scopus were excluded since Scopus was used to identify citations that each selected article received (more below). It was noted in the Scopus web site (info.scopus.com) that Scopus is the largest abstract and citation database of research literature containing 29 million abstracts from about 15,000 peer-reviewed journal titles in all fields along with 265 million citations. Abstracts and citations go back to 1966 and 1996, respectively.

The total number of articles published in OA journals in 1999, 2001 and 2003 were identified for selected nine disciplines. A sample of 30 articles was selected to represent each discipline, thereby making a total of 270 articles for all nine disciplines. Needless to say, sampling intervals were different for each discipline. As the number of OA journals in each discipline varied, articles in the samples for some disciplines came from a few journals (e.g., anthropology). Similarly, the number of articles published in some disciplines were much higher (e.g., physics), thereby making the sampling rates uneven across fields (Table 1).

Subjects	# of journals in DOAJ	# of journals in the sample	# of total articles in OA journals	# of OA articles taken from the sample journals	sample rate
Physics	23	6	2,543	30	1.2
Mathematics	77	16	1,092	30	2.7
Chemical Engineering	6	3	818	30	3.7
Economics	36	2	113	30	26.5
Environmental Sciences	12	3	247	30	12.1
Biology	50	7	690	30	4.3
Psychology	45	4	271	30	11.1
Sociology	33	3	97	30	30.9
Anthropology	22	2	111	30	27.0
Total	304	46	5,982	270	4.5

Table 1: Sampling statistics

All 270 articles were searched on Scopus for citations (March 2007). Retrieval results were entered into SPSS, a statistical analysis software. The number of citations, citing authors and journals along with years, and self-citations were recorded for each article. The citation age of each article was calculated. Various statistical tests were run using SPSS.

5 Findings

Table 2 provides descriptive statistics about citations that 30 OA articles in each subject discipline received. All OA articles ($N = 270$) were cited 761 times ($\bar{X} = 2.8$, $SD = 4.7$). The average number of citations per OA article ranged between 0.8 (Sociology) and 6.4 (Biology), although the distributions of citations for all disciplines were rather skewed (note the standard deviations being always higher than the averages). OA articles in Biology and Economics received almost half of all citations (25.2% and 20.2%, respectively) whereas the ones in Psychology and Sociology did much fewer (3.7% and 3.2%, respectively).

Subjects	# of OA articles	# of citations	%	\bar{X}	SD	# of OA articles with zero citations	median	max
Physics	30	95	12.5	3.2	3.7	9	2	16
Mathematics	30	44	5.8	1.5	1.9	11	1	7
Chemical Engineering	30	63	8.3	2.1	3.2	12	1	16
<i>Subtotal</i>	<i>90</i>	<i>202</i>	<i>26.5</i>	<i>2.2</i>	<i>3.1</i>	<i>32</i>	<i>1</i>	<i>16</i>
Economics	30	154	20.2	5.1	7.5	6	2.5	39
Environmental Sciences	30	63	8.3	2.1	2.8	12	1	13
Biology	30	192	25.2	6.4	7.4	2	4.5	38
<i>Subtotal</i>	<i>90</i>	<i>409</i>	<i>53.7</i>	<i>4.5</i>	<i>6.5</i>	<i>20</i>	<i>2.5</i>	<i>39</i>
Psychology	30	28	3.7	0.9	1.4	17	0	5
Sociology	30	24	3.2	0.8	1.3	20	0	5
Anthropology	30	98	12.9	3.3	5.3	6	2	26
<i>Subtotal</i>	<i>90</i>	<i>150</i>	<i>19.7</i>	<i>1.7</i>	<i>3.4</i>	<i>43</i>	<i>1</i>	<i>26</i>
Grand Total	270	761	100.1	2.8	4.7	95	1	39

Note: The percentage is not equal to 100% due to rounding.

Table 2: Citation statistics of open access articles in different fields

OA articles in the second group of fields received more than half (53.7%) of all citations, followed by the first group (26.5%) and the third group (19.7%). The second group of fields (Economics, Environmental Sciences, and Biology) that have both hard and soft components scored a much higher research impact than either the first group of fields (hard, convergent and urban) and the third group of fields did. The number of citations for each field within groups also differed. For instance, OA articles in Biology and Economics in the second group received much higher citations than that in Environmental Sciences. The difference was even more substantial for OA articles in Anthropology in the third group: they received about four times more citations than that in Sociology and Psychology.

The average self-citation rate for all subjects was 28.4% (216/761). Self-citation rates were much higher in Mathematics (45.5%) and Physics (43.2%) than that in Psychology (7.1%) and Economics (13.6%). More than one third (35%) of OA articles (95/270) were never cited at all. OA articles in Sociology and Psychology had the highest zero citation rates (67% and 57%, respectively) whereas only two out of 30 articles (7%) in Biology went uncited. About 17% (or 45 articles) were cited only once, 15% (40 articles) twice, 7% (20 articles) three times, and a further 26% (70 articles) four or more times. Two OA articles in Economics and Biology received the highest number of citations (39 and 38, respectively). The most-cited 10 OA articles collected 27% (209/761) of all citations (Table 3).

Rank	Authors (Publication Year). Article title. <i>Journal</i> .	# of times cited in Scopus	Subject
1	Berg, A., & Pattillo, C. (1999). Are currency crises predictable? A test. <i>IMF Staff Papers</i> .	39	economics
2	Lyubarsky, A.L. et al. (2001). RGS9-1 is required for normal inactivation of mouse cone phototransduction. <i>Molecular Vision</i> .	38	biology
3	Nishida, T., Kano, T., et al. (1999). Ethogram and ethnography of Mahale chimpanzees. <i>Anthropological Science</i> .	26	anthropology
4T	Plascak, J.A. et al. (1999). Phenomenological Renormalization Group Methods. <i>Brazilian Journal of Physics</i> .	16	physics
4T	Ishida, H. et al. (1999). New hominoid genus from the Middle Miocene of Nachola, Kenya. <i>Anthropological Science</i> .	16	anthropology
4T	Miura, M. (1999). Detection of chromatin-bound PCNA in mammalian cells and its use to study DNA excision repair. <i>Journal of Radiation Research</i> .	16	biology
4T	Yu, Q. et al. (2001). Retinal uptake of intravitreally injected Hsc/Hsp70 and its effect on susceptibility to light damage. <i>Molecular Vision</i> .	16	biology
4T	S.P. Asprey & Naka, Y. (1999). Mathematical Problems in Fitting Kinetic Models—Some New Perspectives. <i>Journal of Chemical Engineering of Japan</i> .	16	chemical engineering
9T	Blanchard, O. & Shleifer, A. (2001). Federalism with and without political centralization: China versus Russia. <i>IMF Staff Papers</i> .	13	economics
9T	Casey, T.G. et al. (1999). Metabolic behaviour of heterotrophic facultative aerobic organisms under aerated/unaerated conditions. <i>Water SA</i> .	13	Environmental sciences

Table 3: The 10 most-cited open access articles

Articles in the sample came from 46 different OA journals across the fields. Fifteen articles that appeared in 7 OA journals in different fields (Environmental Sciences, Mathematics, Physics, and Psychology) received no citations while 7 articles appeared in 7 OA journals (6 in Mathematics, 1 in Physics) received only one citation each (see Appendix). In addition to the Scopus database, half (23) of those OA journal titles were also listed in Thomson Scientific's Web of Science (WoS) citation database. There was no difference, however, between the articles listed in the Scopus database only and that listed in both Scopus and WoS databases in terms of the number of citations they received ($\chi^2_{(21)}=.382, p = .396$).

More than 60% of all citations to OA articles were received within the first three years after their publication (Figure 1). OA articles got cited in the literature less often after three years. The "half-life" (the time it takes to receive half of all citations) was 2 years for OA articles in Physics, Mathematics, Biology, and Psychology, and 3 years in Chemical Engineering, Economics, Environmental Sciences, Sociology and Anthropology.

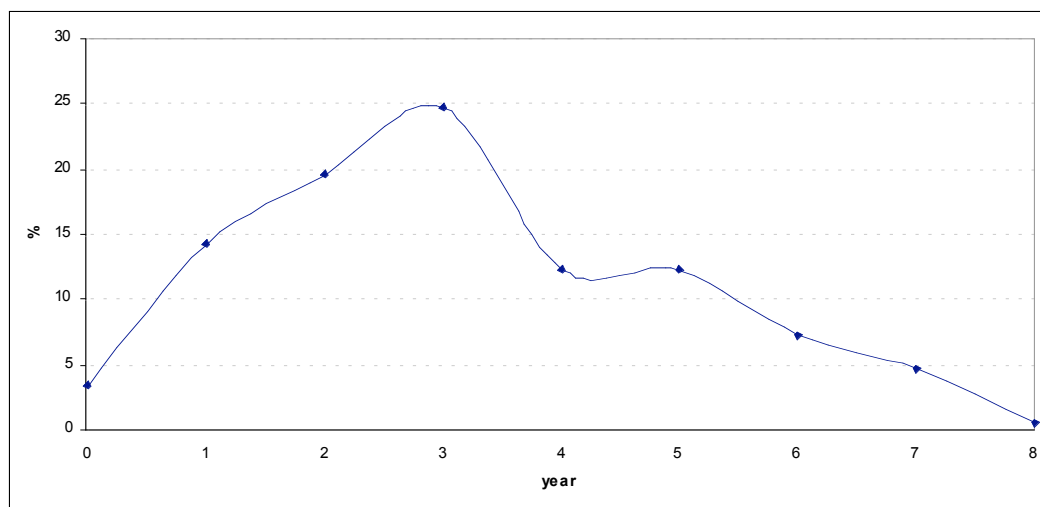


Figure 1: Temporal distribution of citations to open access articles after publication (in years)

6 Discussion

This study confirmed the findings of earlier ones in that the research impact of OA articles differ across the fields. Some subtle differences were observed, however, in terms of the research impact of certain disciplines (e.g., mathematics and anthropology). Antelman [11] found that mathematics had a greater research impact than some social science disciplines (e.g., political science). Yet, OA articles in Mathematics received much fewer citations in the present study and almost half of them were self-citations. Usually, articles in social sciences and humanities get cited much less often. OA articles in Economics and Anthropology were among the most heavily cited ones (after those in Biology).

Such variations in research impact across the fields may be susceptible to the small sizes of samples (30 articles) for each subject discipline and the uneven distribution of sampled articles to journals in respective fields. For instance, OA articles in Mathematics came from 16 different journals, more than half of which received either zero or one citation only (average being 1.5 citations). On the other hand, those in Economics and Anthropology came from two journals in each subject and they collected relatively higher number of citations per article (averages being 5.1 for Economics and 3.3 for Anthropology). This may perhaps be explained by the research impact of articles that appeared in prestigious OA journals in Economics (IMF Staff Papers, Asian Development Review in Economics) and Anthropology (Anthropological Science, and Journal of Physiological Anthropology and Applied Human Science).

The main objective of this paper was to explore if there is a relationship between the research impact of OA articles and the characteristics of the subject fields (e.g., hard/soft, urban/rural, and convergent/divergent). Findings do not seem to indicate any discernible pattern between these two variables. In other words, OA articles in hard, urban and convergent fields such as Physics, Mathematics, and Chemical Engineering did not necessarily have higher research impact than those that have both hard/soft and urban/rural components such as Biology and Economics. In fact, it was just the opposite: OA articles in the second group (Economics, Environmental Sciences, and Biology) received twice as many citations than those in the first group did. OA articles in soft and divergent fields concentrating on rural issues (e.g., Sociology and Psychology) had lower research impact as expected. Although in the same group with Sociology and Psychology, OA articles in Anthropology had higher research impact than all the subjects in the first group (Physics, Mathematics, and Chemical Engineering) and Environmental Sciences in the second group.

Recall that the research question in this study emerged from Antelman's [16] findings on self-archiving rates in different social science disciplines (higher in convergent and urban fields such as Economics, and lower in divergent and rural fields such as Anthropology, Geography, Sociology and Psychology). We hypothesized implicitly that OA articles in hard, urban and convergent fields receive more citations (hence higher research impact) than those in soft, rural, and divergent ones. It appears that the research impact of OA articles in Economics, Sociology and Psychology resembles the behavior of self-archiving. The research impact of OA articles in Anthropology is quite different, however. Moreover, the research impact of hard, urban and convergent fields (Physics, Mathematics, and Chemical Engineering) have no resemblance whatsoever to self-archiving practices. It may well be that self-archiving and research impact measured by the number of citations are two completely different things. It is also highly likely that, as we indicated earlier, the small sample sizes of OA articles in each subject did not allow any trends to emerge. The hypothesis needs to be tested using much larger samples with carefully designed studies.

7 Conclusion

We investigated the research impact of OA articles across the subject disciplines in this exploratory paper and found that it varies from discipline to discipline. OA articles in hard, urban and convergent fields do not seem to have higher research impact as measured by the number of citations than mixed (hard/soft, urban/rural, and convergent/divergent) ones. OA articles in Biology and Economics behaved like hard sciences in terms of research impact. Findings are inconclusive, however. Explanatory studies need to be replicated in order to test the hypothesis that OA articles in hard, urban and convergent fields receive more citations (hence higher research impact) than those in soft, rural, and divergent ones.

References

- [1] MEYER, R.W. (1997). Monopoly power and electronic journals. *Library Quarterly*, 67(4): 325-349.
- [2] HOUSE OF COMMONS. (2004). Select Committee on Science & Technology Tenth Report. (Online). Retrieved, 13 April 2007, from <http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/39902.htm>.
- [3] PROSSER, D. (2004). The next information revolution-How open access repositories and journals will transform scholarly communications. *LIBER Quarterly*, 14 (1), (Online). Retrieved, 13 April 2007, from <http://webdoc.gwdg.de/edoc/aw/liber/lq-1-04/prosser.pdf>.
- [4] WILLINSKY, J. (2003). Scholarly associations and the economic viability of open access publishing. *Journal of Digital Information*, 4(2), Article No. 177. (Online). Retrieved, 13 April 2007, from <http://jodi.ecs.soton.ac.uk/Articles/v04/i02/Willinsky/>.
- [5] BETHESDA STATEMENT ON OPEN ACCESS PUBLISHING. (2003). (Online). Retrieved, 13 April 2007, from <http://www.earlham.edu/~peters/fof/bethesda.htm>.
- [6] EUROPEAN COMMISSION. (2006). Study on the economic and technical evolution of scientific publication markets in Europe. (Online). Retrieved, 13 April 2007, from http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf.
- [7] SUBER, P. (2006). Predictions for 2007. *SPARC Open Access Newsletter*, No. 104. (Online). Retrieved, 13 April 2007, from <http://www.earlham.edu/~peters/fof/newsletter/12-02-06.htm>.
- [8] LAWRENCE, S. (2001) Free online availability substantially increases a paper's impact. *Nature*, 411 (6837): 521. (Online). Retrieved, 13 April 2007, from http://www.copernicus.org/EGU/acp/Nature_ad_1.pdf.
- [9] HARNAD, S.; BRODY, T. (2004 June). Comparing the impact of open access (OA) vs. non-OA articles in the same journals. *D-Lib Magazine*, 10(6). (Online). Retrieved, 13 April 2007, from <http://www.dlib.org/dlib/june04/harnad/06harnad.html>.
- [10] HARNAD, S.; BRODY, T.; VALLIERES, F.; CARR, L.; HITCHCOCK, S.; GINGRAS, Y.; OPPENHEIM, C.; STAMERJOHANN, H.; HILF, E. (2004). The access/impact problem and the green and gold roads to open access. *Serials Review*, 30(4): 310-314. (Online). Retrieved, 13 April 2007, from <http://dx.doi.org/10.1016/j.serrev.2004.09.013>.
- [11] ANTELMAN, K. (2004). Do open-access articles have a greater research impact? *College & Research Libraries*, 65(5): 372-382. (Online). Retrieved, 13 April 2007, from http://eprints.rclis.org/archive/00002309/01/do_open_access_CRL.pdf.
- [12] SUBER, P. (2004). Promoting open access in the humanities. (Working paper). (Online). Retrieved, 13 April 2007, from <http://www.earlham.edu/~peters/writing/apa.htm>.
- [13] BECHER, T.; TROWLER, P.R. (2001). *Academic tribes and territories: intellectual enquiry and the culture of disciplines*. 2d ed. Buckingham: SRHE and Open University Press.
- [14] KLING, R.; McKIM, G. (2000). Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication. *Journal of the American Society for Information Science*, 51(14), 1306-1320.
- [15] FRY, J.; TALJA, S. (2004). The cultural shaping of scholarly communication: explaining e-journal use within and across academic fields. In *ASIST 2004: Proceedings of the 67th ASIST Annual Meeting*, Vol. 41, p. 20-30. Medford, NJ: Information Today. (Online) Retrieved, 10 April 2007, from http://people.oii.ox.ac.uk/fry/wp-content/uploads/2006/03/FryTalja_asistfinalsubmission17May.pdf.
- [16] ANTELMAN, K. (2006). Self-archiving practice and the influence of publisher policies in the social sciences. *Learned Publishing*, 19, 85-95. (Online). Retrieved, 13 April 2007, from http://eprints.rclis.org/archive/00006023/01/antelman_self-archiving.pdf.
- [17] EYSENBACH, G. (2006). Citation advantage of open access articles. *PLoS Biology*, 4(5): e157. DOI: 10.1371/journal.pbio.0040157.
- [18] McVEIGH, M.E. (2004). Open access journals in the ISI citation databases: Analysis of impact factors and citation patterns. (Online). Retrieved, 10 April 2007, from <http://scientific.thomson.com/media/presentrep/essayspdf/openaccesscitations2.pdf>.
- [19] MEHO, L.I.; YANG, K. (in press). A new era in citation and bibliometric analyses: Web of Science, Scopus, and Google Scholar. *Journal of the American Society for Information Science and Technology*.

- [20] BAUER, K.; BAKKALBASI, N. (2005). An examination of citation counts in a new scholarly communication environment. *D-Lib Magazine*, 11, 9. (Online) Retrieved, 10 April 2007, from <http://www.dlib.org/dlib/september05/bauer/09bauer.html>.
- [21] BAKKALBASI, N.; BAUER, K.; GLOVER, J.; WANG, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries*, 7. (Online) Retrieved, 10 April 2007, from <http://www.bio-diglib.com/content/pdf/1742-5581-3-7.pdf>.
- [22] JACSO, P. (2005, November 10). As we may search – Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*, 89(9): 1537-1547. (Online). Retrieved, 10 April 2006, from <http://www.ias.ac.in/currsci/nov102005/1537.pdf>.
- [23] WHITLEY, R. (2000). *The intellectual and social organization of the sciences*. New York: Oxford University Press.
- [24] KEKÄLE, J. (2002). Conceptions of quality in four disciplines. *Tertiary Education and Management*, 8(1): 65-80.

Appendix I: Number of Articles and Citations in 46 Open Access Journal Titles

Journal	Subject	# of times cited	# of articles in the sample	\bar{X}	Indexed in WoS
Acta Physica Polonica B	physics	56	14	4.0	Yes
Brazilian Journal of Physics	physics	21	4	5.3	Yes
Entropy: international and interdisciplinary journal of entropy and information studies	physics	3	2	1.5	No
New Journal of Physics	physics	0	2	0.0	Yes
Pramana: Journal of Physics	physics	14	6	2.3	Yes
Turkish Journal of Physics	physics	1	2	0.5	No
Balkan Journal of Geometry and Its Applications	mathematics	0	1	0.0	No
Bulletin (new series) of the American Mathematical Society	mathematics	1	1	1.0	Yes
Electronic Journal of Differential Equations	mathematics	9	8	1.1	No
Electronic Journal of Linear Algebra	mathematics	0	1	0.0	Yes
Electronic Journal of Qualitative Theory of Differential Equations	mathematics	1	1	1.0	No
Electronic Research Announcements of the American Mathematical Society	mathematics	0	1	0.0	Yes
Electronic Transactions on Numerical Analysis	mathematics	6	1	6.0	Yes
Homology, Homotopy and Applications(HHA)	mathematics	1	1	1.0	Yes
Journal of Graph Algorithms and Applications	mathematics	3	1	3.0	No
Journal of Inequalities and Applications	mathematics	1	1	1.0	Yes
Journal of Integer Sequences	mathematics	4	2	2.0	No
Lobachevskii Journal of Mathematics	mathematics	2	2	1.0	No
Missouri Journal of Mathematical Sciences	mathematics	1	2	0.5	No
The Electronic Journal of Combinatorics	mathematics	6	4	1.5	Yes
The New York Journal of Mathematics	mathematics	1	1	1.0	No
Theory and Applications of Categories	mathematics	8	2	4.0	No
Brazilian Journal of Chemical Engineering	chemical engineering	5	5	1.0	Yes
Iranian Polymer Journal	chemical engineering	4	5	0.8	Yes
Journal of Chemical Engineering of Japan	chemical engineering	54	20	2.7	Yes
Asian Development Review	economics	14	5	2.8	No
IMF Staff Papers	economics	140	25	5.6	Yes
Electronic Green Journal	environmental sciences	2	3	0.7	No
Park Science	environmental sciences	0	5	0.0	No
Water SA	environmental sciences	61	22	2.8	Yes
Biological Procedures Online	biology	12	2	6.0	Yes
Cell Structure and Function	biology	14	2	7.0	Yes
Experimental and molecular medicine EMM	biology	27	5	5.4	Yes
In Silico Biology	biology	3	1	3.0	No
Journal of Biosciences	biology	21	7	3.0	Yes
Journal of Radiation Research	biology	22	3	7.3	Yes
Molecular Vision	biology	93	10	9.3	Yes
Current Research in Social Psychology	psychology	6	5	1.2	No
Dynamical Psychology: an international, interdisciplinary journal of complex mental processes	psychology	0	3	0.0	No
Journal of Technology in Counseling	psychology	0	2	0.0	No
PSYCHE: An Interdisciplinary Journal of Research on Consciousness	psychology	22	20	1.1	No
Journal of Criminal Justice and Popular Culture	sociology	15	12	1.3	No
IDEA: a Journal of Social Issues	sociology	0	4	0.0	No
Journal of Memetics - Evolutionary Models of Information Transmission	sociology	9	14	0.6	No
Anthropological Science	anthropology	58	14	4.1	Yes
Journal of Physiological Anthropology and Applied Human Science	anthropology	40	16	2.5	No