

Proceedings of the 11th International Conference
on Electronic Publishing

Openness in Digital Publishing: Awareness, Discovery and Access

Organised by the Vienna University of Technology (Vienna, Austria)

Vienna
June 13-15, 2007

Editors:

Leslie Chan

University of Toronto at Scarborough (Canada)

Bob Martens

Vienna University of Technology (Austria)

IRIS-ISIS Publications, 2007

**Proceedings of the 11th International Conference on Electronic Publishing
Vienna, 2007**

Vienna University of Technology (Vienna, Austria)
<http://www.tuwien.ac.at>

Edited by:

Leslie Chan, University of Toronto at Scarborough (Canada)
Bob Martens, Vienna University of Technology (Austria)

Published by:

ÖKK-Editions, Vienna

ISBN: 978-3-85437-292-9

First edition

All rights reserved.

© 2007 Leslie Chan, Bob Martens – Editors

© 2007 For all authors in the proceedings

Disclaimer

Any views or opinions expressed in any of the papers in this collection are those of their respective authors. They do not represent the view or opinions of the Vienna University of Technology, University of Toronto at Scarborough, the editors and members of the Programme Committee, nor of the publisher IRIS-ISIS Publications and conference sponsors.

Any products or services that are referred to in this book may be either trademarks and/or registered trademarks of their respective owners. The Publisher, editors and authors make no claim to those trademarks.

Members of the ELPUB2007 Programme Committee

Apps, Ann, University of Manchester (UK)
Baptista, Ana Alice, University of Minho (Portugal)
Bentum, Maarten van, University of Twente (The Netherlands)
Borbinha, Jose Luis, INEC-ID (Portugal)
Chan, Leslie, University of Toronto at Scarborough (Canada)
Cetto, Ana Maria, IAEA (Austria)
Cooper, Graham, University of Salford (UK)
Costa, Sely M.S., University of Brasilia (Brazil)
Delgado, Jaime, Universitat Pompeu Fabra (Spain)
Diocaretz, Myriam, MDD Consultancy (The Netherlands)
Dobрева, Milena, Inst. of Mathematics and Informatics, Acad. Sciences (Bulgaria)
Engelen, Jan, Katholieke Universiteit Leuven (Belgium)
Gradmann, Stefan, University of Hamburg (Germany)
Guentner, Georg, Salzburg Research (Austria)
Hedlund, Turid, Swedish School of Economics and Business Administration, Helsinki (Finland)
Heinrich, Klausjuergen, Donau-Universitaet Krems (Austria)
Ikonomov, Nikola, Institute for Bulgarian Language (Bulgaria)
Iyengar, Arun, IBM Research (USA)
Jezek, Karel, University of West Bohemia in Pilsen (Czech Republic)
Knoll, Adolf, Czech National Library (Czech Republic)
Krottmaier, Harald, Graz University of Technology (Austria)
Kreulich, Klaus, Munich University of Applied Sciences (Germany)
Linde, Peter, Blekinge Institute of Technology (Sweden)
Martens, Bob, Vienna University of Technology (Austria) – CHAIR
Moens, Marie-Francine, Katholieke Universiteit Leuven (Belgium)
Moore, Gale, University of Toronto (Canada)
Mornati, Susanna, CILEA (Italy)
Nisheva-Pavlova, Maria, Sofia University (Bulgaria)
Paepen, Bert, Katholieke Universiteit Leuven (Belgium)
Perantonis, Stavros, NCSR - Demokritos (Greece)
Savenije, Bas, Utrecht University Library (The Netherlands)
Schranz, Markus, Presstext Austria (Austria)
Smith, John, University of Kent at Canterbury (UK)
Tonta, Yaşar, Hacettepe University (Turkey)

Keynotes

Infrastructure and Policy Framework for Maximising the Benefits from Research Output
Jeffery, Keith G. 1

Scientific Publishing in the Digital Era
Kroó, Norbert 13

Open Access and New Publishing Models

Open Access Publishing in High-Energy Physics
Mele, Salvatore 15

Importance of Access to Biomedical Information for Researchers in Molecular Medicine
Roos, Annikki; Hedlund, Turid 25

Representing and Coding the Knowledge Embedded in Texts of Health Science Web
Published Articles
*Marcondes, Carlos Henrique; Rocha Mendonça, Marília Alvarenga; Malheiros, Luciana Reis;
Cruz da Costa, Leonardo; Paredes Santos, Tatiana Christina; Guimarães Pereira, Luciana* 33

Automatic Content Syndication in Information Science: A Brazilian Experience in the
Creation of RSS Feeds to e-journals
Lopes de Almeida, Robson 43

Emerging Business Models for e-Content Delivery

Changing Content Industry Structures: The Case of Digital Newspapers on ePaper Mobile Devices
Van Audenhove, Leo; Delaere, Simon; Ballon, Pieter; Van Bossuyt, Michael 53

Introducing the e-newspaper - Audience Preferences and Demands
Ihlström Eriksson, Carina; Åkesson, Maria 65

Centralized Content Portals: iTunes and the Publishing Industry
Leendertse, Matthijs; Pennings, Leo 75

Enabling Accessibility

The Open Document Format and its Impact on Accessibility for Persons with a
Reading Impairment
Engelen, Jan; Strobbe, Christophe 85

Multimedia Modular Training Packages by EUAIN
Crombie, David; Ioannidis, George; McKenzie, Neil 91

File-Formats for Preservation: Evaluating the Long-Term Stability of File-Formats
Rauch, Carl; Krottmaier, Harald; Tochtermann, Klaus 101

Development of Repository and Publishing Tools

Beyond Publication - A Passage Through Project StORe
Pryor, Graham 107

Challenges in the Selection, Design and Implementation of an Online Submission and
Peer Review System for STM Journals
Best, Judy; Akerman, Richard 117

A Bachelor and Master Theses Portal: Specific Needs and Business Opportunities
for the DoKS Repository Tool
Baccarne, Rudi 129

The FAO Open Archive: Enhancing the Access to FAO Publications Using International
Standards and Exchange Protocols
Nicolai, Claudia; Subirats, Imma; Katz, Stephen 141

Assessment of Open Access and Enabling Frameworks	
Five Years on - The Impact of the Budapest Open Access Initiative <i>Hagemann, Melissa R.</i>	153
Openness in Higher Education: Open Source, Open Standards, Open Access <i>Kelly, Brian; Wilson, Scott; Metcalfe, Randy</i>	161
Peer-to-Peer Networks as a Distribution and Publishing Model <i>De Boever, Jorn</i>	175
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 <i>Johnson, Ian M.</i>	189
Open Access Journals and Regional Perspectives	
Expectation and Reality in Digital Publishing: Some Australian Perspectives <i>Martin, Bill; Deng, Hepu; Tian, Xuemei</i>	199
Libraries as Publishers of Open Access Digital Documents: Polish Experiences <i>Nahotko, Marek</i>	209
Use of Open Access Electronic Journals by Chinese Scholars, and an Initiative to Facilitate Access to Chinese Journals <i>Li, Ruoxi; Rowland, Fytton; Xiong, Zichuan; Zhao, Junping</i>	221
Managing Expectations for Open Access in Greece: Perceptions from the Publishers and Academic Libraries <i>Banou, G. Christina; Kostagiolas, A. Petros</i>	229
Ontology and Meaning Extraction	
Towards a Semantic Turn in Rich-Media Analysis <i>Bürger, Tobias; Güntner, Georg</i>	239
Towards an Ontology of EIPub/SciX: A Proposal <i>Costa, Sely M.S.; Gottschalg-Duque, Claudio</i>	249
On the Evolution of Computer Terminology and the SPOT On-Line Dictionary Project <i>Hynek, Jiri; Brada, Premek</i>	257
Digital Heritage and Access	
Scientific Heritage in Bulgaria Makes First Digital Steps <i>Dobрева, Milena; Ikonomov, Nikola</i>	269
Digitisation and Access to Archival Collections: A Case Study of the Sofia Municipal Government (1878-1879) <i>Nisheva-Pavlova, Maria; Pavlov, Pavel; Markov, Nikolay; Nedeva, Maya</i>	277
The Digital Scholar's Workbench <i>Barnes, Ian</i>	285
Evaluating Digital Humanities Resources: The LAIRAH Project Checklist and the Internet Shakespeare Editions Project <i>Warwick, Claire; Terras, Melissa; Galina, Isabel; Huntington, Paul; Pappa, Mikoleta</i>	297

Impact Analysis and Open Access Journals

Feasibility of Open Access Publishing for SSHRC Funded Journals
Chan, Leslie; Groen, Frances; Guédon, Jean-Claude 307

The Research Impact of Open Access Journal Articles
Tonta, Yasar; Ünal, Yurdagül; Al, Umut 321

Sharing the Know-how of a Latin American Open Access only e-journal: The Case of the Electronic Journal of Biotechnology
Muñoz, Graciela; Bustos-González, Atilio; Muñoz-Cornejo, Alejandra 331

Open Access Journals: A Pathway to Scientific Information in Iran
Noruzi, Alireza 341

Web 2.0 and Social Media

Automatic Sentiment Analysis in On-line Text
Boiy, Erik; Hens, Pieter; Deschacht, Koen; Moens, Marie-Francine 349

A Comparison of the Blogging Practices of UK and US Bloggers
Pedersen, Sarah 361

Enhancing Traditional Media Services Utilising Lessons Learnt from Successful Social Media Applications - Case Studies and Framework
Bäck, Asta; Vainikainen, Sari 371

The Fight against Spam - A Machine Learning Approach
Jezek, Karel; Hynek, Jiri 381

Interoperability and Open Repositories

The Project of the Italian Culture Portal and its Development. A Case Study: Designing a Dublin Core Application Profile for Interoperability and Open Distribution of Cultural Contents
Buonazia, Irene; Masci, M. Emilia; Merlitti, Davide 393

Building Bridges with Blocks: Assisting Digital Library and Virtual Learning Environment Integration through Reusable Middleware
Chumbe, Santiago; MacLeod, Roddy; Kennedy, Marion 405

Designing Metadata Surrogates for Search Result Interfaces of Learning Object Repositories: Linear versus Clustered Metadata Design
Balatsoukas, Panos; Morris, Anne; O'Brien, Ann 415

Disclosing Freedom of Information Releases
Apps, Ann 425

Workshop 1

EPrints 3.0: New Capabilities for Maturing Repositories
Carr, Leslie A. 435

Workshop 2

DCMI-Tools: Ontologies for Digital Application Description
Greenberg, Jane; Severiens, Thomas 437

Workshop 3

DRIVER - Supporting Institutional Repositories in Europe
Robinson, Mary L.; Horstmann, Wolfram 445

Demonstrations

- Cultural Content Management at a New Level: Publishing Theater and Opera Details by Means of Open Technologies from the Web 2.0
Schranz, Markus W. 447
- Ontologies At Work: Publishing Multilingual Recreational Routes Using Ontologies
Paepen, Bert 451

Posters

- The PURE Institutional Repository: Ingestion, Storage, Preservation, Exhibition and Reporting
Alroe, Bo 455
- Access to Free e-journals via Library Portals: The Experience of the Shahid Chamran Ahwaz University in Iran as a Case Study
Asnafi, Amir Reza 457
- Digital Archives at the University of Pisa
Bucchioni, Cinzia; Pistelli, Zanetta; Pistoia, Barbara 459
- A Survey on magiran.com: A Database for the Magazines of Iran
Kokabi, Mortaza 461
- Developing National Open Access Policies: Ukrainian Case Study
Kuchma, Iryna 463
- Digitization of Scientific Journals in Serbia
Mijajlovic, Marko; Ognjanovic, Zoran; Pejovic, Aleksandar 465
- DRIVER - Digital Repository Infrastructure Vision for European Research
Robinson, Mary 467
- The Inclusion of Open Access Journals in Academic Libraries: A Case Study of Bioline International
Sweezie, Jen; Caidi, Nadia; Chan, Leslie 469
- Developments in Publishing: The Potential of Digital Publishing
Tian, Xuemei 471

Index of Authors 473

Index of Keywords 475

Preface

Dear readers and delegates at ELPUB 2007,

It is a pleasure for us to present you with this volume of proceedings, consisting of scientific contributions accepted for presentation at the 11th ELPUB conference, organised by the Vienna University of Technology, Austria.

The 11th ELPUB conference keeps alive the mission of the ten previous international conferences on electronic publishing - held in the United Kingdom (in 1997 and 2001), Hungary (1998), Sweden (1999), Russia (2000), the Czech Republic (2002), Portugal (2003), Brazil (2004) Belgium (2005) and Bulgaria (2006) - which is to bring together researchers, lecturers, developers, entrepreneurs, managers, users and all those interested in issues regarding Electronic Publishing in widely differing contexts.

The theme for this year's conference, "Openness in Digital Publishing: Awareness, Discovery, and Access", is devoted to exploring the full spectrum of "openness" in digital publishing, from open source applications for content creation to open distribution of content, and open standards to facilitate sharing and open access to scientific publications. In addition to technical papers, we also encouraged submissions reporting on research on economics of openness, public policy implications, and institutional support and collaboration on digital publishing and knowledge dissemination. The goal is to encourage research and dialogues on the changing nature of scholarly communications enabled by open peer-to-peer production and new modes of sharing and creating knowledge.

In order to guarantee the high quality of papers presented at ELPUB 2007, all submissions were peer-reviewed by the Programme Committee (PC), whose thirty-four highly qualified and specialised experts represent many different countries and cover a wide variety of institutional and knowledge domains. The PC did a great job in selecting the best submissions for ELPUB 2007, and we would like to thank them sincerely for the valuable time and expertise they put into the peer review process.

At the conclusion of the peer review procedure, all selected and confirmed entries for this conference, including full papers for scientific presentations, and shorter papers for workshops and demonstrations, were pre-published in the ELPUB Digital Library at <http://elpub.scix.net>. Potential delegates could therefore see, in advance, what could be expected at the meeting. The same system - SOPS, or SciX Open Publishing System - was also used to set up the submission and review of abstracts.

To assist with the assignment of reviewers, submitters were asked to characterise their entries by selecting 3-5 key areas out of a larger list of subject descriptors. In a similar way, reviewers identified their 3-5 fields of expertise and this allowed the Programme team to map papers to reviewers. Finally, the same system supported the Programme Committee with the scheduling of the sessions and with grouping papers according to common and over-lapping themes. The Table of Content of this volume follows both the themes and the order of the sessions in which they were scheduled during the conference.

As with all previous ELPUB conferences, the collection of papers and their metadata are made available through several channels of the Open Archives Initiative, including Dublin Core metadata distribution and full archives at <http://elpub.scix.net>. It may appear ironic to have a printed proceedings for a conference dedicated to Electronic Publishing. However, the "need" for printed publications is an old and continuing one. It seems that it is still essential for a significant number of delegates to have "something tangible" in their hands and their respective university administrations.

We hope you enjoy reading the proceedings. It is also our pleasure to invite delegates and readers to ELPUB 2008, taking place in Toronto, Canada. The 12th ELPUB conference will be organised by the University of Toronto, and this marks the first time the conference series will be held in North America. Details of the conference will be forthcoming at the ELPUB web site.

Finally, we would like to thank our Keynote Speakers, Keith Jeffery and Norbert Kroó, for their insightful and timely contributions to the conference. Thanks also go to Grace Samuels for checking the references, to Tomo

Cerovsek for maintaining the submission- and review-interface. We would also like to thank the sponsors for their generous contributions.

With our best wishes for a very successful conference,



Leslie Chan
Programme Committee Chair
University of Toronto Scarborough



Bob Martens
General Chair
Vienna University of Technology

Technical Infrastructure and Policy Framework for Maximising the Benefits from Research Output

Keith G. Jeffery

Science and Technology Facilities Council, Rutherford Appleton Laboratory, OX11 0QX UK
e-mail: keith.g.jeffery@rl.ac.uk

Abstract

Electronic publishing is one part of a much larger process. There is a research lifecycle from creation of a programme for funded research through research proposals, projects, outputs (including publications), exploitation (both for further scholarly work and for commercial or quality of life benefits) and creation of the next programme. Throughout this lifecycle information is the lifeblood; publications are used and created at all stages. The vision proposed brings together electronic research publications with associated datasets and software all contextualised by a CRIS (Current Research Information System) which provides information on projects, persons, organisational units, outputs (products, patents, publications), events, facilities, equipment and much more. Via the CRIS, research output can be linked to financial, project management and human resource data: indeed finally the cost of production of a publication can be compared against its benefit. Realising the vision requires advanced IT architectures including GRIDs and ambient computing. Against this vision current debates about subscription-based publishing and gold author-pays open access publishing, about grey literature and green open access self-archiving can be regarded with clarity and objectivity. The way ahead is clear: funders of research should mandate green self-archiving for the benefit of research and of the twin beneficial consequences: wealth creation and improvement in the quality of life. These benefits far outweigh any short-term benefits from the publishing industry in profits or tax-take. There is still plenty of market opportunity for publishers and their doomsday predictions are unsustainable.

Keywords: open access; CRIS; e-infrastructure; repository

1 Maximising the Benefits from Research Output

1.1 The Requirement

1.1.1 Introduction

Let us start with that which is required. This is detailed below by type of user (actor) and role but we can surely agree that the overall aim must be that research output causes wealth creation and / or improvement in the quality of life. It follows therefore that maximising these desirable properties requires maximum access to research output. Provision of maximum access has technical, legalistic and economic implications. It also requires a broader context to ensure the research output material is understood and used appropriately.

1.1.2 The Actors

The researcher requires access to find relevant pre-existing research output and to find possible research collaborators. The research manager requires access to check completeness of recorded outputs from her institution, to compare with that of other institutions and thus to develop strategy for her institution. The funding agency requires access to ensure defined outputs from the funded research proposal are delivered, to compare outputs with those from other funding agencies and to find appropriate referees. The policymaker requires access to compare outputs produced by different continents, countries, institutions and research teams. The innovator requires access to find new ideas which are exploitable for wealth creation or improvement in the quality of life. The educator requires access to obtain teaching material. The student requires access to use learning material. The media require access to obtain information that can be recast as 'stories' which popularise research or raise social, ethical, political or economic issues concerning the research for the public interest.

1.1.3 The Roles

Any competent researcher before starting a new research idea will review the existing research output. The more complete and accessible this is to her, the better the review will be, nugatory effort will be avoided and a better (novel) idea will be formulated. A researcher working in one topic area may find an applicable and appropriate technique –such as an experimental protocol, or a computer program for simulation or statistical reduction - from another topic area. As a result of one of the above, or by an independent search, a researcher may find a potential collaborator or complementary co-worker for a research idea.

One measure of a researcher capability is evaluation of produced output. The more complete and accessible outputs are, the better the quality of the evaluation. The metrics imposed on the raw data (i.e. how one ranks different publication channels such as journals) are a separate issue, but without complete and verifiable raw data evaluations are worthless. Similarly the performance of an organisational unit can be evaluated based on its outputs. Indeed, one could compare inputs (funding) with outputs as evaluated to obtain some idea of effectiveness and efficiency.

One may wish to evaluate the literature in different topic areas of fields of research. This may inform strategic decisions on research funding, or areas of priority in a research institution. The literature provides a source of ideas, usually with associated research to demonstrate their potential use. This is a mine of information for the entrepreneur or innovator who wishes to invest venture capital to create products or services with associated wealth creation (jobs, profits for shareholders).

Today's teaching material is the research output of years ago. As the pace of learning increases, and the volume of research output increases, there is a need for faster and easier access to appropriate research literature by educators. Modern learning is more project-based and less 'chalk and talk'. Students are encouraged to utilise technology to find relevant information.

Journalists and other media professionals need easy access to research outputs in order to find interesting 'stories' for popularising, to research (verify) the background to 'urban myths' about research and to find researchers suitable for appearing on TV programmes or writing articles.

1.1.4 Conclusion

We can conclude that all these actors, in the various example roles discussed, require easy (fast, efficient) access to research output material. Technically this implies the need for excellent descriptive metadata, fast searching of metadata, fast searching of text and multimedia and well-structured results. Furthermore access to heterogeneous distributed repositories should appear homogeneous and local to the end-user. This implies reconciliation to a canonical syntax (structure) and semantics (meaning) which in turn is likely to involve translation of character sets, language and ontological terms. Legally it requires unfettered access although restrictive metadata may document - for software to enforce - claimed rights which should be respected (like attribution) and even may define a price for access. Economically it requires a business model where costs are minimised (ideally zero as seen by the end-user), any income lies where the work is done and costs are borne where benefit is obtained. Furthermore, ideally the actors require the research output material in the context of research project, researchers, organizations involved, facilities and equipment, funding etc.

1.2 A Scenario

All the actors require access anytime, anyplace, anywhere (so-called martini computing) via any appropriate device. The access should be not only to local (job, role or personal) information but, with minimal effort, to the whole world of research information.

A researcher should be able remotely to set up and control experiments (physical experimentation), take and visualise results, access relevant research literature, access datasets and analytical or simulation software (in-silico experimentation) and create new publications (whether academic or project management reports or deliverables) with automated assistance. She should be able to complete research proposals with intelligent assistant software to fill in the standard form fields. She should be able to find suitable research partners in academia or industry. She should be able to utilise computation power, storage and network resources without knowing where they are – only knowing their capabilities are suitable for her task and respecting any restrictions concerning rights acknowledgement or payment. She should also be able to do all the management/administrative tasks comfortably and efficiently within the same environment: completing time sheets, expense claims, purchase requisitions, travel plans etc. The management of research publications must lie comfortably within this environment.

Similarly research managers in research institutions or funding agencies should be able to gain quickly the ‘state of the world’ in any research area to compare their own organisation with others and thence plan appropriate strategies. This implies knowing what other funding agencies or research institutions have currently in terms of projects, persons, organisation units (e.g. research teams, departments), funding programmes, research outputs (products, patents, publications), events, facilities and equipment and so on.

The point to be stressed is that research outputs are part of a much larger environment, all of which must be recorded and accessible for the end-user to appreciate the research output material.

2 Technical Infrastructure

2.1 Introduction

The solution comes in several components: the e-Infrastructure provides the connectivity, computing power and software engineering environment for ease of access and ease of use. CRIS (Current Research Information Systems) provide structured information documenting the context of the research and providing structured metadata. OA repositories (of publications) provide the scholarly research output. e-Research repositories (research datasets and software) provide the detailed underpinning material of the research.

2.2 e-Infrastructure

Over the last few years it has become apparent that the e-infrastructure solution is based on GRIDs and SOA (service oriented architecture) [1, 2]. The original GRID idea provides metacomputing (linked supercomputers) [3]. The original WWW idea provides access to information but without computation. Bringing them together provides a user-invisible platform [4]. Adding self-* properties (self-management, self-composition, self-repairing, self-tuning) [5] makes the platform effective and efficient. Utilising a SOA (Service-Oriented Architecture) based on discoverable reliable services (pieces of software that execute some function and can be composed into larger software structures to perform human-recognisable tasks) increases the reliability and decreases the software cost. The SOKU (Service-Oriented Knowledge Utility) concept [6] shows much promise: each SOKU would be wrapped in metadata to allow its discovery (descriptive metadata) and to control (parameterise) its execution in both functional (how it does what it does) and non-functional (under what conditions e.g. rights attribution, price) it does it, modes (restrictive metadata) (Figure 1).

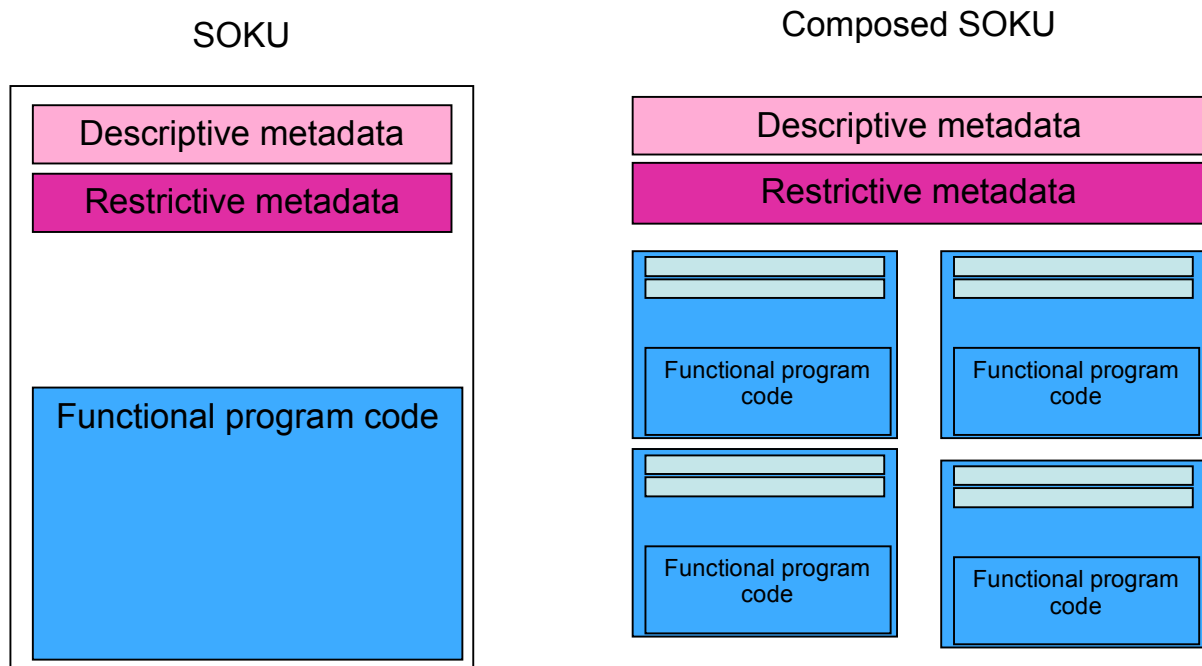


Figure 1: SOKU

The critical requirement for effective e-infrastructure has been recognised internationally. The 'cyberinfrastructure' [7] in North America follows the work on e-infrastructure based on GEANT in Europe [8] itself partly stimulated by the requirements of research facilities [9]. Individual European countries, too, have invested in e-infrastructure; an example is e-Science (applications) utilising the national GRID service (middleware) itself based on JANET (network) in UK [10]. Similar initiatives have been taken elsewhere notably in Australasia, Japan, China, Singapore, India and also in South America.

These e-infrastructures provide fast networking linking supercomputers, repositories and access to experimental facilities. They have schemes for identification, authentication and authorisation of usage. They have middleware to make the base resources invisible to the end-user and to optimize resource allocation. They are developing methods for homogeneous access to heterogeneous resources.

To date the work has largely been academic. The IT companies have been involved in producing components of the solution; e.g. IBM has an autonomic computing product, ORACLE has a clustered database product. Univa [11] offers a commercialised version of the popular open source GLOBUS middleware. However, there are extensive developments underway in many IT companies to produce GRID/SOA-compatible products and some are even basing their future architectures on SOKU.

The challenge posed is how to utilise this emerging e-infrastructure for benefit and specifically how to use it to make more accessible and available the research literature in a form appropriate for the actors performing their roles as outlined above.

2.3 CRIS

CRIS (Current Research Information Systems) have been developed over the last 40 years. Currently an EU Recommendation to member states, CERIF (Common European Research Information Format) is being adopted quite widely and it allows interoperation. A CRIS typically has information on projects, persons, organisational units, funding programmes, research outputs (products, patents, publications), facilities and equipment and events. The novelty of CERIF is its formal data structure, its use of linking relations to allow n:m relationships with role and temporal duration, its use of multiple character sets and provision of multilinguality.

Consider the following case illustrated in (Figure 2) **Fehler! Verweisquelle konnte nicht gefunden werden.** : A person A is an employee of organisation O and a member of organisations M and N both of which are parts of O. She is author of X in which O claims the IPR (intellectual property right) and project leader of P. In CERIF the following records would be in base tables: Person: A; OrgUnit: O,M,N; Publication: X; Project: P. The link tables would be: Person-OrgUnit: A-employee-O, A-member-M, A-member-N; OrgUnit-OrgUnit: M-partof-O; N-partof-O; Person-Publication: A-author-X; OrgUnit-Publication: O-IPR-X; Person-Project: A-projectleader-P. In fact, the link tables include, as well as role, the temporal information concerning start and end date-time. In this example it may be that when A authored X she was no longer a member of M. This, relatively simple, example illustrates the power of CERIF as a data model.

CERIF is maintained by the not-for-profit organisation euroCRIS (www.eurocris.org) from whence details are available. Commercial CRIS offerings are available from uniCRIS [12] which is fully CERIF-compatible, Atira [13], and Avedas [14]. Many funding agencies and research institutions have some form of 'home-brew' CRIS, the majority are more-or-less CERIF-compatible. The provision of CRIS in a modern e-infrastructure environment has been discussed in [15].

2.4 Repositories

Repositories store and provide access to the detailed information. It is usual to separate repositories of research publications from repositories of research datasets and software (e-Science or, better, e-Research repositories) because of their different access patterns and different metadata requirements. The e-Research repositories require much more detailed metadata to control utilisation of the software and datasets in addition to metadata to allow discovery of the resources. At present they tend to be specific to an individual organisation because of their novelty and the differing requirements on metadata imposed by different (commonly international) communities e.g. in space science, atmospheric physics, materials science, particle physics, humanities or social science. Publication repositories typically use some form of Dublin Core Metadata [16] and most are OAI-PMH (Open Access Initiative – Protocol for Metadata Harvesting) [17] compliant for interoperation and are indexed by Google Scholar. Example software systems are ePrints [18], Dspace [19], Fedora [20] and ePubs [21].

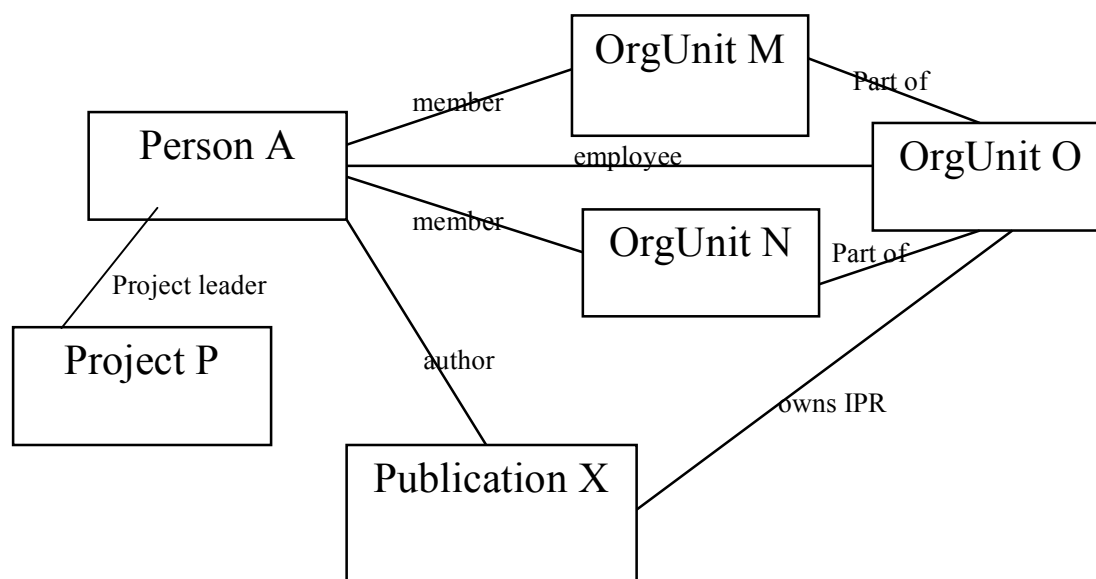


Figure 2: Example of CERIF

2.5 Metadata and Curation

Digitally-created articles rely heavily on both the metadata record and the articles themselves being deposited. International metadata standards and protocols must be applied to repositories so that retrieval may be consistent with appropriate recall (precision) and relevance so that harvesting (or homogeneous retrieval access) across repositories can take place. A model for formalising metadata [22] is required.

The current DC metadata standards DC [16] and OAI-PMH [17] for interoperability are insufficient for scalable, automated retrieval with appropriate relevance (precision) and recall. DC is machine-readable but not machine-understandable. One basic problem is that a formalised syntax and semantics (vocabulary) for each relevant DC element was not specified in ‘simple DC’ and has only partially been overcome by the use of namespaces in ‘qualified DC’. A second problem concerns the element set tags ‘contributor’, ‘creator’ and ‘publisher’ which are actually roles of a person or organisational unit and should be represented by a relationship (between the article and the person or organisational unit) where the role value belongs to a namespace and is temporally limited. A third problem is the tag ‘relation’ which is extremely general; the real world is much better modelled through typed relations with role and temporal validity. Other problems include the tag ‘coverage’ which only recently has been separated into temporal and spatial aspects yet these are fundamental retrieval criteria for much material. A formalised version of DC overcoming these limitations has been suggested [23] and defined [24] to form also part of the CERIF model allowing tight integration with CRIS. Recently the DC community has recognised these problems and with more recent work [25, 26] is attempting to address them.

To ensure that research output material is available for future generations, curation and preservation issues must be addressed. There is current work to define metadata standards to achieve this [27] but a major problem concerns maintaining the articles on current (i.e. usable) media.

2.6 Integration

The linking together at an institution of a ‘green’ OA repository of articles, a CRIS (to provide contextual information) and an OA repository of research datasets and software [28] (Figure 3) ensures that an institution can manage its IP for benefit whether that benefit is in innovation and investment, in educational resources, in stimulation of future research or in publicity. Furthermore, the formalised structure of the CRIS allows a reliable workflow to be engineered which in turn encourages deposit of research outputs. Such a system is being implemented progressively at STFC Rutherford Appleton Laboratory where the CERIF-CRIS is named the Corporate Data Repository, the OA repository is ePubs and the e-research repository is the e-Science repository.

Linking together these institutional CRIS systems - which have a formal structure and hence can be interoperated reliably and in a scalable way [29] - provides a network of access to institutional OA repositories (of articles) or e-research repositories via the CERIF-CRIS gateways enhancing and controlling the access using the CERIF-CRIS information as formalised, structured and contextual metadata which is more detailed than DC and suitable for intelligent (machine-understandable) interoperation (Figure 4). Interoperation of CERIF-CRIS has been

demonstrated, most recently for euroHORCS (European Heads of Research Councils) in October 2006. However, as yet, the whole architecture has not been demonstrated.

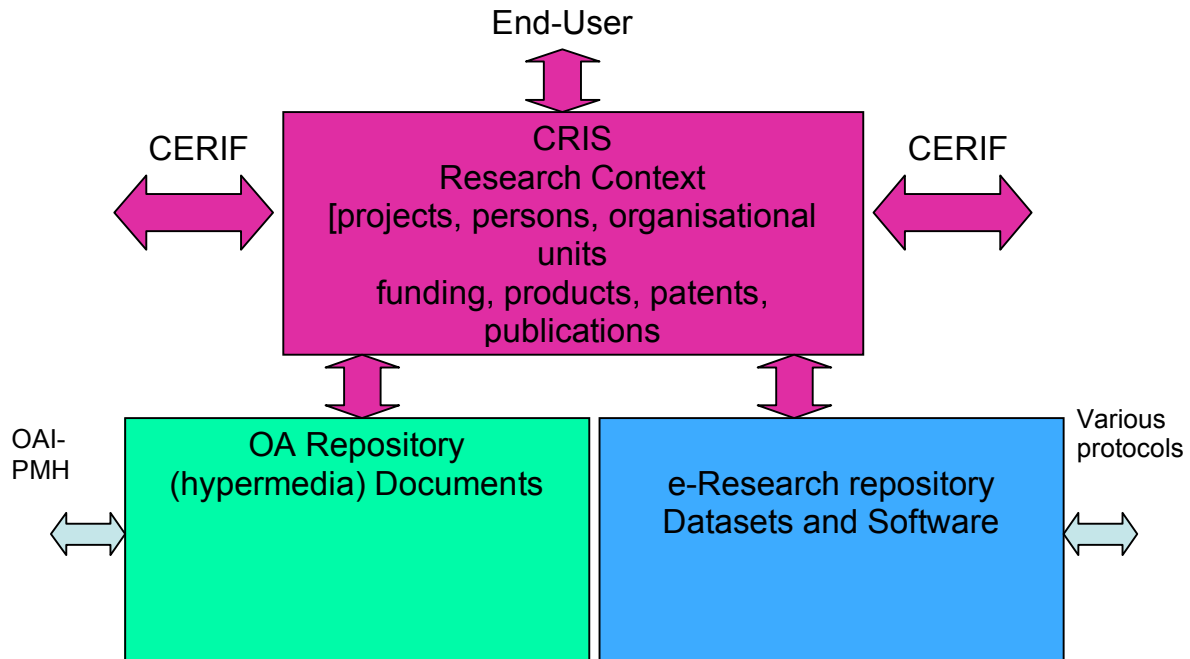


Figure 3: Architecture for an Institution

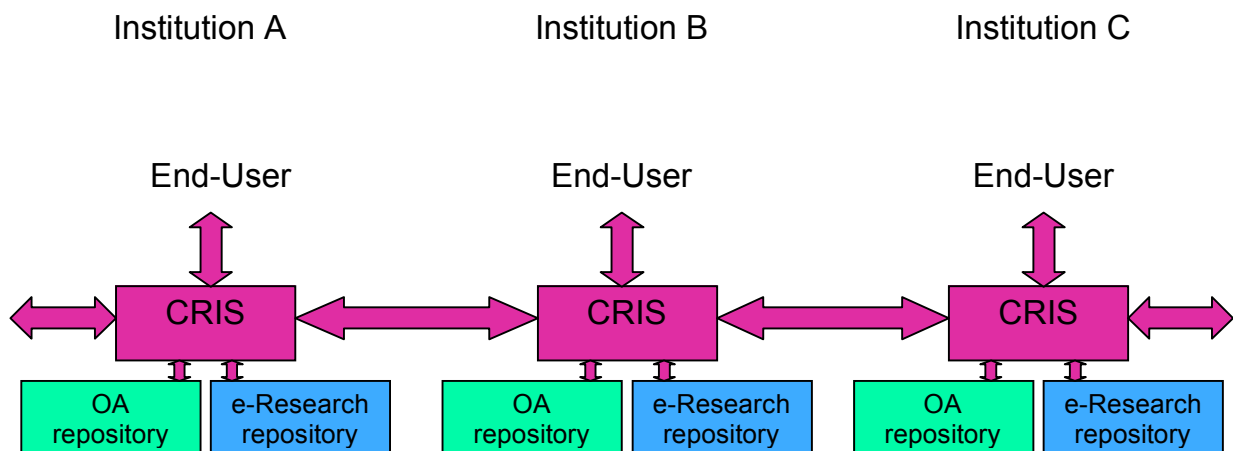


Figure 4: Architecture for OA

3 Policy Framework

3.1 Introduction

The contentious area of discussion in this subject area is open access to research output publications. A brief overview is at [30]. We assume the conference attendees have a good knowledge of OA to distinguish the dimensions of the topic: 'green' and 'gold'; thematic (central) and institutional (distributed); peer-reviewed or not. Furthermore we assume a general knowledge of the differences between white and grey literature. This section discusses motivations and barriers and then concentrates on the two major topics to overcome the barriers: metadata and mandates and finally concludes with speculations on the future.

3.2 Motivations

Open Access (OA) means that electronic scholarly articles are available freely at the point of use. The subject has been discussed for over 10 years [31], but has reached a crescendo of discussion over the last few years with various declarations in favour of OA from groups of researchers or their representatives [32-35]. The UK House of Commons Science and Technology Committee [36] considered the issue in 2004, reporting in the summer in favour of OA. This indicates the importance of the issue, and led to statements from large research funding bodies such as the Wellcome Trust [37] and the Research Councils UK [38]. More recently the USA has attempted to move in this direction [39]. What has motivated this interest?

Ethics: There is an ethical argument that research funded by the public should be available to the public. Since research is an international activity, this crosses national boundaries.

Research Impact: The internet provides an opportunity. Modern harvesting techniques and search engines make it possible to discover publications of relevance if they are deposited in an OA repository with a particular metadata standard. If all authors did this then the world of research would be available 'at the fingertips'. There is evidence that articles available in an OA repository have more accesses / downloads (readers), citations and therefore impact [40, 41]

Costs and economic benefit: There is concern over the hindrance to research caused by the cost of journal subscriptions, whether electronic or paper. These costs run well above the rate of inflation with the result that libraries with restricted budgets (i.e. all of them!) are no longer providing many journals needed by researchers [42-44]. Estimates of the costs of 'gold' OA publishing indicate that for a productive institution these costs could exceed by a factor of 3 current subscription costs. Economic benefit of improved (open) access has been studied and 'green' OA is regarded as beneficial [45, 46].

Metrics: measures of utilisation of research publications are used for various statistical purposes – usually to indicate quality which may be used in evaluation to allocate research funding. Articles in an OA repository allow automation of such metrics including measures of impact and – most importantly – at the level of the article (not the channel) and evenly across all disciplines and language encodings in contradistinction to how ISI manages ranking. This aspect links with those of research impact and costs and economic benefit.

Added value: articles in an OA repository can easily be linked to structured data contextualising the research (CRIS) [47, 48] and thence to repositories of research datasets and software [28].

Just reward: There is also concern that in traditional scholarly publishing, most of the work (authoring, reviewing, editing) is done freely by the community and that the publishers make excessive profits from the actual publishing (making available) process.

3.3 Barriers to OA

Despite the positive motivations there are barriers to OA.

Loss of publisher income: The major objection to 'green' self-archiving comes from publishers and learned societies in publisher role (many of which depend on subscriptions to their publications) who fear that 'green' OA threatens their business viability. To date there is no evidence that 'green' archiving harms the business model of publishing [49, 50]. There is evidence that 'green' archiving increases utilisation, citation and impact of a publication [51, 52] and has economic benefits [45, 46]. Whilst the major commercial publishers could provide additional value-added services to offset the impact of OA on current business models, the impact on learned societies may require new business models to be developed.

Copyright: Copyright agreements between authors and publishers may inhibit the 'green' route. However, to date, over 90% of publication channels (the variability depends on exactly what is counted) allow 'green' author deposit although some insist on an embargo period before the publication is available for OA [53]. In contrast some publishers of journals – of which Nature is the most well-known – do not demand copyright from the author but merely a licence to publish, leaving copyright with the author or their institution.

Difficulties in access and utilization: despite the Dublin Core metadata standard [16] and an interoperation protocol [17] there are difficulties in an end-user obtaining appropriate relevance (precision) and recall in retrieval – certainly when compared with a well-structured library catalog system using e.g. [54]. This indicates that the metadata is insufficient for the purpose. Similarly, if the end-user wishes easy access from the article to research context or associated research datasets and software this is currently extremely difficult. However, linking a repository of articles to a CRIS provides structured metadata which improves greatly relevance (precision) and recall and also provides a link through to e-research repository information.

Completeness: there is great difficulty in persuading researchers to deposit their material in OA repositories. Estimates indicate an 8-15% fill of OA repositories [55] although when a funding organization or institution applies a mandate this rises rapidly to 60%-90% eventually approaching 100% [56]. Additionally, an institution may – following the mandate – assist in automating the process with a workflow such that there is minimum (re)keying of metadata [57]. Again this works best if there is a CRIS with structured metadata.

3.4 Mandates

Both the EU [58] and the USA (proposed US Federal Research Public Access Act [39]) have moved towards mandating that output of publicly funded research should be OA. Neither has (as yet) enacted the mandate. For a summary see [59]. The EU went against the results of its own commission study possibly as a result of the ‘Brussels Declaration’ from the STM (Science, Technology and Medicine) Publishing community [60] despite EURAB (EU Research Advisory Board) recommending green OA [61]. Various funding organisations have mandated open access for the outputs of research that they fund, based on the arguments in 3.2 above. The vast majority mandate ‘green’ OA (parallel self-archiving in an institutional repository) and some (Wellcome, Hughes) agree to fund in parallel ‘gold’ (author funder pays) with preferred publishers. More recently CERN (a research institution, not a funder) has proposed to go the ‘preferred publisher gold’ route [62]. This is surprising since CERN and the particle physics community pioneered ‘green’ OA with arXiv [63].

The preferred, optimal and recommended procedure is immediately upon acceptance for publication the metadata and full article are deposited in an institutional repository. If the publisher does not demand an embargo period both are set to open access; if an embargo period is demanded then only the metadata is made visible until the end of the embargo period. Of course, associated with the metadata record there can be (and ePrints [18] provides) a ‘request button’ so that the material can be sent automatically to any researcher who requests it under the usual ‘fair use’ conditions.

3.5 Integration

What is required now is for all funding agencies to mandate green OA in institutional repositories of research output articles, and for all research institutions to maintain such a repository linked to a CRIS and thence to a repository of research datasets and software. This would provide universal open access and allow researchers, research managers, innovators, policymakers, the media and others to access the research knowledge of the world easily, quickly and cheaply thus promoting wealth creation and improvement in the quality of life.

Such a move will be resisted by the Learned Societies (acting as publishers) and the publishing industry for business reasons. There are two possible ways forward: (1) press ahead with ‘green’ OA ignoring the opposing interests (2) while pressing ahead with ‘green’ OA also engage in debate with the opposing interests to reassure them that there are business models including OA that can work. Stevan Harnad takes the first view and refutes all needless speculation (a position we admire but with which we cannot agree wholeheartedly); we take the second view with a more pragmatic attitude to securing OA for the future.

Thus, there is a need for engagement with the Learned Societies to develop new methods of peer review which can be paid for in order to preserve those societies and the benefits they bring without requiring them to have a business model based on traditional publishing.

Finally there is a need for engagement with traditional publishers to explore what value-added products they could produce harvesting from a rich world of OA repositories of publications cross-linked via CRISs with associated research datasets and software.

4 The Way Forward

In the world of advanced e-infrastructures the progress of research, with its concomitant benefits in wealth creation and improvement in the quality of life, cannot be hindered by obsolete information availability (i.e. commercial publishing) channels.

4.1 Speculation: Future

Looking to the future speculatively, it is possible to imagine ‘green’ OA repositories becoming commonplace and used heavily. At that point, we argue, one could change the business model so that an author deposits in an open access ‘green’ repository but instead of submitting in parallel to a journal or conference peer-review process, the peer-review is done either by:

- a) a learned society managing a ‘college’ of experts and the reviewing process – for a fee paid by the institution of the author or the author;
- b) allowing annotation by any reader (with digital signature to ensure identification / authentication);

in both cases being alerted by ‘push technology’ that a new article matching their interest profile has been deposited.

The former peer-review mechanism would maintain learned societies in business, would still cost the institution of the author or the author but would probably be less expensive than publisher subscriptions or ‘gold’ (author or author institution pays) open access. The latter is much more adventurous and in the spirit of the internet; in a charming way it somehow recaptures the scholarly process of two centuries ago (initial draft, open discussion, revision and publication) in a modern world context. It is this possible future that is feared by commercial publishers.

5 Conclusion

Despite protests and obstacles to improved access to research material over the centuries from religious, commercial, professional or labour groups, none delayed for long progress to meet the requirement as defined by the research community. The advanced international e-infrastructure provides ‘martini computing’ and invisibility of resources to the end-user. It supports access to structured research information on projects, persons, organisational units, funding, research outputs (products, patents, publications), research facilities and equipment, events and more (CRIS). It supports repositories of articles and of research datasets and software. It supports access to experimental facilities and ‘computational steering’ of experiments whether physical or ‘in silico’. There is a new world of research capability. Electronic research output publications must take their place in this new world of accessibility and utilisation unhindered by outdated prejudices. This will lead to maximum use of - and benefits from - the research output for quality evaluation, for innovation, for further research, for education, for research management and planning and for informing public debate on research issues.

Acknowledgements

This article owes much to others. In the area of e-infrastructure I am indebted particularly to colleagues on the NGG expert groups and the UK e-infrastructure group. In the area of CRIS, members of euroCRIS have provided much insight. In the area of OA and repositories the luminaries include Stevan Harnad and Alma Swan to both of whom I owe much for deep discussions and shared thinking. Stevan Harnad, Leslie Chan and Anne Asserson kindly improved the article with their suggestions.

References

- [1] http://www.ercim.org/publication/Ercim_News/enw45/
- [2] http://www.ercim.org/publication/Ercim_News/enw59/
- [3] FOSTER, I; KESSELMAN, C (Eds). The Grid: Blueprint for a New Computing Infrastructure. Morgan-Kauffman 1998
- [4] ftp://ftp.cordis.europa.eu/pub/ist/docs/ngg_eg_final.pdf
- [5] ftp://ftp.cordis.europa.eu/pub/ist/docs/ngg2_eg_final.pdf
- [6] ftp://ftp.cordis.europa.eu/pub/ist/docs/grids/ngg3_eg_final.pdf
- [7] <http://www.adec.edu/nsf/nsfcyberinfrastructure.html>
- [8] EU reflection group roadmap: <http://www.e-irg.org/roadmap/eIRG-roadmap.pdf>
- [9] ESFRI roadmap report: ftp://ftp.cordis.europa.eu/pub/esfri/docs/esfri-roadmap-report-26092006_en.pdf
- [10] UK national report: <http://www.nesc.ac.uk/documents/OSI/report.pdf>
- [11] <http://www.univa.com/>
- [12] www.unicris.com
- [13] www.atira.com

- [14] www.avedas.com
- [15] JEFFERY, K.G.; 'The New Technologies: can CRISs Benefit' in A Nase, G van Grootel (Eds) Proceedings CRIS2004 Conference, Leuven University Press ISBN 90 5867 3839 May 2004 pp 77-88
- [16] <http://dublincore.org/>
- [17] <http://www.openarchives.org/OAI/openarchivesprotocol.html>
- [18] <http://www.eprints.org/>
- [19] <http://www.dspace.org/>
- [20] <http://www.fedora.info/>
- [21] <http://epubs.cclrc.ac.uk/>
- [22] JEFFERY, K G: 'Metadata': in Brinkkemper,J; Lindencrona,E; Solvberg,A (Eds): 'Information Systems Engineering' Springer Verlag, London 2000. ISBN 1-85233-317-0.
- [23] JEFFERY, K G: 'An Architecture for Grey Literature in a R&D Context' Proceedings GL'99 (Grey Literature) Conference Washington DC October 1999 <http://www.konbib.nl/greynet/frame4.htm>
- [24] ASSERSON, A; JEFFERY, K.G.; 'Research Output Publications and CRIS' The Grey Journal volume 1 number 1: Spring 2005 TextRelease/Greynet ISSN 1574-1796 pp5-8
- [25] <http://dublincore.org/documents/2007/04/02/abstract-model/>
- [26] <http://dublincore.org/documents/2007/04/02/dc-rdf/>
- [27] <http://ssdoo.gsfc.nasa.gov/nost/isoas/>
- [28] JEFFERY, K.G; ASSERSON, A: 'CRIS Central Relating Information System' in Anne Gams Steine Asserson, Eduard J Simons (Eds) 'Enabling Interaction and Quality: Beyond the Hanseatic League'; Proceedings 8th International Conference on Current Research Information Systems CRIS2006 Conference, Bergen, May 2006 pp109-120 Leuven University Press ISBN 978 90 5867 536 1
- [29] JEFFERY, K G CRISs, Architectures and CERIF CCLRC-RAL Technical Report RAL-TR-2005-003 (2005) <http://epubs.cclrc.ac.uk/work-details?w=33728>
- [30] http://www.ercim.org/publication/Ercim_News/enw64/jeffery.html
- [31] HARNAD, S. (1995) A Subversive Proposal. In: Ann Okerson & James O'Donnell (Eds.) Scholarly Journals at the Crossroads; A Subversive Proposal for Electronic Publishing. Washington, DC., Association of Research Libraries, June 1995. <http://www.arl.org/scomm/subversive/toc.html>
- [32] <http://www.soros.org/openaccess/read.shtml>
- [33] <http://www.earlham.edu/~peters/fos/bethesda.htm>
- [34] <http://www.zim.mpg.de/openaccess-berlin/signatories.html>
- [35] http://www.oecd.org/document/15/0,2340,en_2649_201185_25998799_1_1_1_1,00.html
- [36] UK House of Commons Science and Technology Select Committee <http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/3990>
- [37] http://www.wellcome.ac.uk/doc_wtx026830.html
- [38] <http://www.rcuk.ac.uk/access/statement.pdf>
- [39] <http://www.taxpayeraccess.org/frpaa/>
- [40] HARNAD, S.; CARR, L.; BRODY, T.; OPPENHEIM, C. (2003) Mandated online RAE CVs Linked to University Eprint Archives: Improving the UK Research Assessment Exercise whilst making it cheaper and easier. Ariadne 35. <http://www.ecs.soton.ac.uk/~harnad/Temp/Ariadne-RAE.htm>
- [41] HARNAD, S. (2007) Open Access Scientometrics and the UK Research Assessment Exercise. Proceedings of the 11th Annual Meeting of the International Society for Scientometrics and Informetrics. Madrid, Spain, 25 June 2007 <http://arxiv.org/abs/cs.IR/0703131>
- [42] ROWLANDS, I.; NICHOLAS, D. (2005) /New Journal Publishing
- [43] SPARKS, S. (2005) /JISC Disciplinary Differences Report./ Rightscom, London. http://www.jisc.ac.uk/uploaded_documents/Disciplinary%20Differences%20and%20Needs.doc.
- [44] EPS (2006) UK scholarly journals: 2006 baseline report An evidence-based analysis of data concerning scholarly journal publishing, RIN, RCUK and DTI,. Available at <http://www.rin.ac.uk/data-scholarly-journals>.

- [45] HOUGHTON, J., STEELE, C., SHEEHAN, P. (2006) Research Communication Costs in Australia: Emerging Opportunities and Benefits. A report to the Department of Education, Science and Training. http://www.dest.gov.au/NR/rdonlyres/0ACB271F-EA7D-4FAF-B3F7-0381F441B175/13935/DEST_Research_Communications_Cost_Report_Sept2006.pdf
- [46] HOUGHTON, J.; SHEEHAN, P. (2006) The Economic Impact of Enhanced Access to Research Findings. Centre for Strategic Economic Studies Victoria University <http://www.cfses.com/documents/wp23.pdf>
- [47] ASSERSON, A; JEFFERY, K.G.; 'Research Output Publications and CRIS' in A Nase, G van Grootel (Eds) Proceedings CRIS2004 Conference, Leuven University Press ISBN 90 5867 3839 May 2004 pp 29-40
- [48] DIJK, ELLY; BAARS, CHRIS; HOGENAAR, ARJAN; VAN MEEL, MARGA (2006) NARCIS: The Gateway to Dutch Scientific Information ELPUB2006. Digital Spectrum: Integrating Technology and Culture - Proceedings of the 10th International Conference on Electronic Publishing held in Bansko, Bulgaria 14-16 June 2006 / Edited by: Bob Martens, Milena Dobрева. ISBN 978-954-16-0040-5, 2006, pp. 49-58
- [49] SWAN, A.; BROWN, S (2005) Open access self-archiving: an author study. http://www.jisc.ac.uk/uploaded_documents/Open%20Access%20Self%20Archiving-an%20author%20study.pdf
- [50] BERNERS-LEE, T.; DE ROURE, D.; HARNAD, S.; SHADBOLT, N. (2005) Journal publishing and author self-archiving: Peaceful Co-Existence and Fruitful Collaboration. <http://eprints.ecs.soton.ac.uk/11160/>
- [51] HARNAD, S.; CARR, L.; BRODY, T.; OPPENHEIM, C. (2003) Mandated online RAE CVs Linked to University Eprint Archives: Improving the UK Research Assessment Exercise whilst making it cheaper and easier. Ariadne 35. <http://www.ecs.soton.ac.uk/~harnad/Temp/Ariadne-RAE.htm>
- [52] HARNAD, S. (2007) Open Access Scientometrics and the UK Research Assessment Exercise. Proceedings of the 11th Annual Meeting of the International Society for Scientometrics and Informetrics. Madrid, Spain, 25 June 2007 <http://arxiv.org/abs/cs.IR/0703131>
- [53] <http://romeo.eprints.org/stats.php>
- [54] <http://www.loc.gov/marc/>
- [55] <http://www.crsc.uqam.ca/lab/chawki/graphes/EtudeImpact.htm>
- [56] SALE, A. (2007) The Patchwork Mandate D-Lib Magazine 13 1/2 January/February <http://www.dlib.org/dlib/january07/sale/01sale.html> doi:10.1045/january2007-sale.
- [57] JEFFERY, K G; ASSERSON, A; 'Supporting the Research Process with a CRIS' in Anne Gams Steine Asserson, Eduard J Simons (Eds) 'Enabling Interaction and Quality: Beyond the Hanseatic League'; Proceedings 8th International Conference on Current Research Information Systems CRIS2006 Conference, Bergen, May 2006 pp 121-130 Leuven University Press.
- [58] http://ec.europa.eu/information_society/activities/digital_libraries/doc/scientific_information/communication_en.pdf
- [59] <http://poynder.blogspot.com/2007/03/open-access-war-in-europe.html>
- [60] <http://www.stm-assoc.org/documents-statements-public-co/2007%20BRUSSELS%20DECLARATION%20130207.pdf>
- [61] http://ec.europa.eu/research/eurab/pdf/eurab_scipub_report_recomm_dec06_en.pdf 965. Cambridge, Mass. : M.I.T. Press, 1965, p. 219.
- [62] http://ec.europa.eu/research/science-society/document_library/pdf_06/aymar-022007_en.pdf
- [63] <http://arxiv.org/>

Scientific Publishing in the Digital Era

Norbert Kroó

Hungarian Academy of Sciences, Roosevelt tér 9, 1051 Budapest, Hungary
e-mail: kroo@office.mta.hu

Keynote Abstract

The new information technology developments change drastically our life. The same applies to scientific research in general and the publication of findings in particular. It offers the chance for faster dissemination of results and broader access to date. The interests of scientists, financing organizations and libraries on one hand and publishers on the other do not overlap completely. Maximizing the speed of dissemination, broad access and securing quality and long time preservation are fields of overlapping interests. Mandatory deposit in open access repositories and pricing are still debated. The lecture discusses the above issues based partly on the basis of the author's motivation to maximize the benefits of public (and so EC) funded research in Europe, influenced by his experience both in European scientific organizations and advisory bodies of the EC.

Keywords: European Union; information technology; open access; research impact

Open Access Publishing in High-Energy Physics

Salvatore Mele^{1,2}

¹ CH-1211, Genève 23, Switzerland

² On leave of absence from INFN, I-80126, Napoli, Italy

e-mail: Salvatore.Mele@cern.ch

On behalf of the SCOAP³ Working Party

Abstract

The goal of Open Access (OA) is to grant anyone, anywhere and anytime free access to the results of scientific research. The High-Energy Physics (HEP) community has pioneered OA with its “pre-print culture”: the mass mailing, first, and the online posting, later, of preliminary versions of its articles. After almost half a century of widespread dissemination of pre-prints, the time is ripe for the HEP community to explore OA publishing. Among other possible models, a sponsoring consortium appears as the most viable option for a transition of HEP peer-reviewed literature to OA. A Sponsoring Consortium for Open Access Publishing in Particle Physics (SCOAP³) is proposed as a central body which would remunerate publishers for the peer-review service, effectively replacing the “reader-pays” model of traditional subscriptions with an “author-side” funding. Funding to SCOAP³ would come from HEP funding agencies and library consortia through a re-direction of subscriptions. This model is discussed in details together with a quantitative description of the HEP publishing landscape leading to a practical proposal for a seamless transition of HEP peer-reviewed literature to OA publishing.

Keywords: open access publishing; high-energy physics; CERN; SCOAP³

1 Introduction

The goal of “*Open Access*” (OA) is to grant anyone, anywhere and anytime, free access to the results of scientific research [1]. The OA debate has gained considerable momentum in recent years across many disciplines, both in the sciences and the humanities. In High Energy Physics (HEP) this debate is driven mostly by two factors:

- The “serials crisis” of ever-rising costs of journals, which has forced libraries to cancel a steadily increasing number of subscriptions, curtailing the access of researchers to scientific literature. This traditional business has become financially unsustainable for libraries and publishers alike;
- The increasing awareness that results of publicly funded research should be made generally available going past the availability of pre-prints, towards peer-reviewed literature.

HEP pioneered OA through its “pre-print culture”: the mass mailing for four decades of preliminary versions of articles, so to ensure their largest diffusion. With the onset of the Internet, the HEP community spearheaded the culture of “repositories”: online collections of freely accessible pre-prints. Thanks to the speed at which they make results available, repositories have become the lifeblood of HEP scientific information exchange. However, they usually contain the original version of articles *submitted* to journals, and not the final, peer-reviewed, *published* version. Notwithstanding the success of repositories, there is consensus in the HEP community that high-quality journals still play a pivot role, by providing [2]:

- quality control through the peer-review process;
- a platform for the evaluation and career evolution of scientists;
- a measure of the quality and productivity of research groups and institutes.

A powerful synergy can arise between the strong OA culture of the HEP community, which finds its roots in four decades of preprint circulation, and its continuing need for high-quality journals, leading to a unique opportunity for a possible transition to OA publishing of the HEP peer-reviewed literature. The community is now moving towards such groundbreaking transition through the establishment of a consortium, SCOAP³ (Sponsoring Consortium for Open Access Publishing in Particle Physics). This consortium would engage publishers of high-quality peer-reviewed journals in order to cover the costs of the peer-review process with

funds previously used for journal subscriptions. This idea is viable for the HEP community since the author and the reader communities largely overlap, and are mostly funded by the same actors. This article describes the SCOAP³ initiative:

- section 2 puts the HEP publishing landscape into context, and describes the background to OA Publishing in HEP and the steps which led to the SCOAP³ initiative;
- section 3 presents the SCOAP³ model through the roles of the main stakeholders in HEP scientific publishing;
- section 4 illustrates the results of an analysis of the HEP publishing landscape and their consequences on the targets of the SCOAP³ initiative;
- section 5 presents the financial aspects of the SCOAP³ model together with a cost-sharing scenario based on an investigation of the author basis of HEP;
- section 6 concludes the article by presenting the status of the initiative at the time of writing.

2 Background

A recent study analyzed articles submitted in 2005 to the *arXiv.org* repository and classified in the *hep-ex*, *hep-lat*, *hep-ph* and *hep-th* categories and subsequently published. Out of a total of about 5'000 articles, more than 80% appeared in just six peer-reviewed journals from four publishers [3]: *Physical Review* and *Physical Review Letters* (published by the American Physical Society), *Physics Letters* and *Nuclear Physics B* (Elsevier), *Journal of High Energy Physics* (SISSA/IOP) and the *European Physical Journal* (Springer). Almost 90% of the articles were published by just four publishers, two out of which (American Physical Society and SISSA/IOP) are learned society.

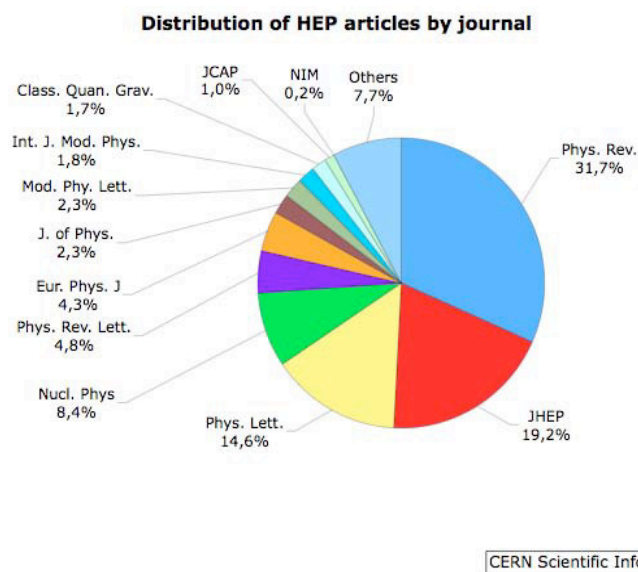


Figure 1: Distribution of the HEP articles submitted in 2005 to the *arXiv.org* repository under the categories *hep-ex*, *hep-lat*, *hep-ph* and *hep-th* and subsequently published in peer-reviewed journals. A total sample of about 5'000 articles is considered. Only journal with a total share above 1% are considered, with the exception of *Nuclear Instrument and Methods in Physics Research* (NIM). The “Others” group comprises 77 remaining journals. From reference [3].

These findings, summarised in figures 1 and 2, spotlight two fundamental points relevant for a possible transition of HEP publishing to OA: the volume of articles is small and these are concentrated in a few core titles, mostly published by learned societies. All HEP leading journals have recently taken a pro-active stance on OA. Journals from the American Physical Society, Elsevier and Springer offer authors an option to pay a fee to make their articles OA, while the *Journal of High Energy Physics* is recently experimenting with an institutional membership fee. The latter appears a more successful scheme, as funding mechanisms in HEP seldom include overhead for scientific publications to be directly used by authors. Moreover, the direct payment for the OA publication of an articles is perceived in very negative terms by the community, reminiscent of the unpopular “page charges” of some journals. This perception might have extended to other journals, such as the *New*

Journal of Physics, which are built on a “pay-per-article” Open Access model, but have so far attracted only a limited HEP content.

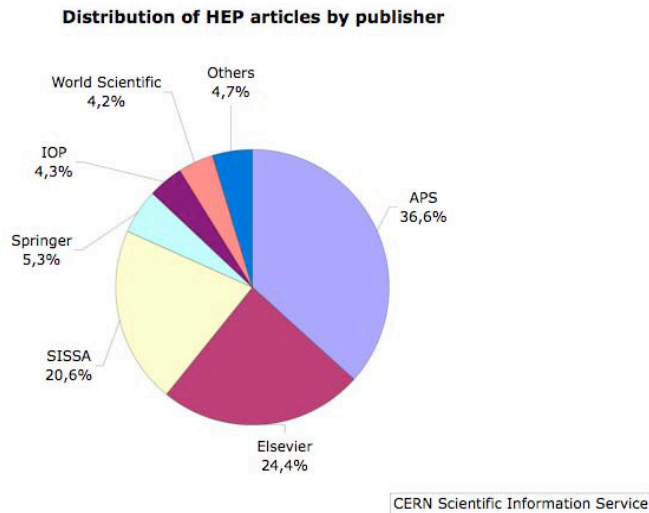


Figure 2: Distribution by publisher of the HEP articles submitted in 2005 to the *arXiv.org* repository under the categories *hep-ex*, *hep-lat*, *hep-ph* and *hep-th* and subsequently published in peer-reviewed journals. A total sample of about 5'000 articles is considered. From reference [3].

The debate on OA publishing in HEP was initiated by CERN. CERN is the leading HEP laboratory, with over half a century of history. Its flagship program, the LHC accelerator, will see four large experimental collaborations probe fundamental questions in our understanding of the Universe. CERN epitomizes cross-border collaboration in HEP: the LHC accelerator and detectors include components built in HEP laboratories and Universities around the world; the largest of the LHC experimental collaborations count as many as 2000 scientists, including about 400 students from 160 universities and laboratories spread over 35 countries. As part of its role to chart the future of HEP, in synergy with HEP funding agencies worldwide, CERN promoted several events focussed on OA publishing in HEP:

- September 2005. *Open meeting on the changing publishing model*. This event brought together representatives of authors, funding agencies and publishers with the aims of first discussing in HEP publishing issues such as publishing costs, competition, fair distribution of costs, opportunities for developing countries, alternative business models and the quality of peer-review [4];
- December 2005. *Colloquium on Open Access publishing in particle physics*. A representative group of authors, funding agencies and publishers indicated a possible way forward to OA publishing based on three pillars: asserting the complementary roles of repositories and peer-reviewed literature, decoupling preservation issues and the publication model, enshrining the importance of peer-review for evaluation and academic credibility [5];
- December 2005 to June 2006. *Task Force on Open Access Publishing in Particle Physics*. This tripartite task force composed by authors, funding agencies and publishers was charged by the main stakeholders to “study and develop sustainable business models for OA publishing for existing and new journals and publishers in particle physics”. In its report [2] it suggested to establish a sponsoring consortium, SCOAP³, as a central body which would remunerate publishers for the peer-review service, effectively replacing the “reader-pays” model of traditional subscriptions with an “author-side” funding;
- November 2006. *Establishing a sponsoring consortium for Open Access publishing in Particle Physics*. Following the task-force report and the acceptance of its model by representatives from major European stakeholders, a Working Party was established to develop a specific proposal for the creation of SCOAP³, which is described in this article [6]. More information is contained in the SCOAP³ Working Party report [7].

3 The SCOAP³ Model

SCOAP³ will act as a single interface between the main stakeholders of the HEP scientific information market: on one side the author and reader communities and on the other side the publishers of high-quality HEP journals. The aim of SCOAP³ is to establish OA to HEP peer-reviewed articles along the lines of the Budapest Initiative [8], namely “*free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself*”. At the time of writing, SCOAP³ is an initiative emanating from:

- several European funding agencies, among which IN2P3 and CEA (France), INFN (Italy), MPG (Germany), PPARC (U.K.), and other funding bodies from Greece, Norway, Sweden and Switzerland;
- the two largest European laboratories, CERN and DESY;
- national and international library consortia such as GASCO (Germany, Austria, Switzerland), INFER (Italy), COUPERIN (France), JISC (U.K.), ABM-uitvikling (Norway).

In the next months SCOAP³ aims to federate similar agents worldwide: the SCOAP³ model will only be successful if all countries contributing to the vast majority of the HEP literature become members of the consortium. Indeed, a pillar of the SCOAP³ model is to ensure OA to all HEP articles appearing in high-quality journals, irrespective of the affiliation of their authors. Manuscripts from authors without academic affiliation or authors from less-privileged countries, which cannot be reasonably expected to contribute to the consortium at this time, will be treated like all other articles. The ethical reason of conserving the access of any author to peer review is obvious. At the same time, this choice has solid financial reasons: restricting OA privileges only to authors affiliated to some countries would simply replace the present toll-access barriers with different barriers, connected to the geographical origin of the articles. Moreover, if only a geographical subset of the HEP scientific literature were available OA, consortium members would still be required to purchase the remaining fraction, with no evident financial benefits from the OA transition.

SCOAP³ will be financed with funds currently used for journal subscriptions by HEP funding agencies, laboratories and libraries. At the same time it will engage other bodies interested in the broad and free dissemination of scientific information. Each country will contribute to SCOAP³ in a “fair” way, according to its share of the worldwide HEP scientific production, as discussed in Section 5. For the SCOAP³ model to be successful, it should represent a stable, viable and sustainable alternative to subscriptions *vis-à-vis* its partners. It is therefore expected that the SCOAP³ operation will follow the financial blueprint of large HEP scientific collaborations, which usually bind over one hundred laboratories and universities in Memoranda of Understanding spanning several decades.

The innovation of the SCOAP³ model with respect to other OA options currently offered by most publishers is that it will centralize all OA expenses, which will not have to be borne by authors and research groups. These other “author-pays” options, of scarce success in HEP, are perceived as an even higher barrier than subscription charges, in particular for theoretical physicists from small institutions, whose articles account for the vast majority of HEP papers.

It is expected that SCOAP³ will contribute to stabilize the rising cost of access to information in the HEP domain by virtue of increasing author awareness to costs and prices, and by fostering new competition in the market, linking quality and price.

A large fraction of the publications on core HEP subjects is published in a limited number of journals [3], as detailed in sections 2 and 4. Among those journals, some carry almost entirely HEP content. SCOAP³ aims to assist publishers in converting these entire “core” journals to OA. It is expected that the vast majority of the SCOAP³ budget will be spent for “core” journals with a “lump-sum” payment: SCOAP³ pays a negotiated price for the peer-review of all articles processed by a journal. Many HEP articles appear in “broadband” journals, which carry just a small fraction of HEP content. It is expected that these articles will be sponsored by SCOAP³ on a “pay-per-article” basis. Conference proceedings and monographs are not within the scope of SCOAP³.

In the SCOAP³ OA model, the publishers will have the prime responsibility of ensure quality of the highest standards through independent editorial boards and the peer review. They will ensure the dissemination of OA articles by posting them onto their web sites and, in addition, feeding them to a SCOAP³ repository.

Publishers will benefit from a more sustainable business model than the traditional subscription scheme, becoming increasingly fragile. They will continue to meet the demand for print subscription, re-print of single articles, color plates in these printed versions, collections of articles in electronic or paper form, citation databases and other “premium” services, which are outside the scope of SCOAP³.

4 The High Energy Physics Publishing Landscape

The definition of HEP is often linked to the theoretical and experimental study of particles produced at accelerators of ever-increasing energy. Both the field and its definition have evolved to include subjects naturally more close to the fields of nuclear physics, of astrophysics and of cosmology. Different authors, different journals and different funding agencies each focus on different parts of the HEP spectrum and therefore have a different definition of the field.

To be successful, SCOAP³ should, at once, aim to convert to OA the subset of scientific literature of common interest to all players, while striving for as wide a scope as possible. A minimal set of common interest to the entire HEP community is constituted by a “core” set of topics such as the phenomenology and experimental investigations of elementary particles and their interactions, quantum-field theory and lattice-field theory. These topics are loosely related to the *hep-ex*, *hep-lat*, *hep-ph*, and *hep-th* areas of the *arXiv.org* repository, which often also carry content in cognate disciplines. Experimental techniques as well as mathematical and numerical methods are also included in this definition of HEP “core” articles. The definition of HEP article covers more loosely other fields of relevance to HEP, such as selected topics in nuclear physics, astrophysics, gravitation and cosmology.

It is important to note that the vast majority of HEP articles concern phenomenology and theory and have on average 2.6 authors [3]. On the other hand, publications on experimental results were often authored by up to 500 scientists in the last decade, while collaborations now publishing their analyses count up to 800 researchers and articles by LHC collaborations will have up to 2000 authors.

Journal	Publisher	IF	N_{tot}	N_{HEP}	N_{core}	f_{HEP}	f_{core}
<i>Phys. Rev. D</i>	APS	4.8	2285	2101	1635	92%	72%
<i>JHEP</i>	SISSA/IOP	5.9	856	856	840	100%	98%
<i>Phys. Lett. B</i>	Elsevier	5.3	957	862	740	90%	77%
<i>Nucl. Phys. B</i>	Elsevier	5.5	522	481	465	92%	89%
<i>Phys. Rev. Lett.</i>	APS	7.5	3836	407	279	11%	7%
<i>Eur. Phys. J. C</i>	Springer	3.2	331	272	234	82%	71%
<i>Mod. Phys. Lett. A</i>	World Scientific	1.2	281	216	138	77%	49%
<i>Phys. Rev. C</i>	APS	3.6	853	298	136	35%	16%
<i>Class. Quant. Grav.</i>	IOP	2.9	491	255	89	52%	18%
<i>Int. J. Mod. Phys. A</i>	World Scientific	1.5	878	143	88	16%	10%
<i>J. Math. Phys.</i>	AIP	1.2	446	108	74	24%	17%
<i>Phys. Atom. Nucl.</i>	Springer	0.9	220	106	72	48%	33%
<i>JCAP</i>	SISSA/IOP	6.7	156	128	57	82%	37%
<i>Gen. Rel. Grav.</i>	Springer	1.6	190	103	20	54%	11%
<i>Nucl. Instrum. Meth. A</i>	Elsevier	1.2	1371	312	16	23%	1%

Table 1: The most popular HEP journals and their publishers, together with their ISI impact factor, IF; the total number of articles published in 2005, N_{tot} ; the number of HEP articles, N_{HEP} ; and the number of articles in the HEP core subject, N_{core} . The journals are ordered in decreasing order of N_{core} . Only journals with $N_{\text{HEP}} > 100$ are shown. The last two columns show the fractions f_{HEP} and f_{core} of HEP and core articles, respectively. From reference [7].

In 2005, about 8’500 HEP articles were published in peer-reviewed journals, as included in the SPIRES database [9]. Of these, 5’200 articles are classified in the core HEP topics discussed above. Table 1 presents the most popular HEP journals and their corresponding publishers, together with their ISI impact factor, IF [10], and the total number of articles published in 2005, N_{tot} . The number of HEP articles, N_{HEP} , is also listed together with the number of articles in the core subject areas of phenomenology and experimental investigations of elementary particles and their interactions, quantum-field theory and lattice-field theory, N_{core} . The journals are ordered in

decreasing order of N_{core} . Only journals with $N_{\text{HEP}} > 100$ are shown. The last two columns show the fractions f_{HEP} and f_{core} of HEP and core articles, respectively [7].

As discussed in section 2, a recent study analyzed core HEP articles submitted in 2005 to the *arXiv.org* repository and classified in the *hep-ex*, *hep-lat*, *hep-ph* and *hep-th* categories and subsequently published. Out of a total of about 5'000 articles, more than 80% appeared in just six peer-reviewed journals from four publishers [3]. Five out of these six journals carry a majority of HEP content, as listed in table 1, these are:

- *Physical Review D* (published by the American Physical Society);
- *Physics Letters B* (Elsevier);
- *Nuclear Physics B* (Elsevier);
- *Journal of High Energy Physics* (SISSA/IOP);
- *European Physical Journal C* (Springer).

SCOAP³ aims to assist publishers in converting these “core” journals entirely to OA. As described in the last column of table 1, these five “core” journals include up to 30% of articles beyond the core HEP topics, particularly in Nuclear Physics and Astroparticle Physics. These articles will also be included in the OA conversion of the journals. This is in the interest of the HEP readership and promotes the long-term goal of an extension of the SCOAP³ model to these related disciplines.

The sixth journal, *Physical Review Letters* (American Physical Society), is a “broadband” journal, which carries only a small fraction (10%) of HEP content. SCOAP³ aims to sponsor the conversion to OA of this fraction on an article-by-article basis. A similar approach holds for another popular “broadband” journal in instrumentation: *Nuclear Instruments and Methods in Physics Research A* (Elsevier), which carries about 25% of HEP content.

These seven journals covered, in 2005, around 4'200 core HEP articles and about 5'300 articles in the wider HEP definition, including all related subjects. The conversion to OA of these five “core” journals and the HEP part of these two “broadband” journals would cover over 80% of the core HEP subjects and over 60% of the entire HEP literature, including all related subjects. The remaining 3'300 HEP articles, not published in the journals mentioned above, are scattered over some 140 other journals. It is important to note that the SCOAP³ model should not be limited to this set of journals but is open to all existing and future high-quality journals which carry HEP content, within budgetary limits.

5 Financial Aspects of the SCOAP³ Model

The price of a journal is driven by the costs to run the peer-review system, by editorial costs for copy-editing and typesetting, by the cost for electronic publishing and access control, and by subscription administration. Some publishers today quote a cost, from reception to final publication, in the range of 1'000 – 2'000 Euros per published article [11]. This includes the cost of processing articles which are eventually rejected, the fraction of which varies substantially from journal to journal.

The annual budget for a transition of HEP publishing to OA can be estimated from this figure and the fact that the five “core” journals, which cover a large fraction of the HEP literature, publish about 5'000 articles per year. Hence, we estimate that the annual budget for a transition of HEP publishing to OA would amount to a maximum of 10 Million Euros per year.

Another indication which corroborates this estimate is that the costs to run a “core” journal such as *Physical Review D*, amount to 2.7 Million Euros per year [11] and it covers about a third of the HEP publication landscape [3].

A “fair-share” scenario for the financing of SCOAP³ is to distribute these costs among all countries active in HEP on a *pro-rata* basis, taking into account the size of the HEP author base of each country. To cover publications from scientists from developing countries, which cannot be reasonably expected to contribute to the consortium at this time, an allowance of not more than 10% of the SCOAP³ budget is foreseen.

The size of the HEP author base in each country is estimated from a recent study [7] which considered all articles published in the years 2005 and 2006 in the five HEP “core” journals, *Physical Review D*, *Physics Letters B*, *Nuclear Physics B*, *Journal of High Energy Physics* and the *European Physical Journal C*, as well as those HEP articles published in the two “broadband” journals, *Physical Review Letters* and *Nuclear Instruments and Methods in Physics Research A*. A total sample of about 11'300 articles was considered and, in each of them, all authors were uniquely assigned to a given country. CERN was treated as an additional country.

In about 5% of the cases, authors were found to have multiple affiliations, often in different countries, reflecting the intense cross-border tradition of HEP. In these cases, the ambiguity in the assignment of authors to countries was solved as described in reference [7]. The results from this study are summarized in table 2 and figure 3.

Country	Share of HEP Scientific Publishing
United States	24.3%
Germany	9.1%
Japan	7.1%
Italy	6.9%
United Kingdom	6.6%
China	5.6%
France	3.8%
Russia	3.4%
Spain	3.1%
Canada	2.8%
Brazil	2.7%
India	2.7%
CERN	2.1%
Korea	1.8%
Switzerland	1.3%
Poland	1.3%
Israel	1.0%
Iran	0.9%
Netherlands	0.9%
Portugal	0.9%
Taiwan	0.8%
Mexico	0.8%
Sweden	0.8%
Belgium	0.7%
Greece	0.7%
Denmark	0.6%
Australia	0.6%
Argentina	0.6%
Turkey	0.6%
Chile	0.6%
Austria	0.5%
Finland	0.5%
Hungary	0.4%
Remaining countries	3.7%

Table 2: Contribution to the HEP scientific publishing of several countries. Co-authorship is taken into account on a pro-rata basis, assigning fractions of each article to the countries in which the authors are affiliated. The last cell aggregates contributions from countries with a share below 0.4%. This study is based on all articles published in the years 2005 and 2006 in the five HEP “core” journals, *Physical Review D*, *Physics Letters B*, *Nuclear Physics B*, *Journal of High Energy Physics* and the *European Physical Journal C* and the HEP articles published in two “broadband” journals, *Physical Review Letters* and *Nuclear Instruments and Methods in Physics Research A*. A total sample of about 11’300 articles is considered. From reference [7].

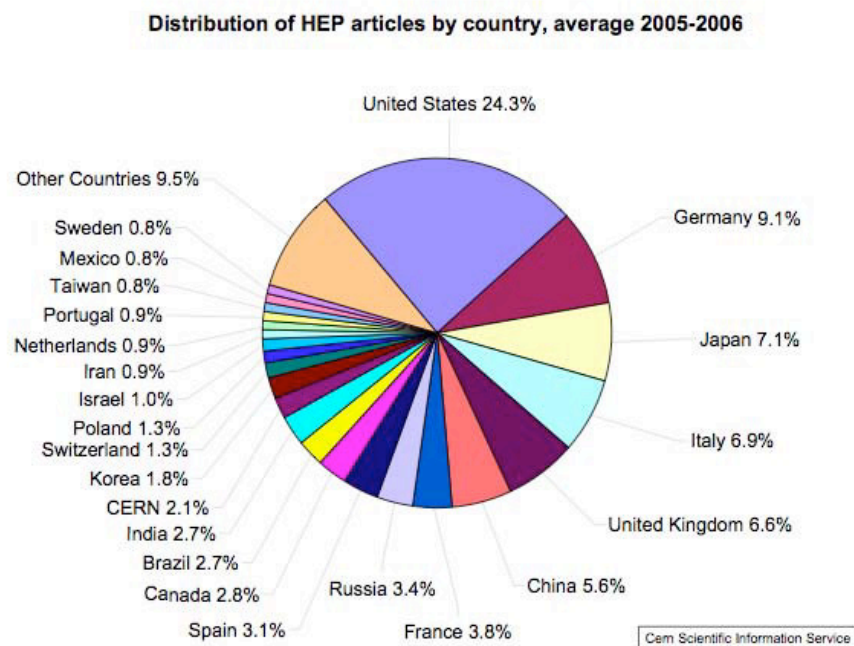


Figure 3: Contribution to the HEP scientific publishing of several countries. Co-authorship is taken into account on a pro-rata basis, assigning fractions of each article to the countries in which the authors are affiliated. The last cell aggregates contributions from countries with a share below 0.4%. This study is based on all articles published in the years 2005 and 2006 in the five HEP “core” journals, *Physical Review D*, *Physics Letters B*, *Nuclear Physics B*, *Journal of High Energy Physics* and the *European Physical Journal C* and the HEP articles published in two “broadband” journals, *Physical Review Letters* and *Nuclear Instruments and Methods in Physics Research A*. A total sample of about 11’300 articles is considered. Contributions from countries with a share below 0.8% are summed in the slice denoted as “Other Countries”. From reference [7].

6 Conclusions and Outlook

At the time of writing, SCOAP³ is an initiative emanating from leading European funding agencies, the two largest HEP European laboratories, and national and international library consortia. The fundamental pillar of the SCOAP³ model is the federation of HEP funding agencies and library consortia worldwide. HEP is the most global of the scientific enterprises and the conversion to OA of its literature, with all the ethical, scientific and financial benefits it implies can only be achieved in a global co-ordinated process. A crucial step towards OA publishing in HEP is therefore the search for a world-wide consensus around the SCOAP³ initiative, aiming to expressions of interest from HEP funding agencies and library consortia in Europe, the United States and beyond.

Once sufficient funds will have been pledged towards the establishment and the operation of SCOAP³, a tendering process involving publishers of high-quality HEP journals will take place. Provided that the SCOAP³ funding partners are ready to engage into long-term commitments, most publishers are expected to be ready to enter into negotiations along the lines presented in this article.

The outcome of the tendering process will allow the complete SCOAP³ budget envelope to be precisely known and therefore the precise contribution expected from each country. A Memorandum of Understanding for the governance of SCOAP³ will then be signed by funding agencies and leading national and international library consortia. Contracts with publishers will be established in order to make Open Access publishing in High Energy Physics a reality at the beginning of 2008, when the first experimental and theoretical publications of the CERN LHC program will appear.

The conversion of the HEP scientific publishing to the OA paradigm, along the lines presented in this article, will be an important milestone in the history of scientific publishing. The SCOAP³ model could be rapidly generalized to other disciplines and, in particular, to related fields such as Nuclear Physics or Astroparticle Physics.

Acknowledgements

This work summarises the report of the SCOAP³ Working Party which, between December 2006 and April 2007, drafted the blueprint for the establishment, financing and operation of the consortium, with vision and dedication. I am grateful to all members of the Working Party for having shared their unique experiences towards making OA in HEP a success: S. Bianco, O.-H. Ellestad, P. Ferreira, F. Friend, P. Gargiulo, R. Hanania, S. Henrot-Versille, A. Holtkamp, P. Igo-Kemenes, D. Jarroux-Declais, M. Jordão, B.-C. Kämper, J. Krause, T. Lagrange, F. Le Diberder, A. Lemasurier, A. Lengenfelder, C. Lindqvist, S. Plaszczyński, R. Schimmer, J. Vigen, R. Voss, M. Wilbers, J. Yeomas, K. Zioutas. We have been constantly inspired by the OA vision of CERN Director General, R. Aymar.

Notes and References

- [1] <http://oa.mpg.de/openaccess-berlin/berlindeclaration.html> [Last visited 4 April 2007].
- [2] VOSS, R., *et al.*, *Report of the Task Force on Open Access Publishing in Particle Physics*, CERN, 2006; http://library.cern.ch/OATaskForce_public.pdf.
- [3] MELE, S., *et al.*, JHEP 12(2006)S01; arXiv:cs.DL/0611130.
- [4] <http://open-access.web.cern.ch/Open-Access/20050916.html> [Last visited 4 April 2007].
- [5] <http://indico.cern.ch/conferenceDisplay.py?confId=482> [Last visited 4 April 2007].
- [6] <http://indico.cern.ch/conferenceDisplay.py?confId=7168> [Last visited 4 April 2007].
- [7] BIANCO, S., *et al.*, *Report of the SCOAP³ Working Party*, CERN, 2007; in preparation. To obtain a copy please contact Salvatore.Mele@cern.ch.
- [8] <http://www.soros.org/openaccess/read.shtml> [Last visited 4 April 2007].
- [9] <http://slac.stanford.edu/spires> [Last visited 4 April 2007].
- [10] <http://scientific.thomson.com/free/essays/journalcitationreports/impactfactor/> [Last visited 4 April 2007].
- [11] BLUME, M. Round table discussion: *Policy Options for the Scientific Publishing System in FP7 and the European Research Area*. Conference on Scientific Publishing in the European Research Area: Access, Dissemination and Preservation in the Digital Age, Brussels 15-16 February 2007.

Importance of Access to Biomedical Information for Researchers in Molecular Medicine

Annikki Roos; Turid Hedlund

Information Systems Science, Department of Management and Organization, Swedish School of Economics and Business Administration, Arkadiankatu 22, 00100 Helsinki, Finland
e-mail: annikki.roos@ktl.fi; turid.hedlund@hanken.fi

Abstract

In this paper, we analyze and describe the information environment of biomedicine from the point of view of the researchers in molecular medicine, which is a sub branch of biomedicine. We shall describe the nature of the discipline and its reflections to the information environment. A survey concerning the most important information resources in one molecular medicine research unit was conducted, and in this paper the main results of the survey is reported. The role of scholarly journals in the research process will also be analyzed. Special attention will be given to the possibilities of open access to the research process.

Keywords: information environment; information resources; databases; research process; molecular medicine

1 Introduction

The aim of this paper is to analyze and describe the information environment of biomedicine from the point of view of the researcher in molecular medicine (MM), a sub branch of biomedicine. Our target group is a research group containing researchers at different stages of their research career and the focus of study is on their daily work using information resources as part of the research process. The discipline is a rapidly growing and developing new research methods and processes which can be observed by the fact that pure laboratory work is to a growing degree transformed to computerized techniques. We argue that the change of the discipline from mainly laboratory based work to data based work has thoroughly changed the research processes. This has natural implications also to the information environment, as well as information retrieval, sharing practices and usage of information.

In this study the focus of research and our main research questions deal with the information environment of molecular medicine and firstly what are the main changes it has undergone. Secondly we investigate by conducting a survey, which are the most important information resources for researchers at different stages of their research career and thirdly what is the role of scholarly journals in the research process? For example, what is the publishing strategy and the criteria for choosing a journal to publish in.

We selected one research unit working in MM in Finland as a case. A web survey was conducted and qualitative information about researchers, their current work tasks, used information resources, publishing strategies and practices were gathered. A presentation and a feedback session concerning the results of the enquiry were given to the researchers. In this session important and explaining comments were given by the researchers in the target group about the use of information resources which have been taken into account when analysing and reporting the results of this study.

The outline of the paper is as follows: In Section 2, we describe the nature of the discipline and its reflections to the information environment. In Section 3, the effects of the changes in the environment will be analyzed against research process and scholarly communication practices. Special attention will be given to the experienced possible effects of open access in its different forms to the process. In Section 4, the results of the study are reported and in Section 5 we come to the conclusions and discussion.

2 Molecular Medicine as a Discipline

The discipline of biomedicine is growing exponentially. There are many factors behind the growth, of which the most important might be substantial increase in government support, the continued development of biotechnology industry, and the increasing adoption of molecular-based medicine. [1]. It has been pointed out in many sources that the nature of biomedicine has changed. It has transformed from laboratory based science to an

information science, science “in silico”. [e.g. 2, 3, 4], which means mainly the computerization of the research process.

Specialization to different research domains, fields and sub-disciplines qualifies biomedicine. As Buetow felicitously remarks each of these “speak its own scientific dialect”. Like in many other scientific fields, “big science” (i.e. big budget, big staff, big machines etc.) is a growing challenge to the discipline. Research equipment and technology are extremely expensive and these are factors which have been leading researchers to work on teams. Biomedicine, according to Buetow is a “team science”. It is typical of biomedical research teams that many research problems in order to be solved have to cross traditional discipline boundaries. [1].

Molecular medicine, a sub-discipline of biomedicine is a practice oriented, applied science and utilizes molecular and genetic techniques in the study of the biological processes and mechanisms of diseases. It is highly reliant upon the development of techniques and technology for acquiring data. [5]. Its final, practical task is to provide new and more efficient approaches to the diagnosis, prevention, and treatment of a wide spectrum of congenital and acquired disorders [6]. The nature of MM, like biomedicine in general is interdisciplinary, it could also be seen as a hybrid of biomedicine and molecular biology. Molecular biology in turn is based on the combination of biochemistry, cell biology, virology and genetics [7].

3 Information Environment and the Changing Research Process

We define information environment in this study as the entity of information objects as well as the tools and services needed to retrieve, manage and analyze them.

A large volume of data in combination with the diversity of data types is typical for MM information environment. The characteristic of the data is that it is rapidly expanding and ever-changing. [1]. Most of the research databases, like genomic and proteomic databases are commonly updated and globally shared. A yearly updated list of online molecular biology databases is found in the website of Nucleic Acid Research [8]. The January 2007 edition contained almost 1000 databases [9]. The amount of data growth could be described by for example the situation of the GenBank, a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation. It doubles in size about every 18 months. At the beginning of 2007, it contained over 65 billion nucleotide bases from more than 61 million individual sequences. [10].

What is even more challenging is that there is a need to integrate different kinds of data, e.g. to move between the biological and chemical processes, organelle, cell, organ, organ system, disease specific, individual, family, community and population. [1]. Like Butler notices, there are some disciplines which already have software that allows data from different sources to be combined seamlessly. For example, a gene sequence can be retrieved from the GenBank database, its homologues using the BLAST alignment service, and the resulting protein structures from the Swiss-Model site in one step. [11]

In parallel with the growth of data, the number of different tools, developed for data retrieval and analysis is growing. An actively maintained directory of bioinformatic links lists over 1000 web servers and other useful tools, databases and resources for bioinformatics and molecular biology research in 2006 [12, 13].

PubMed, the most important bibliographic database in biomedicine consisted in 2006 of 16 million references. The growth rate of the database is about 12 000 references every week, which means yearly over 600 000 new references. The growth curve of Medline, the main database in PubMed is illustrated in Figure 1. These lines describe the growth of traditional, published material, mainly in article format in biomedicine in a condensed way. It seems that inside the growing domain, there are some really “hot topics” where the amount of literature increase is extreme.

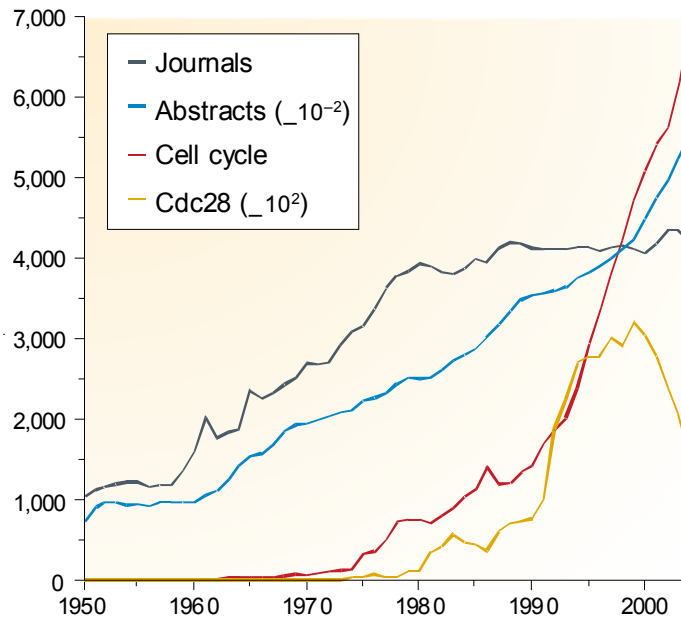


Figure 1: Growth of Medline: the number of journals, abstracts, papers on the cell cycle and papers on Cdc28 [14 published in Nature Reviews Genetics]

The typical features of MM information environment could be concluded as large volume and constantly growing number of data and published material, diversity of data types, great number of retrieval, analysis and other tools, interdisciplinary and globally shared and updated environment and team work. This is a fertile ground for the creation of new knowledge and inventions, but the lack of integration constitutes an increasing challenge to the development.

Cannata et al. have urged the organization of bioinformatics resources; data, knowledge, computational resources and services as a solution to the disintegration. They talk about “bioinformatics resourceome project” which would mean a process of creating a distributed system for describing resources, announcing their availability, and presenting this to the research community in an easy-to-navigate manner. The first step would be creation of an overall, distributed and collaboratively expandable ontology. [15, 16]. Mukherjea [17] has described the possibilities of using the semantic web in integrating the information resources. Grid technology has also been seen as a technical solution to the disintegration of data, information and tools. [1]

4 Results of the Survey

4.1 About the Research Unit and the Current Tasks

The research unit chosen as the case is situated in a Finnish research institute. As their aim, the unit declares to produce top level research in the molecular background of cardiovascular, immunological and neuropsychiatric diseases. At the moment of enquiry (February 2007), the unit consisted of 10 research groups with 83 researchers. From these 58 were PhD students and the rest were graduate students, group leaders and senior researchers. We received totally 63 answers (75.9 %) to our web survey. 43 (68%) of those who responded were students and 20 (32%) were senior researchers, post docs and group leaders.

The research subjects of the groups were quite different, some of the groups concentrating on the genetic background of common diseases (“complex diseases”), some mainly to molecular genetics of monogenic diseases. There was also one bioinformatics group and one which specialized mainly in systems biology, one to quantitative genetics and a couple of groups mainly to the cell and molecular biology of certain diseases. We assume that the diversity of the research subjects caused some variety to reported work tasks between groups.

In the survey, all researchers were asked about their current work tasks and about information resources related to their current project or tasks and some information about usage of resources in general were asked. Respondents did get free spaces to write about their information resources, we gave only some examples for possible answers. We tried to get as broad a spectrum of possible resources, and did not want to limit or direct answers more than necessary. For current work tasks, we gave nine alternatives, from which it was possible to choose as many as were needed. Researchers were also able to add new tasks when necessary.

From the following figure (Figure 2.) the distribution of current tasks and their frequency among researchers is shown. The most common task among researchers was writing a report or an article, about totally 67 % of the researchers were doing it currently, the distribution among seniors and students is 70 % (seniors) and 67 % (students). Two-thirds of researchers were reading, 76 % of them were students. Of those working in the laboratory 74 % were students. It was more common (43 % of the respondents) to search information about literature from databases than data from data collections (25 %). Over one-third of the researchers were doing scientific computing. The researchers, who were studying the genetic background of “complex diseases” were practicing more scientific computing than most of the other groups. In two research groups where two-thirds (over 70 %) of all respondents answered that they were doing scientific computing.

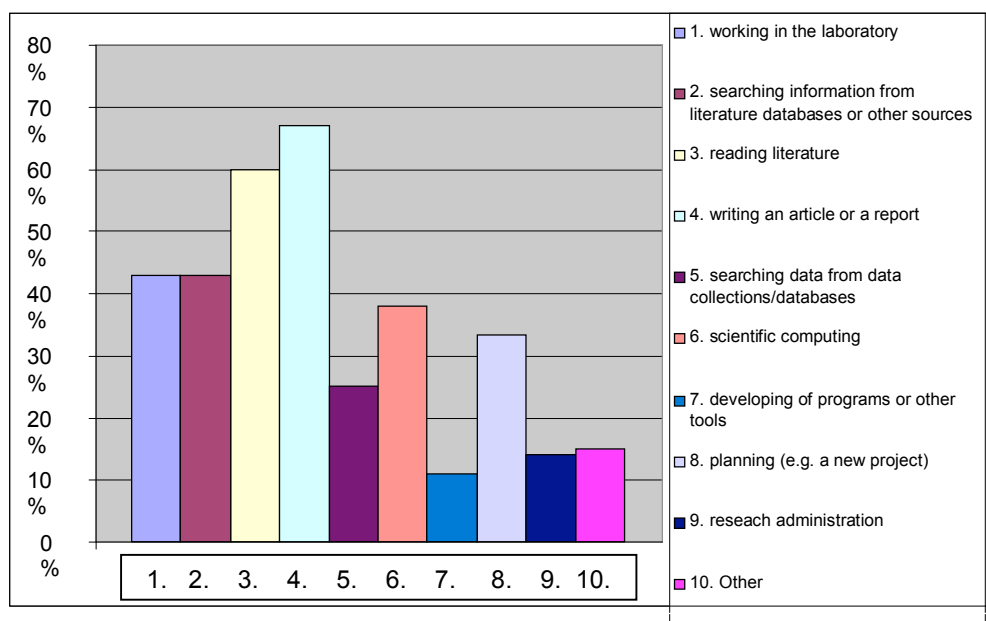


Figure 2: Current work tasks of researchers

4.2 Most Important Resources

When asked to choose at least the three most useful resources for their current research projects, doctoral and graduate (n=43) students named more resources than seniors and group leaders (n=20). PubMed got most references as the most useful resource in both groups. In the student's group UCSC Genome Browser was second and Google third as the number of references are concerned. In the seniors' group the ranking was contrary.

As their first information source 68% of the respondents named intranet/internet and in practice according to their answers, this means mainly PubMed and Google. 27 % of all researchers did prefer to contact a colleague or a supervisor. There seems to be no difference between students and seniors. In the feedback session researchers commented that the first information source depends on the nature of the issue: in practical questions and problems a colleague is preferred. It might also be possible, that some personal characters of the group leaders might at least partially explain the difference. The results indicate that in certain groups more researchers than on average in the groups favoured contacting a colleague in the first place. However, this is a speculation and needs to be observed more thoroughly.

When asked about which published material they use, the majority of respondents (53 %) answered that they use only or dominantly articles. 35 % of the researchers responded that they used articles and books equally and the rest 12 % named articles, databases and also some books.

When asked to name journals that researchers follow regularly, 23 % of the respondents reported that they do not follow any particular titles, rather their own topic from the literature databases. All of these respondents were graduate and doctoral students. Almost all graduate students belonged to this group.

91 % of the researchers said that they had used data collections during their current project. Those who did not use were juniors, who had recently started research work or researchers who were at the moment mainly working in the laboratory and writing articles. The problem with the reported data resources was that, because the question was open, researchers' answers were at very different levels. Some of them named quite general data collections, like "protein databases", or merely services or portals, like Entrez, while there were also respondents who used the detailed names of the databases or services. Totally 43 different data resources or services were named. The most common were NCBI and Entrez databases from National Centre for Biotechnology Information (by NIH and NLM) and UCSC Genome Bioinformatics resources, especially one tool, namely UCSC Genome Browser.

53 % of the researchers replied that they had used some research tools during their current project. The selection of tools and programs was also very diverse, from programs developed in their own laboratory to the commercial products. Totally 67 different tools were named. Students were naming more tools than seniors. The most often mentioned tool was Primer3, which is a PCR (Polymerase chain reaction) primer designer tool. The largest group in our survey as a whole was proteomics and sequence databases and analyzes tools.

It was noteworthy that tools for data mining seem to be common, but none mentioned text mining tools or tools for hypothesis creation. A tool called iHOP was familiar to the researchers. It's interesting, because it integrates gene and protein data from different collections with scientific literature.

Social bookmarking tools like Nature's Connotea were not named, neither any blogs. When asked why not, the answer in the feedback session was that they did not find those useful because their research problems were so specific: "they are only a waste of time". According to some opinions published in Nature researchers in general have not been eager to accept these tools because they might have been afraid of the poor image of the new tools and might have suspected the tools might damage their career [see 18].

4.3 Role of Scholarly Journals in the Research Process

Writing and publishing articles in scientific journals are seen as an important part of the research process in biomedical sciences and molecular medicine. This is shown among others in [14] but also in this present case study of the research group on MM in Finland. When asked about their current work tasks about 67 % of the researchers in the case group answered that they were writing an article or a report.

Since the research group constitutes of senior researchers as well as doctoral and graduate students this can be seen as a high percentage. The amount of work and the importance of article writing is also to be seen in the results presented in Table 1., where we were asking the researchers questions about their publishing strategy for the coming year. All of the senior researchers and group leaders are going to publish at least 1 article, most of them (87.5%) are going to publish at least two articles and 75% of the group leaders and 43% of the senior researchers are planning to publish at least three articles. We have counted as main authors, the first and second author and the last author. In this case study most of the senior researchers and group leaders are acting as supervisors to younger researchers, why it seems appropriate that the last author is counted as important.

	100% minimum 1 article as main author (1 & 2 or last)
Group leaders	100% minimum 1 article as main author
Senior Researchers	83,3% minimum 2 articles as main author
Post doc	88% minimum 1 article as main author
Doct.students	73,3% minimum 1 article as main author
Graduate stud.	

Table 1: Publishing strategy regarding scientific articles of researchers for the coming year of the researchers in MM

When looking at realized results (from 2006) for publications from the research group, 71 research articles in A-class journals and a total of 79 scientific articles were published. Of these 13 articles were in open access hybrid journals (applying some type of embargo) and 2 articles were in purely open access journals.

Regarding the choice of where to publish the researchers were presented the following criteria: impact, the speed of publishing, scope, open access or some other criteria, of which they were asked to name the one they regarded as most important. Impact was named as the most important by 58% of the researchers and scope by 39%. A few of the researchers named a combination of scope and impact. Open access as the main criteria was named by only 3% of the researchers.

The researchers were also asked to name journals with a suitable scope for publishing. On the top of the list of journals with suitable scope (Table 2.) was Nature genetics (named by 15). The impact factor for Nature genetics is also very high (25.797).

Journal title	Number of nominations	Impact factor the journal
Nature genetics	15	25.797
Human molecular genetics	11	7.764
Molecular psychiatry	10	9.335
American journal of human genetics	9	12.649
European journal of human genetics	6	3.251
Nature	6	29.273

Table 2: Top listing of journals with suitable scope for publishing

However, even though journals hold an established position in scholarly communication, there has appeared comments and viewpoints which have suggested that because scientific publications are slow and access to them is limited they act more as barriers to the development of new knowledge and science. [19].

In fact, traditional journals have very seldom made it possible to attach data files containing research data to the article. However, digital publishing and open access initiatives have opened up new possibilities for scientific publishing (Björk 2007). In a study by Hedlund and Roos (2007) on publishing practices among biomedical researchers, the authors found that there is a growing rate of research publications in BioMed Central by Finnish researchers during the years 2003-2004. Cockerill & Tracz (2006) name fields like bioinformatics, genomics and systems biology as possible success fields for open access. The initiative from the open access journal publishers BioMed Central is to put up a structured XML version of each full text article for data mining. There is also an increasing number of institutional repositories that allow researchers to upload data files linked to their published articles, which then serve as a possible source for data mining. Cockerill and Tracz (2006) argue that in the future the potential reader of a research article may not be only human beings but instead software agents looking for data to be extracted and processed for a knowledge base. Therefore open access is important for work that involves multiple disciplines, as for example computer scientists, mathematicians and biologists collaborating in the areas of systems biology and bioinformatics.

5 Conclusions and Discussion

The information environment of researchers in MM could be summarized in the following diagram (Figure 3.)

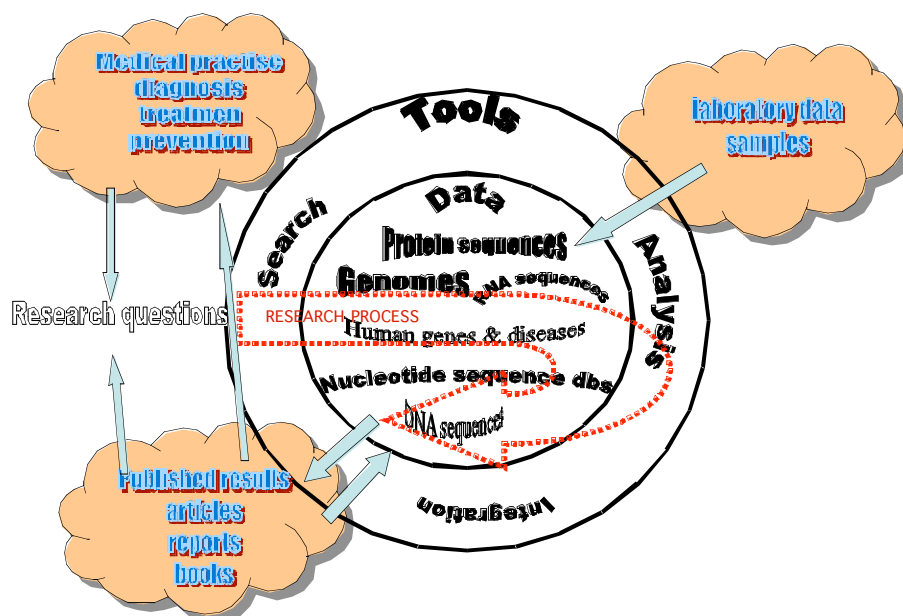


Figure 3: The research process and the information environment of molecular medicine

It can be concluded that access and use of data resources is an important and integral part of the research process in MM. The amount of different data collections, searching and analysis tools is huge. The disintegration of the environment seems also to be quite problematic.

We noticed that a more thorough analysis would be needed to make any conclusions about the relationship between the different work tasks in the research process and the used resources. We assume that many of the tasks might consist of several levels all of which might be worked out via different resources. The reason for this being for example in the varied complexity of the research problems.

The number of published articles is growing exponentially, especially in the “hot topics” of the domain. Researchers might find it difficult to follow even the development in their own research area. Maybe this is the reason why students do not follow particular journals, rather topics. The amount of literature is growing so fast that they are not able to do anything else than to follow the most recent and important articles from reference databases like PubMed. The disinterest to follow particular journals might also be due to the fact that they are not so well integrated into the domain yet, or it could be possible that their research subjects are so interdisciplinary that at least at the beginning of their career they are not able to follow any particular titles.

Journal publishing is still seen as the prominent way of distributing research results in molecular medicine. It has been shown in the case study that writing articles and reports is occupying the researchers as an important part of the research process. Even though many attempts to introduce open access, e.g. by providing institutional and national licences to cover authorship fees in BioMedCentral journals there still seems to be a strong reliance on traditional journals and especially journals with high impact factors. Publishing in journals with high impact factor and the right scope is a strong base in the prevailing publishing strategy. However, it could be possible that the importance of traditional publishing channels and particularly articles might be on their way to change in the future if the text mining and hypothesis creation tools will be developed, and if the technical cyberinfrastructure with semantic web tools will be developed to integrate the environment. Open access will be helpful and a natural part of this development.

Notes and References

- [1] BUETOW, KH. Cyberinfrastructure: Empowering a "Third Way" in Biomedical Research. 2005:821-4.
- [2] LENOIR, T. Shaping Biomedicine as an Information Science. *Conference on the History and Heritage of Science Information Systems*: Information Today 1999.
- [3] LENOIR, T; ALT, C. Flow, Process, Fold: Intersections IN. In: Picon A, Ponte A, eds. *Science, Metaphor, and Architecture*. Princeton: Princeton University Press 2003:314-53.
- [4] HINE, C. Databases as Scientific Instruments and Their Role in the Ordering of Scientific Work. *Social Studies of Science* 2006:269-98.
- [5] MACMULLEN, W. Information problems in molecular biology and bioinformatics. *Journal of the American Society for Information Science and Technology*. 2005;56(5):447-56.
- [6] GOOSSENS; M. Principles of molecular medicine. *New England Journal of Medicine*. 1999;340(20):1601-2.
- [7] MIETTINEN, R; TUUNAINEN, J; KNUUTTILA, T; MATTILA, E. Tieteestä tuotteeksi? Yliopistotutkimus muutosten ristipaineessa. Helsinki: Yliopistopaino 2006.
- [8] Nucleic Acids Research. Oxford Journals | Life Sciences | Nucleic Acids Research | Database Summary Paper Alpha List. 2007 [cited 10 April 2007]; Available from: <http://www.oxfordjournals.org/nar/database/a/>
- [9] GALPERIN, M Y. The Molecular Biology Database Collection: 2007 update. *Nucleic Acids Research*. 2007;35(Database issue):D3.
- [10] BENSON, DA; KARSCH-MIZRACHI I; LIPMAN D J; OSTELL J; WHEELER D L. GenBank. *Nucleic Acids Research* 2007:D21-5.
- [11] BUTLER, D. Mashups mix data into global service. *Nature*. 2006;439(7072):6-7.
- [12] UBIC. NAR Web Server Issue (July 1, 2006) - UBC Bioinformatics Centre. 2007 [cited 10 April 2007]; Available from: http://bioinformatics.ubc.ca/resources/links_directory/narweb2006/
- [13] FOX, J A; MCMILLAN, S; OUELLETTE, B F. A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. *Nucleic Acids Research*: Oxford Univ Press 2006:W3.
- [14] JENSEN, L J; SARIC, J; BORK, P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet*. 2006 Feb;7(2):119-29.
- [15] CANNATA, N; CORRADINI, F; MERELLI, E. A Resourceomic Grid for bioinformatics. *Future Generation Computer Systems*. 2007;23(3):510-6.
- [16] CANNATA, N; MERELLI, E; ALTMAN, RB. Time to Organize the Bioinformatics Resourceome. *PLoS Computational Biology* 2005:e76.
- [17] MUKHERJEA, S. Information retrieval and knowledge discovery utilising a biomedical Semantic Web. *Brief Bioinform*. 2005 Sep;6(3):252-62.
- [18] BUTLER, D. Science in the web age: Joint efforts. *Nature*. 2005;438(7068):548-9.
- [19] INSEL, T R; VOLKOW, N D; LI, T K; BATTEY, J F; LANDIS, S C. Neuroscience Networks. *PLoS Biology*. 2003;1(1):e17.

Representing and Coding the Knowledge Embedded in Texts of Health Science Web Published Articles

Carlos Henrique Marcondes¹; Marília Alvarenga Rocha Mendonça¹; Luciana Reis Malheiros²; Leonardo Cruz da Costa³; Tatiana Christina Paredes Santos⁴; Luciana Guimarães Pereira⁵

¹ Department of Information Science

e-mail: marcon@vm.uff.br; e-mail: mariliaalvarenga@terra.com.br;

² Department of Physiology and Pharmacology

e-mail: malheiro@vm.uff.br

³ Department of Computer Science

e-mail: leo@dcc.ic.uff.br

⁴ Biomedicine student

e-mail: tatianacps.uff@gmail.com

⁵ Library and Information Science student

e-mail: lucianaguipe@yahoo.com.br

Universidade Federal Fluminense

R. Miguel de Frias, 9 – Icaraí

24220-008 - Niterói – RJ Brazil

Abstract

Despite the fact that electronic publishing is a common activity to scholars, electronic journals are still based in the print model and do not take full advantage of the facilities offered by the Semantic Web environment. This is a report of the results of a research project with the aim of investigating the possibilities of electronic publishing of journal articles both as text for human reading and in machine readable format recording the new knowledge contained in the article. This knowledge is identified with the scientific methodology elements such as problem, methodology, hypothesis, results, and conclusions. A model integrating all those elements is proposed which makes explicit and records the knowledge embedded in the text of scientific articles as an ontology. Knowledge thus represented enables its processing by intelligent software agents. The proposed model aims to take advantage of these facilities enabling semantic retrieval and validation of the knowledge contained in articles. To validate and enhance the model a set of electronic journal articles were analyzed.

Keywords: electronic publishing; scientific communication; semantic web; knowledge representation; ontologies

1 Introduction

Nowadays, electronic Web publishing is a common activity to scholars and researchers. However, scientific communication is still a slow social process which largely depends on discourse, text producing, reading/interpreting/inquiring and peer-reviewing by scholars until new knowledge is incorporated into the corpus of Science. The potential of new information technology (IT) has been applied to modern bibliographic information systems to improve scientific communication, providing fast notification and immediate access to full-text scientific documents. But IT is not yet used to directly process the knowledge embedded in the text of scientific articles.

Semantic Web Initiative is a future vision of the Internet which aims to structure today's vast Web content, adding semantic to this content [1]. The technologies and methodologies that have been developed in the context of Semantic Web will enable this content to be understandable not only by people but also by software agents, enabling them to *reason* on this content in achieving different intelligent and useful tasks. In the Semantic Web context, electronic publishing can be a cognitive tool with potential that is far from being explored. Today electronic journals are still based on the print mode. Electronic Web published articles are knowledge bases, but for human reading.

Before the rise of the Web, what constitutes the accented scientific knowledge of humanity was fuzzy, lacks formalization, and was scattered across journals collections throughout libraries. Today there are two main

barriers to a large scale use of this knowledge: the amount of information available throughout the Web and the fact that knowledge is embedded in the text of scientific articles in an unstructured way, not adequate for program processing.

Today, different scientific communities are developing Web ontologies which formally record the knowledge in a domain. W3C [2] defines ontology as “*a knowledge representation*”. According to Jacob [3 p. 200] an ontology is “*a partial conceptualization of a given knowledge domain, shared by a community of users, that has been defined in a formal, machine-processable language for the explicit purpose of sharing semantic information across automated system*”. In a near future, formal ontologies will be developed and recorded in program readable format, containing the accented knowledge in specific domains. Applying Semantic Web technologies to identify and record the knowledge embedded in the text of scientific articles in program-understandable format and compare it to the knowledge recorded in Web ontologies may be a key feature to the development of a future e-Science environment. Both these knowledge resources may be accessed by software agents on behalf of their owners, thus providing scientists with new tools to information and knowledge retrieval, to identify, evaluate and validate new contributions to Science.

The present research is looking for a new paradigm in scientific Web publishing: to publish not only text, for human reading, but also knowledge, formalized as ontologies, able to be processed by software agents. The objective of this research is to develop a Web publishing model which will be the basis for the future development of enhanced scientific authoring, publishing, retrieval and validating tools. These tools will enable the electronic publishing of scientific articles not only as texts for human reading, but also as a knowledge base in program-understandable format. The model aims to identify and record the semantic elements which constitute the knowledge embedded in the text of a scientific article.

What is the nature of scientific knowledge? This knowledge today, although recorded in digital format as Web published articles, are unstructured and not in adequate format for processing by software agents. According to Brookes [4 p. 131]: “*knowledge is a structure of concepts linked by their relations and information is a small part of such a structure*”. Sheth [5 p. 1] states that “*Relationships are fundamental to semantics – to associate meaning to words, items and entities. They are a key to new insights. Knowledge discovery is about discovery of new relationships*”. Miller [6 p. 306] answer these questions as: “*The above remarks imply-that science is a search after internal relations between phenomena*”. Here scientific knowledge is considered as discovering relations between phenomena.

By the 16th century, a mark in the institutionalization of Science is the establishment of the scientific method as a procedure to achieve and communicate true statements in Science. A special element of scientific method is the hypothesis. As Scientific Methodologies handbooks emphasize, the role of hypotheses are central to Science in providing a provisory explanation to a phenomena and thus guiding the scientific inquiry. In the scientific method the hypothesis is the element which expresses a relation between phenomena.

Although a complex phenomena, scientific reasoning as expressed in the text of scientific articles must serve to an essential communicational role to Science as an institution: to validate the knowledge contained in the article, enabling any scientist to reproduce the steps taken by the author in his/her experiment. The need of this rigid protocol when communicating research results is stated by The International Committee of Medical Journals Editors, <http://www.icmje.org>:

“The text of observational and experimental articles is usually (but not necessarily) divided into sections with the headings Introduction, Methods, Results, and Discussion. This so-called “IMRAD” structure is not simply an arbitrary publication format, but rather a direct reflection of the process of scientific discovery”

It is assumed here that knowledge in the text of articles – scientific methodology elements as the Problem, Hypothesis, Results and Conclusions – are all interrelated, constituting the content of the reasoning process developed by the author through which he/she communicates a new discovery. With the support of a Web authoring/publishing tool these semantic elements – the knowledge contained in the article -, can be identified, extracted and recorded in machine-understandable format, as an ontology. Knowledge thus recorded can be processed by software agents thus enabling semantic retrieval, consistence and validate checking. The ontology representing the knowledge extracted from the article can also be compared, matched and aligned to public Web ontologies which more and more represent the corpus of public knowledge in specific domains, thus enabling the establishment of formal relationship between both ontologies. Fails to establish these relationships may be evidences of new discoveries, since it can indicate that the knowledge in the article is not yet represented in the ontology which stores the accented knowledge in a specific domain.

2 Methodology

Building models is an important tool in Science. It enables Science to cope with complex phenomena such as scientific reasoning in communicating new discoveries through the text of scientific articles. An initial semantic model was developed, based on literature on Scientific Methodology, Philosophy and Epistemology of Science. Using the initial framework 53 articles on Health Science were analyzed with the aim of enhancing and validating the model. Articles were chosen from two outstanding Brazilian research journals, 20 articles from the *Memórias do Instituto Oswaldo Cruz*, which scope is mainly Microbiology, <http://www.scielo.br/revistas/mioc>, 20 articles from the *Brazilian Journal of Medical and Biological Research*, <http://www.scielo.br/revistas/bjmb>. Both are international journals using English as primary language. These journals were selected because initially we intended to interview authors personally. 14 additional articles about stem cells were analyzed too. Stem cells as an emerging research area in rapid development, was chosen expecting to find articles reporting important discoveries. Articles analyzed were selected from three recent reviews which present the stem cells research development in a historical perspective, promoting the advances in research, which was of special interest to this research. These reviews are “The Human Embryonic Stem Cell and the Human Embryonic Germ Cell”, the official National Institute of Health (USA) resource for stem cells research, <http://stemcells.nih.gov/>, the article by Bongso et al. [7] and the article by Friel et al. [8].

The analysis simulates the tasks to be performed by an authoring/publishing tool when interacting with an author to identify and record the knowledge embedded in the text of an article. Scientific articles are highly conventional text types, with clear goal shared by authors and readers. Articles in Health Science are chosen for analysis due to their highly standardized structure, the so-called IMRAD – Introduction, Material and Methods, Results and Discussion - structure.

In order to explore the possibilities of using the model to identify new discoveries in Science, it is also verified if concepts found in the knowledge extracted from each article’s text exist in a public knowledge base. DECS – *Descritores em Ciência da Saúde* - <http://www.bireme.br/php/decsws.php>, a Portuguese version of MeSH – *Medical Subject Headings* – <http://www.nlm.nih.gov/MeSH/>, and MeSH itself were both used in this experience in the role of a public knowledge base, with which subject headings found in the article’s corresponding Lilacs (Latin America and Caribbean Literature on Health Science) or Medline database records are compared. MeSH is a component of UMLS - *Unified Medical Language System* -, <http://www.nlm.nih.gov/pubs/factsheet/umls.html>. It is a project of National Library of Medicine, USA, which aims to unify and encompass different medical specialized terminologies, thesaurus and classification schemas. UMLS evolves towards an ontology – the UMLS Semantic Network - in which concepts are organized in 134 classes or “semantic types” and 53 “types of relations”.

The article analysis used the following form:

ARTICLE ANALYSIS FORM	
Journal: Memórias do Instituto Oswaldo Cruz	URL: http://www.scielo.br/revistas/mioc
Reference CAMARA, Geni NL, CERQUEIRA, Daniela M, OLIVEIRA, Ana PG <i>et al.</i> Prevalence of human papillomavirus types in women with pre-neoplastic and neoplastic cervical lesions in the Federal District of Brazil. <i>Mem. Inst. Oswaldo Cruz.</i> [online]. Oct. 2003, vol.98, no.7 [cited 10 March 2005], p.879-883. Available from World Wide Web: < http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0074-02762003000700003&lng=en&nrm=iso >. ISSN 0074-0276	
METHOD OF REASONING	
Deductive: X Inductive: Abductive:	
PROBLEM (extracted from the text)	
As a contribution to the public health authorities in planning prophylactic and therapeutic vaccine strategies, we describe the prevalence of human papillomavirus (HPV) types in women presenting abnormal cytological results in Pap smear screening tests in the Federal District, Central Brazil.(Abstract)	
In contrast to what is observed in developed countries, cervical cancer mortality in Brazil is still high. (Introduction)	

HYPOTHESIS – previous (extracted from the text)
The chronic infection by certain types of human papillomavirus (HPV) is definitely related to the incidence of cervical cancer (Lorincz et al. 1992, IARC 1995) and the HPVs –16, -18, -31, -33, -35, -45, -51, -52, and -58 can now be considered as cervical carcinogenic agents (Muñoz 2000). Squamous carcinomas and adenocarcinomas are the most frequent cervical neoplasias, and may develop from intraepithelial lesions, easily detected in preventive cytological exams (Sherman et al. 1994).
Normalized Relation HPV infection is related to the incidence of cervical pre-neoplastic and neoplastic lesions
Antecedent: HPV, different types / Papillomavirus Humano,
Type fo relation: causes / T147 UMLS Semantic Network
Consequent: cervical pre-neoplastic and neoplastic lesions / Infecções Tumorais por Vírus, Neoplasias do Colo
Mapping to DECS: M (mapped)
DECS Subject Headings Papillomavirus Humano/classificação, Infecções Tumorais por vírus/epidemiologia, Neoplasias do Colo Uterino/virologia, Papillomavirus Humano/genética, Infecções Tumorais por Vírus/patologia Infecções Tumorais por Vírus/virologia, Neoplasias do Colo Uterino/diagnóstico Doenças do Colo Uterino/patologia, Doenças do Colo Uterino/virologia DNA Viral/genética, Esfregaço Vaginal, Reação em Cadeia da Polimerase Polimorfismo de Fragmento de Restrição, Genótipo, Fatores de Risco Prevalência
Citations: (Lorincz et al. 1992, IARC 1995), (Muñoz 2000), (Sherman et al. 1994).
EXPERIENCE
Results
Measure: prevalence
Context: Environment: Place: Distrito Federal, Brazil / Brasil/epidemiologia Time: Group: women / Humano, Feminino, Adulto, Meia-Idade
Methodology:
Conclusions
Observations:

Figure 1: Article Analysis Form

3 Results

We envisage an authoring/publishing software tool which will be available to the author during the process of Web publishing his/her article, and interactively will capture the articles knowledge, recording it in a standard program readable format. This knowledge can then be retrieved and processed by semantic retrieval tools. Validation tools or software agents could also compare the knowledge extracted from articles with that held in public ontologies like the UMLS and thus indicate inconsistencies, faults and even new discoveries. The overall authoring/publishing environment is discussed in Marcondes [9] and illustrated in Figure 1. The authoring/publishing software tool development and how to identify new discoveries using the model proposed are in our agenda and will be object of future research. The present research is conceived only with proposing, testing and validating a model to the knowledge extracted from the article's text by a future authoring/publishing tool to be developed.

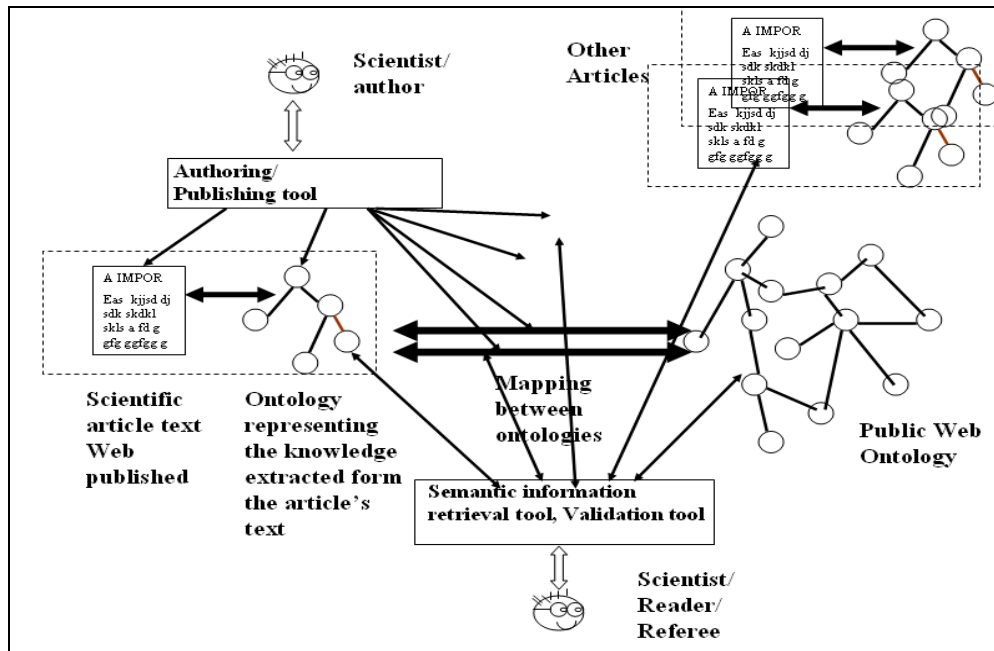


Figure 2: Author's editing/Web publishing environment

What are the methods to achieve the truth in Science? These questions date back to Greek Philosophy with Epistemology, Rhetoric, Dialectics and Sophistic. Aristotle proposed patterns of reasoning from which true statements could be achieved from previous statements. He invented the reasoning method called *deduction*, through which particular statements can be derived from general statements. These patterns were systematized by Medieval Scholastics.

A branch of this discussion with important contributions came at the Modern Age, with the establishment of the scientific method by Francis Bacon [10]. In opposition to Medieval Scholastics, Bacon emphasized the importance of observational experiments to achieve general laws in Science. His reasoning method of deriving general statements from a particular number of observational cases was called *induction*. Besides all criticisms to the bases of the scientific method induction reasoning is still a strong basis to experimental Science.

Pierce adds to deduction and induction the abduction method of reasoning. According to him abduction is essentially the creative process of generation new explanatory hypotheses from apparently unstructured observational data. Pierce also integrated abduction with deduction and induction, proposing a whole method to scientific inquiry: a new hypothesis is abductively generated; its consequences are deductively inferred and inductively tested.

Abduction is considered as the logic of discovery by many researchers as Hoffmann [11], Magnani [12] and Paavola [13]. Pierces' example of abductive reasoning is Kepler discovery that planet orbits are not circles, as believed Copernic, but ellipsis. Abduction has always been associated with new discoveries both by Pierce himself and by researchers working on his legacy. Induction and Deduction are always associated with hypotheses testing and their ratification or refusal, an incremental increase to knowledge stock.

An article's knowledge - or semantic elements - appears according to the reasoning procedure employed by the author. It is important to identify these semantic elements to the development of an ontology which will guide a future authoring/publishing software tool while interacting with the author during knowledge extracting from article's text as a by product of the writing/publishing activity.

The article analysis showed three patterns of reasoning procedures. According to the reasoning procedure employed scientific articles can be classified as *theoretical articles*, which employ abductive reasoning and *experimental articles* which employ inductive or deductive reasoning. The elements complaining the structure of knowledge contained in the text of the article differs depending on the type of reasoning procedure used by the author.

These elements are: the PROBLEM the article is trying to address, the HYPOTHESIS, where the author states a RELATION between phenomena, a possible empirical controlled EXPERIMENT with the aim of observing the phenomena described, specific of experimental articles, divided in RESULTS – tables, figures, numeric data, reporting the observations made -, MEASURE used, a specific CONTEXT where the empirical observations take place, subdivided in ENVIRONMENT – a hospital, a crèche, a high school -, a geographical PLACE where the empirical observations take place, TIME when the empirical observations occurs, a specific GROUP – pregnant women, early born babies, mice - in which the phenomena occurs, and CONCLUSION – a set of propositions made by the author as a result of his/her findings.

Although all these elements are important to reasoning procedure, the hypothesis is the element which has the potential to hold new knowledge. The hypothesis has the form of a RELATION formed by two or more ARGUMENTS linked by a TYPE_OF_RELATION. In every article analyzed concepts found in the ARGUMENTS were tentatively mapped to concepts taken from the UMLS verifying if these concepts correspond to DECS/MeSH subject heading extracted from the article's record in Medline or Lilacs databases.

Theoretical-abductive model of articles are based on synthesis of Gross [14] and Hutchins [15] proposals. *Theoretical-abductive* articles analysis different previous hypotheses, show their faults and limitations and propose a new hypothesis; the reasoning is as follows:

*a PROBLEM is identified, with the following aspects and data;
the previous authors/HYPOTHESES are not satisfactory to solve the PROBLEM due to the following criticism;
so, we propose this new HYPOTHESIS which we consider as a new pathway to solve the PROBLEM.*

Experimental-inductive articles propose a hypothesis and develop experiments to test and validate it; reasoning is as follows:

*a PROBLEM is identified, with the following aspects and data;
a possible solution to this PROBLEM can be based on the following new HYPOTHESIS;
we developed an EXPERIMENT to test this HYPOTHESIS and it comes at the following RESULTS.*

In experimental-inductive articles, a CONCLUSION is one of the following types: or it corroborates the hypothesis, or it refuses the hypothesis or it partially corroborates the hypothesis. However in some cases, the CONCLUSION is neither the former, it just reports intermediate, not conclusive results toward the hypotheses corroboration.

Experimental-deductive articles use hypothesis proposed by other researchers cited by the article's author and apply it to a slightly different context; reasoning is as follows:

*a PROBLEM is identified, with the following aspects and data;
in literature the previous authors/HYPOTHESIS are proposed;
we choose the following previous HYPOTHESIS;
we enlarge and re-contextualize this HYPOTHESIS; we developed a EXPERIMENT to test it in this new context;
the EXPERIMENT shows the following RESULTS in this new CONTEXT.*

Experimental articles also can compare various phenomena or hypotheses, as in a comparative study, a very usual type of article in Health Sciences. The different reasoning procedures can be formalized in an Ontology for Scientific Knowledge in Articles, as illustrated in Figure 2. This ontology has the following Classes and Properties:

Classes: THEORETICAL reasoning and
EXPERIMENTAL reasoning
Subclasses: INDUCTIVE reasoning and
DEDUCTIVE reasoning
Properties: PROBLEM
HYPOTHESIS (previous or new)
Sub-properties: ANTECEDENT
TYPE-OF-RELATION
CONSEQUENT

REFERENCES (just in previous HYPOTHESIS)

EXPERIMENT
 Sub-properties: RESULTS (quantitative data, tables, etc.)
 MEASURE
 CONTEXT
 Sub-properties: SPACE
 TIME
 GROUP

Two Classes of articles were identified: Theoretical and Experimental. Experimental articles in turn have two Subclasses, Inductives and Deductives. The Properties of articles are the following: Theoretical-abductive articles have a PROBLEM, one or more previous HYPOTHESIS, that are discussed, criticized and rejected as solutions to the PROBLEM posed. So, the author proposes a new HYPOTHESIS which may be a solution to the PROBLEM. Theoretical-abductive articles do not present experimental results.

Experimental articles in turn always present experimental results. Experimental-deductive articles have the following Properties: a PROBLEM, one or more previous HYPOTHESIS, by different authors, that are adopted to guide an experiment. Previous HYPOTHESIS are extended, restricted or inserted in a new CONTEXT. An experiment is developed bases in the previous HYPOTHESIS applied to the new CONTEXT and the results of the EXPERIMENT are reported.

Experimental-inductive proposes an original new HYPOTHESIS to address a PROBLEM, develop an experiment to test this HYPOTHESIS and the results of the EXPERIMENT are reported.

HYPOTHESES have an ANTECEDENT, a TYPE-OF-RELATION and a CONSEQUENT. HYPOTHESES hold the knowledge embedded in the article as it proposes a relation between phenomena.

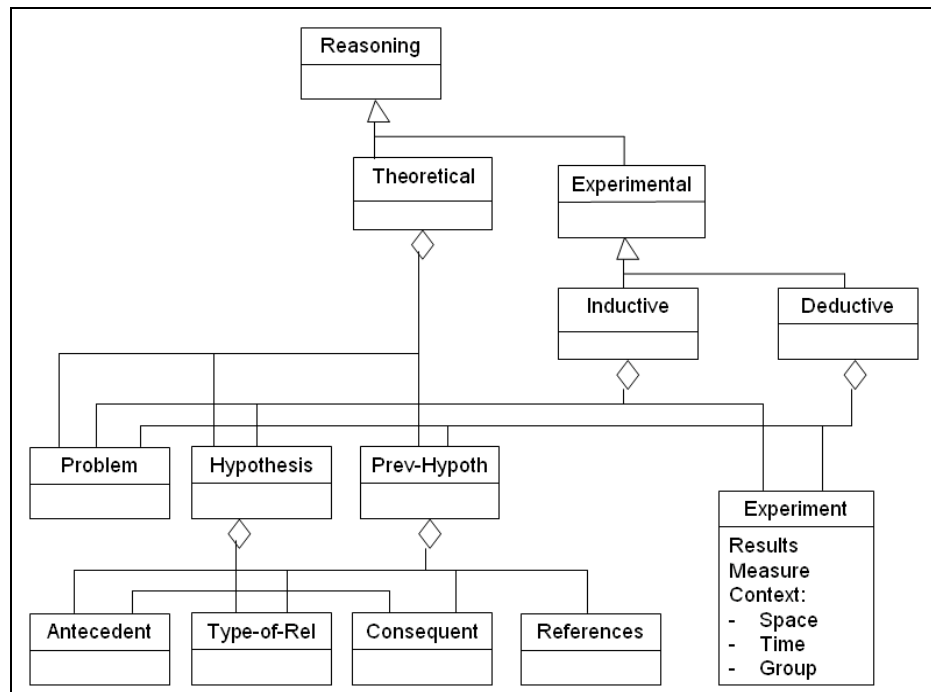


Figure 3: Class diagram of the Ontology for Scientific Knowledge in Articles

We plan to implement the Ontology for Scientific Knowledge in Articles in OWL [16]. The ontology will guide a future authoring/publishing tool in its interaction with an author to extract and record the knowledge embedded in the text of an article. Quantitative results of the analysis done on 53 articles are showed in Table 1. According to the classification proposed the majority of articles are experimental articles, 50 out of 53. Just 3 are theoretical-abductive articles.

Articles analyzed	Exp-inductives	Exp-deductives	Theor-abductives	TOTAL
MIOC	4	15	1	20
BJMBR	4	13	2	19
STEM CELLS	10	4	0	14
TOTAL	17	33	3	53

Table 1: Results of the articles analysis

In all articles the HYPOTHESIS was generally found in the Introduction section, in the Title or in the Abstract. Articles were considered Fully Mapped when concepts in both ARGUMENTs and the TYPE OF RELATION were fully mapped to one or more DECS/MeSH concepts that index the record in databases as Medline and Lilacs and there is a UMLS Semantic Network Relation corresponding to the TYPE OF RELATION. Articles were considered Partially Mapped when concepts in at least one of the ARGUMENTs or in the TYPE OF RELATION were fully mapped to one or more DECS/MeSH concepts and UMLS Semantic Network Relations. Articles were considered Not Mapped when any concept in neither the ARGUMENTs nor in the TYPE OF RELATION were fully mapped to DECS/MeSH concepts and UMLS Semantic Network Relations. The mapping of concepts to the DECS/MeSH is lower - which may be an indicative of new discoveries -, in a research area as stem cells in comparison to the two Brazilian journal. Table 2 shows these results.

Articles analyzed	MIOC	BJMBR	STEM CELLS
Total of articles	20	19	14
Fully mapped	11	4	0
Partially mapped	9	10	11
Not mapped	0 (0%)	5 (25%)	2 (7%)

Table 2: results of the mapping of concepts found in hypotheses to DECS/MeSH

4 Discussion

The majority of articles found are experimental, 50 out of 53. The experimental articles all fit in the IMRAD model, with definite textual parts while the theoretic-abductive articles not. This fact may indicate a pattern of research characterized as “normal Science” according to Kuhn’s [17] theory.

Although foreseen in the literature only three theoretical-abductive articles were found among the articles analyzed. As this is the type of article which reports expressive paradigm changes in a scientific area it is expected that they are not very usual. But their existence is certain. For example, Watson and Crick article proposing a model to the DNA molecule is a typical theoretical-abductive article. All three articles found do not fit into the IMRAD structure. They do not have sections such as *Material and Method* and *Results*. Some review articles and letters to the editor have some traces of theoretical-abductive articles and must be object of future research.

Stem Cells potentialities constitute a new paradigm in cell biology. “*A new era in stem cell biology began in 1998 with the derivation of cells from human blastocysts and fetal tissue with the unique ability of differentiating into cells of all tissues in the body, i.e., the cells are pluripoten.*” (<http://stemcells.nih.gov/>). Since then two problems face the researches in the area: how to maintain stem cells cultures indefinitely undifferentiated in specialized cell types as bone, skin, liver, etc., and how to start and control differentiation into specific cells types. In the Stem Cells articles group there is a predominance of experimental articles reporting culture or control methods, in all of which the TYPE OF RELATION was mapped to relation “method” (UMLS Semantic Network T183). All articles of this group seem to report incremental advances in knowledge. None theoretical-abductive article was found in this group.

Few articles are totally mapped to DECS/MeSH concepts and to UMLS Semantic Network Relations. The process of mapping the concepts found in the ARGUMENTs and in the TYPE OF RELATION of each HYPOTHESIS is just a by-product of the data generated by the analysis process, just an explorative pathway to generate data for future research. In the majority of cases concepts in the ARGUMENTs were too specific in comparison to DECS/MESH concepts used to index the record. On the other hand the majority of TYPE OF

RELATIONS identified was satisfactorily mapped to UMLS “relations”. This fact may be due to the difference in numbers: there are more than 730.000 concepts in UMLS and just 53 “relations”. Relations are more stable across the time and more generic in comparison to concepts in a scientific area. Another explanation to this fact is that there is always a delay to these concepts be incorporated in the UMLS, so it is in dead an indicative of new discoveries. Anyway, operational results enabling software agents to compare the knowledge extracted from the text of articles to the knowledge record in Web ontologies according to the model proposed deserves more research.

The analysis performed shows that the scientific reasoning elements, according to the type of reasoning employed, are structured, forming an ontology, in the sense used in knowledge engineering, as in Sowa [18]. This enables a software agent to perform *inferences* on this structure. Based on the example analysis presented in Figure 1 knowledge extracted from articles, marked up and recorded as described would enable the following queries by a semantic information retrieval system:

- *which other articles have hypotheses suggesting HPV as the cause of cervical neoplasias in women?*
- *which articles have hypotheses suggesting other causes to cervical neoplasias different from HPV in women?*
- *which articles have hypotheses suggesting HPV as the cause of cervical neoplasias in groups different from women?*
- *which articles have hypotheses suggesting HPV as the cause of other pathologies different from neoplasias?*
- *which articles have hypotheses suggesting HPV as the cause of cervical neoplasias in different contexts? (not in women from Federal District, Brazil).*

To publish scientific articles both as text and as machine readable knowledge bases seems to be a promising approach. It will enable the processing of this knowledge by software agents, thus improving critical inquiry, semantic querying and validation of scientific contributions to Science. Experimental Science, as Health Science, offer a solid basis to the development of the model, due to its formalism, derived from the use of the Scientific Method as an reasoning strategy in the text of scientific articles. The model outlined is a semantic model which aims to identify the semantic content of scientific reasoning. It is intended to be the basis to the development of a Web authoring/publishing tool. To reach this objective new research on computational techniques must be developed. We envisage an authoring/publishing tool that offers researchers/authors an interactive Web environment which, through a rich dialogue and using text extraction techniques, interactively identify and extract relevant contents of the article been written/published. This content will then be represented in machine-understandable format as an ontology, using OWL. Scientific articles so published throughout the Web can then be interlinked and linked to the increase number of Web ontologies, forming a rich knowledge network, thus enabling software agents to help scientist identify and validate new discoveries to Science. As the model proposed became more robust, there are plans to test it in other empirical science areas and even in areas as social sciences.

5 Conclusion

In all articles analyzed a relation expressing the mainly findings reported in the article was identified. This seems to indicate that scientific knowledge as expressed in the text of scientific articles can be represented as relations between phenomena. The amount of scientific knowledge now available throughout the Internet is so vast that it can only be processed with the aid of computer power. Here is proposed a standard representation to this knowledge feasible to be processed by software agents. This is essential if the intention is to use software agents to large scale processing of this knowledge in tasks as knowledge validation, semantic retrieval, identification and evaluation of discoveries.

Articles analyzed are very few and restricted to a single scientific area. If we are going to establish a new paradigm in electronic scientific publishing in which articles are published not only to human reading but also to be processed by software agents, this deserves more research. The model proposed is just a starting point to be discussed and enhanced by the scientific community.

Indexing language, as different Thesaurus largely used in information systems, select a set of concepts to describe a document. All knowledge organization effort is oriented toward the organization of systems of concepts. Generally all these concepts play an identical role when representing and retrieving a document. Although relations play a key role in scientific knowledge conventional indexing languages play no attention to them. Indexing language to no express the relations held between the subject headings indexing a document.

Indexing language must include relations between subject headings. There is also a need of the development of a taxonomy of relations used in Science to help indexing/retrieval scientific articles.

The model proposed also points to the standardization of a SkML - Scientific Knowledge Mark up Language - encompassing the semantic content of scientific articles Web published. This article highlights the benefits of a semantically richer format to represent the knowledge in scientific articles. With the aid of adequate software tools, this knowledge can be extracted as a by-product of authoring/publishing an article by the author. This opens an all new perspective in scientific knowledge acquisition, storage, processing and sharing.

Notes and References

- [1] BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. *Scientific American*, May, 2001. Available from Internet: <<http://www.scian.com/2001/0501issue/0501berners-lee.html>>.
- [2] *OWL Web Ontology Language Guide*. Available form Internet: <<http://www.w3c.org/TR/2004/REC-owl-guide-20040210/>>.
- [3] JACOB, E. K. Ontologies and the Semantic Web. *Bulletin of the American Society for Information Science and Technology*, Abril/May, 2003.
- [4] BROOKES, B. The foundations of Information Science. Part I. Philosophical aspects. *Journal of Information Science*, vol. 12, 1980, pp. 125-133.
- [5] SHETH, A; ARPINAR, I. B.; KASHYAP, V. Relationships at the heart of semantic web: modeling, discovering and exploiting complex semantic relationships. In: NIKRAVESH, M. Et al. *Enhancing the power of the internet studies in fuzziness and soft computing*. Springer-Verlag, 2002. Available from Internet: <<http://cgsb2.nlm.nih.gov/~kashyap/publications/relations.pdf>>.
- [6] MILLER, D. L. Explanation Versus Description. *Philosophical Review*, vol.. 56, no. 3, May, 1947. pp. 306-312. doi:10.2307/2181936.
- [7] BONGSO, A; RICHARDS, M. History and perspective of stem cell research. *Best Practice & Research Clinical Obstetrics & Gynaecology*, vol. 18, no. 6, 2004. pp. 827-842.
- [8] FRIEL, R; SAR, S; MEE, P. Embryonic stem cells: Under standing their history, cell biology and signalling. *Advanced Drug Delivery Reviews*, vol.57, no.13, 2005. pp. 1894-1903.
- [9] MARCONDES, C. H. From scientific communication to public knowledge: the scientific article Web published as a knowledge base. In: Proc. 9th ICCC EIPub - International Conference on Electronic Publishing, Leuven, Belgium, 2005, 9, p.119-27. Available from Internet: <<http://elpub.scix.net>> .
- [10] BACON, F. *Novum Organum*. São Paulo : Abril Cultural, 1973.
- [11] HOFFMANN, M. Is there a “Logic” of Abduction? In: Proc. 6th. Congress of the IASS– AIS International Association for Semiotics Studies, Guadalajara, Mexico, 1997. Available from Internet: <<http://www.unibielefeld.de/idm/personen/mhoffman/papers/abduction-logic.html>>.
- [12] MAGNANI, L. *Abduction, Reason, and Science: Processes of Discovery and Explanation*. New York : Kluwer Academic; Plenun Publishers, 2001.
- [13] PAAVOLA, S. Abduction as a Logic and Methodology of Discovery: the Importance of Strategies. *Foundations of Science*, Vol.9, No. 3. November, 2004. p. 267-283. doi: 10.1023/B:FODA.0000042843.48932.25.
- [14] GROSS, A. G. *The Rhetoric of Science*. Cambridge, Massachusetts; London: Harvard University Press, 1990.
- [15] HUTCHINS, J. On the structure of scientific texts. In: Proc. 5th. UEA Papers in Linguistics, Norwich.. Norwich, UK: University of East Anglia, 1977. p.18-39.(Conference Proceedings). Available from Internet: <<http://ourworld.compuserve.com/homepages/wjhutchins/UEAP/L-1977.pdf>>.
- [16] OWL- Ontology Web Language, a W3C standard language to represent ontologies. Available form Internet: <<http://www.w3.org/2004/OWL/>>
- [17] KUHN, T. The structure of scientific revolutions. In: *Foundations of the unity of Science*, vol. 2. Chicago : the University of Chicago Press , 1970.
- [18] SOWA, J. *Knowledge representation: logical, philosophical and computational foundations*. Pacific Grove, CA : Brooks/Cole, 2000.

Automatic Content Syndication in Information Science: A Brazilian Experience in the Creation of RSS Feeds to e-journals

Robson Lopes de Almeida

Department of Information Science, Universidade de Brasília
Campus Universitário Darcy Ribeiro, Brasília, DF - Brasil
e-mail: rlalmeida@unb.br

Abstract

This paper reports the partial results of an exploratory study which intends to develop a methodology for a Web feed-based aggregation content service to electronic journals in Information Science. Ten scientific e-journals were chosen as sample to demonstrate the potential of the Web syndication technology. These e-journals are supported by the Brazilian Electronic Journal Publishing System (SEER), adapted from the Open Journal Systems (OJS), an open source software for the management of peer-review journals, developed by the Public Knowledge Project (PKP). In this context, the present study describes the concepts of aggregation and content syndication in Web environments. Moreover, it discusses the possibilities, advantages and eventual barriers to the implementation of RSS applications concerned with electronic journals in Information Science, specially the ones supported by the OJS Systems.

Keywords: metadata aggregation; content syndication; electronic journal; RSS; web syndication

1 Introduction

With the advent of the so-called Technologies of Information and Communication (TICs), particularly the Internet, which stands out as its main exponent, a significant raise in the amount of information can be observed, and we are exposed to them in our daily life. These pieces of information, when they are not useless, end up leading to a real overload, which is harmful to the absorption of the contents that really interest us. This also causes a sense of discomfort to the majority of people.

In the early 90's, with the advent of World Wide Web – the graphic and multimedia part of the Internet – information started to be even more easily disseminated. Because of that, new contents have been added to the web disorderly. Nowadays, the raising amount of Internet-generated information is not an object for information scientists only, but also for researchers from several different areas of study. They have also been attentive to the effects caused by every kind of information overload.

It is true that the simplicity of the existing web publishing tools has been useful not only to those who produce but also to the ones who acquire information, offering dynamic and low-cost mechanisms in order to communicate new ideas. The expansion of the blogosphere phenomenon [1] is a proof of that. On the other hand, the fast dissemination of digital information has demanded close attention in relation to the quality of the content which is about to be published, and also discernment concerning its use; otherwise, we run the risk of having our precious time drastically wasted.

Although this problem can be considered a natural consequence of our “Information Society”, historically, in the 50's, the first systems able to select relevant information to a certain user, considering his/her profile of interest, appeared. This concept is called Selective Dissemination of Information (SDI), created by Hans Peter Luhn, from IBM Corporation, in order to improve the alert services offered by libraries, documentation centers and specialized centers of documental information.

From this perspective, and considering the current Web chaos, we intend to deal with the concepts of “content syndication” and “content aggregation”, which has become popular from an Internet technology standard that allows users to receive updates to Web-based content of interest, simply called RSS.

The use of RSS began about ten years ago, meeting Internet user's information needs, keeping people up to date with new and revised information without making them feel lost facing Web content overload.

The present study intends to comment on the possibilities of how this resource can be used by electronic journals, especially the ones which already count on resources which make the implementation of RSS feeds [2] easier. In addition, it discusses the advantages to content publishers and to readers/users, and also the possible barriers to this implementation. Finally, this paper describes the progress of the first RSS feeds created by the author from a sample of Information Science e-journals supported by the Brazilian Electronic Journal Publishing System (SEER/OJS).

2 What is RSS?

Basically, the RSS format can be understood as a dialect or part of vocabulary from the XML family [3], meant for automatic capturing and website content distribution, used to publish frequently updated digital content, such as blogs, news feeds or podcasts. RSS allows Internet users to subscribe to websites that provide RSS feeds; these are typically sites that change or add content regularly. However, its applicability is not strict to these domains, once everything which is possible to be described by means of <tags> can be integrated by RSS.

The popularity of RSS technology is due to the agility which this format provides to the reading of new contents, since it does not need any access to websites where the information was originally published. In fact, the main feature of the RSS pattern is to allow a website's frequent reader to track updates on the new issues of an e-journal, for example. Moreover, another advantage to the user is the facility of finding, in one single place, the current summaries of the main publications in the particular area. These characteristics called our attention to a deeper investigation concerning its use in Brazilian scientific journals on Information Science.

The most practical way of benefiting from this technology is having a news aggregator software, a type of application that retrieves syndicated Web content that is supplied in the form of Web feed. Such softwares are generally free, easy to install, and the great majority resembles an e-mail reader. Figure 1 presents a typical screen with one of these applications. It is possible to see in the left column all the chosen and added feeds, which can be read in the right column. On top of the right column, there is a list of headlines; while at the bottom, we can observe part or the complete post text. When double clicked in the headline title, the full content will appear in the inferior window, exactly as it was originally published in the Web. These headlines may be stored or deleted. There is also the possibility of filtering them by subject or date.

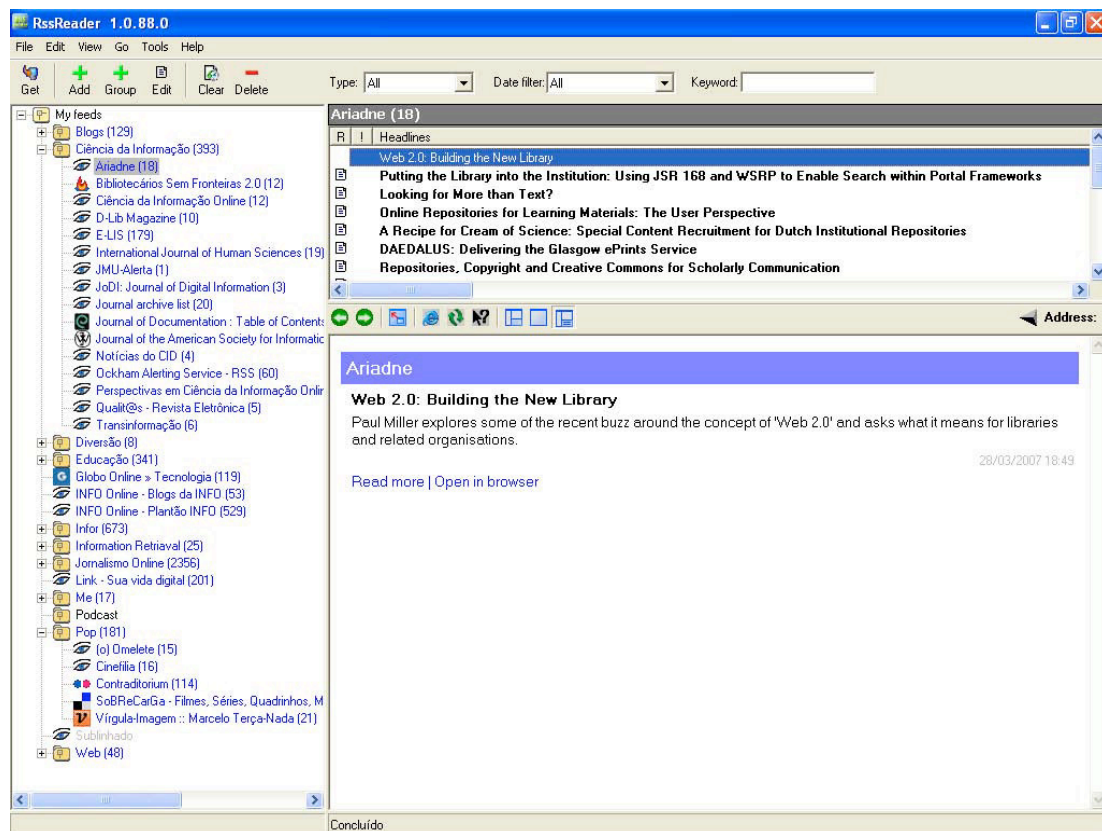


Figure 1: screenshot of RSS Reader, an stand-alone client aggregator

Another way of getting the same function without having to install or configuring any kind of software is by using Web-based aggregator – a remote-hosted service offered by a third party that allows you to subscribe to and read feeds. To use this kind of free service, the user creates an account and then logs in to perform all feed-related activities, like reading a set of news sources in several XML-based formats. The user can find the news bits and display them in reverse-chronological order on a single page.

By means of these RSS “readers”, it is possible to make a kind of subscription of the contents of different sources by themes, and quickly examine the title of the news articles and the summaries of a new issue in a condensed way. When the user finds some information which arouses his/her interest, the only thing he/she needs to do is to click on the title of the article to read its full content.

The name "RSS" is an umbrella term for a format that spans at least two different (but parallel) formats. Then, RSS could stand for “Rich Site Summary”, “RDF Site Summary” or “Really Simple Syndication”, depending on which version you are using. Regardless of what they are called or the version number, feeds are all XML-based languages. That is to say they are written to conform to the XML rules. For those who are familiar with HTML (Hypertext Markup Language) code, the structure of feeds will look familiar as we can see in figure 2. However, differently from HTML, which is limited to provide a universal format to represent information, without making reference concerning the structure and semantics of the data, RSS, as an XML-based language, is able to represent information about resources in the Web. It is intended to represent metadata about Web resources, such as the title, author, date of a webpage, copyright and licensing information about a Web document.

```
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0">
  <channel>
    <title>Example e-Journal </title>
    <description>Short description of the journal </description>
    <link>http://www.cid.unb.br/</link>
    <language>pt-br</language>
    <lastBuildDate>Wed, 27 Mar 2007 14:02:11 -0300</lastBuildDate>
    <managingEditor>rlalmeida@unb.br</managingEditor>
    <pubDate>Wed, 17 Jan 2007 15:47:50 -0200</pubDate>
  <item>
    <title>Title of Article 1</title>
    <link>http://www.cid.unb.br/ e-journal/article_1</link>
    <description>
      Abstract of article 1
    </description>
  </item>
  <item>
    <title> Title of Article 1</title>
    <link>http://www.cid.unb.br/e-journal/article_2</link>
    <description>
      Abstract of article 2
    </description>
  </item>
</channel>
</rss>
```

Figure 2: Example of a RSS feed (2.0 version)

The use of specific tags, such as <title>, <link> and <description> allows us to treat each information unit (the title, the place where the information can be found and the summary) as a distinct object. This enables us to structure information so that it is interpreted and treated by means of computer resources, such as scripts or special softwares. This procedure turns data into qualified objects, such as attributes. This way, there is a possibility of automatic reusing of the information, making easier to share it with users and with other information systems (interoperability), and also organize it into database and do automatic research.

This way, it is simple to create a website reply to the content of another one which has an RSS feed. To do this, it is necessary to insert, in a page of the intended site, a script which points to an original website XML archive. These codes can be easily found in the Internet itself.

3 A Brief History of Web Syndication

In the late 90's, some experiences began, intending to provide required information to the Internet user in the Web context, such as Crayon [4]. Another initiative which was successful at that time was the resource called "Active Channel", developed by Microsoft to its browser Internet Explorer 4.0.

These two projects had in common the mission of joining in one place (aggregating) varied and scattered contents in the Web, by means of a technology called "push", since the idea was to send customized information to their users, instead of waiting for them to visit the websites to "pull" the desired contents. Specialized companies, such as the North American PointCast were pioneers in this type of syndication format, but the problem is that they were not interoperating; in fact they worked independently. Moreover, the softwares were too complex to users who were still getting familiar with the recently-created Web environment. At that time, Ramanathan Guha and other researchers from Apple Computer developed the Meta Content Framework (MCF), a specification of a format for structuring metadata about websites and other data.

At the beginning of March, 1999, when the research project was discontinued, Guha left Apple for Netscape Corporation, where he adapted MCF to use XML. He created the first version of the Resource Description Framework (RDF), which turned to be the basis for the creation of the first Web syndication format, called RSS (RDF Site Summary), 0.90 version.


In July, 1999, Dan Libby, also from Netscape, improved the format and produced a prototype tentatively named RSS 0.91 (RSS standing for Rich Site Summary) to be used in the "My Netscape" portal, as a pattern to the construction of headlines publishing systems in webpages, working as a summary of the content of a site with its respective links to the original information sources.

In the following year, as the group of developers from Netscape decided to leave the portal business, a lower-sized company called UserLand decided to keep on developing the RSS in order to apply it to their virtual electronic diaries tools, which, later, would become popular as weblogs or just blogs.

In August, 2000, another group of independent programmers (RSS – DEV Working Group), led by Rael Dornfest from computer book Publisher O'Reilly and Associates, proposed a new specification named RSS 1.0, according to the RDF metadata format. This one joined most of the preceding versions of RSS.

However, the group from UserLand, led by Dave Winer, continued their work, developing other versions of RSS, such as the 0.92 and 0.93, until they reached the version 2.0, in September, 2002. The abbreviation RSS had, then, another meaning: Really Simple Syndication, once its focus was on the simplicity in content syndication. Nowadays, this version is widely used by thousands of websites, including blogs and podcasts.

As Winer left UserLand, Berkman Center to Internet and Society, from Harvard University, was, then, in charge of the development of RSS 2.0, making this technology available to public domain, under a Creative Commons license, in 2003. In this same year, a group of leading service providers, tool vendors and independent developers, worried about this problem of the development of the RSS specifications, decided to create a new format to content distribution: Atom [5] (originally called Echo). Its aim was to be 100% neutral, open and easily implemented by any developer. Atom is also based on the XML format, but its development is considered more sophisticated. According to specialists, it consists of a proposal of unification of RSS 1.0 with RSS 2.0, and it might be its natural substitute, since it counts on the support of great corporations, such as Google, which has adopted this format to its blog service. The final draft of Atom 1.0 syndication format was published in July, 2005, and was accepted by the IETF (Internet Engineering Task Force) as a "proposed standard" in August of 2005. The work, then, continued on the further development of the publishing protocol and various extensions to the syndication format.

In December, 2005, the Microsoft Internet Explorer team and Outlook team announced that they would be adopting the feed icon  first used in the Mozilla Firefox browser, effectively making the orange square with white radio waves the industry standard for both RSS and related formats such as Atom.

In February, 2006, Opera Software announced they would also add the orange square to their Opera 9 release. Also in 2006, Microsoft decided to incorporate the RSS 2.0 extensions in its operational system Windows Vista, while Google announced the launch of a new content syndicated reader tool – the Google Reader – with support to RSS. Seven years later, the technical developments related to Web syndication seemed to be just beginning, with companies investing in new applications. But what stands out in the moment is the fact that the content

providers and readers have found in Web syndication technology a fast and practical way of distributing and receiving updated contents through the Web [6].

4 Methodology

The sample chosen to carry out this research was the collection of Brazilian electronic journals in the area of Information Science which use the Brazilian Electronic Journal Publishing System (SEER), a tool applied to the administration of the editorial process of electronic journals, adapted from Open Journal System (OJS) to Portuguese language by the Brazilian Institute of Information in Science and Technology (IBICT/MCT), in 2004.

The preference for journals based on OJS is due to the fact that this is a consolidated system for publication and managing of peer-reviewed publishing. In March, 2007, over 900 titles were using OJS [7], including the main titles in Brazil, thanks to the effort of IBICT in offering specialized training to the editors. Another important factor determined the choice for these journals based on OJS: the system already has an RSS/Atom plug-in that produces Web feeds from articles that have been published in journals since the 1.x version. However, this feature is already included in recent releases of OJS 2.x.

Ten of these journals, having met the required criteria, were the object of several analyses, during the months of January and February, 2007. These analyses intended to investigate the main characteristics of the journals, according to the way of divulging its content, and, mainly, if they made Web feeds available to their clients by means of incorporating the RSS format in the publishing OJS tool, or even if it had at least an alert service, through electronic mail to notify the updating of the current edition.

Once the journals selected in this study did not present any kind of feeds, they were hand-created by means of an authorship tool called FeedForAll (<http://www.feedforall.com>) and were, later, hosted by the author's webserver. The initial idea was to create a basis with the contents of those feeds, so that they could easily syndicate.

Although none of the sample journals presented feeds to their users, we could identify that there is at least one national journal that uses the OJS feed plug-in to generate RSS/Atom feeds automatically. It is called Qu@litas, an electronic journal edited by The Center of Applied and Social Sciences of the University of Paraiba.

After a testing period, these journal's feeds were included into a content aggregator application specially created, using Netvibes service (<http://www.netvibes.com>), which provides a personalized page in which the author can manage several modules created from RSS/Atom feeds. Creating this "prototype" was a way of demonstrating the potential of a content aggregator application, starting from the simple process of creating Web feeds.

5 Results and Expectations

The RSS feeds created by the author to the 10 selected journals had as basis the last edition available in the website. The task of creating every feed lasted about 15 minutes, once it counted on the help of an RSS feed creation tool (FeedForAll), which made the creation of documents easier, without being necessary to write down codes which are particular of the RSS format.

Once the software is installed, it was quite simple to create RSS feeds. First, it is necessary to fill in the channel's basic information: title, link (the URL of the webpage that corresponds to the channel) and description (a brief description of the content of the feed). Once the feed is created, it is necessary to add items. This task corresponds to the addition of metadata related to the articles. The indispensable information of each item are the same for creating a feed: the title of the article, the link (location of the page where the article can be found) and description (a summary of the article), as shown in Figure 3, which illustrates the filling of the required fields from "Items".

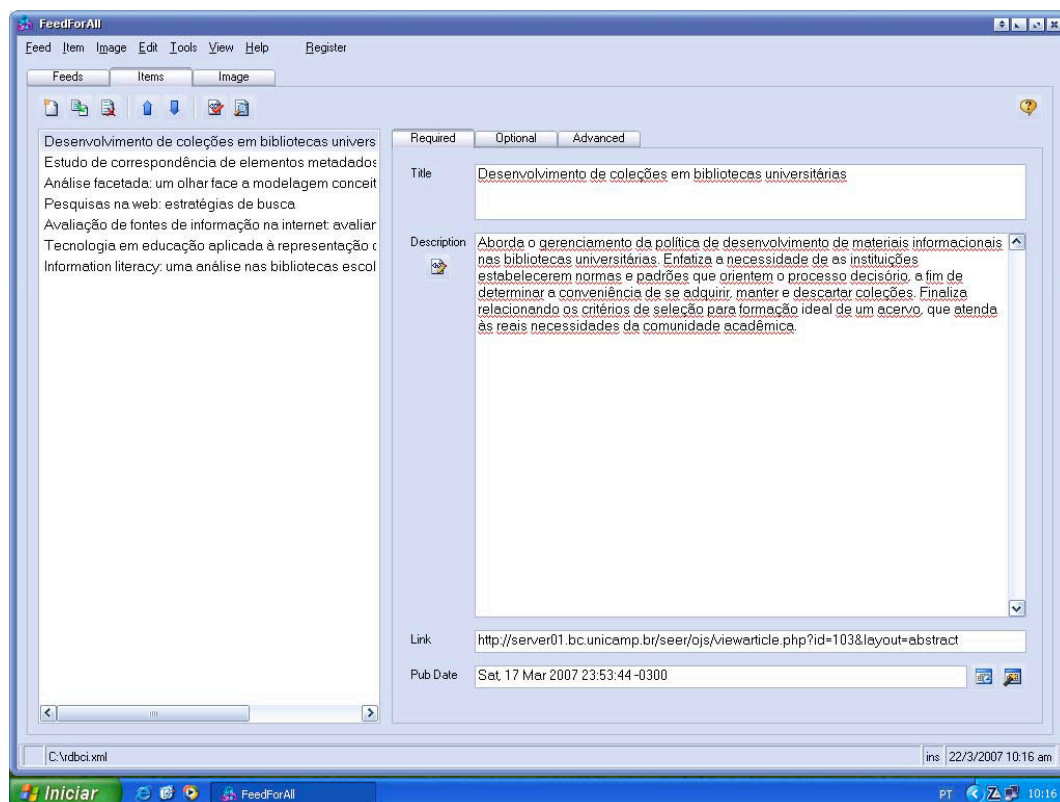


Figure 3: Screenshot of RSSForAll tool

In a second moment, we created a directory list containing the RSS feeds of all the periodicals being worked with, summarized in Table 1.

Journal	Created RSS feed
Arquivística.net	http://www.rlalmeida.correiovip.com.br/arqnet/arqnet.xml
Ciência da Informação	http://www.rlalmeida.correiovip.com.br/cionline/ciinfo.xml
Em Questão	http://www.rlalmeida.correiovip.com.br/emquestao/emquestao.xml
Informação e Sociedade	http://www.rlalmeida.correiovip.com.br/ies/ies.xml
Informação e Informação	http://www.rlalmeida.correiovip.com.br/iei/iei.xml
Perspectivas em Ciência da Informação	http://www.rlalmeida.correiovip.com.br/pci/pci.xml
Pesquisa Brasileira em Ciência da Informação e Biblioteconomia	http://www.rlalmeida.correiovip.com.br/pcbci/pcbci.xml
Revista ACB	http://www.rlalmeida.correiovip.com.br/acb/acb.xml
RDBD	http://www.rlalmeida.correiovip.com.br/rdbci/rdbci.xml
Transinformação	http://www.rlalmeida.correiovip.com.br/transinfo/transinfo.xml

Table 1: Information Science Journals under SEER/OJS and its respective RSS feeds

As a final result of this study, we propose a model of aggregation service on the specific content of Information Science e-journals based on Netvibes, a free service which uses the Ajax technology (Asynchronous Javascript And XML), in order to make the browser more interactive with the user, allowing him/her to create and manage models whose content come from Web feeds. It consists of an online service which was developed by means of XML and JavaScript. Once the feed(s) is(are) added, the application harvests the specific content and brings about, as a result, the titles and summaries of the updated articles. If the user wishes to access the whole content of any article, he/she will be sent to the correspondent page in the periodical itself, accessing the source or the document directly. The screen with all the aggregated periodicals can be seen in Figure 4.

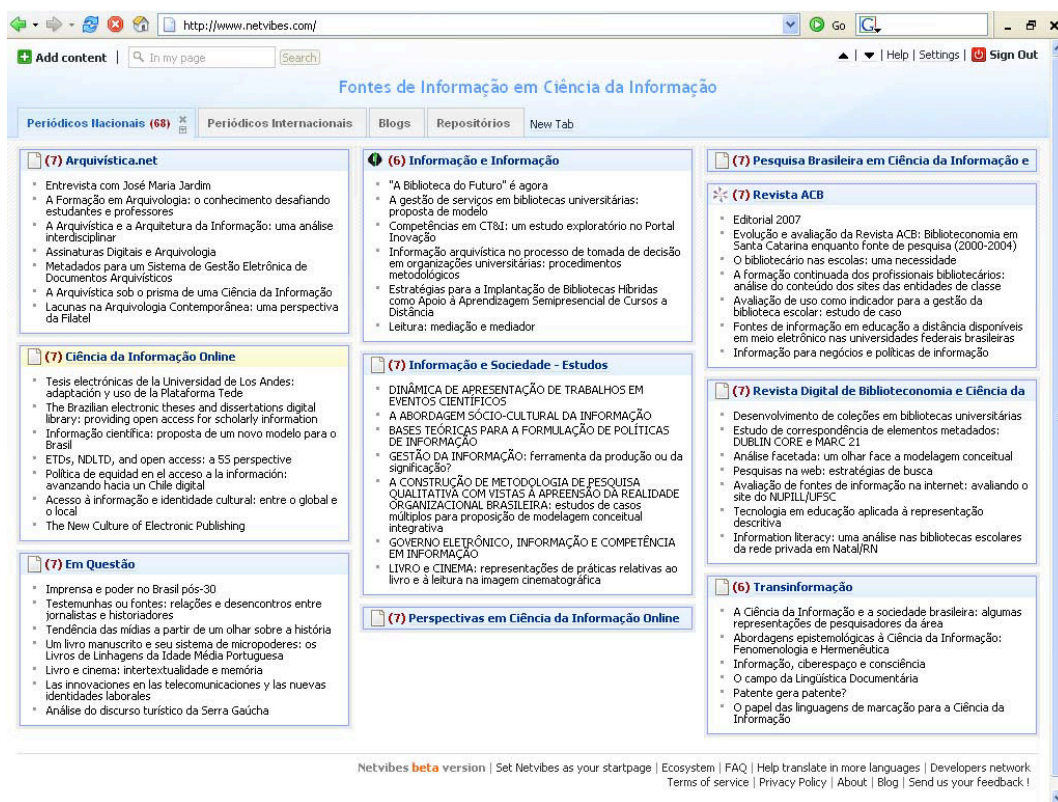


Figure 4: Screenshot of Netvibes application joining the 10 Brazilian Information Science Journals

The model for this service shows one of the main applications for Web feeds nowadays. In this case, we present the dynamics of a mechanism which is able to gather, in one single Web page, the indication of the Brazilian journals articles in Information Science (with their respective summary), by the use of Web Syndication. Since then, this service could be advertised to the community of potential users, mainly researchers, professors and people engaged in post-graduation courses on Information Science. Through these feeds provided by the directory, these users will be able to locate easily or, if they want, to subscribe to receive the updates of the publications they wish by means of the aggregation service they prefer.

Once all this content is gathered, it is possible to search every journal of this collection simultaneously. This way, if the user wishes to research the word “ontology”, for instance, he/she will have as a result all the articles which contain the word “ontology” in its title or in the summary, no matter which journal that may be.

6 Discussion

6.1 Advantages

According to our perception, the advantages to the reader are great. We will be able to count on a powerful tool with which we will be able to keep ourselves up-to-date in relation to several sources of information and, at the same time, make simultaneous researches on relevant content, which enables us to refine this search, raising the relevance of the recovered terms.

Saving time reading practically personalized information is a great advantage of using these kind of services which are able to join, in one single environment, a variety of contents produced by several different sources, with no need to access each site individually. Another characteristic of this system is that when a certain topic is selected by the user, the RSS technology offers the possibility of showing the full content of the document, with direct access to the source. That is, there is no copy of information or inappropriate seizure. The publishers, however, will aggregate value to the content of their publication and, consequently, will gain visibility to their publication, once RSS feed allows their users to read their content without going out of their way to visit. While this may seem a flaw at a first glance, it actually improves the visibility by making it easier for users to see it the way they want to. Because there are so many sources on the Web, most viewers won't come to the same site

every day. By providing a feed, publishers are in front of them constantly, improving the chances for them to click through to an article that catches their attention.

Nowadays, most journals which use OJS publishing system have a notifying service which offers the user the possibility of enrolling to receive, by e-mail, a notice with the summary of the new editions as they are published. If the user wishes to follow publications in a certain area, he/she will have to repeat this procedure for each journal. This means that he/she will receive several different notifications for each updating. Through a Web syndication service, the user does not need to enroll to keep updated, with the advantage of not having his/her mail box full.

6.2 Barriers

The main barrier for the implementation of this type of service seems to be the publishers' and users' lack of knowledge about the Web syndication technology. In Brazil, only the great newspapers which circulate around the country and some other specialized websites make feeds available to their users. Even in the academic context, with the exception of the courses on Computer Science, there is still ignorance of terms such as "feeds", "syndication" or "aggregator". In other countries, the adoption of RSS in information services is more common. However, we notice some resistance from the scientific editors. Even the group responsible for the development of the OJS admits that Atom/RSS feed plug-in is not a very well-known feature for their users, and few OJS e-journals make Web feeds available in their editions.

Differently from reading an e-mail, for instance, the current method of subscribing to a feed – copying the URL from the link (normally with .xml or .rss extension), and pasting it into a reader application – is not obvious to the new user. When doing this, the user normally sees XML codes on the screen. These are, at first, incomprehensible. It is normal for lay users to be confused with such information and, instead of subscribing the channel, they end up not doing that because they think they have committed some kind of mistake.

In the case of the feeds generated automatically through OJS plug-in, the subscription process is even more difficult. Once the desired format is chosen (Atom or RSS 1.0/2.0) with a click on the respective icon on the main page, the browser will ask the user to indicate an application to open an unknown document format. In order to avoid this discomfort, it is necessary to click the right button of the mouse on the icon which represents the chosen format, and select the option "Copy Link Location" (Figure 5) to, later, paste it into the reader application. This important piece of information, however, is not available in the journals analyzed. These are pretty user unfriendly and, probably, will be a barrier to widespread the adoption of RSS by 'non-techies'.

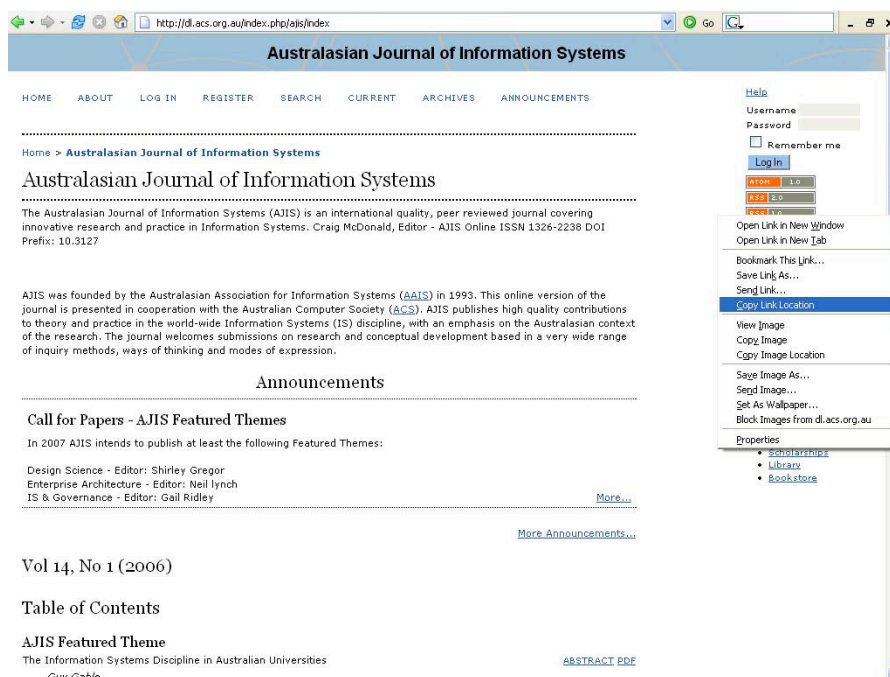


Figure 5: Screenshot of an OJS Journal which makes Web feeds available. The procedure for subscribing the channel is not informed in the website.

6.3 In the Future

In a near future, the Web Syndication technology might be widely known and broadly used by the ones who publish and the ones who acquire information in the Web. The availability of Web feeds in the periodical publications, no matter its nature may be, will be so natural that the user will not find it strange to meet indicative icons that a certain website or blog has an RSS channel. The process of subscription of a feed might be an extremely simple task and also transparent to the user. It will not be necessary to do copy&paste from the link which corresponds to the XML file, once the browsers will have efficient mechanisms to read feeds [8].

Independently of its format and version (RSS, Atom or even another one which may be created), the reader will learn, little by little, how to deal with a “power” that is hard to be imagined up to now: personalize the content of the information he/she wants to acquire, as well as produce new contents, through the use of resources offered by syndication technology, without the intervention of intermediates.

This way, when reading an article about “Semantic Web”, for instance, the user may, if he/she wants to, subscribe to the Web feeds about the same theme and the ones which are syndicated from other articles, data base, repositories, and so on. Everything turns out to be interoperating, thanks to the compatibility provided by the applications based on the XML language and its derivatives. From the user’s point of view, there will not have distinction among journals A, B or C in a certain area anymore. To him/her, it is as if there were one single source of information, easily available in his/her “personal digital library”.

In the next five years, there may be an explosion of new Information Retrieval Systems (IRS), with Web feed resources. This way, once the results have been recovered from a research in a database, users have the option of subscribing the feed related to that expression for search, and, so, keep themselves updated in relation to that specific matter. One of the existing services which follows this model is the one offered by the Agência Brasil (<http://www.agenciabrasil.gov.br>), the Brazilian government news agency which allows the user to create new channels based on their searches, more than offering RSS feeds for more than 120 different subjects.

The scope of this type of service is great and it is a good example that its potential is being tested by Ockham Alerting Service (<http://alert.ockham.org>), a current awareness service based on the National Science Foundation Digital Library content. According to their website, it demonstrates a standard-based method for collecting content, providing access to it and disseminating it on a regular basis in the form of an alerting service. The method includes: a) identifying OAI repositories with content of interest; b) using OAI to harvest content and store it in a central pile; c) indexing the content of the central pile; d) providing an SRU interface [9] to the index; e) allowing users to save the SRU URL's as "profiles" (RSS feeds); f) allowing users to have the profiles executed on a regular basis; g) making the results of searches available as HTML, e-mail, RSS, etc.

7 Conclusions

Despite the barriers identified in this study, we believe that the Web syndication technologies are viable to the integration of any Web-based information system, from search engines to publication systems, such as the case of OJS, presented in this study.

The reach of the Web syndication resources goes beyond the management of Internet content. It can also be a useful way of marketing, for instance, or even serve to notify users of the status of projects, monitor web statistics or otherwise replace the "notifications" that are now sent out as e-mail alerts.

Concerning the electronic journals, this type of technology seems to be welcome, once the simple inclusion of an RSS feed may aggregate a great value to the publication, not only as an intelligent way of divulging, but mainly because of the possibility of integrating with other contents, thanks to the interoperability which exists among the formats compatible with XML language.

This way, it seems that the new products and information services will have even more relationship with RSS/Atom feeds. The OJS feed plug-in is a proof of that. Before, it was available as an external archive to the OJS 1.x, and nowadays it includes recent releases of OJS 2.x.

Finally, after this investigation about Web syndication and the possibility of its use as resource for electronic journals, it is possible to summarize our conclusions based on the following topics:

1. The group of resources based on the Web syndication standards constitutes a technological innovation in the field of new reference services for information units, as well as for the development of potentialities of electronic journals;
2. It can be ascertained that the RSS is a meaningful tool for warning and automation of the content in the Web;
3. The publishers' commitment is essential for the dissemination of new products that use the technology of content syndication;
4. This technology is easily applied to the systems of information backup and of selective distribution;
5. New studies are necessary to widen the discussion about this issue, through new approaches and applications;
6. This technology is based on the information sharing paradigm. Therefore, it contributes to the generation of new knowledge.

Notes and References

- [1] According to Wikipedia, Blogosphere is the collective term encompassing all blogs as a community or social network.
- [2] A "RSS feed" is a XML-based document that usually ends in .xml or .rss and is a slimmed-down version of a website created to be easily syndicated. It contains a list of items or entries of content metadata. News websites and blogs are common sources for RSS feeds, but feeds are also used to deliver structured information, such as articles from periodicals.
- [3] The eXtensible Markup Language (XML) is a W3C-recommended general-purpose markup language for creating specion-purpose languages, capable of describing many different kinds of data.
- [4] Crayon (<http://crayon.net>) is the abbreviation of Create Your Own Newspaper, a personalized information service that offers users the possibility of creating its own journal with links for more than 100 sources of news available in the Web.
- [5] "What is Atom?" (<http://www.atomenabled.org>)
- [6] For a full discussion of the history of web syndication, see Wikipedia. History of web syndication technology: http://en.wikipedia.org/wiki/History_of_web_syndication_technology
- [7] A selected list of OJS journals is available on the PKP website: <http://pkp.sfu.ca/ojs-journals>
- [8] Nowadays, the Firefox browser identifies if a webpage uses RSS showing the feed icon in the browser's status bar. However, it is not an RSS complete client, being necessary the installment of extensions to make its support even more powerful.
- [9] SRU (Search/Retrieve via URL) is a standard search protocol for Internet search queries, utilizing CQL (Common Query Language), a standard query syntax for representing queries. Standards for SRU are promulgated by the United States Library of Congress.

Changing Content Industry Structures: The Case of Digital Newspapers on ePaper Mobile Devices

Leo Van Audenhove; Simon Delaere; Pieter Ballon; Michael Van Bossuyt

SMIT-IBBT, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium
e-mail: leo.van.audenhove@vub.ac.be, simon.delaere@vub.ac.be,
pieter.ballon@vub.ac.be, michael.van.bossuyt@vub.ac.be

Abstract

The proposed paper analyses the changes in business models employed by the stakeholders in the newspaper value network, in the context of a new type of electronic reading device –the ePaper. This PDA-like device uses a new high-contrast, low-power screen technology (eInk), which holds the promise of a digital and mobile reading experience close to that of ‘real’ paper. The potential impact of massive digitally distributed reading content –newspapers, but also magazines, books, and all other material previously printed on paper– on the traditional publishing value chain and its different constituent actors could be significant. For example, content aggregation roles already greatly dispersed by the internet could move further away from the traditional newspaper publishers; using logging data and RSS feeds on the device, newspaper advertising could become personalised and interactive; for newspaper publishers, production and distribution costs could go down and updated content could be sent to the device whenever needed etc. This paper is based on a large scale research project in Flanders/Belgium, which has brought together a device manufacturer, a financial newspaper publisher, a telecoms incumbent and several technological and social science research groups from Flemish universities. To complement the technological development and an extensive field trial with near-market devices, the authors analysed how this new technology might transform the traditional publishing value chain, what the strategic options of the different actors are, and what scenarios are possible and likely to occur in the development of ePaper publishing. To do this, they make use of the theoretical framework for business model analysis. Using literature study as well as empirical data (i.e. face to face interviews with important stakeholders from the newspaper and book publishing sectors), a number of scenarios for the re-definition of roles are outlined. The authors come to the conclusion that the choice for an open versus a closed architecture, along with the technological roadmap of the device, will be crucial in establishing a valid business model for ePaper. In this paper we complement the scenario study with information on the first commercial trials and products using electronic eInk based reading devices.

Keywords: electronic newspaper; mobile newspaper; electronic paper

1 Introduction

The rise of PC from the 1970s and the Internet and mobile communication from the 1990s have lured many self-proclaimed gurus in predicting that we are moving towards a paperless society. However, so far this idea has not materialised. If anything, the use of ICTs and the Internet seem to increase the use of paper, and the publishing industry is performing quite well despite all electronic information available. The main reasons why people still print electronic content on paper are 1) the portability of paper and 2) the high quality of the printed material. Visual displays still cause physical stress on its readers and the quality of the image is lower than on paper [1].

Different companies are searching for electronic alternatives for the traditional paper. One of the most recent additions is called eInk, a screen technology developed by a consortium consisting, among others, of Philips, Toppan Printing, Gruppo Espresso, Hearst Corporation, Motorola and Vivendi. The company’s *electronic ink* – ink that carries a charge enabling it to be updated through electronics– allows for the production of so-called *Electronic Paper Displays (EPD)* possessing a paper-like high contrast appearance, ultra-low power consumption, and a relatively thin and light form factor. Theoretically, these devices could therefore be able to give the viewer the experience of reading from paper, while having the power of updatable information.

This paper analyses how the introduction of a device using this technology might provoke changes in business models, actors and roles in the (newspaper) publishing sector. It is based on the business modelling Work Package within a large scale government funded research project in Flanders (Belgium), called ePaper. The project brought together a device manufacturer (Philips/iRex Technologies), a financial newspaper publisher (De

Tijd), a telecoms incumbent (Belgacom), advertisers (Hypervision–Agency.com)/iMerge and several technological and social science research groups. To complement technological development of an ePaper device based on eInk technology, and an extensive field trial with near-market devices, the authors have analysed within this project how this new technology might transform the traditional publishing value chain, what are the strategic options of the different actors, and what scenarios are possible and likely to occur in the development of ePaper publishing. The potential impact of massive digitally distributed reading content on the traditional publishing value chain and its different constituent actors could be significant. For example, content aggregation roles already greatly dispersed by the internet could move further away from the traditional newspaper publishers; using logging data and RSS feeds on the device, newspaper advertising could become personalised and interactive; for newspaper publishers, production and distribution costs could go down and updated content could be sent to the device whenever needed etc.

In this paper, the results of our analysis will be briefly outlined. The methodological framework for business model analysis is concisely described. The paper focusses on the analysis of the ePaper value chain, and on the empirical elaboration of possible business model scenarios. The empirical basis for this work are expert interviews with representatives of the publishing industry in Flanders [2]. This analysis is complemented with information on the first commercial trials and products using electronic eInk based reading devices.

2 Approach and Methodology

Despite growing interest in business modelling in recent years, no clear definition of the term exists today. Different definitions emphasize diverging aspects such as the architecture of a product or service, a description of the roles of and the relations between companies, the ways in which business can be conducted, the way in which value is created etc. [3]. In this report, we use a definition, which tries to synthesize the most crucial elements in the mentioned literature and definitions [4]. We define a business model as: *'A description of how a company or a set of companies intends to create and capture value with a product or service. A business model defines the architecture of the product or service, the roles and relations of the company, its customers, partners and suppliers, and the physical, virtual and financial flows between them'*.

This definition relates to three levels of the business model: a *functional* level (dealing with the architecture of a product or a service), a *strategic/organisational* level (dealing with the roles and relations between actors and the physical and virtual flows between these actors) and a *financial* level (dealing with the sources of revenue of and the financial flows between the actors involved). In our analysis, we add to this a fourth level, i.e. the *value proposition*. This fourth level, which is the way value is created in the market, can be considered as a logical outcome of the strategic choices made on the other three levels when designing business models.

An important aspect of this definition is that it does not limit the focus of analysis to one specific firm, but instead takes into account a network of actors involved with the production, distribution and consumption of products and services. This reflects the growing complexity of innovation processes in what is called the network economy and society. From a financial perspective, the emphasis is on structuring the revenue streams and on creating models for revenue sharing.

In terms of the value chain, a concept coined by Porter to describe the primary value-adding activities of a firm or of a set of firms, this means looking at the whole chain [5]. In fact, most scholars agree that the increasing complexity and flexibility of business design means that the representation of business processes by a linear value chain has to be replaced by more fluid value networks, in which roles and functions can be combined in different ways by different actors. Business design is therefore increasingly about defining firms' boundaries and the level of horizontal and vertical integration. Taking into account the three basic levels of business modelling and the value proposition that is the outcome of these, a successful business model will emerge when a so-called *strategic fit* occurs between the different firms involved in the production of a product or a service, and on the different levels discussed, as well as between a firm's business model and the consumer [7].

3 The ePaper Value Chain

3.1 Value Chain and Network

We have started our business scenario analysis by analysing the ePaper value chain. This value chain contains the roles that are essential for the production and distribution of content on the ePaper device. It is important to point out that these roles may be taken up by diverging actors. In the ePaper value chain, we discern the roles of *Content Provision*, *Content Aggregation*, *Platform Content Aggregation*, *Platform Provision*, *Network*

Operation as well as *Service Provision*, *Advertising*, *Device Supply* and *Device Manufacturing*. The latter four roles are basically related to the strategies of other actors and to the business scenarios chosen, and are therefore not included in the value chain as such.

3.2 Roles and Actors in the Value Network

Below, we define the different roles in the ePaper value network. We indicate which actors are potentially interested in taking up any of the roles in the network. This implies that, besides looking at the newspaper sector, we also include the news production and publishing sectors in this value network; looking at the present functionalities of the ePaper device, content published on it will—at least initially—be of a written nature.

Content Provision. In the news and newspaper sector many actors take up this role (e.g. independent journalists, national and international news agencies, newspapers delivering syndicated content etc.) The newspaper itself acts as a producer for a lot of content; besides this, ePaper also provides a platform for other written content such as literature, magazines, trade journals, corporate publications etc. coming from a host of different providers.

Content Aggregation. In the news production sector, the newspaper is a typical example of an aggregator of content. Newspapers and magazines make a profession out of bundling content, services and advertising in a coherent editorial concept. These actors strongly believe that this aggregation function will remain an important task in the digital age. However, the digitisation of content and the subsequent creation of new communication platforms such as the Web, i-mode, iDTV etc. have spurred the development of alternative content aggregators.

Platform Content Aggregation. It is important to make a distinction between *Content Aggregation* and *Platform Content Aggregation*: the former relates to the filtering, editing and branding of content in an editorial concept, the latter points to the assembling of already aggregated content (e.g. newspapers, books, magazines, etc.) of different *Content Providers* and *Aggregators* onto an electronic platform. For example, *Newsstand.com* offers a broad selection of digitised international newspapers and magazines from different publishers on the Internet Platform. A crucial point of discussion surrounding ePaper is the degree to which content from newspapers and other providers will be offered in an aggregated or a disaggregated manner. In constructing business scenarios for the ePaper platform, a central variable will be who takes up the role of *Content Platform Aggregation*.

Platform Provision, i.e. the provision of a technical platform that links content and technology. This role is significant because it determines, to a large extent, the control of who publishes on the device and what is possible on it. This role can be divided into a server-side and a software/DRM function. The server-side function assures communication between the content provision and the ePaper device and therefore constitutes a potential bottleneck. The uncertainty on which actor will take up this function, renders the function into a possible source of conflict within the value network.

Network operation. This is the domain of telecommunications operators, whose services might be considered as substitutable commodities. In such case, *Network Operation* is reduced to the provision of a *pipeline* for the content. However, network operators worldwide are trying to broaden the scope of their operations from pure transmission to the offering of content-related services. Within ePaper, these actors might have the ambition to take up the roles of *Platform Content Aggregation* and *Content Aggregation*.

Service Provision. This is a crucial role in the ePaper value network, relating to who maintains the customer relationship and effectively markets the service. For the time being, this role cannot be identified in the value chain, since its positioning within this value chain depends from which actor takes up this role. The newspaper or its overarching publisher seems to be well-placed to do this, because—especially in subscription models—it has a unique relationship with its customers. However, when looking at the technological functionalities of ePaper, other actors—for example *Platform Content Aggregators*—could also take up this role.

Device Supply. The question here is by whom and in which way the device is marketed. Again, this role cannot be identified in the value chain for the moment because it is dependent upon the business scenario chosen. Taking into account the cost of the device, we expect that this role will often coincide with the offering of content and services, and that the device will be offered in some sort of subscription model. However, other options, among which an ePaper reader as a simple consumer device remain possible.

Device manufacturing. At the moment there are only a few commercial eInk based devices on the market. iRex technologies—the company involved in the trial this paper is related to—developed the Iliad reader and Sony

developed a Librie and Sony Reader. Both can display different content, but the Iliad was specifically developed with electronic newspapers in mind, whereas Librie and Sony Reader were developed to display ebooks.

Advertising. This role is already fully part of the traditional newspapers' value chain, with newspaper publishers in the role of *Content Aggregators* integrating advertisements coming from other parties. However, ePaper offers new opportunities for advertising, e.g. interactive and personalised ads, on the level of the electronic newspaper (*Content Aggregation*) as well as on the level of the device (*Platform Content Aggregation*). The *Advertising* role will therefore be dependent upon the business scenario chosen. Initially it is not foreseen that the advertisers will play a central role in the ePaper value network: our interviews with the newspaper and magazine sector in Flanders have shown that these sectors are rather skeptical about highly personalised content and advertising.

4 About the Potential Scenarios for ePaper

The above discussion of the ePaper value network has made clear that this network contains several roles which can be taken up by different actors. Question is how these roles are complementary with the interests and strategies of existing actors. The digitisation of content implies that the role of *Content Aggregation* –which, in the offline world, is a clear prerogative of the newspaper editors– could shift towards the platform itself by means of *Platform Content Aggregation*. The roles of *Service Provision* and *Device Supply*, for their part, are closely linked to the business scenario chosen.

In order to gain insight into potential and probable business models, we use the scenario method, in which two uncertain variables are defined, along which four potential futures can be outlined. In the present context, many of these uncertainties are surrounding the ePaper device and possible business scenarios; based on the interviews and on our literature review, we were able to define two uncertainties which can be considered as crucial:

Aggregation vs. Desaggregation, i.e. the degree to which content is offered on the platform in an aggregated or disaggregated manner, defined from the perspective of the newspaper. *Aggregated* signifies that the newspaper can offer its content *as such* on the platform, whereas *desaggregated* means that the content on the device originates from different content providers and is more *fragmented*, i.e. less edited, packaged and branded.

Open vs. Closed, i.e. the degree to which the device is accessible for content originating from different content providers. A crucial question for determining this variable is whether –and if yes, to what degree– an exclusive link exists between the offering of content and the display of that content on the ePaper device.

It is striking that the different actors interviewed and studied have pronounced often conflicting opinions about the necessity of an open or a closed model and about the inevitability of the evolution of media towards a disaggregated model. Both variables may be used to create a co-ordinate system comprising four quadrants, with each quadrant representing a potential business scenario. We discern these scenarios: (1) *Newspaper model (Aggregated–Closed)*; (2) *Kiosk model (Aggregated–Open)*; (3) *iTunes model (Desaggregated–Closed)*; (4) *Web model (Desaggregated–Open)*. Below, we shall describe four generic scenarios and analyse their potential.

5 Scenario 1 – The Newspaper Model on ePaper

5.1 Business Scenario Outline

In this scenario one party, the Content Aggregator, offers a particular service on the ePaper device. This scenario is largely similar to the experimental IBBT ePaper project, in which De Tijd publishes an electronic version of its newspaper onto the device. In principle this can be done in two ways: (1) the newspaper can be uploaded to the device as is, without any major adaptations to the structure; (2) the newspaper may, as Content Provider and Content Aggregator, make use of the new capabilities of this medium. In the latter case it can alter its service by (1) publishing up-to-date content multiple times per day, (2) offering specific information aimed at particular audience segments, (3) personalising content, (4) integrate personalised advertisements into the content etc. Whatever option is picked, the newspaper remains the primordial provider of content on the device.

In the figure below we have displayed the value network of this scenario in a generic fashion. Besides the newspaper's role of Content Provider and Content Aggregator, the ePaper device offers new opportunities to put content on the device originating from third party providers. In this scenario, we make the assumption that the newspaper itself might play a potential role; in other words, the newspaper could take up the role of Platform Content Aggregation –or part of that role (see figure). Two options exist for doing this:

- 1) The newspaper could complement its own content with content from its own publishing group, thereby enhancing the attractiveness of its own service and possibly also increasing revenues of its entire group. An important condition for this is the availability of a sufficiently large and complementary offer within this publishing house that can appeal to the targeted audience;
- 2) In case the newspaper wishes to offer content originating from third parties outside its own publishing group, then this content can be expected to be mainly complementary; other newspapers will have little inclination to publish their product on a competing platform. This hypothesis is confirmed by the Content Aggregators interviewed for this study, who clearly indicate that they are only prepared to provide content for a device which is administered by a neutral party.

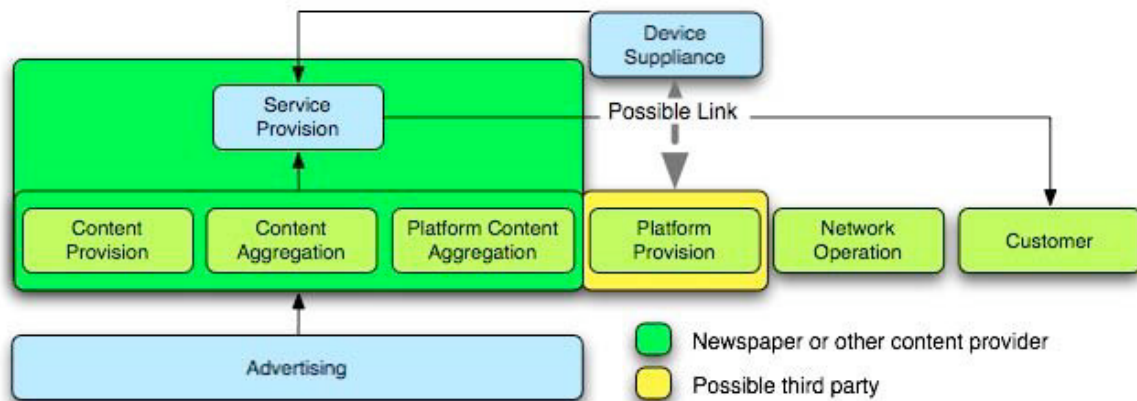


Figure 1: Newspaper model value network

If a newspaper integrates the roles of Content Provision, Content Aggregation and Platform Content Aggregation, then it is clear that this actor will market the service. It has considerable advantages over other parties in doing this: (1) an existing customer relationship, (2) content for which customers are prepared to pay and (3) a certain market intelligence.

The role of Platform Provisioning may be taken up by the newspaper itself or by a third party. Newspapers might well be interested in doing this, since a number of parties indicate that newspapers are, in a digital environment, prone to handle distribution themselves. Other potential actors are the Device Manufacturer, the Device Supplier or the Network Operator. The Device Supplier has a certain control over the device configuration, the standards used, the capabilities and limitations imposed by DRM etc. In the Flemish case, iRex is taking up this role by having developed a client as well as a server component, and is able to simultaneously offer tailored services to different parties; the functionalities of the architecture are negotiated with the newspaper in its different roles.

For marketing the device, two main options exist: (1) the customer may individually purchase an ePaper device and subsequently take a digital subscription to a newspaper; (2) the newspaper may offer the ePaper device as part of a subscription to the digital paper. In this project, it is clear that iRex, as a Device Supplier, has chosen the second model. The argument for this is that the ePaper device, unlike the iPod for example, does not have an unambiguous, easily recognisable functionality for the consumer, and that it is rather expensive at the moment. The device therefore seems easier to integrate into the market when being part of a subscription model. However, this also implies that the newspaper will need to carry the financial burden of pre-ordering the devices. As for the Device Supplier, this actor could create an additional revenue stream by also taking up the role of Platform Provider. In its turn, the Platform Provider could be inclined to shift towards the role of Platform Content Aggregator and publish services on the device itself. However, as it is the newspaper who markets the devices itself, this scenario seems rather implausible.

In case the actors choose to make use of personalised or more directed advertising, an exchange of information will need to take place between the Platform Provider, the Platform Content Provider (being the newspaper in this scenario) and the Advertiser. Firstly, the Advertiser will be interested in obtaining information about (1) the use of the platform and the characteristics of the user, and (2) which user has seen/clicked on which

advertisement. This information is also important for the newspaper itself since clicking through on advertisements usually generates higher revenue.

5.2 Evaluation

In this scenario the newspaper plays a dominant role. It has a number of important advantages: a large reader base, a good customer relationship and content that customers are willing to pay for. The newspaper may address this reader base in order to try to make a large group of readers use ePaper as quickly as possible. In making this effort, marketing the ePaper device as part of a subscription offers a number of additional advantages. Firstly, readers will be more easily persuaded to switch to the technology; secondly, in the longer term this strategy might have a cost-reducing effect for the newspaper; and finally, the newspaper would be able to monitor the reading behaviour of its customers in order to better tune the content to reader preferences.

However, the functionality of ePaper as a digital reading platform for content originating from a large array of producers is threatened, particularly if the platform is too strictly protected by DRM and proprietary standards. In this case, this scenario might become alienated from the actual wishes and demands of the targeted audience (in this case, business professionals). In this sense, the use of ePaper as a mere digital substitute for the newspaper could be considered as a rather conservative reflex by newspapers in order to maintain readership in the digital era. Moreover, an initiative launched by only one newspaper or publishing house, might be boycotted by other players in the market.

The first commercial products with ePaper readers are examples of this scenario. At the end of 2006 the Yantai Daily Media Group started publishing its main newspaper on the Iliad in China. In May of 2007 two newspaper of the Dutch PCM Group *De Volkskrant* and *NRC Handelsblad* will become available on the Iliad. There is so far little known about the projects and the agreement between PCM and iRex, but PCM is in discussion with other groups to extend services. This might indicate that PCM might also be interested in the kiosk model [7].

6. Scenario 2 – The Kiosk Model on ePaper

6.1 Business Scenario Outline

We call this the kiosk model by analogy with the newspaper kiosk. Currently, kiosks offer –besides a selection of national and foreign newspapers– a wide array of magazines, books etc. Transposed to the ePaper device, the user of this device has, in this scenario, access to a wide choice of textual media originating from different publishers. However, these publishers mainly continue to provide content in aggregated format. For the user, this scenario adds value because he can use the ePaper reader as a mobile platform for a large selection of content.

In the realm of the *audiobooks*, a platform similar to this one exists which is called *audible.com*. Audible is a platform for digital audiobooks which has a library of over 27,000 titles originating from 318 *Content Providers/Aggregators*. After installing a piece of software –either *iTunes* or *Audible Software*– files may be purchased and downloaded to a computer and subsequently to an mp3 player. Audible makes use of DRM to prevent files from being copied, but does not link its software to one particular device for using these files. According to the company, more than 200 devices are able to deal with the format used. In the realm of ebooks similar initiatives exist such as ebooks which brings together 80.000 titles from different publishers and mobipocket with 39.000 premium titles. eBooks distributes books in three standards i.e. Microsoft reader, Adobe reader and Mobipocket reader. Mobipocket uses its own standard.

In this scenario, an *intermediary* is a central actor in the value network. This intermediary takes up the role of *Platform Content Aggregation* and brings together content from diverging *Content Providers* en *Content Aggregators*. The main advantage for an intermediary is that it unites two markets, namely that of information providers and that of information users. If the intermediary succeeds in bringing a large segment of both markets to its platforms, significant network externalities occur on both these markets: the *Content Providers* gain access to a potentially larger customer base, while users have a much larger selection of content [8]. Following this strategy, Audible for example has succeeded to use the internet to create a *one-stop shop* for English language, digital audiobooks and has been able to further diversify into spoken newspapers, magazines, radio programmes and talk shows, which were distributed to 278,000 paying customers in 2006 [9].

In this scenario, it seems logical that the *Platform Content Aggregator* maintains the customer relationship or, put differently, that it takes up the role of *Service Provision*. The *Content Provider* or *Aggregator*, be it a newspaper or a publisher, uses the *Platform Content Aggregator* as an alternative distribution channel. In that

case the newspaper could lose its relationship with the subscribed readers to the *Platform Content Aggregator*. In an online environment the latter actor could create a relationship with its customers, even if they don't take a newspaper subscription. A potential alternative to this model is that the newspaper, as a *Content Aggregator*, retains the role of *Service Provision*, but uses the platform to grant users access to a larger array of content.

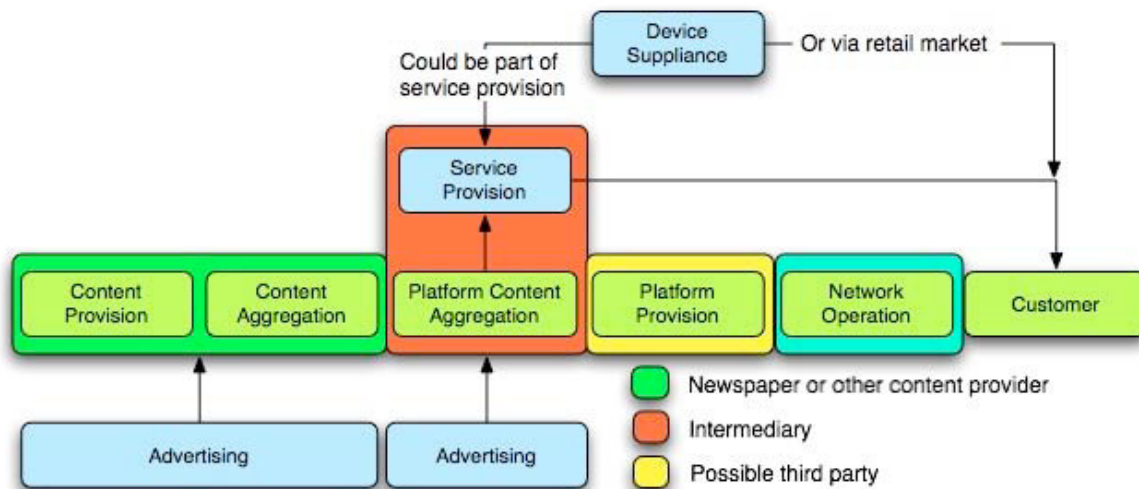


Figure 2: Kiosk model value network

It remains an open question who takes up the role of *Platform Provision* within this scenario. This role can be exerted by the *Platform Content Aggregator* itself, by the *Network Operator* or by a third party. In case the roles of *Platform Content Aggregation* and *Device Supply* are not combined, the *Platform Content Aggregator* –in this case the intermediary– faces two crucial challenges. On the one hand, this actor wishes –partly under pressure from the *Content Providers*– to prevent the copying of content, among other things by including DRM; on the other hand he wishes to offer his content on as much devices as possible. On the level of functional architecture, this party will therefore strive towards (1) the use of open standards that allow publication on multiple devices, or (2) the development of a proper solution that is subsequently supported by multiple producers. The latter strategy can only work if the intermediary has a sufficiently strong market position. A central question remains the role of the device manufacturers. Do they wish to sell their device as a piece of hardware with a number of technical service components, or do they also wish to take up other roles in the value chain, namely that of *Platform Content Aggregator*? (cf. next scenario). When transposing the scenario to the newspaper sector, the question is which party will take up the intermediary function. The establishment of a region- or nationwide intermediary could be a possibility that different actors seem to prefer –as was shown by the interviews.

In this scenario, advertisement might in principle play a role on two levels, namely that of the *Content Aggregation* (by a.o. newspapers and magazines) and that of the *Platform Content Aggregation*. As for the first level, an important issue here again is whether agreements can be reached and information exchanged between the *Platform Content Aggregator* and the *Content Aggregator* to allow personalised advertising on the level of the newspaper. After all, in the proposed scenario it will particularly be the *Platform Content Aggregator* which has disposal of a large amount of data concerning the user and content consumption behaviour. As for the second level (*Content Platform Aggregator*), advertisements might be possible here as well. However, experience has shown that this only occurs in a limited way; the main reason for this is that the *Platform Content Aggregator* is deemed to remain a neutral party. Both *iTunes Music Store* and *Audible* –two intermediaries on the internet– do not allow publicity on their platforms, and have strict editorial guidelines as regards the presentation of products. Our interviews have clearly shown that advertisements on the level of the *Content Platform Aggregator* would not be readily accepted by *Content Aggregators*.

In this scenario the two options for marketing the device are open, and lot depends on the payment options used. In our example Audible offers several of these payment options: (1) a one-off payment per title, (2) a subscription granting a year long reduction on titles, (3) a subscription giving access to one title per month for a one year period or (4) a similar subscription allowing access to two monthly titles. In this case, the device is part of the *Service Provision*. However an ePaper device could also be marketed as a consumer device. The examples of payment methods for products and services mentioned above could also be implemented for the newspaper

and (book) publishing sectors. In this scenario, it will likely be the *Platform Content Aggregator* which bundles services and device. However this is not a necessity: one of the interviewed *Content Aggregators* indicated that it was prepared to subsidise the device as part of a subscription *and* to grant access to third party content.

In this scenario, price-fixing and revenue sharing between *Platform Content Providers* on the one hand and *Content Providers* and *Content Aggregators* on the other hand, will be a difficult exercise and a possible source of conflict. The *iTunes* case in the music sector (cf. sub) constitutes a nice example of this: while a price of USD 0.99 per downloaded song is generally assumed to be too high, this price has to a large extent been imposed by the music industry [10]. A possible solution for avoiding conflict is the establishment of a *Platform Content Provider* within the sector in which the different actors participate.

6.2 Evaluation

This scenario offers interesting opportunities to stimulate the ePaper device as a mobile platform for different types of content originating from different parties while, from the publishers' perspective, the products offered retain their editorial function. It is less clear whether this scenario also contributes to the innovative use of the interactive capabilities of the device; this will require clear agreements between the *Platform Content Aggregator* and the *Content Providers* and *Aggregators*.

The introduction of an intermediary party as *Platform Content Provider* offers major advantages in terms of network externalities related to two-sided markets. However it also holds some threats: taking into account the economies of scale and network advantages created by internet and ICT-based platforms, this party could in time become a powerful actor, in particular if it maintains the customer relationship and if it has data on user preferences at its disposal. An additional threat is that the intermediary would shift toward *Content Aggregation* and *Content Provision*. In our example, Audible offers audiobooks that it has produced itself. Besides this, the launch of a new intermediary also implies larger necessary investments and limited brand awareness.

Setting up a totally new intermediary platform might prove to be a difficult exercise. Although all Flemish newspapers and publishers indicated to be in favour of the kiosk model actually setting up such a platform is another issue. Competition and mistrust might easily prevent this scenario. However, in other countries umbrella organisations representing or serving the newspaper industry already exist. E.g. the Joint Purchasing Association of the Danish Newspapers is an umbrella organisation aggregating demand for and purchasing paper for the different Danish newspapers. Such organisations might be the basis for an intermediary platform.

7 Scenario 3 – iTunes for ePaper

7.1 Business Scenario Outline

At first sight, the iTunes model seems to resemble the preceding model: here too, an intermediary partner takes up the role of Platform Content Aggregator, bringing together content from Content Providers and Aggregators. However, the scenario differs in two crucial points. Firstly, there is a certain degree of desaggregation. On the iTunes Music Store, users are able to download a single song. Transposed to the newspaper and publishing sector, this implies that separate articles could be purchased. We immediately need to add to this that desaggregation of newspapers will be trickier because the advertisements inserted are an important source of revenue. Secondly –and fundamentally differing– the same party (i.e. Apple) takes up the role of Platform Provision and of Device Supply, for Apple controls, via its software, the interaction between the iTunes Music Store and its device –the iPod– and songs downloaded via iTunes can only be played on the iPod.

A similar scenario can also be elaborated for the newspaper and publishing sector. Sony is currently aiming to do this for eBooks by using its new Sony eReader. This device can only access content from Sony's own content site Sony Connect. For this content, the Japanese firm has concluded agreements with a number of big publishing houses in the United States. In this scenario, the user still has access to a large offer originating from a number of Content Providers and Aggregators, but is forced to watch this content via a specific device, i.e. an ePaper reader. By analogy with the iTunes software, it would however be possible to print a selection [11].

As in the preceding scenario, the intermediary fulfils a crucial role in terms of uniting offer and demand. However, in this scenario the intermediary integrates even more roles, i.e. that of Platform Content Aggregation, Platform Provision, Service Provision and Device Supply (as well as Device Manufacturing). Especially in the iTunes case, where Apple has reached a US market share of more than 70 percent of mp3 players with its iPod, the combination of Platform Provision and Device Supply results in a fairly dominant position [12]. In this

scenario too, there is a certain danger that the Platform Content Aggregator gradually shifts towards Content Aggregation and Content Provision; through the desaggregation of content coming from Content Providers and Aggregators, the Platform Content Aggregator is able to personalise its service to users even better.

In the iTunes case, a link exists between the iTunes Music Store, iTunes software and the iPod. The iTunes software gives access to the iTunes Music Store and takes care of file transfers to the iPod. The files on the iTunes Music Store are protected by DRM and Apple uses a proprietary encoding standard for its files, i.e. AAC. This way, files can only be transferred to four different iPods; however the software does allow content from third parties to be loaded onto the device in mp3 or AAC. For ePaper a similar—or even stricter—scenario could be chosen, in which the device itself (and not the PC) acts as the interface between the store and the platform. Moreover, the publishing sector could use a strong push-model, in which up-to-date content is pushed towards a device after the user has indicated which content is of interest to him or her.

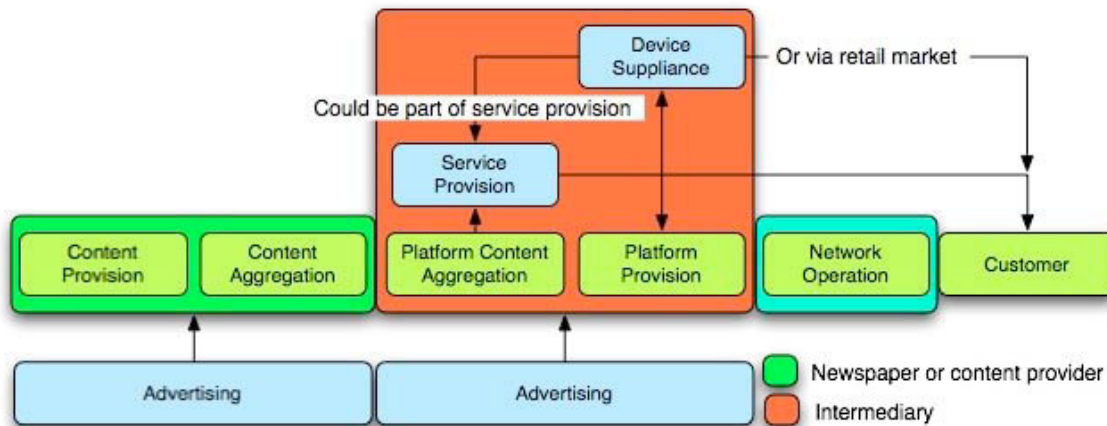


Figure 3: iTunes model value network

Taking into account this integration, it seems obvious that the Platform Content Aggregator is also responsible for Service Provision and thus maintains the relationship with the customers. Here too one can wonder about the plausibility of a scenario in which the newspaper, as Content Provider and Aggregator, takes up its own part of Service Provision. Finally, the Advertising role can be exerted on the same two levels as in the previous scenario, so the same issue apply.

In this scenario, different payment methods are equally possible; in that sense, it largely resembles the previous scenario. As it is assumed here that content can be accessed in a desaggregated format, separate articles from different Content Providers may be purchased. This necessitates new ways for paying this content, among which micro-payments. In case the Network Operator takes up the role of Platform Provisioning—or part of that role—it may be well placed to take care of billing in this model.

A particularity in this scenario is that a larger number of roles are combined, among which Platform Content Aggregation, Platform Provision, Service Provision and Device Supply. This gives the opportunity, for the actor taking up these roles, to generate revenues on different levels: (1) as a percentage on sold content or subscriptions, (2) on the basis of devices sold or (3) on the basis of a service component aimed at Content Providers and Aggregators. Option (1) and (3) may eventually be combined as one percentage on content sold, including service provision. The price that can be asked by an intermediary for selling content depends on the negotiations with the Content Providers and Aggregators and what the bargaining power of these latter actors is. The intermediary could also strategically opt to position itself between these two revenue streams. Although little is officially known about this, it is generally assumed that Apple only generates limited profit out of its iTunes Music Store and instead focuses mainly on iPod sales.

Within this scenario, it is again possible to insert advertising on two levels, i.e. on the newspaper level (or even within a separate article), and on the level of the platform. Because access to desaggregated content is possible, it seems more logical within this scenario to administer at least part of the advertising on the platform level. Besides this, it is also the intermediary which possesses the knowledge about device and platform use as well as user preferences, which it could exploit as a third revenue stream. However, it seems unlikely that newspapers

and publishing houses would hand over an important portion of their advertising revenues to the intermediary without any compensation.

7.2 Evaluation

In this scenario, the user has access to desaggregated content, i.e. individual articles from newspapers, magazines etc. This type of service clearly fits closer to the changes in reading behaviour of modern newspaper readers, as well as to changes in users' experiences with other ICT devices.

The intermediary party which integrates the roles of Platform Content Aggregation, Service Provision and Device Supply, threatens to become dominant within this scenario, which might render the publishing sector reluctant towards participating in it. Moreover, this sector traditionally attributes high value to the editorial concept with which it links its brand names, and possibly fears that excessive desaggregation will turn their content into an easily substitutable commodity. Finally, if the intermediary party protects content and devices by using DRM and proprietary standards, the user will in turn be rather reluctant to purchase such a device.

8 Scenario 4 – The Web on ePaper

8.1 Business Scenario Outline

In this scenario the ePaper device may be considered as a new gateway to the Web. The device has little or no protection by DRM or proprietary standards, so the user can upload any content –coming from the Web or produced by him/herself– onto the device. In a sense, the role of *Content Aggregation* shifts to the user by becoming that of *Content Selection*: the user actively searches for information from newspapers, weblogs, government websites, discussion forums, newsgroups, entertainment companies etc. This *prosumer* can also create information himself and make that information available to others.

All this does not necessarily mean that the user is not prepared to pay for content. He/she can still purchase certain types of content, albeit directly from the *Content Providers/Aggregators* and *Platform Content Aggregators*. Thus, while these latter roles continue to exist, the user has access to a large number of actors which individually make content available; the user is not necessarily tied to one actor.

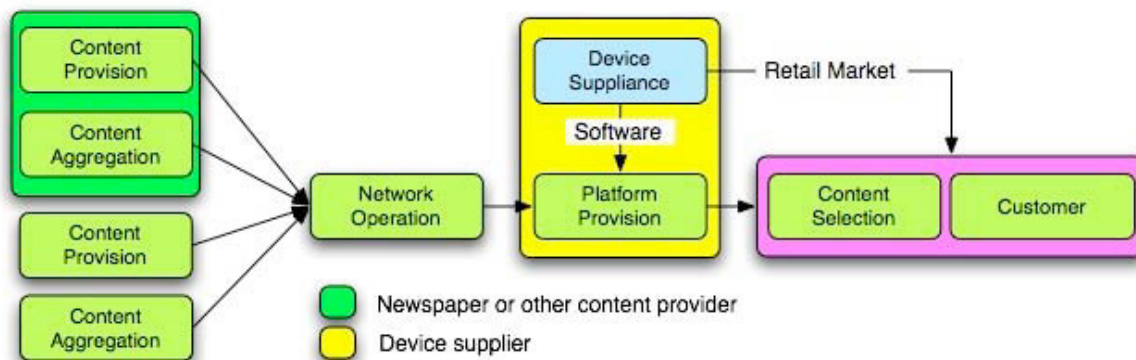


Figure 4: web model value network

The value network of the web model strongly differs from the other scenarios. Firstly, in this model *Content Provision*, *Content Aggregation* and *Platform Content Aggregation* are vertically aligned. The consumer has individual access to the content of one or more of these actors and newspapers, as *Content Aggregators*, directly compete with other *Content Aggregators* such as Google News, Newsstand etc. as well as with individual *Content Providers*. Secondly, the role of *Platform Content Aggregation* (at least at the device level) no longer exists; on the one hand, this role largely taken over by the user, while on the other hand one could argue that *search engines* also take up part of it. Thirdly, *Platform Provision* can still occur in the shape of software making up the interface between the internet and the device. Although this software could protect part of the content using DRM, the *Device Supplier* will not be inclined to consider this option. To the extent that *Content Providers* are only willing to publish their content on devices that protect this information, it is possible that

pressure is exerted in order to include DRM solutions on these devices. The same goes for standards: as device sales are crucial for the *Device Supplier* in this scenario, he will be prone to support multiple and open standards.

In this scenario, it is more difficult to monitor the use of the device. Every *Content Provider* is able to track which of its content is downloaded, but the possibilities to gather information on what the user does with this content, are rather limited. These functionalities could be incorporated into the interfacing software of the device (as *adware* or *spyware*); however, these types of monitoring are usually strongly disapproved of by the user.

In this model, it seems fairly implausible that one party would market the device as part of a subscription; the consumer will rather buy such a device by itself. Although iRex has indicated that it would primarily focus on the B2B market, it is not inconceivable that another manufacturer would brand a similar device as a consumer product. This scenario becomes more plausible if multiple *Device Manufacturers* compete with each other on a device level. On the *Content Provision* and *Aggregation* levels, the revenues are generated by the individual actors.

8.2 Evaluation

This scenario probably fits in best with the desires and expectations of the user; he or she potentially gets access to a very broad range of content. However, it remains to be seen whether the different parties are willing to realise this scenario. Newspapers are primarily interested in finding new distribution channels for their product, and not in a device that offers desaggregated content and on which they have to face full competition from free internet services. The device manufacturers for their part possibly face a *chicken-and-egg* dilemma if they cannot link the sale of devices (with the inherent distribution and marketing costs) to the guaranteed availability of content for the user.

9 Conclusion

In this study we have elaborated scenarios that describe *possible* roads towards a business model for ePaper. For doing this, we have used two fundamental uncertainties, being (1) the degree of aggregation versus desaggregation from the perspective of the newspaper, and (2) the degree to which the device is open for content originating from different providers. The combination of these variables has resulted in four scenarios: the newspaper model, the kiosk model, the iTunes model and the web model. To contextualise the scenarios we have conducted interviews with actors within the Flemish newspaper, publishing and telecommunications sector. Furthermore we have complemented the analysis with information on the first commercial trials currently running.

The described models are generic and represent only one type of business model. Besides the crucial uncertainties used in this study, too many variables exist –hence our choice for the scenario methodology. The eventual model depends on the strategic choices made by the different actors; in this regard, our interviews have already shown major differences in opinion between the actors involved. We have generically integrated these insights into the scenarios. The combination of the interviews, the literature review and the scenarios drawn up, has led to a number of strategic considerations:

Both newspapers and publishers in general will continue to believe in the importance of editorial concepts and guidelines. They will therefore have little inclination to give this up in favour of a completely desaggregated system. The fact that a large number of customers is still prepared to pay for this service (be it in paper or for the online version of newspapers), certainly proves its relevance. In each of the scenarios, the newspaper's customer database offers a major advantage for marketing ePaper.

The newspaper has –much more than other media– a relationship with its customers. This is particularly the case for subscription readers –which form a large part of the audience in Flanders, but also in many other countries. Therefore, newspapers will mainly consider new distribution channels as a way to diversify their services, but will not be willing to give up this customer relationship, especially since the possibilities for monitoring news consumption offered by ePaper allow these newspapers to further deepen their knowledge about their customers.

Taking into account these arguments, scenario 1 seems to be an important plausible option. This is confirmed by the first commercial initiatives with ePaper devices. Both the Yantai Daily Media Group in China and the PCM Group in Holland have started with offering titles on the Iliad on an individual basis. Nevertheless, platforms such as iTunes, Audible, Rhapsody, Amazon etc. show that intermediaries in two-sided markets –aggregating *Content Providers/Aggregators* on the one hand and users of content on the other hand– can become a big

success. Two-sided markets have significant network externalities that may be of particular benefit to users by creating a much broader offer of information. As newspaper markets are to a large extent delineated by language and national boundaries it will remain to be seen whether intermediaries will develop at this national level. In the present context, the position of the *Device Manufacturer* and the roles it will take up, constitute important and uncertain variables. For the moment, the actors involved seem to opt primarily for a B2B strategy. In the short term, this renders scenario 4 less plausible.

As mentioned, the question which scenario –or which derivative of such as scenario– will eventually become reality, largely depends on the strategies of and the negotiations between actors. Two final important remarks need to be made in this regard. Firstly, the scenarios *are not mutually exclusive*: it is perfectly possible for a newspaper and a *Device Manufacturer* to strive, in the short term, towards a newspaper model (scenario 1) while leaving room for elaborating other scenarios, such as a kiosk model (scenario 2). Secondly, it is not inconceivable that, as time passes, a shift occurs from scenario 1 to scenario 4. Particularly if eInk or similar technologies become more broadly adopted and multiple devices are launched, the pressure for creating open systems might increase. On the one hand it is important for newspapers to take this into account *a priori* and to avoid investing in systems and technology that create too much path dependency or that are not adaptable. On the other hand it remains to be seen whether this ‘conservative’ sector will grab the new opportunities this technology offers or whether it will be the Internet or electronics sector who will drive the initiatives.

Notes and References

- [1] SHAVER, D.; SHAVER, M. A. (2003). Books and Digital Technology. A new industry model. *Journal of Media Economics*, 16(2), 71–86
- [2] Interviews with: iRex Technologies, Philips, Hypervision, Uitgeversbedrijf De Tijd, I-Merge, Belgacom, Lannoo, Magnet Magazines, Concentra, De Standaard Uitgeverij, De Standaard
- [3] WEILL, P.; VITALE, M. (2001). *Place to Space: Migrating to eBusiness Models*, Boston: Harvard Business School Press; OVANS, A. (2000). E-Procurement at Schlumberger. *Harvard Business Review*, 78(3): 21–23; TIMMERS, P. (1998). Business Models for Electronic Markets. *EM– Electronic Markets*, vol. 8, no.2; SLYKOTSKY, A. J. (1996). *Value Migration – How to Think Several Moves Ahead of the Competition*. Boston: Harvard Business School Press.
- [4] BALLON, P. (ed.) (2005). *Best Practice in Business Modelling for ICT Services*. Delft: TNO
- [5] PORTER, M. (1985), *Competitive Advantage: Creating and Sustaining Superior Performance*, New York: Free Press.
- [6] METHLIE, L.; PEDERSEN, P., (2001) *Understanding business models in mobile commerce*, Paper presented at WWRP 3, Stockholm, September. BOUWMAN, H. (2002). *Business Models for Innovative Telematics Applications*, Enschede: Telematica Instituut; FABER, E., BALLON, P., BOUWMAN, H., HAAKER, T., RIETKERK, O., STEEN, M., (2003) *Designing business models for mobile ICT services*, Paper presented at E-commerce workshop, Bled, June 9–11.
- [7] HOOFT VAN HUYSDUYNEN, M. (2007) Volkskrant en NRC op elektronisch papier, *FEM Business online*, <http://www.fembusiness.nl/fembusiness/content/nieuws/55059/article.html>
- [8] CORTADE, T. (2006). A Strategic Guide on Two-Sided Markets. *Communications and Strategies*, No. 61, 1st Quarter, 17–37, EISENMANN, T.; PARKER, G.; VAN ALSTYNE, M. (2007) Strategies for two-sided markets, *Harvard Business Review*, oktober, 92-101.
- [9] MACKENZIE, K. (2006). Audio books open a new chapter in digital age. *FT.COM Financial Times*, May 26
- [10] KUSEK, D.; LEONHARD, G. (2005). *The Future of Music. Manifesto for the digital music revolution*. Berklee: Berklee College of Music.
- [11] VAN AUDENHOVE, L. (2004) *The business scenario behind the iTunes Music Stores and the iPod*. B@Home Working Paper, Delft: TNO-STB
- [12] SONY (2006) Sony and Borders to sell digital reading device, *Sony Electronic News and Information*, from: news.sel.sony.com (Accessed 5/16/2006)

Introducing the e-newspaper – Audience Preferences and Demands

Carina Ihlström Eriksson; Maria Åkesson

School of Information Science, Computer and Electrical Engineering, Halmstad University
P.O. Box 823, S-301 19 Halmstad, Sweden
e-mail: carina.ihlstrom_eriksson@ide.hh.se; maria.akesson@ide.hh.se

Abstract

This paper adds to the overall understanding of new media adoption in general and the promotion of the e-newspaper in particular by empirically studying the preferences and demands of the potential users. The e-newspaper is a newspaper published on e-paper technology. The findings in this paper is based on the results from two studies, i.e. an online questionnaire with 3626 respondents and an evaluation in real life settings with 10 families over a two week period. Our initial hypothesis was that: *users confronted with a vision of new technology and services are more positive to adopt than users with actual use experience of technology and services in an early stage of development with inherent technology problems.* The research question of the paper is: *How does use experience influence perceptions of preferences and demands for the e-newspaper?* The findings showed that the hypothesis proved to be false, the test persons that have an actual use experience of the e-newspaper, despite the shortcomings in the device and service, were more positive to adopt than the respondents that have experienced concept movies and prototypes with more advanced functionality and interface.

Keywords: new media adoption; user experience; e-newspaper

1 Introduction

New mobile devices are constantly being introduced to the market offering new opportunities for publishing mobile media content and services. It is very difficult however, for content providers to predict m-commerce markets due to the uncertainties related to adoption of new mobile technology and services [1]. Moreover, this situation is new not only to content providers, it is also new to the audience. The rapid introduction of mobile technology and new services has led to a situation where users are constantly trying out new appliances and new services. This in turn changes use patterns as well as creates new preferences and demands, which leads to uncertainty about what people want [1].

Mobile service adoption has been studied by many scholars, e.g. drivers for adoption and intentions to adopt mobile services [2], factors influencing adoption [3, 4], adoption patterns [5-7], and attitudes towards using mobile services [8]. Much of this research has been focused on the adoption of mobile devices as such, as without adoption of devices there is not any prospect for successful m-commerce [9]. On the other hand, we argue that without attractive mobile content and services there is no incitement for m-commerce. This is indicated by the fact that in spite of the high penetration of mobile phones, which in Sweden and Italy were as high as 110% in 2006 [10], m-commerce has not taken off as hoped for [11, 12].

In the DigiNews and UbiMedia projects we have studied the potential of a new innovation for the media sector, i.e. the e-newspaper published on e-paper technology. As the e-newspaper introduction concerns both a new device as well as new content, it makes it an interesting case to study from an adoption point of view. As argued by Sarker and Wells [9], there is a need to understand adoption from the perspective of the consumers themselves. We have studied the potential willingness to adopt a future e-newspaper by presenting an online questionnaire resulting in 3626 respondents. However, it is very uncertain to trust what people think they want before having an actual experience of a product or service, we therefore also performed a user evaluation of an actual e-newspaper over a two week period with 10 families.

The research question in this paper is: *How does use experience influence perceptions of preferences and demands for the e-newspaper?* The aim is to contribute to the understanding of new media adoption as well as to contribute to the newspaper organizations preparations for launching the e-newspaper. This challenge will be

studied using the e-newspaper case described below and by testing the following hypothesis: *users confronted with a vision of new technology and services are more positive to adopt than users with actual use experience of technology and services in an early stage of development with inherent technology problems.*

The structure of this paper is as follows. In section 2 the e-newspaper case is presented followed by a description of the research method in section 3. The theoretical framework is presented in section 4. In section 5 the findings are presented and section 6 discuss the findings and conclude the paper.

2 The e-newspaper Case

This research has been conducted within two projects, i.e. DigiNews (ITEA 03015) and UbiMedia (Designing Ubiquitous Media Services through Action Research). The research started within the DigiNews project, which was a two year project including partners from Belgium, Spain, Netherlands, France and Sweden and consisted of several major technology firms, media houses and universities. The overall goal was to explore research and development issues for the future e-newspaper, i.e. a newspaper published on e-paper technology. After the DigiNews project ended in mid-year 2006, the research continued within the UbiMedia project, which is a Swedish project with partners from 9 Swedish newspaper, the Swedish Newspaper Publishers' Association and Stampen. This two-year project targets the challenge of designing ubiquitous media services for a multitude of devices and contexts to be consumed anytime and anywhere.

Electronic paper (e-paper) is the common term for several different technologies that can be used to produce screens with a number of specific characteristics. The e-paper is reflecting, giving the same reader experience as paper (such as high contrast, good color representation and the possibility to read in sunlight). The e-paper is thin, flexible and non-sensitive. In addition, it does not require high battery performance – ultimately, the screen image is stable and fix even when there is no electrical voltage applied.

The e-newspaper is predicted to combine the readability and overview from the printed newspaper with the possibilities of online media such as constant updates, interactivity and video [13], and is even predicted to replace the printed edition in the long run [14]. The potential replacement of the printed newspaper with the e-newspaper would dramatically reduce production and distribution costs for the newspaper companies.

The introduction of the e-newspaper has already begun, during 2006 two experiments with e-newspapers in real life settings has been performed, the first with the financial paper De TIJD in Belgium [15] and the second with Sundsvalls Tidning in Sweden which is one of the studies presented in this paper. In China the Yantai Daily Media Group started to publish an e-newspaper in October 2006. In all these examples the device iRex iLiad (Figure 1) was used, which is one of the two available “reading devices” on the market today, using e-paper technology.



Figure 1: iRex iLiad [16]

iRex Technologies BV, a spin-off from Royal Philips Electronics, launched the iLiad, a first generation electronic reader product in April 2006. The iLiad includes an 8.1 inch screen with 16 levels of grey and 160 dpi resolution, Wi-Fi, USB ports and MP3 capabilities [16]. Using a special marker, readers can comment on articles and scribble their notes on the screen.

The other device on the market is the Sony Reader (Figure 2), which was launched during the fall of 2006 on the U.S. market. The Sony Reader has a 6-inch screen, weight is less than 9 ounces and one can do 7.500 page views for each charge by an AC adapter. It can hold up to 80 eBooks at the same time, and allows PDFs, personal documents, newsfeeds, blogs and JPEGs. Sony offers books for the device on a new web site called Sony Connects [17].



Figure 2: Sony Reader [18]

Just to show how big this industry is expected to be, we give an example of analysts from IDTechEx who forecast plastic electronics will be a \$30 billion industry by 2015, and could reach as much as \$250 billion by 2025 [19].

3 Research Method

In the DigiNews and UbiMedia projects, described above, we have conducted several studies concerning audience preferences and demands of the future e-newspaper. In this paper we report from two of these studies, i.e. a survey with 3626 respondents and an evaluation of an early version of an e-newspaper with 10 families over a two week period.

The Survey

The survey was done through a web-based questionnaire. We presented the questionnaires at the news sites of the three Swedish newspapers that we have collaborated with in developing e-newspaper prototypes within the DigiNews project, i.e. Aftonbladet, Göteborgs-Posten and Sundsvalls Tidning (Table 1). Aftonbladet is a tabloid with the most visited news site in Sweden, Göteborgs-Posten is a local morning paper covering Göteborg (the second largest city in Sweden) and its surroundings, and Sundsvalls Tidning is a local morning paper in the north of Sweden.

<i>Newspaper</i>	<i>URL</i>	<i>Unique visitors/day</i>	<i>No. of respondents</i>
Aftonbladet	aftonbladet.se	1.200.000	3757
Göteborgsposten	gp.se	41.500	135
Sundsvalls Tidning	st.nu	14.500	447

Table 1: Newspapers hosts for questionnaires and number of respondents

The questionnaire was divided in four parts concerning background data, business models for electronic news, preferences for future electronic news and use of mobile media services. In total, 127 questions were asked and 4339 respondents answered the questionnaire. The respondents that had given an age under 15, those who did not complete or answered the questions included in this study contradictorily were excluded from the data set, resulting in a dataset containing 3626 respondents. In this paper we report from the background questions and questions regarding preferences for future electronic news.

We choose to use online questionnaires because that allowed us to show concept videos and prototypes for the respondents to obtain an understanding of the e-newspaper concept. Moreover, Buchanan and Smith [20] have argued that web samples can be as representative as or more representative than traditionally collected samples because of the heterogeneity of the online population. Although, admittedly there are inherent problems in controlling whom responds to online questionnaires. Control for cases with multiple submissions from the same IP number was handled in the data collection. Since the e-newspaper concept was not known to all potential respondents we provided them with the possibility to read more about e-paper technology on a separate page which consisted of a simplistic picture of the concept as well as links for further reading.

Further, we provided three concept videos of future e-newspaper scenarios in conjunction to the questionnaire for the respondents to watch. The movies envisioned the benefit of the e-newspaper for three different personas: the business women, the student and the senior citizen. Close ups on the designed user interface together with examples of functions showed the future e-newspaper in detail. The scenarios were based on the assumed preferences of the three personas and showed how a future e-newspaper could support their media consumption in different contexts. Watching these videos provided the respondents with an idea of what functionality the future e-newspaper could provide.

During the DigiNews project different prototypes (Figure 3) were developed for PC:s and tablet PC:s to be able to test conceptual ideas. These prototypes were developed together with newspaper designers and used contents from the newspaper partners. The prototypes also served as a way to explain how a future e-newspaper may look like and where presented with the introduction to the questionnaire. The respondents could download and test the prototypes on their own computer before they answered the questions.



Figure 3: Interactive e-newspaper prototype

The questions about preferences for a future e-newspaper regarded the e-paper device as well as the content and services. Some of the questions were statements with a 7-grade Lickert scale and others were multiple choice questions. The responses to the questionnaire were analyzed using SPSS v14.0. The analysis focused on calculation of mean scores and standard deviations for each statement and on frequencies and percentages for multiple choice questions. The goal was to generate an overview of what that was comparable to the results from the e-newspaper test persons.

The Evaluation

The evaluation was conducted with 10 families who tested an early version of an e-newspaper (Figure 4) published on the iRex iLiad in real-life settings over a two week period in the autumn of 2006. The e-newspaper of Sundsvalls Tidning was published twice daily, at 6 pm and 1 am, and was downloadable via Internet. The respondents were foremost selected to represent different types of households such as singles, couples, families with children, and senior citizens, to secure different use patterns, but we also tried to get differences in gender, ages, occupation and education. In two of the families both adults participated in the evaluation resulting in a total of 12 respondents (but only one of the extra family members answered the questionnaire – giving a total of 11 respondents to that part of the evaluation).

The two-week evaluation started with a meeting in Sundsvall, where the respondents were introduced to the device, and got a questionnaire about their media and reading habits. After two weeks of e-newspaper use, the respondents were visited in their homes for an interview about their experiences and preferences of the e-newspaper. A semi-structured interview approach [21], with an interview guide was used. These interviews were recorded and transcribed. Finally, they received a questionnaire consisting of 14 questions, partly matching the questions in the survey above.



Figure 4: E-newspaper prototype

However, as the e-paper technology in the iRex iLiad is in an early stage of development, there are some limitations compared to the e-newspaper prototypes described above. For example, the iLiad only presents 16 grayscale and have several limitations in the navigation system regulated by the device.

4 Theoretical Framework

One of the major topics in m-commerce research is user's adoption of mobile devices and services. Most of this research has been related to mobile telephony and services in 3G networks using theoretical frameworks like Innovation and Diffusion Theory [22] and Technology Acceptance Model [23]. Roger's [22] Innovation and Diffusion Theory explains among other things the Innovation-Decision Process, which contains five stages. In the second stage, the persuasion stage, the general perception of the innovation is developed which is explained by the perceived attributes, relative advantage, compatibility, complexity, observability and trialability. The later two are related to how users can experience the new technology before adoption which is the topic of this paper [3]. Rogers [22] define *observability* as the degree to which the result brought by the technology and the

technology itself is visible before adopting the technology and *trialability* as the degree to which a technology can be experimented with before adoption.

There are several studies addressing use experience and exposure in adoption processes. Sarker and Wells [9], for example, took an approach grounded in users actual practice and designed a framework for studying key issues related to mobile device use and adoption, including aspects of mobility. This framework is built as an input-process-output model. The inputs are user characteristics, communication/task characteristics, technology characteristics, modality of mobility, and surrounding context. The process consists of *exploration and experimentation* as one sub process and *assessment of experience* as another. Output refers to actual adoption behaviors, such as continuity of use over time. As this framework is based on study of motivations and circumstances surrounding individual's adoption and use of mobile devices, some of the issues described in the framework are related to communicational tasks such as voice communication, SMS, e-mail not applicable in this study.

In another study, the role of exposure to the adoption process was studied [3]. *Exposure* is defined as the degree to which an individual has acquired or exchanged information about the technology and its usage. The suggested model included exposure in form of *trial, communication, and observation*. The findings in this study suggest that the level of exposure of a new technology has an effect on the user's attitude towards that technology and thereby strengthens or weakens the user's intention to adopt. Trial and communication proved to be more effective than observation. The conclusion drawn is that exposure is likely to facilitate adoption of m-commerce.

In this study we are addressing adoption of a new technology together with its content. This was also the case in a study concerning adoption of services in mobile phones it was found that mobile services are adopted according to several patterns despite having the same technology base [5]. Allowing user to try the services rather than the technology was the focus in these tests. There are also studies indicating that *use situation and mobility* has a significant effect on intention to use a mobile service since user perceive services differently in different situations [24].

A summary of this literature review on factors influencing intention to use relevant for this study, i.e. related to use experience, are presented in Table 2.

<i>Factor</i>	<i>Reference</i>
Observability and trialability	Rogers (1995)
Exploration, experimentation and assessment of experience	Sarker and Wells (2003)
Exposure (trial, communication, and observation)	Khalifa and Cheng (2002)
Trial of service (not technology)	Carlsson, <i>et al.</i> (2005)
Use situation and mobility	Mallat <i>et al.</i> (2006)

Table 2: Summary of factors related to use experience

In this study we have let respondents experience the e-newspaper in two different ways. In the first study the respondents could observe visions of future e-newspaper usage in different situations watching concept videos. Further the respondents could try and experiment with e-newspaper prototypes. However, they were not exposed to the actual e-paper technology. In the second study, users tried the e-newspaper prototypes implemented in an e-paper device with its constraints in their everyday situations such as at home, commuting to work, at work etc. They were allowed to experiment with the e-newspaper prototypes for two weeks. In the following we describe and compare how these differences have affected the audience preferences and demands in the two studies.

5 Findings

In section we first present the background data to the respondents from the survey (Table 3) and the test persons in the evaluation. Thereafter we compare the results on preferences and opinions regarding the e-newspaper from the two studies. The test persons in the evaluation were 3 women and 9 men with the average age of 39,7. Their educational level was: elementary (3), grammar (5), and university level (5). 9 of the test persons work full time, 2 were students, and 1 a senior citizen. 9 of them subscribe to printed newspapers and all 12 read online news. Finally, 5 regularly use mobile services.

Demographic data		<i>All</i>	<i>Men</i>	<i>Women</i>	
<i>No of</i>		3626	2216	1410	
<i>%</i>		100	61,1	38,9	
<i>Mid age</i>		37,1	37,9	35,7	
Background data in percentage of total data set					
<i>Occupation</i>	Full time	55	<i>Income</i>	Low	30,0
	Part time	7,7		Medium	41,9
	Unemployed	5,8		High	28,1
	Senior citizen	6,3	<i>Education</i>	Elementary	9,1
	Student	19,0		Grammar	45,5
	Sick leave	3,6		University	44,0
	Other	2,6		Other	1,5
	<i>Newspaper subscriber</i>	<i>Read online news</i>	<i>Possession of mobile phone</i>	<i>Use mobile services</i>	
Yes	48,8	99,4	97,5	53,8	
No	51,2	0,6	2,5	46,2	

Table 3: Demographic and background data of the questionnaire respondents

The first topic for comparison regards the reason for considering exchanging the traditional printed newspaper with an e-newspaper. This question was asked in both studies and the respondents were given statements that they answered on a 7-grade Likert scale. The mean scored from both studies are presented in Table 4. *Availability anywhere* and *Added value such as new services* are the most important reasons for both groups. The least important reason in the survey was *Environmental reasons* and in the evaluation *Time savings*. Notable is that all reasons were rated higher by the test persons apart from *Time saving*. Some of the test persons mentioned that the e-paper device should not only contain their morning paper but also support all their reading, as illustrated by one of the test persons: “*I want to read everything on this device...it has to be good enough for that to make up for the inconvenience of downloading and updating. It has to be as good as what I have today.*”

<i>What reasons are critical if you sometime in the future would exchange your traditional printed newspaper to an e-newspaper?</i>	<i>Survey</i>		<i>Test persons</i>
	<i>Mean</i>	<i>Std dev</i>	<i>Mean</i>
Environmental reasons	3,6	2,59	4,8
Cost savings	4,0	2,64	5,6
Time savings	3,9	2,76	3,9
Availability anywhere	4,8	2,69	6,3
Satisfaction with new technology	4,0	2,64	6,1
Added value such as new services	4,2	2,60	6,0

Table 4: Reasons for exchanging traditional newspaper

The second question asked in both studies concerned what added services that the respondents regarded as interesting to include in an e-newspaper (Table 5). Both studies show that *Archive*, i.e. the possibility of saving newspapers from previous days, is the most interesting added service followed by *Personalization* and *Community information*. *Personal information* was regarded as the least interesting in both studies. Interestingly, the test persons scored all added services to be more interesting than the respondents in the survey. During the interviews the test persons gave additional input to preferred added services, e.g. the possibility of cutting out and save items from the printed edition was seen as an aspect that needed to be transferred to the e-newspaper. Some also mentioned the possibility of e-commerce as an attractive add-on. Other aspects mentioned were environmental, e.g. not cutting down trees and less decontamination due to less distribution by vehicles, and as one of the test persons said: “*I do not know how much time it would take to get used to it, but I think it is better than recycling old newspapers*”.

<i>Apart from reading the news, what added services do you think should be included in the e-newspaper?</i>	<i>Survey</i>		<i>Test persons</i>
	<i>Mean</i>	<i>Std dev</i>	<i>Mean</i>
Personalization	4,4	2,65	5,3
Community information	4,4	2,64	5
Personal information	2,9	2,40	3,4
General information	3,7	2,51	4,6
Archive	5,1	2,62	6,2
E-commerce	3,5	2,50	3,6
Entertainment	4,0	2,58	4,5

Table 5: Added services preferences

Next, the respondents were asked about the acceptable cost level compared to the printed newspaper (Table 6). In general, the test persons from the evaluation are more inclined to pay the same price or higher than the respondents from the survey. Some of the test persons mentioned that the e-newspaper had to be cheaper than the printed edition due to the saved printing costs for the publishers. Others mentioned that the price had to be lower than the printed edition, due to less content in the e-newspaper.

<i>What cost level is acceptable for you to change to an e-newspaper?</i>	<i>Survey</i>		<i>Test persons</i>	
	<i>No</i>	<i>%</i>	<i>No</i>	<i>%</i>
Cheaper than the traditional printed newspaper	2119	58,4	5	45,4
Same price	371	10,2	4	36,4
Can be more expensive if there is added value	281	7,7	2	18,2
Price is unessential	171	4,7	0	0
Missing	684	18,9	0	0
Total	3626	100	11	100

Table 6: Acceptable cost level

Thereafter the respondents were asked how they think the e-paper device should be financed (Table 7). The most preferred model in both studies is inclusion in subscription.

<i>How do you think the e-paper device should be financed/ paid for?</i>	<i>Survey</i>		<i>Test persons</i>	
	<i>No</i>	<i>%</i>	<i>No</i>	<i>%</i>
Hire-purchase	281	7,7	2	18,2
Buy the device	711	19,6	1	9,1
Included in subscription	1708	47,1	8	72,7
Other	592	16,3	0	0
Missing	334	9,2	0	0
Total	3626	100	11	100

Table 7: Finance of device

Further, they were asked about their willingness to exchange the traditional newspaper with the e-newspaper in the future (Table 8). Surprisingly, all test persons answered yes compared to two thirds of the respondents from the survey.

<i>Would you consider to, some time in the future, exchange your traditional printed newspaper for the e-newspaper?</i>	<i>Survey</i>		<i>Test persons</i>	
	<i>No</i>	<i>%</i>	<i>No</i>	<i>%</i>
Yes	2431	67,1	11	100
No	1070	29,5	0	0
Missing	125	3,4	0	0
Total	3626	100	11	100

Table 8: Willingness to exchange the traditional newspaper

Moreover, they were asked within which time frame they would be ready to read their newspaper on e-paper (Table 9). In both studies there were surprisingly many that were prepared to exchange today or within five years. However, 25% of the respondents in the survey did not answer this question indicating that it was difficult to decide.

<i>Within which time frame are you ready to read your newspaper on e-paper?</i>	<i>Survey</i>		<i>Test persons</i>	
	<i>No</i>	<i>%</i>	<i>No</i>	<i>%</i>
Today	1521	41,9	6	54,5
Within 5 years	683	18,8	5	45,5
Within 10 years	209	5,8	0	0
Within 20 years	88	2,4	0	0
Never	223	6,2	0	0
Missing	902	24,8	0	0
Total	3626	100	11	100

Table 9: Time frame

Finally, the survey respondents were asked about the factors that influenced their decision to read their newspaper on e-paper (Table 10). *Stable technology* and *Easy to find content* were the factors that scored the highest. The least influencing factor was *Observability of use*. The test persons were asked similar questions in the interviews.

<i>How important are the following factors for you choosing to read your newspaper on e-paper?</i>	<i>Survey</i>	
	<i>Mean</i>	<i>Std dev</i>
The appearance of the device	4,0	2,72
Continuous updates of news	4,8	2,80
Added functions such as chat	3,1	2,46
Easy to use and handle	4,6	2,81
Stable technology	4,9	2,81
Observability of use	1,9	1,87
Environment friendly	3,8	2,72
That it is the latest technology	3,0	2,41
Easy to find content	4,9	2,82

Table 10: Influencing factors

One of the major concerns of the test persons was the navigation of the e-newspaper that was very constrained by the device. Almost everyone mentioned that the navigation needed to be improved if they should consider exchanging the printed edition with the e-newspaper. The other most mentioned issue that needed to be addressed was the refresh rate of the display, i.e. it needed to update faster in order to create a pleasant reading experience. Some but not all test persons regarded color as essential for exchanging the e-newspaper. One of the test persons expressed: "*Color would be fun, but it is nothing that I prioritize as essential, I would exchange it even if it is not in color*". News updates during the day was also found important by several of the respondents.

6 Discussion and Conclusion

In this paper we have addressed the research question: *How does use experience influence perceptions of preferences and demands for the e-newspaper?* We did so by testing the hypothesis that users confronted with a vision of new technology and services are more positive to adopt than users with actual use experience of technology and services in an early stage of development with inherent technology problems. This hypothesis proved to be false, in this case it was the other way around. In spite of the e-paper device technical constraints and the early prototypes, the test persons in the evaluation were more positive to the e-newspaper. On the one hand, the respondents in the survey were not able to experience the actual e-paper technology, they could only read about it and watch concept videos of the future e-newspaper vision and interact with e-newspaper prototypes to get an understanding of the concept. On the other hand, the test persons in the evaluation who experienced the e-paper technology first hand, also experienced the bugs and limitations in this early stage of technology. The e-newspaper prototypes available for online experience in the survey were all in color and had well functioned navigation systems and interaction possibilities whereas the Sundsvalls Tidning in the evaluation was presented in 16 grey scales and had limited navigation options and interaction possibilities due to limitations with the technology. When we set out to test the hypothesis we believed that the respondents in the survey that were confronted with a vision of a multimedia e-newspaper in color with added services would be more positive. Even so, there are similar patterns in preferences and opinions in both studies. The relations between reasons and importance of different added services were very alike, indicating the relevance of the results in both studies. We can conclude that our findings are in line with previous research, i.e. that experiencing technology and services in different ways and in different situations have impact on intentions to use. Even though the respondents in the second study were exposed to technical and navigational difficulties they were very positive towards adopting the e-newspaper. We believe that trying the e-newspaper during two week in their everyday setting as well as being able to experiment with the services in the actual e-paper technology have contributed to this positive attitude.

By empirically testing potential adoption of the e-newspaper with two different approaches, we have contributed to the overall understanding of new media adoption in general and the promotion of the e-newspaper in particular. To summarize the findings according to newspaper organizations preparations for launching the e-newspaper, the following can be derived: The e-newspaper need to contain archive functions. Providing added value by personalization and by offering community information would increase the potential adoption. If the newspaper organizations offer added value according to the audience preferences, they could expect the same willingness to pay as for the printed edition. However, before launching the e-newspaper, the navigation has to be improved as well as the refresh rate of the display. As most respondents preferred to have the device financed by inclusion in the subscription, this could be considered as the initial alternative. Finally, as almost half of the respondents stated that they were ready to start reading the e-newspaper already today, it is time for the newspapers to start preparing for the e-newspaper introduction.

Further research includes a major e-newspaper test in real life settings with 5 Swedish newspapers and 50 families at different locations in Sweden. This test will build on the results from the studies presented in this paper and will include more added value and more services.

References

- [1] TILSON, D.; LYYTINEN, K. AND BAXTER, R. A Framework for selecting a Location Based Service (LBS) Strategy and Service Portfolio, in Proceedings of the 37th Annual Hawaii International Conference on System Sciences, Hawaii. (CD-ROM), Computer Society Press (10 pages), 2004.
- [2] ANCKAR, B.; AND DÍNCAU, D. Value-Added Services in Mobile Commerce: An Analytical Framework and Empirical Findings from a National Consumer Survey, in Proceedings of the 35th Annual Hawaii International Conference on System Sciences, Hawaii. (CD-ROM), Computer Society Press (10 pages), 2002.
- [3] KHALIFA, M., AND CHENG, S.K.N. Adoption of Mobile Commerce: Role of Exposure, in Proceedings of the 35th Annual Hawaii International Conference on System Sciences, Hawaii, (CD-ROM), Computer Society Press (7 pages), 2002.
- [4] HONG S-J.; TAM K. Y.; KIM J. Mobile data service fuels the desire for uniqueness, Communications of the ACM, Vol 49 No 10, 2006, pp. 89-94.

- [5] CARLSSON, C.; HYVÖNEN, K.; REPO, P.; WALDEN, P. Asynchronous Adoption Patterns of Mobile Services, in Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Hawaii, (CD-ROM), Computer Society Press (10 pages), 2005.
- [6] MALHOTRA, A.; SEGARS, A. H. Investigating Wireless Web Adoption Patterns in the U.S. Communications of the ACM, Vol. 48, No. 10, 2005, pp. 105-110.
- [7] KNUTSEN, L. A.; CONSTANTINOU, I.D.; AND DAMSGAARD, J. Acceptance and Perceptions of Advanced Mobile Services: Alterations during a Field Study, in Proceedings of International Conference on Mobile Business, Sydney, 2005, pp. 326-332.
- [8] KNUTSEN, A. M-Service Expectancies and Attitudes: Linkages and Effects of First Impressions, in Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Hawaii, (CD-ROM), Computer Society Press (10 pages), 2005.
- [9] SARKER, S.; WELLS, J.D. Understanding mobile handheld device use and adoption. Communications of the ACM, Vol 46, No 12, 2003, pp. 35-40.
- [10] 30 Countries Passed 100% Mobile Phone Penetration in Q1. Available at: http://www.telecommagazine.com/newsglobe/article.asp?HH_ID=AR_2148 (January 30th, 2007)
- [11] CARLSSON, C.; CARLSSON, J.; HYVÖNEN, K.; PUHAKAINEN, J.; WALDEN, P. Adoption of Mobile Devices/Services – Searching for Answers with the UTAUT, in Proceedings of the 39th Hawaii International Conference on System Sciences, Hawaii. (CD-ROM), Computer Society Press (10 pages), 2006.
- [12] HAMMOND, K. B2C e-Commerce 2000-2010: What Experts Predict. Business Strategy Review, Vol. 12, 2001, pp. 43-50.
- [13] IHLSTRÖM, C.; ÅKESSON, M.; NORDQVIST, S. From Print to Web to e-paper - the challenge of designing the e-newspaper, in Proceedings of ICCO 8th International Conference on Electronic Publishing, ELPUB 2004, Brasilia, 2004, pp. 249-260.
- [14] IHLSTRÖM, C. The e-newspaper innovation - converging print and online, in Proceedings of the International Workshop on Innovation and Media: Managing changes in Technology, Products and Processes, Stockholm, 2005.
- [15] Belgium: e-paper test launch. Available at: http://www.editorsweblog.org/news/2006/02/belgium_epaper_test_launch.php (February 3rd, 2007)
- [16] What is the iLiad? Available at: <http://www.irextechnologies.com/products/iliad> (April 10th, 2007)
- [17] E-paper E-merging. Available at:
- [18] <http://www.signonsandiego.com/news/computing/personaltech/20060123-9999-mz1b23epaper.html> (January, 24th, 2007)
- [19] Sony Reader. Available at:
- [20] http://products.sel.sony.com/pa/prs/reader_features.html (February, 16th, 2007)
- [21] Oak Investment Partners invests in Plastic Logic. Available at: <http://www.idtechex.com/printelecreview/en/articles/00000399.asp> (December 19th, 2006)
- [22] BUCHANAN, T.; SMITH, J. L. Using the Internet for psychological research: Personality testing on the World-Wide Web. British Journal of Psychology, 90, 1999, pp. 125-144.
- [23] PATTON, M. Q. Qualitative Research & Evaluation Methods (3 ed.). Sage Publications, Inc. California, 2002.
- [24] ROGERS, E. M. Diffusion of innovations. New York, The Free Press, 1995.
- [25] DAVIS, F. D.; BAGOZZI, R. P.; AND WARSHAW, P. R. User acceptance of computer technology: A comparison of two theoretical models. Management Science, 35(8), 1989, pp. 982-1003.
- [26] MALLAT, N.; ROSSI, M.; TUUNAINEN, V. K.; ÖÖRNI, A. The Impact of Use Context and Mobility on the Acceptance of Mobile Services, in Proceedings of the 39th Hawaii International Conference on System Sciences, Hawaii. (CD-ROM), Computer Society Press (10 pages), 2006.

Centralized Content Portals: iTunes and the Publishing Industry

Matthijs Leendertse; Leo Pennings

TNO Information and Communication Technology, Department: ICT & Policy
Brassersplein 2, 2600 GB Delft, the Netherlands
e-mail: matthijs.leendertse@tno.nl; leo.pennings@tno.nl

Abstract

This paper addresses new questions around media performance as a result of the rise of centralized content portals such as iTunes or MySpace. We first describe the rise of centralized content portals in different media industries, and discuss how these portals are creating a dominant position for themselves by using lock-in strategies. Then we describe the concept of media market performance, and discuss two important media performance concepts: access and diversity. Using scenario analysis, this paper describes three learning scenarios that outline the effects of different configurations of centralized content portals on behavior of publishers, users and advertisers, and through that content on access to and diversity of content.

Keywords: content portal; iTunes; access model; scenario analysis

1 Introduction

Centralized content portals are platforms that allow third party content suppliers to offer their digital content products to consumers. These platforms are typically developed by firms that are not rooted in the content industry, and as a result are based on business models that do not necessarily revolve around content itself. Mobile operator portals for instance have been developed by mobile operators to facilitate content owners to sell content to users of mobile devices, with the intention of promoting their mobile internet services. NTT DoCoMo has been very successful with this concept in Japan with over 47 million subscribers in February 2007. The i-mode ecosystem allows mobile phone users to access a mobile portal where third party content providers that – abiding the standards of NTT DoCoMo – offer their content to end users. This content is charged on the mobile phone bill and NTT DoCoMo takes a cut of the revenues. Within Europe and the US, mobile operators have tried to mimic NTT DoCoMo's offering, but never succeeded in attracting significant groups of users.

One of the most prominent examples of centralized content portals in our part of the world has been Apple's iTunes Store. The stellar success of the iPod and its easy to use music software iTunes prompted Apple to start selling music through its iTunes Store. This centralized content portal is integrated in the iTunes software. With this integration between hardware, software and a centralized content portal, Apple proved to record companies that digital music could be monetized. One of the strengths of the iTunes Store has been the seamless integration of software (iTunes) and hardware (the iPod), allowing for purchases in the iTunes store to be immediately transferred to the iPod when the user connects its iPod to the PC. The sale of music is not particularly profitable for Apple as most of the revenues have to be transferred to the right holders, mostly to record companies. Apple's profits come from selling the hardware (iPods).

So what do these centralized content portals have in common? First and foremost they increase the ease of finding, selecting, purchasing and distributing digital content for end users. Second, these portals are usually not designed around the requirements of content suppliers. Third, these portals tend to exclude rival services. In other words, users of NTT DoCoMo's i-mode service or iPods are disabled or discouraged to access rival centralized content portals. The integration of hardware, software, payment mechanisms and in the case of mobile operator portals also network connectivity, effectively locks users into a centralized content portal.

2 Natural Tendency Towards Dominant Platforms

“What you see on the internet is very much the 'Highlander Theory'. There can be only one. There's only one search engine, there's only one big book retailer, one big online auction house, and so on. That's not necessarily a bad thing, as long as it's able to supply a super hot service at a reasonable price” (Greg Eden of AIM Digital in the Guardian Newspaper of August 17, 2006).

Media markets have a natural tendency towards concentration. The Dutch Media Authority labelled this tendency as the law of 3; in each media market (e.g. television, news or book publishing) in the Netherlands there were 3 large media companies competing [1]. Within the digital media sphere, there is not so much a tendency towards concentration in terms of content creators or packagers, but more in the sphere of facilitating services. Microsoft Windows is by far the dominant operating system; Google by far the dominating search engine (in particular in Europe) and iTunes dominates the paid for music downloads market. As Eden notes, indeed there can only be one.

Media technology that is used to distribute digital media to consumers can be divided into 4 main categories: PC based online services, digital television based services, Game Consoles and Mobile Platforms [2]. In all 4 categories we find evidence for the Highlander theory: a natural tendency towards a dominant standard:

- PC based online services: here content is downloaded over the open internet, and is typically accessed on a PC but also on audiovisual equipment using the PC as the central server. Apple pioneered this market with its iTunes Music Store and established a dominant position on the market for paid for music downloads. We can see the publishing industry is following this strategy. In 2006, Sony launched the 'Sony Reader', a device that allows users to access digital text through a device that provides an experience similar to paper. Sony copied Apple's strategy for digital music, and bundled their Sony Reader with their so called 'Sony Connect Store' offers access to (premium) e-publishing products. This store is integrated in the software that synchronizes the content on the end user's PC with the Sony Reader, again similar to iTunes.
- Digital Television based services: the digital television platform is increasingly used to distribute content other than video. Subscribers to a digital television service can only use the set top box of their supplier, and hence only access the (often third party) content these DTV suppliers offer. In other words, the digital television platform of the supplier (e.g. a cable or satellite company) is effectively the dominant standard for content for subscribers of that platform.
- Game consoles: all three so-called next generation game platforms – Microsoft's Xbox 360, Sony's PlayStation 3 and Nintendo's Wii – feature an internet connection and have some sort of content offering beyond traditional gaming. Nintendo's Wii for instance, offers a news channel where stories are linked to a map of the world and end users can search for content by scrolling the globe. Content for now is provided by the Associated Press and the service is offered for free. On its Xbox 360 platform, Microsoft offers a vast array of television and movie content to its US based clients and outlined plans to distribute the service in other territories as well. As owners of game consoles typically do not own game consoles of other brands, these users can only use the platform of Sony, Nintendo or Microsoft.
- Mobile platforms: mobile devices are increasingly sophisticated and have increasingly access to mobile data networks such as UMTS or WiFi 802.11. One of the main advantages of mobile products is that it can easily recycle existing content into a market where people are more willing to pay for access to information. However, yet again there are dominant standards emerging. For instance, many operators have mobile operator portals through which third party content can be searched, purchased and downloaded. The strongest asset of mobile phone companies with regards to mobile content is their payment system and payment relation with their consumers. Third parties that want to are dependent on the mobile phone company, that holds a near monopoly. Vodafone for instance discourages its subscribers from using other mobile content than offered through their Vodafone Live! mobile operator portal by charging for all data traffic outside the Vodafone Live! domain.

That begs the question as to why the rise of such dominant Highlander-esque portals is an important topic.

3 The Power of Lock-In

“The European Commission can confirm that it has sent a Statement of Objections to major record companies and Apple in relation to agreements between each record company and Apple that restrict music sales: consumers can only buy music from the iTunes' on-line store in their country of residence. Consumers are thus restricted in their choice of where to buy music, and consequently what music is available, and at what price. The Commission alleges in the Statement of Objections that these agreements violate the EC Treaty's rules prohibiting restrictive business practices (Article 81)” [3].

As the quote above demonstrates in the case of the European Commission against Apple and several record labels, dominant content portals can damage the interests of citizens. Suppliers of these content portals – be they Apple or Vodafone – can for instance determine pricing, availability, ranking, disclosure and usage restrictions on the content sold through their portals. This case of the EC against Apple is one of many. The Norwegian

Ombudsman for instance declared on January 24th 2007 that the iTunes Music Store is illegal because it only allows purchased music to be played on Apple's iPod devices. Rivalling MP3 players cannot be used to play the purchased content, effectively locking-in customers to their proprietary system. Such a lock-in strategy intends to prevent buyers from turning to alternative suppliers. For suppliers such strategies seem advantageous because lock-in allows them to raise prices without having to invest in innovation or product quality. Consumers however are dependent on one supplier, and their interests are potentially endangered. In addition, competitors of Apple are effectively restricted from engaging in competition at all, raising questions with regards to competition policy.

Lock-in strategies have existed for a long time within the media industry, mostly in the form of subscriptions, and are on the increase due to new media technologies. The most common lock-in strategies are contractual agreements, taking advantage of durable purchases that demand complementary compatible purchases in a later stage, supplying products that demand brand-specific training, developing propriety information and database standards that are not compatible with other databases, becoming the specialized supplier for specific products and offering loyalty services. [4, p. 117]. Most of these strategies involve increasing the costs for consumers to switch to an alternative supplier. This is especially apparent in markets where suppliers have the exclusive rights to a particular technology or system. Because digital media products involve many different layers of the communication system, suppliers can develop proprietary technology to take advantage of this. It is therefore not strange that within digital media, we see this tendency towards centralized content portals that effectively lock-in their users. The next natural question is how we can assess the effects of these portals.

4 Assessing Consequences of Centralized Content Portals

So how can we assess the effects of such centralized content portals on digital media markets? Within media economic theory, these effects can be studied by using media market performance criteria. The concept of market performance comes from welfare economics, where performance is traditionally assessed by economic indicators such as allocative efficiency and industry profitability ratios [5, 6]. Media economic scholars have adapted the notion of market performance to assess media markets. Rather than focusing solely on economic indicators, media market performance is assessed by social, political and cultural indicators such as media diversity, freedom of expression, access to media outlets and services are the most prominent [5, 7-9]. Media market performance could be described as an assessment of mass media from a public interest perspective [5, p. 62].

It is important to outline that market performance refers to the outcome of the total market, and should not be mistaken with the performance of individual firms or other actors. "Performance is, first and foremost, appraised with reference to a market, which comprises all the interacting buyers and sellers as a whole, rather than to individual economic agents such as firms" [10, p. 4]. The normative approach to market performance proposes several performance indicators to assess whether markets deliver what society wants [6].

Media policy based on media performance is based on media performance assessment, and typically concentration within the media is not greeted with great enthusiasm. The rise of centralized content portals that effectively lock-in consumers raises new questions around familiar media performance criteria, most notably access and diversity.

5 Access

Accessibility has also been an important criterion for media regulators, and has become more prominent in debates on the future of media policy. In contemporary media policy, access to communications is an central concept [8]. Access to communications can be defined as "the possibility for individuals, groups of individuals, organizations and instructions to share society's communications resources" [11, p. 204]. Access can be looked upon from different perspectives, such as access to markets or consumers or access to content by users. An important access performance indicator for users is affordability of content [8]. With an abundance of content available, access to that content is becoming increasingly important. Issues such as media education for groups that do not have the skills to access this content, households that do not have access to new media devices and infrastructures, but also economic accessibility as some content can only be accessed through payment.

Also, the availability of content is an important indicator of accessibility. For content suppliers, access to markets is an important factor. When for instance Apple would prevent some record companies from selling their content through iTunes, a large part of the digital music buying audience is shielded off from this content. This is not only detrimental to the record company in question, but also to the audience as it limits their access to

content. With regards to publishing products, in particular news, educational and professional content, accessibility is an even more important issue. In light of the democratic and socio-cultural functions that media have apart from economic functions [12], it is important that citizens of democratic societies have access to information. When for instance a centralized portal for news content would be as dominant as the iTunes Store, restricting access of content suppliers to this platform could seriously undermine the democratic process.

In addition to publishers of information, access to these platforms is also important for advertisers. We see that many platforms are replacing traditional advertising outlets as the main facilitator of advertisements. Online, we see that Google is now dominating the advertising industry using its AdSense and AdWords advertising network. Within the digital television domain, we see that cable companies and IPTV providers moving towards the advertising market as well, taking over the roles of traditional broadcasters [13].

6 Diversity

Diversity is perhaps an even more important media market performance criterion in Western countries. The adjectives to diversity in government reports usually reflect the desired media performance: cultural diversity, opinion diversity, regional diversity, genre diversity, ethnic diversity etc. The concept of media diversity could be defined as the heterogeneity of the media [14, 15]. McDonald and Dimmick [15] argue that diversity is a two dimensional construct: [1] a set of categories within a given distribution (e.g. content categories) and [2] the allocation of elements to these classifications (e.g. how many programs are devoted to the content category news).

The concept of media diversity can be deconstructed into three distinct forms of media diversity: source diversity, content diversity and audience exposure diversity [16]. Source diversity refers to the number of media outlets (TV channels, newspapers) and the ownership structures of these outlets and is traditionally measured using economic measures for competition such as the HHI index or Competition Ratios. Content diversity refers to actual media supply, and is mostly assessed by content analysis studies. These studies typically classify media content into predefined categories, for instance into content categories. Most policy research around diversity assumes that audiences provided with a diversity of content options also consume a diversity of content. However, the mere availability of diverse information does not necessarily equal exposure to diverse opinions and information. Without audience exposure to diverse content, availability of content has no effect on the political and socio-cultural functions of media. Exposure diversity is defined as “the diversity of content or sources consumed by audience members” [16]. Many indicators for media diversity have been formulated, as listed below:

- Media should reflect the various social, economic and cultural realities of the societies in which they operate, more or less proportional.
- Media should offer more or less equal chances of access to the voices.
- Media should serve as a forum for different interests.
- Media should offer relevant choices of content at one point in time and also variety over time. [17].

With the increased importance of centralized content portals, it is important to reassess to what extent these diversity indicators are being met.

7 Scenario Methodology

To hypothesize the effects of different configurations of content portals, we conducted a scenario exercise. The main purpose to develop scenarios is that they should paint distinct different pictures of the future with unique implications for strategic decision-making [18, 19]. We use secondary sources and media economic theory to develop the scenario lines [20]. In our case, we wanted to develop scenarios that explain the effects of open or closed and commercial and non-profit content portals on access to and diversity of e-publishing products. Therefore, we developed three sets of conditions for the scenarios, each with their own unique configuration of content portals:

1. One dominant content portal similar to the iTunes Music Store, owned by a provider of e-publishing hardware (often referred to as e-readers);
2. Several interoperable open content portals that have been developed by commercial search engines and software companies;
3. A web full of Wikis with freely accessible co-created content, in tandem with securely sealed off walled garden of traditional publishers.

In each of the scenarios we established the so-called rules of interaction [21]. These rules outline how the important actors respond to the above mentioned conditions. The actors that we included in the scenarios are based on the value chain: content creators, content packagers (publishers) and content distributors [12]. In addition, we included advertisers as an important actor for the simple reason that many consumer based publishing products are dependent on advertising revenues. Based on this, we developed a linear story line that depicts not only the conditions and rules of interaction, but also the consequences for access to content and the diversity of content. In order to keep the scenarios clear, the broader conclusions are discussed in a separate concluding paragraph. We choose a time path of 5 years for the scenarios, placing them in the year 2012. This time path was selected because the technological progress is advancing at such a rapid pace that looking further into the future (e.g. 10-15 years) would be near impossible.

These scenarios aim to help policy makers to assess the performance of future e-publishing markets based on content portals, and support the policy-making process. Policy is based on beliefs around the future benefits of content. That is why policy makers often conduct ex ante policy assessments, whereby the effects of different policy options are weighed [22]. Managers within e-publishing companies can use these scenarios to evaluate the strategic position of their organizations under different configurations of content portals, and can hence formulate counterstrategies [21].

8 Scenarios

Please find below three scenarios, in which we discuss the effects of different configurations of content portals on the behavior of the main market actors, and hence on accessibility and diversity of content. The scenarios do not try to describe utopist visions, or intend to prescribe one future as the best. Assessment of these scenarios is a normative matter for policy makers and an economic / strategic question for commercial publishers.

9 IPUB (Proprietary Standard)

The year is 2012. After the stellar success of the iPod, Apple has ventured into other content markets as well. In 2008, the company launched a hard disc based text reader dubbed the iPub. Within 4 years, this device has captured 85% of the market for digital reading devices. Yet again, Apple proved that user friendly hard- and software can significantly increase end users' appetite for digital content. In all major cities, people are reading their iPubs in public transport, in restaurants and increasingly also in schools and offices. Traditional publishers are now selling their content directly to the iPub, making use of its Wimax wireless connection. Digital newspapers or magazines are directly downloaded to the iPub when the user has a subscription. The iTunes store can be accessed on both the PC / Mac and on the device itself and offers the largest collection of e-publishing products in the Western world. In terms of sales, the iTunes store has overtaken Amazon.com, Barnes & Noble and Bol.com as the largest reseller of e-publishing products in both the US and Europe. The iTunes store provides copyright protected e-publishing products that can only be viewed on an iPub. In terms of prices, Apple has set fixed prices for different product types so that users are confronted with a simple to understand pricing mechanism. Books for instance can be priced at €5, €10, €15 or €20 in the German iTunes store. These prices are set by Apple, and individual suppliers that want to sell their products using the iTunes store have to abide their pricing scheme. Apple gets a fixed share of 20% on all sales. In exchange the company deals with all the handling, the platform and the payment mechanisms. Because of the dominant position of Apple's iPub on the e-reader market, Apple controls the dominant platform for premium e-publishing products. Other OEMs have significantly less power, and are therefore not able to attract content owners, let alone dictate pricing and format standards to them.

Traditional publishers dominate the premium e-publishing products in iTunes. They have signed agreements with Apple to distribute their products, and sometimes demand certain special features such as additional promotion in iTunes and exclusion of certain rival products. Smaller or even individual content creators such as journalists or writers lack the bargaining power of the large publishers. As a result, they pay higher royalties to Apple for selling their e-publishing products and find it difficult to negotiate special deals or promotional activities. Although Apple promotes itself as a corporate responsible company, it remains a commercial company and money logic dictates that Apple will give priority to the large publishers, and put the squeeze on their smaller rivals.

Publishers can also use the iTunes store to provide ad supported content for free. Because Apple does not earn money on freely distributed content by its cut on each purchase, it charges for delivery of ad supported content to end users through its platform.

The dominant position of the iTunes Store for e-publishing products makes it a walled garden in which Apple dictates standards with regards to content packaging, price setting, copyrights technology and disclosing and distributing content. Large publishers have preferred access to the iTunes Store and are more heavily promoted by Apple. Smaller suppliers find it more difficult to access this platform. As a result, many niche e-publishing products are not present or hard to find in the iTunes Store. It is primarily the mainstream content that is promoted, similar to the music offering in the iTunes Store. Already in 2006, the Guardian newspaper reported that: "(...) iTunes does sell a reasonable volume of niche music, but as a mainstream music retailer, it markets to and mostly attracts mainstream music fans" [23]. Four years after the introduction of the iPub, the same has happened for e-publishing products. Sure, the iTunes store contains a diverse range of content, but effectively the content offering contains of mainstream mass media e-publishing products. Consumers that do not own an iPub or do not have a iTunes Store account are barred from many e-publishing products as these are exclusively available on the iTunes Store. Many publishers use the iTunes Store as their sole or prime distribution platform for e-publishing products, not in the least because Apple's pricing policy allows them to set relatively high prices for their e-publishing products. The relatively high prices combined with the enormous reduction in printing and distribution costs make an interesting business case for publishers.

10 Interoperable Content Portals

It took some time, but in the three years from 2009 up till now, e-publishing has finally matured. Several open content portals have been developed that are used by a myriad of content suppliers. These platforms are truly cross-medial, i.e. they can be accessed on the open internet, on mobile devices and on e-readers with network connectivity. Search engines and software providers are the main drivers behind these open platforms. These facilitators use their content portals to push their payment systems and advertising networks. Google, Paypal (eBay) and Nokia Software are the three leading providers of such platforms in Europe, and offer a payment system for purchases of e-publishing products. These payment systems can also be incorporated in other websites, so content providers can also offer, sell and distribute content through their own portals.

International open source standards are used for content packaging, copyrights and metadata in order to optimize interoperability and retrievability of e-publishing products over various portals. Content owners can easily distribute their content to the different portals, and the open character of these content portals ensures that also small publishers can find their way to end users. As content is organized along the lines of standard metadata sets, the size of the publisher does not influence the degree of retrievability on the large portals. However, these larger publishers also have an extensive online presence themselves, which makes up almost 40% of their turnover in the e-publishing domain. For these publisher websites they often prefer to use the payment system and advertising networks of facilitating companies over proprietary systems.

The large content portals for their part are used by a large variety of publishers, from individual authors to multinational media conglomerates. Standardized open platforms with standardized copyright protection reduce transaction costs [24, 25]. Since the size of a firm depends foremost on the question whether it will pay to bring an extra exchange transaction under the organizing authority of a firm [26], lower transaction costs as a result of standardization would dampen the economic case for organizing transactions within a firm. In other words, the *raison d'être* for large publishing houses as content packagers is becoming rather questionable. More and more content creators bypass the publishing houses and directly market their products using open content portals.

Because these platforms are open, several non profit organizations are also accessing the market. This results in a negative price spiral. For instance, the rivalry between the free news e-publishing products of public broadcasters such as the BBC and ZDF and those of commercial newspaper has driven most quality papers to an advertisement supported free e-newspaper model.

The large content portals use the traditional motto of telecommunication providers that they are not interested in the message, but only in the facilitation of the message. This content agnostic view has allowed them to incorporate a myriad of content suppliers from various backgrounds, such as political parties, student organizations, individual authors or educational publishers. As a result, end users have access to a very diverse pallet of e-publishing products.

The high level of competition between suppliers drives down prices of e-publishing products, which in turns increases dependency on advertising revenues. This benefits the advertisement networks of the companies behind these content portals, such as Google's AdSense and Nokia's Ad Accelerator. Especially the smaller publishers do not have the means or the know-how to sell advertisements for their products, and can easily tap into the systems of the large advertisement networks. We predict that in April 2017, 5 years from now,

advertisements will make up 70% of the income of e-publishers. Because the suppliers of centralized content portals monitor user behaviour on the portals, but also their reactions to certain types of information, they can greatly increase the effectiveness of advertisements in e-publishing products, which has a positive effect on demand for and prices of advertisements in e-publishing products.

11 iWiki-publishing

It is hard to imagine that only 5 years ago, Wikipedia was primarily an online encyclopedia. We have come a long way since then. The managing board of Wikimedia declared on January 4th 2012 that the Wiki format has now become the dominant method of publishing for digital text, photo and audiovisual content. Although the large publishing companies still play an important role with their print and online propositions, in terms of digital content Wikis are the undisputed market leaders in Europe, the US and East Asia in all major digital content categories, but especially when it comes to news content. Recognizing the importance of the Wiki movement, iWiki-publishing has been added to the Wiki and Oxford Dictionaries as:

“i-wi-ki-pub-lish-ing Pronunciation [ai- wee-kee-pub-lish-ing] –derived from a verb. Meaning: an e-publishing product that has been created on or for an open platform and is continuously subject to changes from visitors”.

As there is no central organization that creates or packages iWiki publishing products, there are no business models around this type of content. When Wikis were still relatively unimportant in the media industry, many non-profit organizations started to use this mechanism. Public broadcasters were particularly instrumental in promoting the Wiki concept, and en masse started to use Wiki from 2008 on. As a result, there is a large variety of iWiki publishing products and these are not only free, but also void of advertisements. The servers and sites that host the Wikis are maintained by donations from individuals. These servers and sites are non-profit foundations with a democratically elected managing board. Wikis are always non-profit, because it is impossible to divide the proceedings of co-created content over all the co-creators.

This leads us to the problem of iWiki-publishing. As there are no business models, it is hard to find sufficient funding for more specialized or professional content. Investigative journalism for instance requires relatively large investments, with journalists sometimes having to infiltrate organizations for prolonged period of time. The traditional print publishers are aware of this problem and offer their often more specialized content for free to subscribers of their printed material. They have created online walled gardens where subscribers can access the digital version and extras around the printed product. By doing so, these traditional publishers effectively defend their positions in print and lure users with a need for specialized content to their printed products. Product innovation for these more specialized forms of digital content such as investigative journalism are dependent on innovation in the portfolio of printed material that these traditional publishing houses have. Because they have a semi-monopoly on specialized content, prices for printed material are increased in order to make up for costs of their digital offering. This started at the beginning of the century with educational publishers that increased prices of text- and workbooks to cross subsidize investments in ‘free’ digital content [27], and has now spread to all e-publishing segments.

For advertisers it has become difficult to reach audiences of e-publishing products. They are restricted to the walled gardens of the traditional publishers, and these only give access to a large minority of the total audiences. Although these audiences are interesting for many advertisers (the average subscriber to printed publications are more affluent than the average reader of free iWiki-publishing products), e-publishing has become less interesting to advertisers than other media. This also benefits the traditional publishers, as they are often part of a larger media conglomerate that can utilize alternative media to lure advertisers to their platform. Nevertheless, access to high quality, specialized e-publishing content has become more difficult and content innovation is dependent on end users.

All in all, Wikis have greatly benefited access to content distribution platforms for individual authors and provide a diverse supply of digital content. However, the counter reaction is that specialized digital content is now more disclosed than ever and less affluent consumers might have difficulties to access this content.

12 Conclusions

The answer to the question to what extent different centralized content platforms are beneficial or detrimental to the e-publishing sector, is dependent on the subjective evaluation of the person who asks. Publishers have different interests than politicians, and we can be sure that politicians rooted in different political ideologies also

differ in their evaluation criteria. However, it is safe to say that centralized content portals will most likely change the way e-publishing products are packaged, disclosed and distributed.

In the first scenario, we found benefits for publishers that are allowed access to the iTunes store. Prices are relatively high, whereas the percentage that has to be paid to iTunes in return for facilitating the sale of e-publishing products small in comparison with for instance traditional bookstores. Consumers might evaluate this scenario less favorably, as they are locked into hardware and can choose from a pre-selected set of content for relatively high prices. Governmental agents might also be less than happy, especially given the lack of competition.

The second scenario might be less interesting for publishers, as it remains difficult for them to develop profitable services in a very competitive environment. Publishers that want to attract users must invest in quality and innovation, because only then users will be persuaded to pay for access to e-publishing products. Smart use of the networks of the facilitators might help smaller publishers to develop profitable business models based on the advertising market. Consumers have the best deal in this scenario, because they can select content from an infinite number of sources, and prices are very subdued. Regulators might also look favorable upon this scenario, as competition is high.

The final scenario is a more scientific approach, whereby knowledge is not created by any organization but by a collective of co-creators. Publishers won't cheer for this scenario, since it is near impossible to create profitable e-publishing products. Rather, e-publishing has become a by product for printed material. There is no such thing as a free lunch, so money has to come from alternative sources. For consumers it is also perhaps a less positive scenario as access to specialized content is somewhat restricted and the relevance and authenticity of content can be questionable. Governments should also consider to what extent this type of market, where digital content is either free or a by product, is desirable for their policy goals.

These scenarios intend to help policy makers and strategists within publishing companies think about the future impact of centralized content portals. Going forward more empirical research into the effects of these portals on media market performance is required. Especially the reaction of e-publishers to these portals should be further analyzed. The next step of the research would be to develop quantitative models that predict reactions of e-publishers to the rise of centralized content portals. This enables us to empirically test the validity of the predictions made in the scenarios.

Acknowledgements

This paper has been written within the framework of the FLEET project, a multidisciplinary research project on Flemish e-Publishing Trends that is financed by IWT.

References

- [1] Commissariaat voor de Media, *Mediaconcentratie in Beeld. Concentratie en Pluriformiteit van de Nederlandse Media 2001*. 2002.
- [2] VAN WOLFINKEL, R.; LEENDERTSE, M. *Deploying Broadband Services in a Competitive Environment*. in *Broadband Europe*. 2006. Geneva, Switzerland.
- [3] Commission of the European Communities, *Competition: European Commission confirms sending a Statement of Objections against alleged territorial restrictions in on-line music sales to major record companies and Apple*. URL: (<http://europa.eu/rapid/pressReleasesAction.do?reference=MEMO/07/126&format=HTML&aged=0&language=EN&guiLanguage=en>) [Accessed: 06-04-2007].
- [4] SHAPIRO, C.; VARIAN, H.R., *Information Rules - a strategic guide to the network economy*. 1999, Boston: Harvard Business School Press.
- [5] HENDRIKS, P., *Communications Policy and Industrial Dynamics in Media Markets: Toward a Theoretical Framework for Analyzing Media Industry Organization*. *The Journal of Media Economics*, 1995. 8(2): p. 61-76.
- [6] SCHERER, F.M.; ROSS, D., *Industrial Market Structure and Economic Performance*. 1990, Boston: Houghton Mifflin Company.

- [7] VAN DER WURFF, R.; VAN CUILENBURG, J., *Impact of moderate and ruinous competition on diversity: the Dutch television market*. The Journal of Media Economics, 2001. **14**(4): p. 15-26.
- [8] VAN CUILENBURG, J., *On competition, access and diversity in media, old and new: some remarks for communications policy in the information age*. New Media & Society, 1999. **1**(2): p. 183-207.
- [9] MCQUAIL, D., *Media Performance*. 1992, London: Sage.
- [10] WAYNE FU, W. *The S-C-P Framework: applying the structure-conduct-performance framework in the media industry analysis*. in *AEJMC Annual Convention*. 2003. Kansas City.
- [11] VAN CUILENBURG, J.; MCQUAIL, D., *Media Policy Paradigm Shifts*. European Journal of Communication, 2003. **18**(2): p. 182-207.
- [12] BARDOEL, J.; CUILENBURG, J.V., *Communicatiebeleid en Communicatiemarkt*. 2003, Amsterdam: Otto Cramwinckel.
- [13] LEURDIJK, A.; LEENDERTSE, M.; DE MUNCK, S., *Reclame 2.0 De toekomst van reclame in een digitaal televisielandschap (Advertising 2.0. The future of advertising in a digital television landscape)*. TNO: Delft.
- [14] VAN CUILENBURG, J., *On monitoring media diversity, media profusion and media performance: Some regulator's*
 a. *notes*. Communications, 2005. **30**: p. pp. 301-308.
- [15] MCDONALD, D.; DIMMICK, J., *The Conceptualization and Measurement of Diversity*. Communication Research, 2003(February): p. 1-10.
- [16] NAPOLI, P.M. *Television station ownership characteristics and commitment to public service: an analysis of public affairs programming*. in *the Association for Education in Journalism and Mass Communication*. 2002. Miami, USA.
- [17] MCQUAIL, D., *Mass Communication Theory*. 2000, London: Sage Publications.
- [18] CHERMACK, T.J.; MERWE VAN DER, L., *The role of constructivist learning in scenario planning*. Futures, 2003. **35**: p. 445-460.
- [19] COURTNEY, H.; KIRKLAND, J.; VIGUERRIE, P., *Strategy under uncertainty*. Harvard Business Review, 1997(November - December): p. 67-79.
- [20] LEENDERTSE, M. *Balancing Business and Public Interests - Theoretical scenarios on how standardization & copyrights regimes impact the structure, conduct and performance of the market for learning objects*. in *World Media Economics Conference*. 2004. Montreal, Canada.
- [21] SCHOEMAKER, P.J.H., *Scenario Planning: A tool for strategic thinking*. Sloan Management Review, 1995. **Winter 1995**: p. 25-40.
- [22] HOOGERWERF, A., *Beleid, processen en effecten*, in *Overheidsbeleid*, A. Hoogerwerf, Editor. 1998, Samsom Uitgeverij: Alphen aan de Rijn.
- [23] The Guardian, *A musical tail of hits and misses*. URL: <http://arts.guardian.co.uk/netmusic/story/0,,1852005,00.html> [Accessed: 25-03 2007]. 2006.
- [24] FUNK, J.L.; METHE, D.T., *Market- and committee-based mechanisms in the creation and diffusion of global industry standards: the case of mobile communication*. Research Policy, 2001. **30**: p. 589-610.
- [25] WILLIAMSON, O.E., *Strategy Research: Governance and Competence Perspectives*. Strategic Management Journal, 1999. **20**: p. 1087-1108.
- [26] COASE, R.H., *The firm, the market and the law*. 1988, Chicago: The University of Chicago Press.
- [27] LEENDERTSE, M. *Policy & Performance of the Market for Digital Educational Content*. in *ICA*. 2005. New York City, USA.

The Open Document Format and its Impact on Accessibility for Persons with a Reading Impairment

Jan Engelen; Christophe Strobbe

Kath. Univ. Leuven, Research Group on Document Architectures
Kasteelpark Arenberg 10, B-3001 Leuven (Belgium)
e-mail: Jan.Engelen@esat.kuleuven.be; Christophe.Strobbe@esat.kuleuven.be

Abstract

It has become very common in the current information society to talk about “open” and to use this term as a quality mark. Open standards, open source software, open archives, open formats etc. are all very much promoted. In this contribution, we would like to focus on the file structure of documents such as texts, spreadsheets and presentations, and more specifically on the Open Document Format. ODF is becoming increasingly popular for many reasons, but it also is the first document format for use in office suites that has unique features built in (as of version ODF 1.1) for persons with a reading impairment such as low vision, blindness or dyslexia.

Keywords: open document format; accessibility; ODF; OOXML

1 What is a Document Format?

When we produce computer documents, be it a text document, a spreadsheet, a presentation etc., the result of our work has to be stored in computer files. Sometimes this is just one file (a Microsoft Office Word document, a slideshow presentation...), sometimes the contents of a document can only be restored from several files (a well known example are HTML webpages where text and images are stored in separate files).

When we look at the availability of documentation for document formats, it is possible to categorize them as “open” or “closed”. Over the years, commercial software makers have come up with document formats that were tied to their own products. Examples of proprietary formats for text processing include Microsoft Word’s binary format and WordPerfect’s WP format. These formats often changed when new versions of the software products were released, and forced users to migrate to a newer version of the product. This upgrade cycle has to do with several things. On the one hand, a software maker may want to incorporate new features into the next version of their product, and this may necessitate changes to the format (commercial argument). On the other hand, users feel forced to buy the newer version of the product because they fear that they may no longer be able to exchange documents with users who have migrated to the newer version. Users may also feel frustrated because the changes to the document format and the document format itself are not well documented, so they cannot judge the impact of the changes.

At the other end of the spectrum are so-called “open” formats. These formats may have been created by a de facto or official standardization organization and through an open process, for example the HyperText Markup Language (HTML) of the World Wide Web Consortium (W3C). The availability of the W3C’s Extensible Markup Language (XML), which is not a document format but a generic syntax for document formats and other data, has led to the creation of hundreds of formats, by standardization organizations as well as companies, individuals and online communities. With the advent of XML, and with the availability of many free or open-source XML parsers, creating, reading and implementing document formats came within the reach of many more users. Document formats passed through standardization or were based on standards, or for which documentation was available to everyone (free or for a fee) gained popularity for several reasons, one being that these characteristics appear to guarantee long-term readability and usability.

It would not be correct, however, to equate proprietary with “closed”. The specification for Microsoft’s Rich Text Format, for example, is available on Microsoft’s web site¹. Corel offers software development kits (SDKs) for the WordPerfect file format². Portable Document Format (PDF) was created by Adobe Systems but its specification is publicly available, and subsets of the format have undergone or are undergoing standardization by the International Organization for Standardization (ISO). Open formats may also change, and may force users to get new versions of software products, just like changes to proprietary formats.

2 The Open Document format

The development of the Open Document Format (ODF) was initiated by developers of word processing and other software who wanted to make their product available in the public domain (open source software). In this particular case the major stimulus came from OpenOffice.org, an office suite that can be used freely by everyone. But as we will see later, there are nowadays many more products supporting this format.

The ODF format has the following characteristics:

- several XML files are produced to describe one document; minimally one has four XML files for any document (cf. table 1);
- the content description of these files is well documented and was the result of a public (“open”) standardization procedure (more details about this later);
- the different files are usually (but not necessary) bundled together in one single ZIP file. Zipping files is nowadays a de facto standardized procedure for compressing and joining files and folders together. This action is always transparent for the user.

meta.xml	information about the document (author, time of last save, ...)
styles.xml	styles that are used in the document
content.xml	main document content (text, tables, graphical elements)
settings.xml	document and view settings (such as magnification level and selected printer); these are usually application specific

Table 1: The four basic building blocks of an ODF document

More details can be found in the relevant Wikipedia page³ or in chapter 17 (Packages) of the complete ODF standard⁴.

3 Why is this Format Important for Persons with Disabilities?

The ODF format is based on XML technology, which is promoted through the World Wide Web Consortium (W3C), and reuses formats whose accessibility has been verified through W3C's Web Access Initiative. Also when the ODF standard was developed, under the umbrella of the OASIS consortium, accessibility requirements were taken on board.

However, serious accessibility-related problems showed up when the US State of Massachusetts adopted the use of ODF as the only admissible interchange format for official documents in 2005. That decision has provoked a lot of criticism by groups that do not believe in open source solutions but also by organizations of handicapped persons fearing that they would be forced to use software with less accessibility provisions than their current, Microsoft-based, tools. That is why OASIS set up a special ODF accessibility subgroup in 2005.

Anyhow, there remains still a lot of confusion on the topic of accessibility to information. It is much more important for users with a visual or other impairment that the software they are using is accessible and usable than that the resulting document formats are accessible. In practice these files will never be read by human beings but only by machines. Despite this, the file format is important because it may or may not contain the data needed for an accessible reproduction.

At this point in time, general computer accessibility to Microsoft Windows-based software is quite good, especially for persons with a visual impairment. They can efficiently use their special hardware and a special computer program, called a *screenreader*, gives them access to the information and the commands on the computer screen. This is not because of Microsoft cared for this but because a whole group of external companies is building screen enlargement and screenreader software to be used on Windows platforms.

The software packages that support the ODF format currently are less accessible than the Microsoft office products although they are catching up rapidly. Promoters of Unix/Linux systems are especially convinced that this is only a matter of time because, for example, the Gnome Unix desktop is at the same time a so-called Recommended Engineering Accessibility framework⁵. These are frameworks that permit intimate interaction between general applications and accessibility software.

4 ODF Accessibility Guidelines

Current status

The ODF document format has, right from the beginning, been developed with accessibility in mind. For this process one could rely heavily on the long term experience gathered around web accessibility via the WAI guidelines.

The details can be found in “Accessibility Guidelines for Implementations of Open Document Format v1.1. Draft 19, 14 March 2007”⁶ and the major items are:

a) About the ODF format itself:

- Descriptive texts should be used for anything that is not text (graphs, pictures, sound inserts etc.). All necessary tagging is available.
- Tables and especially column and row headers must be marked up as such. This permits screenreaders to speak out table information together with the cell location information.
- A strict scheme of document divisions and corresponding headers should be maintained, using named stylesheets.
- There is a provision for logical description of navigation inside drawing layers.

b) About the software used for ODF production or conversion:

- The software must check the use of the accessibility features and stimulate authors to use them as much as possible.
- When converting a document in another format, all the accessible information must be kept and must remain available for further conversion, e.g. back into the original format.
- Users must be able to have the layout following general rules (e.g. on font size or color schemes) set up at the level of the operating system. Personal layout wishes (e.g. for persons with low vision) must always adhered to (stylesheet priority management).

Most of the above items have been incorporated into ODF 1.1

Future work

The ODF accessibility sub-committee has a number of work items planned for the next release of ODF. These are:

1. Background images. Access to any information contained in images used as backgrounds.
2. Navigation. The way in which people with disabilities can navigate round an individual slide in a presentation. Improving access to tabular data as is found in spreadsheets.
3. Multi-modalities. The provision of access in alternate modalities. For example, improving access to charts and graphs.
4. Reviewing ODF support for a wider range of disabilities.
5. More detailed support for spreadsheets. Easier access to header information, cell labels and formulas.

5 Accessibility Testing

At the CSUN 2007 conference, Jonathan Whiting and Aaron Anderson (WebAIM) gave a presentation on “Creating Accessible Content in OpenOffice.org”⁷. Another recent evolution is the development of software packages that audit the accessibility features of ODF documents. In 2006 IBM and the U.S. Department of Education organized a contest to produce such testing software. The winning solution (“ODF accessibility validation tool”⁸) was developed by two American students (from Capitol College and Oklahoma University) and a Chinese student from Tsinghua University (Beijing). This was also announced at the well-known Technology & Persons with Disabilities Conference (CSUN 2007). The winning application and several others are given to the open source community via Sourceforge.org⁸. An online Open Document Format (ODF) Accessibility Evaluator is also available on the website of the Illinois Center for Information Technology Accessibility⁹.

6 Why is ODF Readily Accepted by so Many Authorities and Companies?

One of the major goals of the Open Document Format is to guarantee access to content on very long time scales and this without technical legal barriers. In other words, efficient archiving with guaranteed future retrieval possibilities and the wish to become independent of Microsoft’s business strategy are among the main drivers of adoption.

The fact that the early adopters are mainly public authorities has definitely increased the visibility of accessibility aspects in ODF as these authorities nowadays often have the legal obligation to consider the needs of all the citizens.

It is expected that ODF will slowly gain momentum mainly through acceptance by authorities.

The Massachusetts case had shown the weak point: very few software packages that natively use ODF were available in late 2005. And the incident also led to the creation of the OASIS subgroup on Accessibility.

One of the early adopters is the Belgian government that has decreed that only open formats are acceptable as an exchange format between the Belgian authorities, and this from 2008 onwards. If realized in time, Belgium would be the first country to prohibit the use of closed document formats. As could be expected, Microsoft has reacted strongly.

7 How Popular is ODF in Reality?

By the end of 2006 there were eleven word processors, six spreadsheet programs and 5 presentation managers (Powerpoint-like programs) with support for ODF available. Furthermore, three groups are active in the development of conversions from Microsoft Word into the ODF format and vice versa. They are SUN Microsystems, the Open Document Foundation and the public domain Sourceforge.net project, "ODF Add-in for Microsoft Word"¹⁰. Within the UK Royal National Institute of the Blind, a project has been set up to turn ODF documents into the Daisy format, the new, worldwide accepted standard for talking books and multimedia documents. The ODF format is also used in the online text processing facilities of **docs.google.com**. Documents produced online can be stored in Microsoft Word, Microsoft RTF, and OpenDocument formats. As it is possible to upload and to download the online documents, the **docs.google.com** facility can in fact be used for file format changes too: upload in one format, download in another. PDF can be used as output, not as input.

8 Standardization

The ODF format v1.0 became an international ISO standard in 2006. After having been developed within a working group of the OASIS foundation, it was passed through the International Organization for Standardization, ISO, where it became ISO standard ISO/IEC 26300. The proponents of ODF have created several organizations for discussion and exchange of information. Two of them are very well known:



Figure 1: Logo of the ODF Alliance¹¹



Figure 2: Logo of Opendocument-xml.org¹²

Meanwhile Microsoft has launched a counterattack by creating and promoting XML version of its proprietary office formats, and called them Office Open XML format (OOXML)¹³. This format will be used in Microsoft Office 2007.

In the beginning of 2007, this led to a very controversial issue on the standardization of XML-based open document standards. The Open Document Standard became ISO 26300 (700 pages) through a very formal process typical for ISO work. Microsoft's alternative, OOXML (Office Open XML) was produced in one year by a technical committee¹⁴ chaired by two Microsoft persons, contained many references to specific behavior of Microsoft software that were not documented and counted 6000 pages. It was ratified as ECMA-376 by ECMA International¹⁵, which was consequently described as "a private association that drafts standards on demand"¹⁶.

In late 2006, Microsoft wanted to put the OOXML specification on a fast ratification track within ISO because the document had already been ratified by ECMA. This created a lot of dismay within organizations that had been developing and promoting ODF. People were even asked to lobby with national ISO delegates to cancel the fast-track procedure. Websites listed arguments against the fast-track process, for example “EOOXML objections” on the Grokdoc website¹⁷.

In spite of this, it was recently (April 2, 2007) announced¹⁸ that ISO-Joint Technical Committee 1 has started the voting period for ISO/IEC standard DIS 29500...

9 And now?

ODF still seems to attract quite a lot of organizations. It is public domain and vendor independent, it is well defined (and not too complex) and it is accessible. However, what is even more important for reading impaired users is the fact that the software producing open office documents is made accessible. The existence of different types of converter plug-ins has taken away the major objection against the use of ODF as people can, for example, stay with the more traditional Windows-based platforms.

Acknowledgment

The authors wish to acknowledge the support of Kris Van Hees (PhD student at K.U.Leuven) and Peter Korn (SUN Microsystems) during the compilation of this contribution.

Notes and References

- [1] Rich Text Format (RTF) Specification 1.6: [http://msdn2.microsoft.com/en-us/library/aa140277\(office.10\).aspx](http://msdn2.microsoft.com/en-us/library/aa140277(office.10).aspx).
Word 2003: Rich Text Format (RTF) Specification:
<http://www.microsoft.com/downloads/details.aspx?familyid=AC57DE32-17F0-4B46-9E4E-467EF9BC5540&displaylang=en>.
- [2] http://apps.corel.com/partners_developers/csp/wordperfect_fileformatsdk.htm
- [3] <http://en.wikipedia.org/wiki/OpenDocument>
- [4] ODF standards:
The ISO version:
<http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=43485&scopelist=PROGRAMME>
The freely available OASIS version can be downloaded from:
<http://www.oasis-open.org/committees/download.php/19274/OpenDocument-v1.0ed2-cs1.pdf>
(OpenDocument 1.0, Second Edition ; 722 pages) or
<http://docs.oasis-open.org/office/v1.1/OS/OpenDocument-v1.1.pdf> (OpenDocument 1.1).
- [5] Other examples are the Java Accessibility API, the Apple Accessibility API (with limited possibilities) and WAI ARIA for web based applications.
Info from Korn & Schwerdtfeger: *Accessibility guidelines for ODF* (cf. ref. #7 below)
- [6] KORN, P.; SCHWERDTFEGGER, R. eds. *Accessibility Guidelines for Implementations of Open Document Format v1.1*. Draft 19, 14 March 2007. http://www.oasis-open.org/committees/download.php/22957/ODF_Accessibility_Guidelines_19_14Mar2007.odt (only available in Open Document Format).
- [7] <http://webaim.org/presentations/2007/CSUN/ooo.htm>
- [8] Meanwhile there are already a number of ODF validator projects on Sourceforge.net. On April 8, 2007 we found: ODF_Accessibility_Tester, ValidODF, Accessibi Add-on component of OpenOffice, ODF Accessibility Validator, Check the accessibility of ODF file, **ODF accessibility validation tool**, odfav, av4odf, Validator for ODF Accessibility, SalihiODF and ODFable.
More details by filling in their name in the search project page of www.sourceforge.net.
- [9] <http://odf.cita.uiuc.edu/>
- [10] http://sourceforge.net/project/showfiles.php?group_id=169337

- [11] <http://www.odfalliance.org/>
- [12] <http://opendocument.xml.org>
- [13] http://en.wikipedia.org/wiki/Office_Open_XML
- [14] ECMA International TC45: <http://www.ecma-international.org/memento/TC45.htm>
- [15] Formerly known as European Computer Manufacturers Association; <http://www.ecma-international.org/>.
- [16] http://press.ffii.org/Press_releases/FFII_opposes_Fasttrack_adoption_of_Microsoft_OOXML_format_as_ISO_standard
- [17] http://www.grokdoc.net/index.php/EOOXML_objections
- [18] http://www.ecma-international.org/news/PressReleases/PR_TC45_April2007.htm

Multimedia Modular Training Packages by EUAIN

David Crombie¹; George Ioannidis²; Neil McKenzie¹

¹ Research and Development Department, Dedicon
Molenpad 2, Amsterdam, The Netherlands
e-mail: dcrombie@dedicon.nl; nmckenzie@dedicon.nl

² Image Processing Department, TZI (Technologie Zentrum Informatik), University of Bremen
Postfach 33 04 40, D-28334 Bremen, Germany

Abstract

The European Accessible Information Network (EUAIN) was established to support the move to incorporate accessibility within mainstream content processing environments. EUAIN has brought together a considerable base of knowledge that has now been structured into a series of training modules and curricula which are intended to meet the real needs at this point in time. In this paper we outline how the EUAIN training and learning framework is primarily intended to provide support for everyone who is directly involved in digital content creation and document distribution channels. This target audience requires general courses and training materials as well as domain-specific materials. These general training materials include information about digital document standards and formats, accessibility guidelines and different kinds of publishers and distribution channels. Also important is knowledge about accessibility and alternative forms of presentation that fulfil special requirements for print impaired people. The curricula are illustrated by good practices of accessible content publishing and good examples of accessible digital documents. The specific training materials are addressed to different branches of publishing (books, newspapers, magazines, etc.) and content creators (multimedia content designers, web designers, authors of e-learning content). A significant part of the materials are curricula that demonstrate tools and techniques for accessible content processing. Additionally, the training materials are in modular form to allow them to be adopted within courses and programs to meet the requirements of particular groups. These modular materials are also extensible and scalable, and it is our intention that many new curricula will be developed using this ever-growing resource base. Indeed, the newly-established PRO-ACCESS project is disseminating this information across the publishing industries.

Keywords: visually impaired; e-learning; EUAIN; digital publishing

1 Introduction

Structured information is the first step in the accessible information process. A document whose internal structure can be defined and its elements isolated and classified, without losing sight of the overall structure of the information, is a document that can be navigated.

Most adaptive technology allows the user to access a document, and to read it following the "outer" structure of the original. But if the same information also has an "inner" structure that allows the adaptive device to distinguish between a phrase and a measure, between a paragraph and a sentence, highlighting particular annotations, then the level of accessibility (and therefore usability) of the whole document will be greatly enhanced, allowing the user to move through it in the same way as those without impairments do when looking at a printed document, and following the same integral logic.

In an ideal world, all documents made available in electronic formats should contain this internal structure that benefits everyone. Highly-structured documents are becoming more and more popular due to reasons that very seldom pertain to making them accessible to people with disabilities. The move to XML related formats and associated standards for metadata has provided an impetus for far greater document structuring than before. Whatever the reasons behind those decisions are, the use of highly-structured information is of great benefit to anybody accessing them for any purpose.

In recent years, the market for accessibility and assistive technologies has started to gain recognition. It is clear that the integration of accessibility notions into mainstream technologies would provide previously unavailable opportunities in the provision of accessible multimedia information systems. It would open up modern information services and provide them to all types and levels of users, in both the software and the hardware

domain. Additionally, new consumption and production devices and environments can be addressed from such platforms and this would provide very useful information provision opportunities indeed, such as information on mobile devices with additional speech assistance.

It is equally clear that we remain at the very beginning of the move to incorporate accessibility within mainstream content processing environments. The EUAIN consortium has brought together a considerable base of knowledge that has been structured into a series of training modules and curricula which we believe meet the actual needs at this point in time. These materials are also extensible and scalable, and it is our hope that many new curricula will be developed using this ever-growing resource base.

2 The EUAIN Project

The EUAIN project [2] is now nearing completion, and as such much interesting information has been brought together concerning the provision of published information for visually impaired end users. In order that the information brought together by the consortium can have a maximum effect on stakeholder communities in Accessible Information Processing the EUAIN network has created a comprehensive set of instructional training materials. These flexible materials can be used in different environments and work is now underway to translate them into multimedia materials. This paper is a presentation of these developed materials.

3 Training Materials

The training and learning framework was primarily constructed with the intention to provide support for everyone who directly effects digital content creation and decides about document distribution channels. This group requires general courses and training materials as well as domain-specific training materials.

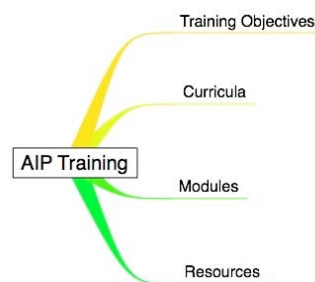


Figure 1: Accessible Information Processing(AIP) Training

The general training materials include information about digital document standards and formats, accessibility guidelines and different kinds of publishers and distribution channels. Also important is knowledge about accessibility and alternative forms of presentation that fulfill special requirements for print impaired people. The curricula are illustrated by good practices of accessible content publishing and good examples of accessible digital documents.

The specific training materials are addressed to different branches of publishing (books, newspapers, magazines, etc.) and content creators (multimedia content designers, web designers, authors of e-learning content). A significant part of the materials are curricula that demonstrate tools and techniques for accessible content processing. Additionally, the training materials are in modular form to allow them to be adopted within courses and programs to meet the requirements of particular groups.

In general, there are three themes. The first is related to different types of digital documents and their accessibility issues for print impaired people. The subject of the second theme is to discuss and demonstrate workflows for authoring tools and techniques that allow people to create documents accessible for all. The last theme addresses the processes that must be considered regarding content distribution and digital rights management.

The EUAIN training materials consist of:

- Practical examples of good practice;
- Illustrated explanations of good process management for accessible information production;

- Detailed explanations of approaches, technologies and tools;
- Detailed explanations and examples of benefits and weaknesses of different formats;
- Step by step, modular instructions for producing accessible information in different formats.

Furthermore, the educational process and especially the course materials are themselves a good example of accessible content creation.

After detailed consideration and advice from industry, there is also a requirement that the training materials should operate on several levels. These levels will become increasingly detailed and complex. In this way, different people can choose the level of detail that is required for their situation, or their position in the decision-making chain. In essence, the three levels are as follows:

- **Level 1: Descriptive** - Should teach actors to think about the issues and finding the solutions for their situation. There will only be simple explanations, and not detailed or over-technical information.
- **Level 2 : Decision making** - Should teach actors how to make the right decisions to implement accessible information processing. There will only be relatively simple explanations, and not detailed or over-technical information . These descriptions will link directly to level 3 detailed information.
- **Level 3 Training packages** - At this level, the detailed and more technical information is provided. This level essentially provides the answers to detailed questions and applications.

To this end we have constructed our training materials in such a way that people can choose exactly the most appropriate training packages for their local environments.

3.1 The Curricula

In order to target the EUAIN modular training packages at the correct market segments, it is important to understand the various targets of the curricula presented. The most relevant modules can then be presented to these audiences. As a starting point, the WAI Resources on Developing Web Accessibility Training and Presentations[1, 7, 8] have been used and adapted to be more specific to Accessible Information Processing:

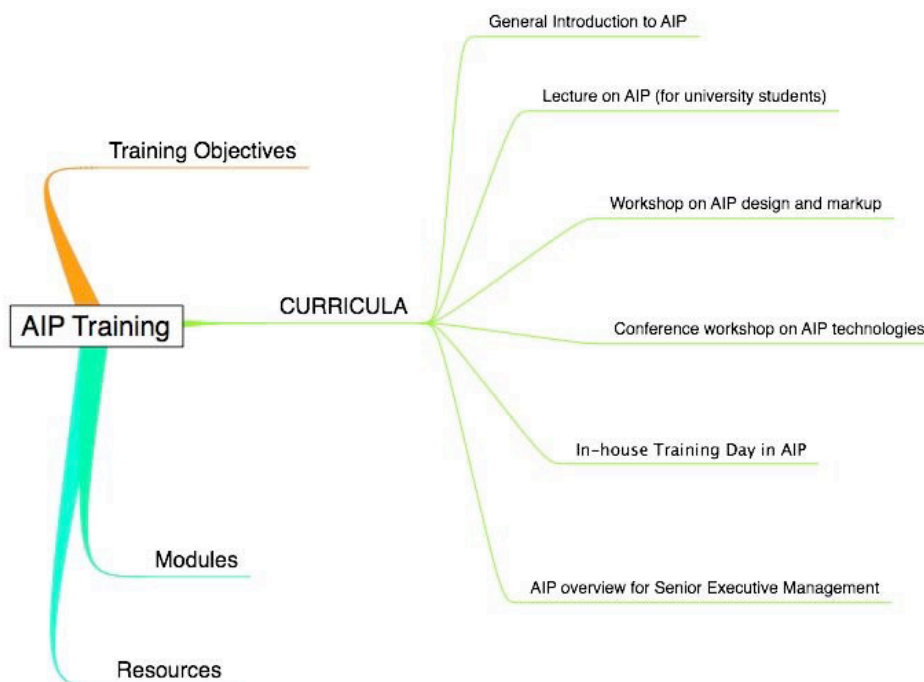


Figure 2: An expansion of the Curricula section of the AIP training

The example curricula currently available are:

- **General introduction to Accessible Information Processing:** A general introduction to a heterogeneous audience provides the background information required to understand why Accessible Information Processing is important in whatever environment the recipients of the training are involved in. The curricula is designed to take 40 minutes to present in order that it can be easily incorporated in other curricula and training sessions.
- **Accessible Information Processing Lecture:** a single two-hour lecture/presentation on AIP as part of a full semester's introductory course on general web design skills.
- **Workshop on AIP design and markup Context:** Hands-on workshop on Accessible Content design and mark up, for a class (~10-20 people), of content creators. The workshop assumes some knowledge of the Business case and the Market for Accessibility, and is taught with computers for learning assistance. The class has to be taken by someone who has a reasonable experience in Accessible Information Processing as the workshop requires a lot of interaction with the subject matter.
- **Conference Workshop on AIP technologies:** This curriculum specifies a ninety minute workshop which can be given at a conference or trade event. It is aimed at IT workers who have some knowledge of software design and development. It is assumed that the audience is familiar with the need for accessible information processing.
- **In-House Training Day on Accessible Information Processing:** In-house training at a publisher, content creation company, or software development company. The audience is assumed to have some level of knowledge of Accessible Information Processing. The training session requires considerable preparation (Possibly with the help of an organisation contact point) by the facilitator of the training in order that the training is relevant to the organisations specific field, workflows and authoring tools.
- **Accessible Information Processing Overview for Senior Management:** A brief presentation around a conference table during a senior management meeting. The focused delivery of this training aspect requires a familiarity with the material.

These curricula are aimed at different audiences and market segments. The materials for these curricula draw on a body of topic specific modules which have been brought together for use in training. Each sample curriculum highlights several objectives and learning outcomes for the topic and provides a list of resources relevant to those learning objectives. An estimated time frame for each curriculum is provided. They are designed in such a way that the curricula can be altered and personalised for more specific needs and situations.

3.2 Modular Training Packages

The Modules for the EUAIN training materials have been modeled on the structure of the WAI Resources on Developing Web Accessibility training and presentations. , The structure is as follows:

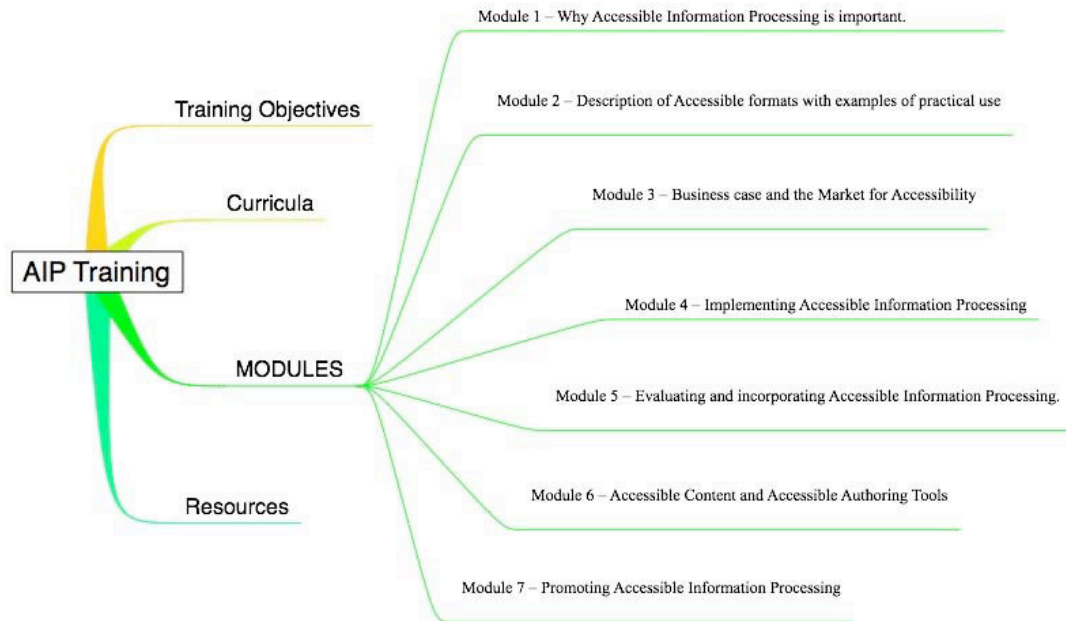


Figure 3: The Modules section of the AIP training packages.

These modules consist of materials which can be used in the curricula described above. Each module focuses on a particular aspect of Accessible Information Processing and provides reusable materials such as hand-outs, presentations, software and other useful materials:

- **Module 1 – Why Accessible Information Processing is important.** This module provides an overview of Accessible Information Processing. This includes an overview of the topics in the other modules and should be seen as an introduction module to all the EUAIN training modules. The resources presented in this module are relevant to all the other modules and relate directly to all the curricula.
- **Module 2 – Description of Accessible formats with examples of practical use.** This module is aimed at providing a solid understanding of the formats relevant to Accessible Information Processing and how they are used in processes and workflows. As such it presents descriptions of how these formats used in both mainstream environments and environments that are more focused on print impaired users. An understanding of these formats is essential before conversion processes can be built out of the formats such that they can be incorporated into workflows and processes relevant to the stakeholders which these resources and training materials are presented to.
- **Module 3 – Business case and the Market for Accessibility.** It is important to understand not only the technical perspectives for formats and conversations required for Accessible Information Processing but also the business angle and the market relevance for incorporating Accessible Information Processing within existing industrial environments. This module targets decision makers and executives within stakeholding communities who have to assess the cost and benefits of implementing Accessible Information Processing.
- **Module 4 – Implementing Accessible Information Processing.** Accessible Information processing is a series of processes., but very few of these processes stand alone, in most cases, the accessibility component will be one in a chain of processes with the input and outputting feeding from and to other processes. This module describes how Accessible Information processing ties in with these other processes and workflows already in place in mainstream environments.
- **Module 5 – Evaluating and incorporating Accessible Information Processing.** In order to successfully incorporate Accessible Information Processing within existing workflows, it is important to first evaluate the workflows for accessibility. This relates to the formats, authoring tools and standards already in place in the processes within these workflows. This module focuses

on evaluating this accessibility and how these evaluations can help point top answers on the best way to implement further accessibility.

- **Module 6 – Accessible Content and Accessible Authoring Tools.** In order to implement accessible information processing within organisations, clear understanding is required of how to create, modify and process content using both the tools available within mainstream organisations and also the accessibility conversion tools and assistive technologies. This module provides resources such to make this possible.
- **Module 7 – Promoting Accessible Information Processing.** This module ties the previous 6 modules together in order that participants of EUAIN learning packages can reuse their knowledge and understanding of Accessible Information Processing within heir organisation and further promote accessibility. This module ensures that accessibility is reverberated through the organisation and can be promoted from the top down.

Each module is intended for use as information which feeds into specific curricula but they have specific objectives, resources and learning outcomes such that they can be used as a stand alone information package.

3.3 Resources and Additional Materials

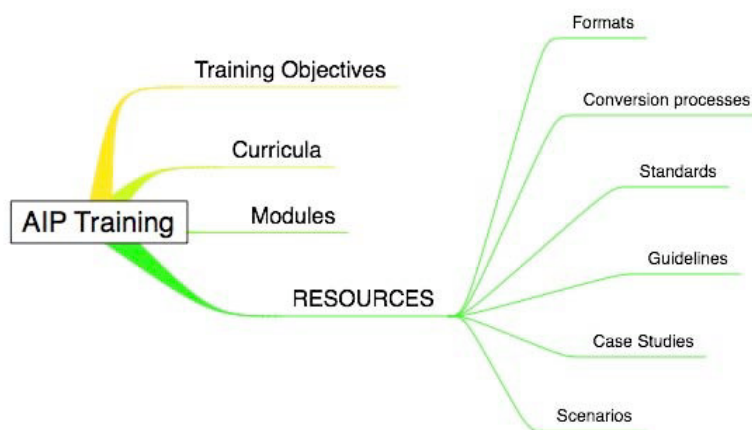


Figure 4: Overview of AIP training resources

Formats

In order to look at the various processes involved in Accessible Information Processing, it is essential to build up a finite list of the formats which are commonly used during these processes and interactions within supply chains. After careful consideration of several specialist organisations, publishers and users, we came up with the following list of formats.

- Printed paper
- Printed Braille
- Audio(Wav)
- ASCII Text
- HTML
- XML
- Multimedia Packages

It was felt that these descriptions covered all areas. There is a specific focus on formats for the print impaired, so formats such as bitmap or JPEG are considered to be components of more complex multimedia packages, as they are rarely dealt with without some sort of surrounding information or multimedia package.

Conversion Processes

Given that we have a finite list of formats used for accessible information processing, a conversion from every format to every other format provides us with a list of accessible information conversions. This provides us with a list of 42 conversion processes.

Each conversion process is dissected to provide:

- a description of the conversion from a accessible information processing perspective
- examples of the conversions use in real life case studies (see below)
- examples of the conversions use in hypothetical scenarios (see below)
- related guidelines and best practices
- A flow chart description of the process

For example:

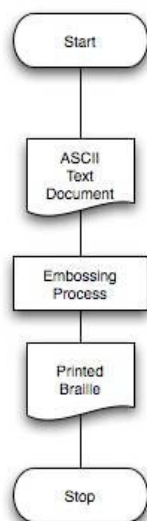


Figure 5: Flow chart of ASCII to Braille Process

Standards

As part of the work of EUAIN a deliverable entitled “Standards for Accessible Information Processing” was created. The deliverable is public and available on the EUAIN website. This information is used as a resource in many of the modules and curricula.

Guidelines

EUAIN is not the first project to tackle the issues of Accessible Information Processing and there are several sets of Guidelines and Best Practice already in existence. However, until now these have not been brought together in a systematic manner. EUAIN has collated this information in order to focus stakeholders on specific information based on their specific requirements. This information is available on the EUAIN website, the EUAIN wiki and it is also fed into the resources for the training materials.

Case Studies and Scenarios

Based on real-life examples of accessible content processing, we have prepared a number of Case Studies and scenarios to illustrate different aspects of accessible content processing. These Case Studies are drawn from different publishing sectors and address a variety of different issues. Each Case Study provides an in-depth examination of key factors and provides practical explanations of how the various processing stages were addressed to achieve accessible content. The case studies are constructed out of the same conversions which were described above:

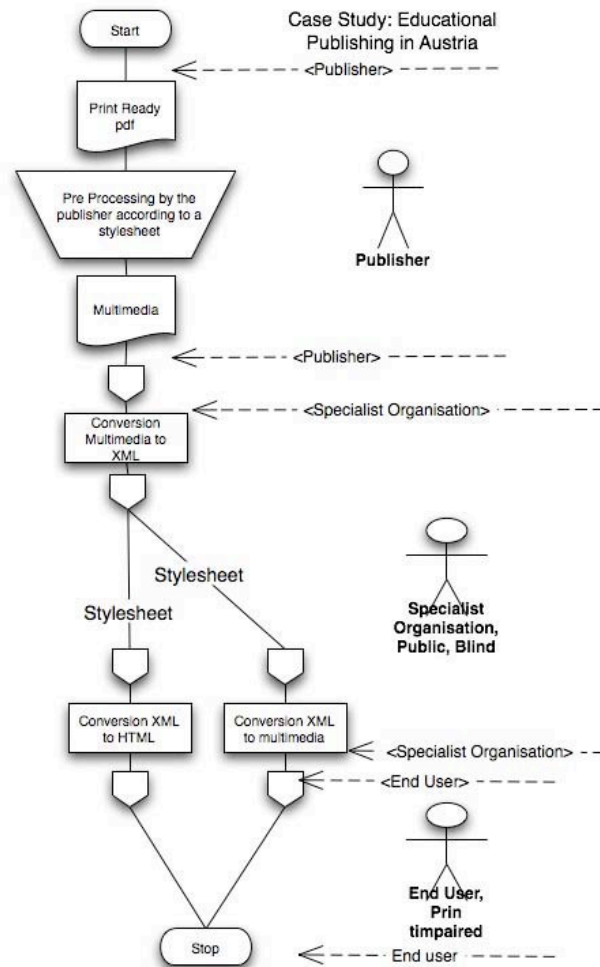


Figure 6: Example of a case study flow description

This case study is then dissected in terms of:

- Actors involved in the information processing chain
- Conversions and processes used
- Standards used
- Guidelines used.

This body of information is also available on the EUAIN wiki[ref].

4 Pro-Access

Following on from EUAIN, the consortium intend to explore these themes further by taking part in several more practical implementations which tackle that which EUAIN raised. One such endeavour is the PRO-ACCESS[4] project which started recently.

The project will provide practical tools for publishers and content providers to address the targeted audience of primary and secondary students with specially formatted and accessible course materials on a timely basis with total respect of copyright.

This main objective of the project will be achieved following these steps:

- evaluate the actual situation in the involved countries, analysing on one side the needs of the disabled people and on the other the problems and the concerns arose from these request in the

publishing sector, involving key schoolbook publishers and printed disabled people representatives in the process;

- define the production process needed to create accessible documents, starting from the achieved results of the EUAIN project;
- promote the results as wide as possible in the publishing sector;
- analysis of the content value chain in the education sector (authors, publishers, intermediaries, schools, students) to define a set of shared rules to managing rights.

Expected results:

- a set of ISO 9001 compliant Certification guidelines for publishers to create an accessible school materials in a standard way. The publishers who will follow these guideline will be appointed with a specific process certification;
- a standard license for publishers to be used to manage the relations between the publishers and the students, or the schools asking for special formatted materials;
- a set of materials devoted to create awareness in the print disable people environment and in the school environment, teachers and educational authorities in particular;
- a standard module for blended training courses for publishers and content providers in order provide them with all the information needed both on the technical and legal solutions defined in the project itself.

As these results come to the fore, they will be disseminated through similar channels to EUAIN.

5 Conclusion

The EUAIN network has provided some practical training solutions and it is now important to create broader awareness on these topics in the content producers market (i.e. publishers, Learning Object producers, digital content and software developers) and promote the adoption of collaborative and practical solutions to allow them quickly to make available these accessible materials. New projects such as PRO-ACCESS can help to achieve these goals.

The coherent and sustainable provision of accessible information cannot be tackled in isolation by individual actors in the information provision chain. While examples of good practice are emerging in the production sphere and in new collaborative distribution models, a European-wide approach offers far greater potential. In particular, a collaborative approach involving content producers and users' associations allows us to approach key aspects like rights clearance, definition of standard formats for exchanging content files, and finally actual increase of accessibility. As noted in the recent report produced for WIPO:

“This [EUAIN] is perhaps an example of a way forward more generally and work of this nature should perhaps be promoted more widely by governments and international agencies. It seems to be in everyone’s interests that a desire to build in access from the start is both encouraged and facilitated by ensuring that what this requires in practice is widely understood and adopted.”[6]

By focusing closely key issues in this area (rights management, production processes, content value chain, and standard information exchange), we can make an important and lasting contribution to the Accessibility For All initiative and help to provide the educational building blocks needed to help make consumer needs more explicit to the designers of products and services for print impaired people.

References

- [1] <http://www.w3.org/WAI/training/>
- [2] <http://www.euain.org>
- [3] <http://wiki.euain.org>
- [4] <http://www.euain.org/proaccess>
- [5] <http://www.ormee.net>
- [6] SULLIVAN, J., (2007) Study on Copyright Limitations and Exceptions for the Visually Impaired, SCCR15/7, WIPO, Geneva
- [7] <http://www.w3.org/WAI/>
- [8] <http://www.w3.org/TR/WAI-WEBCONTENT/>

File-Formats for Preservation: Evaluating the Long-Term Stability of File-Formats

Carl Rauch¹; Harald Krottmaier²; Klaus Tochtermann³

¹ Styria Media AG, Schönaugasse 64, 8010 Graz, Austria
e-mail: carl.rauch@styria.com

² Institute for Computer Graphics and Knowledge Visualization, Graz University of Technology
Infeldgasse 16c, 8010 Graz, Austria
e-mail: h.krottmaier@cgv.tugraz.at

³ Knowledge Management Institute, Graz University of Technology, and Know-Center Graz
Infeldgasse 21/II, 8010 Graz, Austria
e-mail: klaus.tochtermann@tugraz.at

Abstract

While some file-formats become unreadable after short periods, others remain interpretable over a long-term. Among the over 1.000 file-formats, some are better and some are less suited for long-term preservation. A standardized process for evaluating the stability of a file-format is described in this paper and its practical use is shown with file-formats for 3D-objects. Recommendations to users of 3D-applications are given in the last section of this article. Some of the results are used in PROBADO, a sophisticated search engine for non-traditional objects (such as 3D-documents, music etc.).

Keywords: digital preservation; evaluation metric; file-formats

1 Introduction

In file-format registries like PRONOM, filext or MyFileFormat, over 1.000 file formats are registered. Even when removing all depreciated formats and even when setting the focus on one type of digital records only, e.g. 3D-objects, the number of available file formats is big (in this case among others dxf/dwg, iges, 3ds/max, 3dm, obj). While some file-formats depreciate over time, other file-formats are evolving. Formats, which were frequently used 10 years ago, are unreadable now as will many today's formats in ten years. But even slight modifications in the representation of digital objects can have major influences on their significance. An example would be a computer game with a slightly higher processing speed - it would become many times more difficult to play.

When a digital object needs to be available over a long-time period, users face the question, which file-format to choose for long-term preservation. Based on the concept of Utility Analysis [12] and on work done by Rauber, Strodl and Rauch [11], an evaluation process is described in this paper for analyzing and ranking file formats in terms of long-term reliability.

An evaluation of file-formats for 3D-objects is used for showing the process in practice. The remainder of this paper is organized as follows: Section 2 provides an overview over related work. In Section 3 the workflow and parameters for evaluating file-formats is described. In Section 4 the criteria for evaluating file-formats are shown in detail. A practical implementation for 3D-objects shows the feasibility of the described approach in Section 5.

2 Related Work

The work described in this paper is based on three research areas. The first basis is the area of digital preservation, where methods and workflows for comparing various preservation alternatives are developed and implemented. The second area are already existing initiatives to examine a file format's preservation risk. The third are file-format registries.

In the research area of digital preservation, several processes for evaluating preservation strategies were presented in the last couple of years. Among them are the test-bed workflow of the Dutch Preservation Test-bed [9] and the Utility Analysis workflow of the Vienna University of Technology [8]. As part of the DELOS

Network on Excellence project, these two workflows were combined to the DELOS digital preservation Test-bed's workflow [11], which is shown in Figure 1.

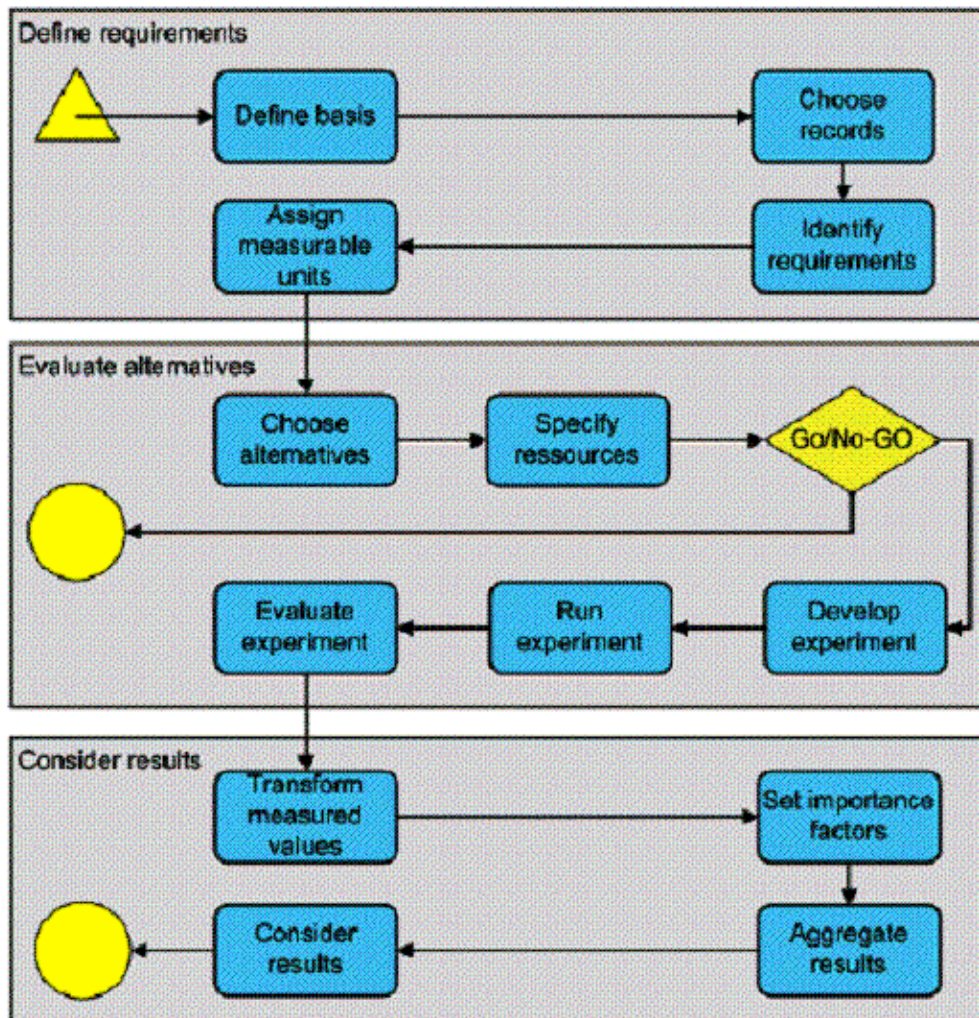


Figure 1: Overview of DELOS Digital Preservation Test-bed's workflow [11]

The DELOS workflow consists of three main parts: At the beginning, the requirements of an institution for a digital preservation strategy are defined. Here the record set, which is to be preserved, is selected, a list of criteria for evaluating the strategies defined and measurable units are assigned to each criterion. In the second part the evaluation takes place. After defining alternatives and resources to be tested, an experiment is developed and different preservation strategies are applied to the chosen objects. In the third part finally the results are examined by aggregating the performance of each alternative for the different criteria. This workflow forms the basis for the evaluation of the file-formats.

Another research is methods for evaluating the preservation risk. During the last couple of years two initiatives were started to evaluate the preservation risk of a file-format. First the INFORM system of the Online Computer Library Center [10]: There the durability of file-formats in a specific environment is evaluated, considering not only the reliability of a file-format itself, but also of the opening software, the hardware, of associated organizations, the digital archive and migration and derivative-based preservation plans. The main disadvantages of this system are, that for the assignment of a risk-factor to one of the six risk-areas, a high level of expertise is required for each individual environment. Thus the process needs highly qualified officers and cannot be standardized easily. The here proposed workflow suggests an alternative solution to these drawbacks.

A second initiative is the 'Virtual Remote Control' project of the Cornell University [4]. VRC focuses on the preservation of web pages. If the VRC-web-crawler detects a page with dysfunctional hyperlinks, longer downtimes or older server-software, the VRC-administrator is notified about the preservation risk of the web

page. VRC provides some interesting insights on evaluating the preservation risk, however it is only focusing on web pages and the file-format itself plays a minor role.

The last research area on which this paper is based is file-format repositories. Several repositories exist, where different aspects of file-formats are stored. The best-known example is the PRONOM-database of the UK National Archives. In this archive the following information are stored (among others) about a file-format [7]:

- Name, Version and other Names
- Identifiers
- Family, Classification and Orientation
- Byte Order and Related File-Formats
- Release date and support end date

A second file-format registry is FILEExt. In FILEExt [3] the external and internal signatures of a file-format, the software programs able to interpret the format, the MIME types, the main producing company, the file-formats name and a description is given for each file-format.

Neither of the registries contains a specific measure on the reliability of a file-format. For both the information given needs to be interpreted by a file-format expert to evaluate the appropriateness of a format for digital preservation.

3 The File-Format Evaluation Process

Based on the workflow shown in Figure 1 a process for evaluating the reliability of file-formats is presented in this section. Due to the smaller scope - the DELOS workflow is designed for comparing whole preservation strategies including appearance, process characteristics and costs - the here shown process consists of less steps than the DELOS workflow. Most of these steps are standardized for all file-formats.

1. Review Requirements: The requirements for a reliable file format are structured in a criteria-tree. The criterion focuses on two areas: on technical characteristics and on the integration of the format within the marketplace. The criteria tree described in detail in Section 4 is the same for all file-formats in order to allow comparability;
2. Assign measurable categories: The second step is to assign measurable categories to each criterion. A metric is defined describing, how to convert the measured numbers into a zero-to-five scale (e.g. number of users between 10.000 and 100.000 is equal to '3' for the market penetration criterion). These conversion tables are described in more detail in Section 4 and are standardized for every evaluation run;
3. Choose alternatives: In this step file-formats are chosen, which are evaluated during a session of the workflow. In the here presented work, six file-formats for 3D-objects are evaluated as a proof-of-concept;
4. Evaluate file formats and transform values: Based on the seven sub-criteria of the criteria tree and on the measurable categories the file-formats are evaluated and a value between zero and five (five is the best) is assigned to every criterion of each file-format. These evaluation results do typically not change over time and are stored as a basis for the final aggregation;
5. Set importance factors: After the evaluation, each criterion is ranked with a percentage value according to the user's priorities; the sum of all percentages has to be 100 %. Each user can determine the importance of certain criteria for individual circumstances with values from 0 % (is not interesting at all) to 100 % (is the only relevant criterion);
6. Aggregate results: A final value per file-format is found by multiplying the value per criterion with its weight and summing these values up. The higher the value, the better a file-format is suited for long-term preservation. By aggregating the final values of several file-formats or by taking earlier evaluations as a reference, a clear ranking can be created. A measure suggested for file-formats is the preservation risk, which is calculated by dividing the final value per file-format by the maximum value possible (in the here described metric, the maximum possible number is five). This fulfillment percentage-value has then to be subtracted from one. The higher the preservation risk, the lower the probability of being able to interpret the file-format after a couple of years.

From the above listed steps, the requirement review and the assignment of measurable categories is standardized for every evaluation run. When evaluating file-formats, a user has to do the steps three to five for each run; the aggregation of results follows again a standardized scheme.

4 The File-Format Evaluation Tree

In this section the tree of requirements and the assignment of measurable categories are described. In order to compare and evaluate file-formats in terms of long-term reliability, criteria were defined and structured in a criteria-tree. The tree is based on a discussion process with the Department for Software Technology, Vienna University of Technology, the Austrian National Library, the Austrian Phonogrammarchiv and the Dutch Nationaal Archief; it is structuring all criteria, which are seen as important to measure the long-term reliability of a file-format. The tree is shown in Figure 2. The tree consists of two branches, the technical and the market characteristics.

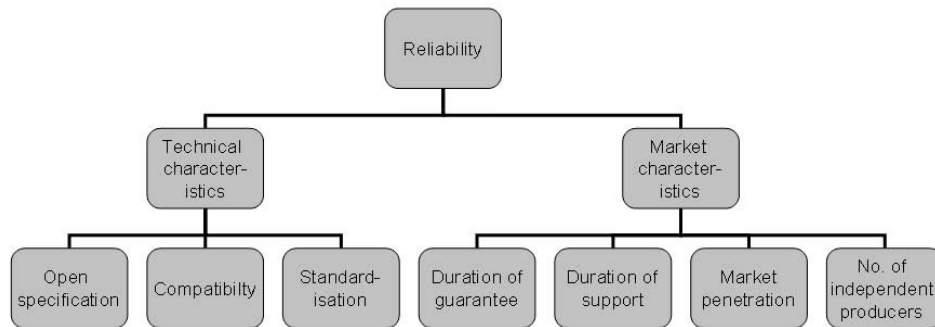


Figure 2: Criteria-tree for evaluating the long-term reliability of file-formats

The technical characteristics focus on the specification of a file-format. It consists of the following three sub-criteria:

- Open Specification: Is the specification of the file-format publicly available?
- Compatibility: Is the file-format supported and maintained by one or several software companies?
- Standardization: Is the file-format standardized by a recognized standardization agency, such as DIN or ISO?

The market characteristics focus on the acceptance and position of the file-format in the market. It is divided into the following sub-questions:

- Duration of guarantee: How long does the main producing software company guarantee to repair bugs in the interpreting software?
- Duration of support: How long does the main producing software company supports the interpreting of the file-format with its software?
- Market penetration: How many users are working with the file-format at the current time?
- Number of independent producers: How many software products exist, which are able to interpret the file-format?

In order to transform measurable units into values from zero to five the following transformation tables are suggested. The intervals are chosen in a way, which should bring a maximum distinction between typical software formats. By targeting the range of values, which typical software formats have, the differences between formats can be shown explicitly:

- Open Specification: Yes = 5, Partly available = 3, No = 0
- Compatibility: Number of software systems compatible with the format: 1 system = 1, 2 systems = 2, 3 systems = 3, 4 systems = 4, > 4 systems = 5
- Standardization: Yes = 5, Partly standardized = 3, No = 0
- Duration of guarantee: 0 years = 0 (no guarantee), > 0 years and <= 1 year = 1, > 1 year and <= 3 years = 2, > 3 years and <= 5 years = 3, > 5 years and <= 10 years = 4, > 10 years = 5
- Duration of support: 0 years = 0 (no support), > 0 years and <= 1 year = 1, > 1 year and <= 3 years = 2, > 3 years and <= 5 years = 3, > 5 years and <= 10 years = 4, > 10 years = 5
- Market penetration: < 100 users = 0, > 100 users and <= 10.000 users = 1, > 10.000 users and <= 100.000 users = 2, > 100.000 users and <= 1.000.000 users = 3, > 1.000.000 users and <= 10.000.000 users = 4, > 10.000.000 users = 5
- Number of independent producers (that support the software): 0 producer = 0, 1 producer = 1, 2 producers = 2, 3 producers = 3, 4 producers = 4, > 4 producers = 5

5 Evaluating Digital Objects for 3D-Data

As a proof-of-concept, file-formats for 3D-objects were evaluated and ranked according to their preservation risk. The steps three to six of the evaluation process are described in detail in this section.

The choice of alternatives is the first step, which needs to be done before an evaluation run. The following file-formats were selected, based on inputs from the PROBADO project [6]: Drawing Exchange Format DXF/DWG, Initial Graphics Exchange Specification IGES, 3D Studio 3DS/MAX, 3D Model 3DM and Object OBJ .

Based on publicly available sources, such as Internet queries and producer information, the file-formats were evaluated. Please note that the proof-of-concept is primarily done to show the functionality of the evaluation process and can not be seen as a final judgement on the performance of every file-format.

Criterion	DXF/DWG	IGES	3DS/MAX	3DM	OBJ
Open Specification	5	5	3	0	5
Compatibility	5	5	5	5	5
Standardization	0	5	0	0	0
Duration of guarantee	0	0	0	0	0
Duration of support	0	0	0	0	0
Market penetration	3	1	5	1	3
No. of independent producers	1	5	1	1	5

Table 1: Evaluation results per file-format

Some of the results are exemplarily described in more detail to clarify the evaluation process:

- Duration of guarantee / duration of support: No information was publicly available for these two criteria, so these criteria are always evaluated with zero (since all file-formats have the same value here, the ranking is not influenced). Data like these are typically given by software companies during sales negotiations.
- Open specification: Open specifications exist for the DXF/DWG [1], IGES [5] and 3DS/MAX [2] file-format. 3DS only gets three points, since the last found specification is from 1997, although 3DS is still under development by Autodesk.
- Compatibility of IGES: At the time of its creation IGES was compatible with most available software products. Meanwhile in PRONOM only one compatible software is listed: Adobe FrameMaker 2002; in a web-search additional software products, such as ModelPress Desktop, CrtIView or 3D Shop ModelScan are named (see <http://www.programurl.com/>, Date of Download: 09.03.2007). Additionally a conversion tool for Autodesk exists.
- Standardization of IGES: IGES has been standardized by the Department of Defense and the National Institute of Standards and Technology [5].
- Market penetration of 3DS MAX: Wikipedia [13] lists 42 software companies, which use the 3DS MAX format, among them major producer of computer games and animated movies.
- No. of independent producers of OBJ: According to Wikipedia, the OBJ file-format has been adopted by several software vendors and can be imported and exported to a number of software programs.

As can be seen, the above shown evaluations rely on Internet-sources only. We recommend a detailed clarification with software vendors before deciding for one format or another.

Rank	File-Format	Preservation Risk
1	IGES	40.00 %
2	OBJ	48.57 %
3	DXF	60.00 %
4	3DS	60.00 %
5	3DM	80.00 %

Table 2: The final evaluation result

After the evaluation step importance factors are set for each criterion. These factors indicate how the end-user values certain criteria. In the here shown example, all seven criteria get the same weight – 14.29 %. The evaluation results are multiplied with the weight of its criterion and summed up per file-format. By taking the percentage value from the maximum possible value (which is five) and by subtracting it from 100, the preservation risk can be obtained. The final result is shown in Table 2. The differences between the file-formats in terms of preservation risk are significant and IGES is ranked top as a format for long-term preservation.

6 Conclusion

In this paper a methodology for evaluating file-formats in terms of reliability for long-term preservation is presented. In the first part the steps of the evaluation process are described in detail. In the second part of the paper a proof-of-concept is done for 3D-file-formats to show the functionality and details of the process in practice.

After evaluating several file-formats, a file-format list can be created, where all selected formats are ranked according to their preservation risk. Such a list could be maintained by a research institution or a library and could be continually updated. By including software companies and the open-source community into the evaluation process, the evaluation results can on the one hand become more precise and on the other hand become a motivation for improving the preservation reliability of file-formats. Additionally such a ranking could be added to existing file-format registries, such as PRONOM or the Global Digital Format Repository.

Notes and References

- [1] AutoCAD2006. *DXF Reference*, July 2005. URL <http://www.autodesk.com/>, Date of Download: 04.02.2006.
- [2] Autodesk Ltd. *3D-Studio File Format*, January 1997. URL <http://www.martinreddy.net/gfx/3d/3DS.spec>, Date of Download: 04.02.2006.
- [3] *FILEExt - The File Extension Source*, 2007. URL <http://FILEExt.com>, Date of Download: 31.01.2007.
- [4] MCGOVERN, N. Y.; KENNEY, A. R.; ENTLICH, R.; KEHOE, W. R.; BUCKLEY, E. *Virtual Remote Control, Building a preservation risk management toolbox for web resources*. D-Lib Magazine Volume 10, Number 4 (2004).
- [5] National Institute of Standards and Technology. *Initial Graphics Exchange Specification (IGES)*, April 1996. FIPS PUB 177-1.
- [6] *PROBADO - Prototypischer Betrieb fuer Allgemeine Dokumente*, 2007. URL <http://www.probado.de>, Date of Download: 05.02.2007.
- [7] *PRONOM, the technical registry*. URL <http://www.nationalarchives.gov.uk/pronom/default.htm>, Date of Download: 07.07.2006.
- [8] RAUCH, C.; RAUBER, A. *Preserving digital media: Towards a preservation solution evaluation metric*. In Proceedings of the 7th International Conference on Asian Digital Libraries, Shanghai, ICADL 2004 (December 2004), Springer-Verlag Berlin, Germany, pp. 203–212.
- [9] SLATS, J.; VERDEGEM, R. *Practical experiences of the Dutch Digital Preservation Testbed*. VINE, The journal of information and knowledge management systems, Volume 34, Number 2 (2004), 56–65.
- [10] STANESCU, A. *Assessing the durability of formats in a digital preservation environment*. D-Lib Magazine 10, 11 (2004). URL <http://www.dlib.org>, Date of Download: 14.03.2005.
- [11] STRODL, S.; RAUBER, A.; RAUCH, C.; HOFMAN, H.; DEBOLE, F.; AMATO, G. *The DELOS testbed for choosing a digital preservation strategy*. In Proceedings of the International Conference on Asian Digital Libraries, ICADL (2006), Springer-Verlag, Berlin, Germany.
- [12] WEIRICH, P. *Decision Space: Multidimensional Utility Analysis*. Cambridge University Press, 2001. URL <http://www.missouri.edu/weirichp>, Date of Download: 03.08.2005.
- [13] *WIKIPEDIA, The free Encyclopedia*, 2007. URL <http://en.wikipedia.org>, Date of Download: 20.02.2007.

Beyond Publication – A Passage Through Project StORe

Graham Pryor

University of Edinburgh
Digital Library Division
George Square, Edinburgh, Scotland
e-mail: graham.pryor@ed.ac.uk

Abstract

The principal aim of Project StORe is to provide middleware that will enable bi-directional links between source repositories of research data and the output repositories containing research publications derived from these data. This two-way link is intended to improve opportunities for information discovery and the curation of valuable research output. In immediate terms, it is expected to improve citation rates as a consequence of increasing the accessibility of research output. A survey of researchers in seven scientific disciplines was used to identify workflows and norms in the use of source and output repositories, with particular attention being paid to the existence of common attributes across disciplines, the functional enhancements to repositories considered to be desirable and perceived problems in the use of repositories. Cultural issues were also investigated. From the results of the survey, a generic technical specification was designed and a pilot environment created based upon the *UK Data Archive* (source repository) and the London School of Economics' *Research Articles Online* (output repository). A further link to a prototype institutional repository at the University of Essex was used as a control mechanism. The StORe middleware was designed using a Web 2.0 approach similar to existing FOAF (Friend Of A Friend) services such as Flickr and MySpace, but incorporating a federation of institutional, source and output repositories rather than one central area where digital objects are deposited. Researchers can deposit digital material in various formats at their institutional repositories until the data and publications are made available at linked source and output repositories. An enabling central portal provides an OAI-based aggregator service, which harvests the contents of the federation's repositories and provides a simple search facility. Whilst all digital objects are title visible, a key feature of the middleware is the Flickr-like option for regulating access, which gives researchers control over who can see objects they have designated 'non-public'. Using the StORe middleware, it will be possible to traverse the research data environment and its outputs by stepping seamlessly from within an electronic publication directly to the data upon which its findings were based, or linking instantly to all the publications that have resulted from a particular research dataset. It has already been endorsed by participating researchers as having the potential for integrating multiple data sets from different publications. Following completion of the pilot demonstrator, an independent evaluation undertaken by the National Centre for e-Social Science found it effective and easy to use. It may also be said to have broadened the meaning of the terms *publish* and *publication*.

Keywords: interoperability; research publications; institutional repositories; middleware

1 Introduction

Project StORe is an initiative funded by the UK's Joint Information Systems Committee within its 2005-7 Digital Repositories Programme.^[1] StORe's principal aim is to attach new value to published research through the provision of two-way links between the output repositories that contain research publications and the source repositories of original and processed data from which those publications originated. Hence the project name, which is an acronym of **S**ource to **O**utput **R**epositories. This bi-directional linkage is predicted to increase opportunities both for information discovery and the curation of valuable research data. Specifically, it will provide members of the research community with the means to navigate directly from within an electronic article to the source or synthesised data from which the article was derived; conversely, direct access will also be provided from source data to the publications associated with those data. Researchers will benefit from this linkage through an enhanced capacity to track the use and influence of their published research, as well as to engage in the more comprehensive dissemination of research and scholarship, which it is anticipated will increase the citation rate for research papers linked to their sources. Scientific researchers involved in the development phases of the project have already identified other advantages, such as the ability to conduct a reanalysis of source data as new methods emerge, a feature that should lead to improvements in the integrity of

published results, whilst the potential for integrating multiple data sets from different publications has been perceived as promising time saved and more productive research.

On the subject of reanalysis, an incident reported in *Science* late last year^[2] underwrites the potential value from being able to take a critical look at a published paper alongside its data. In a September 2006 paper in *Nature*, Swiss researchers cast serious doubts on a protein structure described in a 2001 *Science* paper by Geoffrey Chang's pioneering group at the Scripps Research Institute, San Diego. Upon investigation, Chang found that his homemade data-analysis program had inverted the electron-density map from which he had derived the final protein structure. Consequently, Chang and his colleagues had to retract three *Science* papers and report that two papers in other journals also contained erroneous structures. If his original paper and its data had been published together perhaps this mistake would have been discovered earlier.

Having referred here to the dual *publication* of a paper with its data opens up a more controversial realm than is first suggested by the design of a piece of functional middleware, since one may speculate that the provision of a mechanism for accessing not just electronic publications but also their underlying data raises fresh questions about the nature and meaning of the terms *scholarly* or *scientific publishing*. Reflecting on the open access publishing and repository movements, one detects a strong current of opinion that making data available does not constitute publication, which benign strategy contributes of course to the avoidance of unhelpful quarrels with publishers; but greater flexibility of interpretation and less defensiveness would be both appropriate and defensible, since the publication of scholarly papers and the dissemination of data are necessarily distinct acts, each being defined by their particular purpose. Any set of data selected specifically for inclusion in, or as the basis for a scientific paper is chosen with the principal purpose of helping to persuade the reader to accept a hypothesis or theory as proven, and its value is gauged by the degree to which it supports the effectiveness of the set piece of rhetoric that is the paper. The larger collection of data from a research programme, possibly archived in a source repository, does not serve that same purpose of persuasion. Indeed, it may be argued that by making this broader cache of data accessible via a link from a scientific paper to its source repository could even subvert the arguments in the paper, should there be weaknesses in the data or the research, although from a different perspective this does strengthen the case for the bi-directional link as a mechanism for ensuring the integrity of the source data. So whether making data from a source repository publicly available is an act of dissemination or publication, the answer is probably irrelevant. What is more to the point is the impact from enabling dual accessibility.

The impetus for Project StORe came from a belief held by members of the research library community that an achievable set of functional enhancements to both source and output repositories could be identified and built, on a generic basis, as a piece of middleware, and that this might be approached in a manner similar to the way in which digital library technologies have produced generic tools in other heterogeneous environments, such as 'metasearch' interfaces to publisher and local databases, metadata harvesters and link resolvers. These tools are based upon recent digital library protocols and standards such as OAI-PMH,^[3] qualified Dublin Core^[4] and OpenURL.^[5] Project StORe was therefore conceived as a vehicle for undertaking the essential groundwork preparatory to building a production system solution that would meet the requirements for permitting useful interoperation between the two repository types, and it would be undertaken using the systems, standards and metadata protocols developed and used in other JISC projects, where appropriate, to ensure the widest possible interoperability. Its rationale would be that of a proof of concept, but from the start there was a firm aspiration to deliver an authentic pilot infrastructure capable of translation across multiple disciplines.

2 Methodology

In the first phase of the project a survey of researchers was conducted across seven scientific disciplines in the UK to understand their workflows and working philosophies, as well as to identify norms in the use of source and output repositories. The disciplines investigated were archaeology, astronomy, biochemistry, the biosciences, chemistry, physics and social sciences. The astronomy survey had a broader base, including members of the astronomy research community in the USA, in recognition of the internationally collaborative work undertaken by astronomy research teams at Edinburgh and Johns Hopkins universities and the discipline's separate Mellon-funded analysis of repositories and applications. The survey, which was carried out over four months in 2006, first through an online questionnaire and subsequently by one-to-one interviews, addressed such issues as the existence of common attributes across disciplines (in terms of the data formats employed, the quality and method of metadata assignment, and the volume of data produced), the functional enhancements to repositories that were considered to be desirable, and the nature of problems experienced in the use of repositories. Cultural and organisational issues were also investigated, ranging from attitudes towards the concept of open access publishing to the measures employed for sharing and protecting data. Invitations to

participate in the online questionnaire were sent to 3,700 scientific researchers and produced a return in excess of 10%, whilst the in-depth interviews were held with between 10 and 15 respondents per discipline, selected to ensure an equitable representation from all stages of the academic/research career path. Each individual discipline survey produced a published study that described the source and output repositories used by members of that discipline, including a brief history and statistical information on their use, with a detailed analysis of responses to the questionnaire and the structured interviews. These reports, which have been archived in the Edinburgh Research Archive (ERA), also incorporate scenarios and use cases.^[6]

Project partners at university libraries identified staff to undertake the discipline surveys, with a view to exploiting their knowledge and the effectiveness of their relationships with researchers 'on the ground'. The libraries responsible for the survey work and the disciplines they surveyed are shown in Table 1.

Surveying University Library	Subject
Edinburgh (lead) / Johns Hopkins	Astronomy
Birmingham	Physics
Imperial College	Chemistry
London School of Economics	Social Sciences
Manchester	Biosciences
University College London	Biochemistry
York (for the White Rose Partnership)	Archaeology

Table 1: Project partners & survey disciplines

Whilst it was important to the design of a relevant and appropriate solution that actual research working practices and environments would be identified and understood, the survey team's principal role was to address the requirements for new functionality within source and output repositories that would permit interoperability from the point of ingest, so that authors of papers could insert links to data and to published/unpublished papers, associating newly deposited publications with data held in data repositories. It was anticipated that a number of new operations could be supported within the two types of repository, both for academic submitters and for repository users, including automatic link creation, automatic embedding of source repository metadata, and a facility to run operations upon data. The desirability of these features was explored in depth during the interviews.

Upon completion of the survey, a business analysis of the survey reports was undertaken by staff at the UK Data Archive (UKDA).^[7] This analysis was used as the foundation for a generic technical specification of the proposed bi-directional link, with the aim of translating real requests for 'missing' functionality into a structured technical architecture. The assumptions and deductions made in the business analysis were then tested with research active staff at the University of Essex and with library professionals from the London School of Economics (LSE), leading to further refinements to the specification.

In the final phase of this development process, the generic technical specification has provided the platform for the pilot implementation of a working bi-directional link. This has featured social sciences data and publications exclusively, using the *UKDA* as the source repository and the LSE's *Research Articles Online* as the test output repository, augmented by a further link to a prototype institutional repository at the University of Essex, which served as a control mechanism. It should be emphasised that limiting the pilot to only one of the original seven disciplines has been necessary to meet the logistical constraints of a test environment, but in building that environment the full set of requirements established by the survey of seven disciplines has been incorporated with a view to proving the middleware as a generic, non-discipline specific tool.

Throughout, the rationale of Project StORe has been to anchor technical and user aspirations to the pursuit of practical benefits. During the pilot implementation, a critical element of the process has been user testing, involving members of the original cohort who responded to the survey, and at its conclusion the pilot demonstrator has been subject to a rigorous, independent evaluation by the National Centre for e-Social Science,^[8] which has depended for its legitimacy upon user participation in a series of workshops.^[9]

3 Survey and Analysis

A majority (85%) of respondents to the StORe survey judged the provision of a bi-directional link as likely to prove advantageous to the research process, with a small preference overall for an output to source link. Key benefits were described as an opportunity to access the large data sets it is not possible to reproduce in an article;

and more specifically, an output to source connection would enable the comparison of results, thereby providing the means to authenticate claims made, which was deemed to be of particular value where claims are considered controversial.

By selecting from prepared lists, respondents were asked to identify the data types and their formats that might be generated during research, with the range of data types given in the lists appearing to satisfy the majority as being representative, and with no data type receiving a nil response. A further 32 *Other* types were also declared but were found to describe either a sub-type of items from the lists, the name of experimental equipment or process-specific data sets. Nonetheless, across and within the seven disciplines, the volume, range and diversity of data produced was confirmed as considerable. Whilst generic types such as drawings, plots, images and text-based files scored highly, each showing in excess of 150 responses, noteworthy scores were attributed to more specialised types such as radiographic data (11), remote sensing surveys (15) and gene/protein sequences (42).

In terms of data format, image files, spreadsheets and word processed files comprised the majority, with around 200 responses each. In the next tier, plain text, database files, portable document format and tables/catalogues all scored more than 100 responses. Of the 76 *Other* formats volunteered by respondents, those that were not species from the main selection list tended to be proprietary and linked to specific discipline processes or equipment. Of greater significance to the design and maintenance of links from publications to their source data is that almost three quarters of the survey's respondents were found to generate and use complex data sets (i.e. data produced and held in combinations of data formats and files).

All of the seven disciplines identified barriers to the deposit of data or publications in repositories, citing time constraints, the bureaucracy imposed by repository administration and structures, or constraints arising from their own or others' intellectual property rights. A perceived inconsistency across all repositories was also reported in terms of content coverage and in the standards and methods used for keywords, metadata and data formats. It was in this latter area that the most powerful consensus was found amongst the survey cohort, with the appropriate assignment of metadata being roundly acknowledged as critical and demanding, both intellectually and in the time required to do it well. Perversely, this consensus on the need for good metadata did not necessarily translate into good practice, there being a high level of self-assignment and with limited evidence that standard schema or thesauri were being employed. Perceived responsibilities for metadata assignment are illustrated by the following table from the StORe questionnaire.









I decide which terms to use and I assign them		212
Research colleagues assign metadata on the team's behalf		55
Research support staff assign metadata on the team's behalf		22
Metadata are assigned by library/information services staff		4
Metadata are assigned by the repository administrators		37
Metadata are generated automatically		63
It is not known who assigns metadata		68
Other (please specify)		37

Table 2: The Assignment of Metadata to Research Data

In order to establish whether there is a core set of metadata that might satisfy the needs of researchers in the seven disciplines, respondents were invited to identify key terms from a predefined list and to suggest their additional requirements. A large majority subscribed to the list as representing a functional generic suite of metadata, selecting such terms as project title, description and reference numbers, together with keywords, project and publication dates and format. Only 58 *Other* terms were suggested, and these were found to be highly discipline specific (e.g. archaeological period, celestial object, position and observation date, chemical entity, protein sequence).

As shown in Table 2, the subject of metadata provision revealed a broad spectrum of awareness and response amongst the survey cohort that was sustained when they were asked to indicate the point at which metadata are assigned. Assignment 'during file saving' attracted the highest score of 142, but there was insufficient evidence to deduce whether such a practice represented a properly structured activity or merely the casualty of afterthought. More reassuring were responses to the options 'Prior to data creation' (82), 'As part of the indexing

process' (98) and 'When submitting data to the repository' (89). Of some concern were the 35 respondents to this question who believed no metadata were being assigned to their research output, with a further 75 admitting they were not sure at which stage metadata are assigned.

The disjunction between aspiration and practice in the assignment of metadata is perhaps explained by tensions between the prevailing research culture and embedded attitudes towards the support services. It was made clear during the StORe survey that researchers from all disciplines favoured self-reliance in matters associated with data management and the use of repositories, as opposed to the provision of institutional support from the library or other areas of professional expertise. The inherent culture of self-sufficiency within research groups or programmes, where normal practice is to manage all aspects of the research lifecycle internally, was evident from statements submitted during the StORe survey. Whilst this culture has given rise to the development of some highly effective data repositories focused on serving specific disciplines, the general effectiveness of a self-sufficient approach to accessing, organising, promulgating and curating data was not demonstrated across the scientific research spectrum.

National and international strategies for data deposit and preservation are of course already emerging. One can point, for example, to the Wellcome Trust's flagship initiative to mandate the deposit of research publications in the biosciences, which mandate is anticipated will extend to the deposit of data; or to the astronomy community's Virtual Observatory, an initiative to make all the astronomy data in the world easy to access.^[10] They are not isolated examples, but when one considers the research milieu as a whole their considerable progress was found not to be typical. At the level of the individual researcher, whether asked about metadata assignment in particular or data management in general, responses such as "it's my problem, I'll deal with it" were commonplace. Whilst libraries have conducted advocacy campaigns on behalf of open access publishing and repositories, in some cases providing technical expertise to support the use of repositories, researchers canvassed by the StORe survey in most cases perceived there was no support available, they had little confidence in what support was known to be provided, and they claimed sufficient familiarity with information technology to consider themselves self-reliant. Yet at the same time as declaring they would not normally associate the management of research data with librarians, and evincing little apparent demand for assistance in seeking and navigating information, there was evidence of a clear requirement for information intermediaries to assist not only in the construction and maintenance of metadata but also in the preservation and curation of data. This dichotomy was reflected in a further aspect of the survey, which concerned researchers' attitudes towards making data available, and would prove a singular force in the design of the StORe middleware.

With few exceptions, respondents to the survey supported the statement that it should be a requirement for data from publicly funded research to be made freely available, but generally with the caveat that access should be restricted until results are published in a paper, in order to prevent data scavenging. Others noted that whilst this might be a creditable aspiration, without a data administrator it represented a potentially large burden from editing, compiling and sanctioning the release of data. In fact, both the provision of access and the sharing of data were found to be constrained by a lack of confidence in processes, and it was difficult to conclude whether some practices were deliberately designed to frustrate accessibility. For example, the storage of unique and original research on PCs and laptops was found to be common practice, and the failure to take a more relaxed approach to access was influenced by a perceived absence of adequate protection in networked systems. As one respondent described his data management regime: "data is held on secured CDs in encrypted format with only an identifying code. The codebook is kept physically separate".

The StORe survey revealed a range of diversity in practice and attitude, both within and between the seven disciplines, but with a consistently firm body of consensus when it came to explaining fundamental needs. When searching for information, a universal preference for simple keyword searching was declared and browsing amongst library shelves appears to have been replaced by browsing within repositories and other online resources. This practice is of course only effective when enabled by the functional efficacy of application and metadata structures, designed by system and data experts to meet the clamour for a 'Google-type' approach to searching.

4 The Generic Model

The business analysis that followed the StORe survey revealed sufficient shared ground between the disciplines to suggest the basis for a common model. To recap, an examination of the discipline-specific reports produced a majority in every discipline favouring two-way links between data repositories and publications, but with barriers to the actual deposit of data or publications found to be a consequence of time constraints, organisational bureaucracy or concerns over intellectual property rights, although the concept of data sharing was considered

fundamental and important. A perceived inconsistency across all repositories in terms of coverage, standards and data formats was reported, with a simple 'Google type' approach to searching being preferred. Researchers from all disciplines also seemed to exhibit self-reliance in matters of data management and in the use of repositories, whilst recognising the need for assistance in the provision of some common minimum metadata.

Taking this level of consensus, the design of the model for a bi-directional link has adopted a Web 2.0 type approach, similar to existing FOAF (Friend Of A Friend) services such as Flickr or MySpace, but incorporating a federation of institutional, source and output repositories rather than one central area where digital objects are deposited. Articulation of a Web 2.0 rationale for the middleware has been a deliberate decision aimed at meeting cultural aspirations for self-determination and those individual anxieties concerning data ownership that were revealed during the survey, since it places control firmly in the hands of the researchers. In this model, objects deposited in federated repositories would be referenced by persistent identifiers that include domain identifiers, with researchers depositing digital material in various formats at their institutional repositories until the data and publications are ready to be made publicly available at linked source and output repositories. This focus on the institutional repository environment is predicted to have further value in providing a context for future implementations of asset-based research data repositories, in cases where global services from established discipline platforms such as astronomy's Virtual Observatory or the social sciences' UKDA are not provided, and discipline needs could be met instead by a regime of institutional data curation.

What may be described as the central StORe portal has been designed as an OAI-based aggregator service that will harvest the contents of a federation's repositories and provide a simple search facility based on centralised indexes. This basic level of searching can be enhanced for individual disciplines by the inclusion of domain ontologies, reflecting the need highlighted in the survey to enable discipline-specific terminology. All digital objects will be title visible to all, but researchers can restrict access to non-public objects to communities of project-specific colleagues, institutional colleagues, personal colleagues, or all of these. This is similar to the option for restricting access to family and/or friends in Flickr, in order to bar public access to private photographs, and is again a direct attempt to satisfy the demands of researchers to remain in command of their data.

Access management has proved to be a defining feature of the StORe middleware. Some data repositories are open to all enquirers, while others are password-protected, and in a scenario where users of open access research publications wish to view data in repositories to which access is normally controlled, a validation process will be required in order to allow temporary access rights. In this context we have investigated the authentication and authorisation issues involved with reference to the developing international work on Shibboleth, a federation-based architecture that enables organisations to build single sign-on environments for accessing Web-based resources.^[11] Whilst it is not yet in place, it is planned that a production version of the central StORe portal will authenticate through Shibboleth, using a simple deposit interface to request the minimum amount of mandatory metadata for each object, identify the group or individual to which it is accessible and check whether it is a candidate for public submission. Until Shibboleth is adopted, we are applying a dummy Shibboleth mechanism for allocating user names and passwords. This will trigger an automatic process for setting up user accounts when legitimate users log in for the first time.

The minimum metadata required for any individual item is a title, provided the item is being associated with project data in a repository already assigned the metadata elements *author*, *title*, *geography*, *time*, *keywords* and *abstract*. The digital object will be deposited in the researcher's institutional repository, whilst the metadata and access conditions will be stored centrally; in turn, the search indexes will be built up from the centrally held metadata and harvesting from the objects themselves. This harvesting can also be used in the creation of the discipline-specific ontologies needed to satisfy metadata requirements that are not met by the generic core. Both source and output repositories in the federation will regularly trawl for potential acquisitions and, if a publication or data are accepted, the repository will supply a public link to a peer-reviewed version of the publication or to the data.

Hence, the generic model planned to be tested by the pilot demonstrator combines informal networking and sharing of data with a public access system that supports stronger links between data sources and publications. A user entering a StORe generic portal would log in to authenticate and the system will respond by determining his/her organisation, recorded preferences and known colleagues. Options would then be made available to browse any new activity of colleagues; to browse any objects available to the user (i.e. the user's own and other colleagues' objects); to search all discipline-specific or all repository-specific objects, with a further option to filter on a temporal basis; to deposit an object; to create a new project; to make an object available to another user; to request that an object be made available; to submit an object to an output repository for publication; to

submit an object to a source repository for preservation; to download a repository object; and to edit, delete, organise or manage the user's own objects.

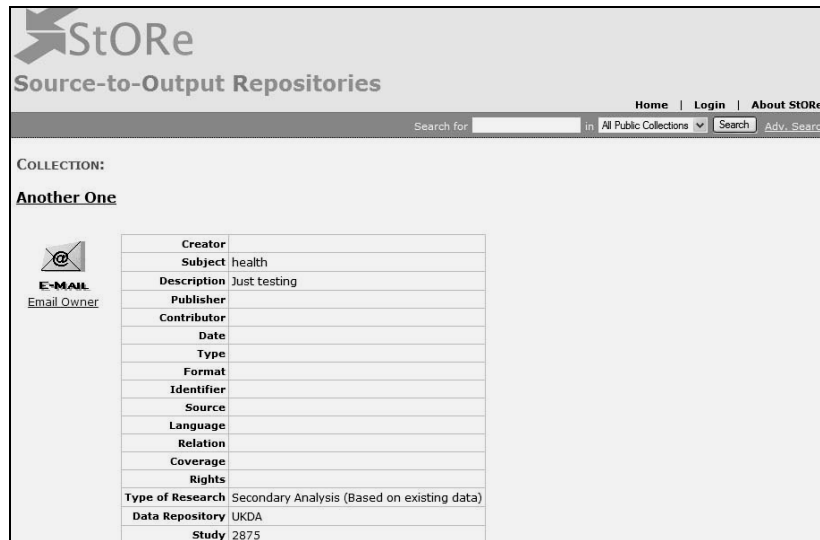
It was clear from the outset that the success of this model would be determined by three factors. Researcher acceptance of Web 2.0 technologies was essential, and we have been actively encouraged here by the younger members of the user testing cohort who already work routinely within that environment. Persuading researchers to use a third party portal for deposit into their local institutional repository was also acknowledged to be challenging, whilst the third and possibly most difficult obstacle lay in the resolution of potential security and policy objections to sharing sensitive data across institutions. Eventually, it was decided that these barriers could be broken down in a stage by stage approach that would embed a federation in the established publishing process and restrict the sharing of non-public data to institutional colleagues. A demonstration of simplicity would be the key to stage one, with the objects stored required to be identifiable only by title, discipline, project, file type and format, employing minimum Dublin Core metadata elements. In the second stage, each individual institutional repository would act as a portal to itself and all the domain specific source and output repositories in its federation, thereby preserving familiarity of the working environment but allowing the addition of Web 2.0 and FOAF features. Only at stage three would the concept of a StORe subject or domain portal be openly introduced to the discipline-specific elements of the federated repositories. Here, one solution to security concerns would be the temporary copying of protected objects to the portal for download within a prescribed period.

Looking beyond the pilot environment, this approach offers wider coverage, more choice of source and output repositories and more scope for Web 2.0 service features. There could even be a common interface for deposit to individual institutional repositories, and it was envisaged that listing of forthcoming conferences, wikis, and other networking facilities might encourage use. The final stage would see the full generic solution implemented, comprising the entire federated institutional, source and output repositories that have adopted the approach outlined in stage one. This solution is well placed to encourage cross-disciplinary research, a key driver in the modern research environment, although metadata mappings will have to be employed and even more additional features devised to encourage the use of such a universal portal.

5 A Passage Through Project Store

StORe's pilot demonstrator was built for a test federation using the UKDA as source repository and the LSE's *Research Articles Online* as the output repository, complemented by a prototype institutional repository at the University of Essex.^[12] Options for linking to a commercial publisher had also been explored but were considered logistically too ambitious for a pilot implementation. The pilot was designed and implemented between November 2006 and April 2007, and what follows is an abridged system walkthrough showing how items (data and publications) are managed.^[13] This description is of a standalone system, but in a live working environment access could be initiated within an electronic article in an output repository or from a source repository having an association with the federation.

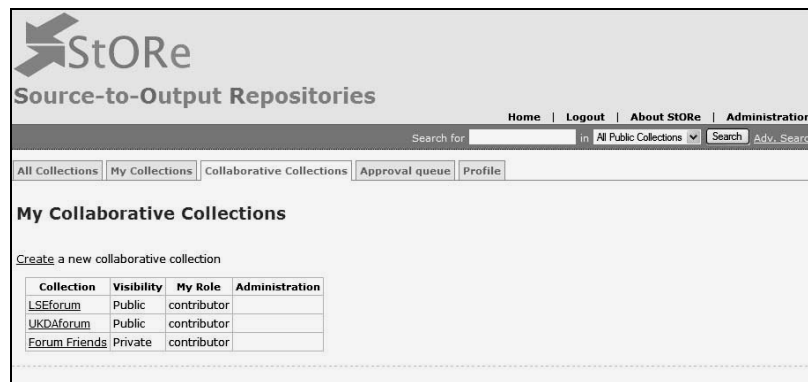
In the pilot, as in a working system, it is possible for an unregistered user to search or browse across all or specific research collections in the federation, but any titles marked as private will not function as a hyperlink to their content. Collection metadata can, however, be seen via a *View Collection* link. If the research project from which the target collection was generated involved the secondary analysis of existing data, a link to the underlying data will already exist, and will take the user to the relevant Web page of the supporting source repository. If the collection owner has agreed, then a further link will appear, allowing users to send an email requesting additional details or to be granted access to items in the collection.



Creator	
Subject	health
Description	Just testing
Publisher	
Contributor	
Date	
Type	
Format	
Identifier	
Source	
Language	
Relation	
Coverage	
Rights	
Type of Research	Secondary Analysis (Based on existing data)
Data Repository	UKDA
Study	2875

Figure 1: View Collection Metadata Screen

Registered users logging in to the pilot federation can view the content of all items in their public and private solely-owned collections. They can also see those items in public collaborative collections with the UKDA or *Research Articles Online* where they are a contributor, and may view public or private collaborative collections made with project colleagues or ‘friends’, where they are identified as either contributor or administrator. Collaborative collections are linked via a unique *LinkID*. In the example below, the user (identified as Forum) is a member of a private collaborative collection created by another researcher in order to share documents with Forum. Each collaborative collection is distinguished as a collection type, either source/archive, output/publisher or user/researcher. The logged-in and authenticated user has access to full functionality and can create private or public, solely-owned or collaborative collections, including an option to allocate other registered users to a collaboration.



Collection	Visibility	My Role	Administration
LSEforum	Public	contributor	
UKDAforum	Public	contributor	
Forum Friends	Private	contributor	

Figure 2: Collaborative Collections

Figure 5, overleaf, shows how the process of adding metadata to a publication has been kept simple. At collection level, apart from the collection name and description, only subject terms and the type of research and study (if secondary research) are mandatory. All other Dublin core fields are optional. The subject terms can be directly typed into the box or chosen from a list of tags displayed at the right-hand side of the page. The type of research is selected from a drop-down menu (Figure 3),

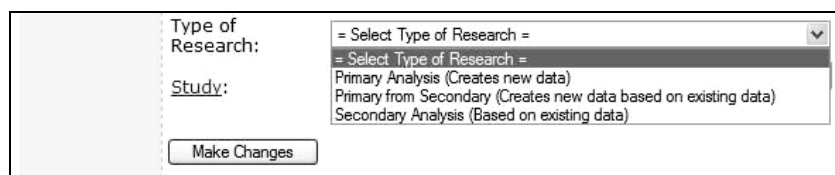


Figure 3: Allocating Research Type

and a study number corresponding to the number assigned to the corresponding data within the source repository is chosen from a further drop-down list (Figure 4).

The screenshot shows a form with a 'Study:' label, a text input field containing '4800', and a dropdown menu labeled '= Select Repository ='. The dropdown menu is open, showing options: 'UK Data Archive', 'ICPSR (Michigan)', and 'LSE (eprints)'. There is also an 'Add Study Field' link and a 'Make Changes' button.

Figure 4: Selecting The Study Number

The study number then becomes a link to the appropriate page within the repository's web site.

The screenshot shows the 'EDIT' page for a collection named 'Edinburgh'. The page has a header with the StORe logo and navigation links. The main content area includes an 'EDIT' section with various metadata fields: Creator (Ken), Subject (pilot), Description (A collaborative collection that I might use in the pilot demonstration in Edinburgh), Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights, and Type of Research (Secondary Analysis). The 'Study' field is set to '4800' and has a dropdown menu for repository selection (UK Data Archive). A 'Unique link' annotation points to the 'Study' field. There is also an 'ADD SUBJECT TAGS' section on the right with a list of tags and an 'Add Selected' button.

Figure 5: Assignment of Metadata

Moving a data item to the UKDA collaborative collection is a two-tier process. First, the data's identity is verified (which will enable publications based on this data to be moved and approved in *Research Articles Online*) and, where required, an embargo can be set by the data owner. Once verified, an acquisition number is assigned to the data item in the UKDA collaborative collection, which as already intimated will subsequently be assigned to any associated publications moved to a *Research Articles Online* collaborative collection. Upon approval by the UKDA this acquisition number is replaced by the actual research study number, which will function as a link to the data from its publications in *Research Articles Online*.

In StORe, individual items or folders are added to a collection either singly or bundled. Only the provision of an additional title and file name (or URL) is required, since each item adopts all the metadata associated with the

collection itself. Files in different formats (Word or PDF documents, URLs, image files, etc.) are associated to an item or folder, and the Dublin Core fields may be edited if required to produce a more specific metadata record. When a scientific paper ready for publication is moved from a researcher's institutional repository into a collaborative collection owned by *Research Articles Online*, all the metadata associated with it moves as well. Simultaneously, the middleware automatically assigns a metadata term to identify the collection of origin, and confirms that corresponding data exists in the UKDA collaborative collection. It also provides functionality enabling the addition of further files or URLs to the item, or to add additional metadata.

6 Conclusions

The StORe pilot has demonstrated the feasibility of a bi-directional link within the specific context of a single discipline. However, despite the level of consensus identified by the survey, discipline variations would need to be managed during export of the StORe model across other domains. Individual institutional repositories will also contain different file types and formats, and will apply different metadata standards. For certain disciplines data interpretation, manipulation and methodology are as, if not more significant than access to the raw data, and although a simple search might cross disciplines, more advanced discipline-specific searches would be more in demand, with the resulting hit lists, relevance ranking and sorting being different for each discipline. Consequently, both subject and global portals will require different Web 2.0 features for each discipline.

Recognising the key preferences and practices of researchers interviewed during the StORe survey, the solution developed showed that traditional practices for the informal networking and sharing of data could be combined with a public access system supporting stronger links between data sources and publications. The StORe solution gives researchers the means to manage a level of privacy and access defined by themselves, countering expressions of apprehension towards full open access, which some saw as a threat to data ownership. It also offers a simple Google type search, preferred amongst the majority of those surveyed, and viewed by many as an effective tool for replacing the option of browsing amongst shelves in a library, although Boolean operators and wildcard functionality are made available for more advanced searches. Using the StORe middleware, researchers can move seamlessly around the research data environment and its outputs, stepping from within an electronic publication directly to the data upon which its findings were based, or linking instantly to all the publications that have resulted from a particular research dataset. By intrinsically connecting the process of publishing scientific papers with the provision of their underlying data, StORe has also broadened the connotation of the terms *publish* and *publication*.

References

- [1] http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories.aspx
- [2] MILLER, G. *A Scientist's Nightmare: Software Problem Leads to Five Retractions*. Science, 22 December 2006, pp. 1856-1857
- [3] An explanation of the Open Archives Initiative (OAI) and the OAI protocol for Metadata Harvesting may be found at <http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/faq/oai/>
- [4] The Dublin Core metadata element set is explained at <http://www.ukoln.ac.uk/metadata/resources/dc/>
- [5] An OpenURL demonstrator can be found at <http://www.ukoln.ac.uk/distributed-systems/openurl/>
- [6] The individual discipline reports, together with the survey overview, may be examined as documents within the Edinburgh Research Archive, <http://www.era.lib.ed.ac.uk/handle/1842/1412> or at the project wiki, <http://jiscstore.jot.com/SurveyPhase>
- [7] UK Data Archive at the University of Essex, <http://www.data-archive.ac.uk>
- [8] <http://www.ncess.ac.uk>
- [9] The NCeSS evaluation plan can be examined at <http://jiscstore.jot.com/EvaluationOfPilot>
- [10] <http://www.virtualobservatory.org/>
- [11] http://www.athensams.net/federations/shibboleth_intro.aspx
- [12] Both institutional output repositories have been constructed using open source software: *ePrints* at the LSE and *Fedora* at Essex.
- [13] A full walkthrough of the StORe middleware can be accessed at the project wiki, <http://jiscstore.jot.com/PilotDemonstrator>

Challenges in the Selection, Design and Implementation of an Online Submission and Peer Review System for STM Journals

Judy Best; Richard Akerman

Technology and Research; Canada Institute for Scientific and Technical Information
National Research Council of Canada, 1200 Montreal Road, Ottawa, Ontario K1A 0R6 Canada
e-mail: judy.best@nrc-cnrc.gc.ca; richard.akerman@nrc-cnrc.gc.ca

Abstract

Two international scientific publishers collaborated to develop an Online Submission and Peer Review System (OSPREY) for their journals. Our goals were to meet market demand, increase editorial efficiency and streamline the transition from peer review to publishing. One of the publishers (National Research Council (NRC) Research Press, Canada) had previously purchased a third-party system that was subsequently discontinued by the vendor. Because of this experience and its complex requirements, we decided to build rather than buy a new system. The collaboration with the second publisher, Commonwealth Scientific and Industrial Research Organisation (CSIRO) Publishing, Australia, allowed sharing of resources within a common vision and goals. Agile development through the use of iterations allowed us to continuously add functionality, make improvements and incorporate new requirements. The development team included technical staff as well as stakeholders, future users, business analysts and project managers. The architecture chosen was based on open source technologies, with Java servlets and Java Server Pages for the Web interface. OSPREY currently supports 32 journals at the two publishers. Users accomplish all regular tasks in peer review (submission, selection and invitation of reviewers, submission of review, recommendations and decision) through the software. Editorial staff verifies submissions, sends correspondence and assigns customizable roles and tasks. All tasks are accomplished through a Web browser accessing the application on central servers at the publisher, with no special software or configuration required for any users. Currently, the system integrates with the publishing system by generating manuscript metadata in an XML format, although closer integration with a workflow management system is planned. Since OSPREY implementation, the number of submissions has risen, although marketing and higher ranking of the journals are also factors. For the future, we plan to add new functionality for business tasks and for parsing, tagging and linking of article references.

Keywords: on-line peer review; open source technologies; software architecture; workflow transition

1 Introduction

Our Online Submission and Peer Review System (OSPREY) is a web-based manuscript submission and peer review system used by scholarly publishers and societies to automate and streamline the publication process. It supports the submission of articles and the subsequent peer review process within a configurable automated electronic environment.

Communication with authors, reviewers and editors is handled by e-mail using customizable templates within the system. This is one of many features customizable by publisher or journal; others include copyright and reviewer forms and branding. Authors can upload a single file or multiple files consisting of many file types, and an Adobe Portable Document Format (PDF) file is created immediately. Metadata of accepted manuscripts in an XML format is integrated into the publishing system, however; some manual intervention is still required.

OSPREY is developed and maintained in collaboration between two leading Canadian and Australian scientific publishers. These are Commonwealth Scientific and Industrial Research Organisation (CSIRO) Publishing, Australia, and National Research Council (NRC) Research Press, part of the Canada Institute for Scientific and Technical Information (CISTI).

The objectives were to meet market demands, reduce turnaround times, increase efficiency within the editorial offices and to streamline the transition between peer review and publishing.

1.1 Background

Publishing

CISTI is a science library and a world leader in document delivery for all areas of science, technology, engineering and medicine. CISTI's publishing arm, NRC Research Press, has been a traditional publisher since 1929 and currently publishes 16 international print and online STM journals. With its resources and expertise in place, NRC Research Press began offering its print and electronic publishing services to other Canadian publishers in the late 1990s; as a result, NRC Research Press also publishes 15 client journals.

CSIRO Publishing operates as an independent science and technology publisher with a global reputation for quality products and services. Its internationally recognized publishing programme covers a wide range of scientific disciplines, including agriculture, plant and animal sciences, and environmental management. CSIRO Publishing publishes content in print and online. CSIRO Publishing is an autonomous business unit within CSIRO.

NRC Research Press moved into the electronic publishing world by first publishing content in PDF format on the Web for its subscribers in 1996 and later implementing a process to generate SGML metadata for searching, distribution to aggregators and dynamic generation of table of contents and abstract HTML pages on the Internet. NRC Research Press has since implemented an XML publishing system (Fig. 1) in which content is tagged according to a very rich custom Document type Definition (DTD) and published in print, PDF and HTML formats.

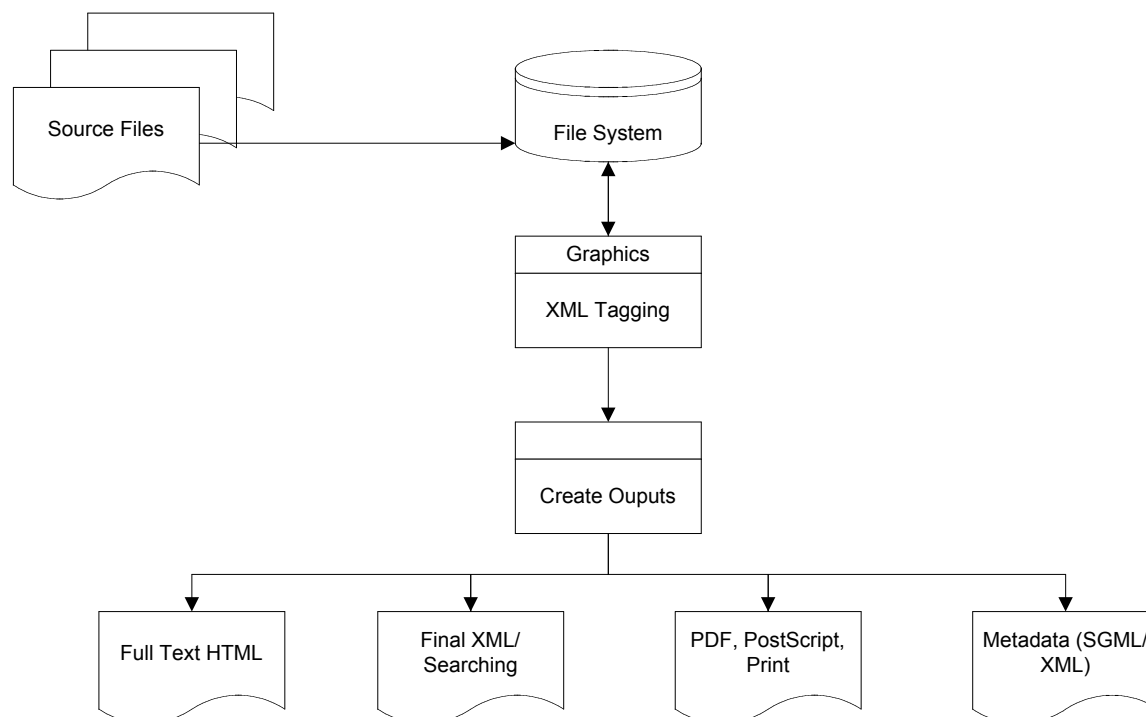


Figure 1: NRC Research Press Publishing Process

Transition to Electronic Peer Review

Traditionally, the submission of manuscripts, the peer review process and the management of the editorial process were paper-based and manual. Authors, reviewers and the journal offices would rely on mail, fax and courier services to deliver manuscripts, reviews and decisions throughout the workflow. Gradually the journal offices moved toward using e-mail for quicker transmission of manuscripts.

In early 2000, NRC Research Press purchased its first online submission and peer review system (PaperPath 2000). This new technology would help bridge the gap in the digital world between the peer review process and publication. It allowed authors and reviewers to submit manuscripts and reviews using a web browser and editorial offices to manage the workflow using third-party client software installed on their desktops. PaperPath 2000 was implemented in 15 journal offices over a 10-month period. In late fall of 2001 the vendor discontinued PaperPath 2000, leaving NRC Research Press with an unsupported product.

NRC Research Press supported PaperPath 2000 for another year and then began its search for a new online peer review system. Many Commercial-Off-The-Shelf (COTS) or licensed products were evolving, however; they did not adequately support a single sign-on, our diverse editorial workflows or our need for a multi-language (English and French) user interface with the potential to expand to other languages. There are many other factors to take into account when deciding whether to purchase or build, after evaluating each option, a decision was made to build in-house. Factors in making the decision were:

- available features in COTS or licensed product
- associated costs (one-time costs and maintenance and support)
- our diverse editorial workflows
- requirement for an English and French interface
- CISTI's technical skills and infrastructure
- storage of confidential data off site
- accessibility to data for customer relationship management
- integration to publishing and subscription management systems
- past experiences with the purchase and implementation of its previous online peer review application

Given the factors listed above we determined that the best approach was to leverage our internal capabilities and ensure our continued access to the source code and systems.

A significant effort was put into developing a solution that would ensure ease of use, minimal overhead and support costs, flexibility, scalability and future growth of features. The system was developed to current standards, using Java and XML for the application and Oracle as its database engine. The system architecture uses a component-based design methodology that enables flexibility and the potential for growth. It supports loose coupling and high cohesion, not only from a functional point of view, but also in the underlying data architecture.

2 Methodology

2.1 Collaboration

A key benefit to a collaborative approach is the ability to share resources (i.e. people and money) and to gather a broader set of requirements. Our experiences have shown us that identifying the roles and responsibilities, following sound project management principles and effectively communicating among team members, users and stakeholders are critical for success.

The two organizations shared a common vision, priorities and goals, helping us to develop ways to work collectively and to communicate effectively. Working together enabled us to draw on the skills and experiences of two scientific publishers. The stakeholders were instrumental in giving the project the priority and support required to develop OSPREY.

2.2 Development Approach

The international collaboration required rapid response to requests and iterations in the development of the user interface. For these reasons, an agile development methodology was chosen [1]. Iterations, which included use cases, analysis, design, implementation and tests, were 6 weeks long. That meant that both sides of the partnership could see new working functionality frequently. This approach made it possible to continuously add new functionality and make improvements. Each iteration could deliver minimal functionality. As new requirements were uncovered, a new iteration would replace the previous one.

Although each organization would install and support OSPREY, the technical development was completed at CISTI by two developers. An infrastructure was put in place to manage concurrent versions of source files and a centralized build function of the application.

The technical project team consisted of two developers, a system administrator, a database administrator and an application architect from CISTI. Representatives from CISTI and CSIRO rounded out the team and included stakeholders, users, business analysts and project managers.

2.3 OSPREY Architecture

There were several options available to deliver a web-based application to clients: Java Applets, Visual Basic, or server-side solutions such as Perl, ColdFusion, ASP or Java Servlets. CISTI's previous implementation of PaperPath 2000 required specific client-side software and hardware which added an additional burden to our technical support team. It was determined early in the project that the best means to meet the objectives was to implement a web-based application, requiring only a standard web browser as the client interface and no additional software on the client side. A Windows platform was chosen for development, however; the application could be ported to a Linux environment if required.

2.3.1 Open Source Technologies

Java, which is platform independent, was used as the programming language. The web interface was implemented with Java servlets and Java Server Pages (JSP), allowing loose coupling with the client-side.[2] CISTI had previously demonstrated the power of these technologies through the development of other successful web applications. The Model View Controller (MVC) design pattern was selected to facilitate rapid development, ease of presentation and consistent application behaviour. MVC is useful in achieving a separation of the business logic, the system control and presentation layer of the application.

The Data Access Object (DAO) pattern is used to allow abstract Enterprise Information System (EIS) independent data access.[3] The OSPREY DAO implementation was designed in the simplest form to allow for maximum speed of development and minimal knowledge to maintain. Consideration was taken to ensure the basic structure of the DAO framework would allow it to be easily extended in future and allow even greater flexibility in selecting DataSources.

Tomcat is used to serve dynamic servlet and JSP pages, while Apache is the web server, serving HTML pages to the user. Using Tomcat for development helped to ensure that the code would be portable and would not use proprietary packages, libraries and classes that are not otherwise available. It has the further advantage of integrating relatively seamlessly with the Apache web server and being open source.

OSPREY was designed to allow a single user to be logged into multiple journals or multiple instances of the same journal from the same HttpSession.[4] This design challenge prompted the creation of a very simple JournalSession framework. A user has one JournalSession for each authentication to an OSPREY journal. Each JournalSession is uniquely identified, holds a reference to the user's name information, and has the basic capacity to store and retrieve attributes.

Reviews and editor decisions are captured in XML and are translated into HTML or plain text format. Metadata in an XML format is exported and imported into the publishing system.

2.3.2 Conversion Service and Software

This service forwards requests to the appropriate servers, performs the transformation, and returns the created PDF to the user. Java Remote Method Invocation (RMI) is used to communicate between the application and the conversion service [5]. It was essential for us to support multiple files and multiple file types in one submission and create a single PDF immediately. To enable this speed and flexibility of conversion, it had to be scalable. The architecture is designed so that the software used to perform the conversions can be easily replaced.

PDF files are created by two different software packages, depending on their file format. LaTeX manuscripts are converted using open source software, MikTeX.[6] Adlib Express Server (third-party licensed software) processes all other file types. Adlib supports up to 300 different file formats and is upgraded frequently to meet the demand of converting newer versions of source files.[7]

2.3.3 Other Technologies

Manuscript data (names, addresses, manuscript data, correspondence, manuscript tracking dates) are stored in an Oracle 9i database, while all versions of the generated PDF and original submitted files are stored in a central Network Attached Storage (NAS) system.

2.3.4 Implementation

Data Conversion

To maintain review and manuscript history, we wanted to migrate as much data as possible to OSPREY. The data were analyzed, and a mapping between the Paperpath 2000 database and OSPREY was created in combination with scripts to extract, validate and import the data. The validation and testing were very time-consuming and should not be underestimated in projects of this type. A trial data conversion was completed prior to moving to production.

The Production Environment

As previously discussed, to fit into the current technical infrastructure of each organization, Windows 2000 was chosen as the operating system and Apache and Tomcat as the web/application server. The OSPREY interface is web based and supports leading web browsers. The file-conversion programs (Adlib and MikTeX) convert several different file types into a single PDF file for reviewing purposes which are not visible to the user.

Four servers and a central storage device support the production environment (Fig. 2) for OSPREY at CISTI. Three additional servers are in place to support development, testing and failover:

- Web/application server
- Database server
- Two conversion servers
- Network attached storage

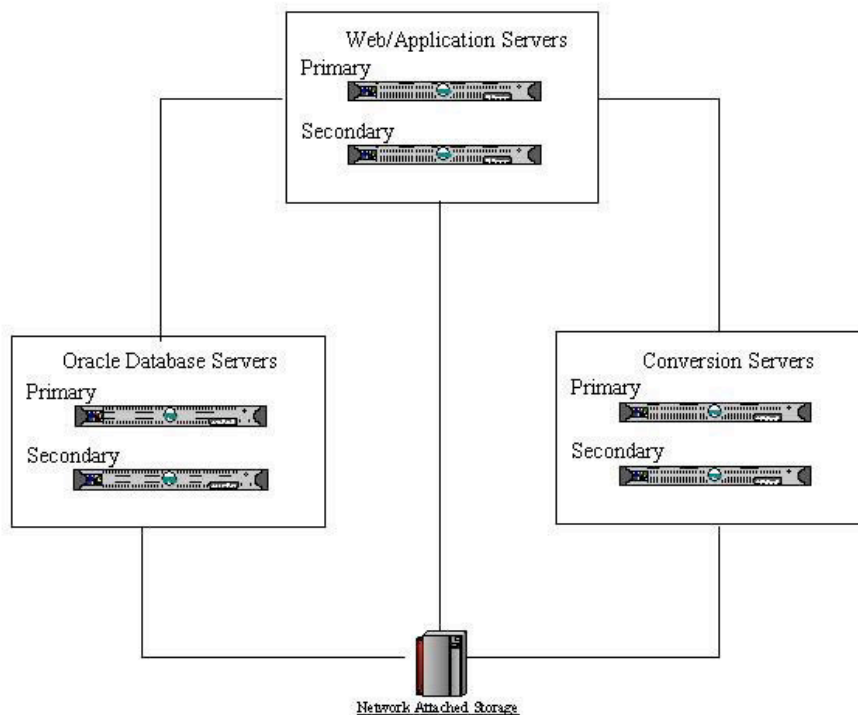


Figure 2: NRC Research Press Production Environment

Integration with NRC Research Press Publishing System

Upon acceptance of a manuscript, an XML file is created and sent to the publishing system (Fig. 3). Some processes are handled manually, while the manuscript, text, figures and supplementary data files are moved automatically to the appropriate file folders on the NAS.

```

<?xml version="1.0" encoding="UTF-8"?>
<EXPORT>
<ITEM>
<ITEM_INFO>
<EXPORT_DATE>13/09/2005</EXPORT_DATE>
</ITEM_INFO>
<MANUSCRIPT>
<TITLE>Life history variation among populations of Canadian Toads in Alberta, Canada</TITLE>
<NUMBER>4727</NUMBER>
<TYPE>Article</TYPE>
<DATE_SUBMITTED>31/08/2004</DATE_SUBMITTED>
<DATE_FINALIZED>13/09/2005</DATE_FINALIZED>
<OUTCOME>Accepted</OUTCOME>
<ABSTRACT>Development of appropriate conservation plans relies on life history information and how
life history traits vary across populations of a species. Such data are lacking for many amphibians,
including the Canadian Toad (Bufo hemiophrys Cope, 1886). Here we use skeletochronology to estimate
size-at-age, growth rates, age at maturity, and longevity of toads from nine populations along a latitudinal
gradient in Alberta, Canada. Size of individual toads within each year class was highly variable, but age and
size (measured as

```

Figure 3: Excerpt of Metadata exported from OSPREY

Training and Support

Training was offered to all users prior to implementation with the exception of authors and reviewers. The length of training was dependent upon their user role. Editorial Office Assistants (who coordinate the peer review process for one journal and use the software extensively in their day-to-day work) received up to 4 days of training, while Editors and Associate Editors received 2 to 3 hours. A helpdesk is in place and is supported by one individual on a full-time basis with backup when required. The helpdesk communicates with users by phone or e-mail. The skills required include knowledge of business processes, browser functionality and PDF conversion, as well as an in-depth knowledge of the application and its configuration options and their use.

3 Results

OSPREY has been installed in Canada and Australia, where it is currently supporting a combined total of 32 journals. NRC Research Press began its implementation in 2004, and over several months its journals began to accept online submissions using OSPREY. OSPREY is also used by 5 client journals through a licensing agreement with NRC Research Press.

OSPREY's interface is available in English and French and allows authors to upload manuscripts (Fig. 4), tables and figures, to create a single PDF file, and to check the status of their submissions. Upon the creation of the PDF file, authors are asked to review and approve the PDF file (Fig. 5).

NRC Research Press Canadian Journal of Zoology

OSPREY Online Submission and Peer Review [Français](#) [Contact Us](#) [Support](#) [About](#) [Log Out](#)

Logged in as: Louis Lafleur

Manuscript Submission [Previous Step](#) [Next Step](#)

Cancel Submission
Help

Step 1 Add Details and Authors **Step 2 Attach Files** Step 3 Create PDF Step 4 Print Copyright Step 5 Complete Submission

[View submission summary](#)

Attached Files
You must have 1 file attached for the following: **Cover letter, Manuscript Text**, before moving to the next step.
No files have been attached for this manuscript.

Files
You must enter at least one type. Enter the number of files you will attach, then click 'Browse for files'.

File	Number to attach
Cover letter	<input type="text" value="0"/>
Manuscript Text	<input type="text" value="0"/>
Figures For Peer review, we recommend uploading low-resolution graphics files.	<input type="text" value="0"/>
Tables	<input type="text" value="0"/>
Supplementary Data	<input type="text" value="0"/>

You may save this submission at this point and complete it later. Important Note: Only the Corresponding Author will be able to complete the submission.

OSPREY Online Submission and Peer Review [Français](#) [Contact Us](#) [Support](#) [About](#) [Log Out](#)

Canada

Figure 4: Upload files

Reviewers enter their reviews online or upload files. Editors select reviewers, vet the reviews and make recommendations or decisions on manuscripts, while editorial staff interacts with the system to verify submissions, send correspondence and assign roles and tasks.

NRC Research Press **Canadian Journal of Zoology**

OSPREY Online Submission and Peer Review [Français](#) [Contact Us](#) [Support](#) [About](#) [Log Out](#)

Logged in as: Louis Lafleur

Manuscript Submission [Previous Step](#) [Next Step](#)

[Cancel Submission](#)
[Help](#)

Step 1 Add Details and Authors	Step 2 Attach Files	Step 3 Create PDF	Step 4 Print Copyright	Step 5 Complete Submission
-----------------------------------	------------------------	------------------------------	---------------------------	-------------------------------

[View submission summary](#)

Create PDF File(s)
Your manuscript (including tables and figures) will proceed through the editorial process in PDF format.

After attaching or deleting files, click on 'Create PDF'. The process of creating PDF file(s) may take several minutes. Upon completion, the page will refresh and a link to the PDF file(s) will be displayed below.

Please click once. Create PDF may take a few moments to initiate. Do not re-click.

Created PDF Files
Please review the PDF file(s) to ensure that they are of adequate quality. Adobe Reader is required to view the PDF files.

File Name

*I have viewed the PDF file(s) and I approve their quality. If the quality of the PDF file(s) is not satisfactory, [please let us know](#). The Journal Office will contact you upon receipt of your message.

OSPREY Online Submission and Peer Review [Français](#) [Contact Us](#) [Support](#) [About](#) [Log Out](#)

Canada

Figure 5: Create, Review and Approve PDF file

The system is role-based, which allows journals to limit functions to appropriate users and restrict access to sensitive data. Users have a single sign-on; once logged in, they are presented with links to assigned tasks for each role [Fig. 6].

Each journal may be separately configured and many options are available including branding, workflow and selection of roles and tasks. To meet a journal's workflow requirements, each journal chooses tasks, and the order in which they occur, from a predefined list. In addition, OSPREY provides the flexibility to customize the term used to identify editorial staff roles (editor, editor-in-chief, associate editor, section editor, co-editor).

NRC Research Press Canadian Journal of Zoology

OSPREY Online Submission and Peer Review [Français](#) [Contact Us](#) [Support](#) [About](#) [Log Out](#)

Logged in as: Louis Lafleur

Journal
[Return to Main Page](#)
[Instructions to Authors](#)
[Submit manuscript](#)

Your Work Areas
[Editor \(1 tasks\)](#)
[Reviewer \(1 tasks\)](#)
[Author \(1 tasks\)](#)
[Editorial Assistant \(0 tasks\)](#)

Your Account
[Change Your Details](#)

Reviewer Work Area

Actions
[View Completed Reviews](#)

Task	Manuscript Number	Title	Date Assigned	Date Due
Submit Review	5881	The use of multiple den sites by Eurasian badgers, Meles mel...	10/04/2007	25/04/2007

OSPREY Online Submission and Peer Review [Français](#) [Contact Us](#) [Support](#) [About](#) [Log Out](#)

Canada

Figure 6: Roles and Assigned Tasks

3.1 Impact

Since the implementation of OSPREY, the number of submissions has increased significantly at NRC Research Press. Overall, they have risen over 32% (Fig. 7). Improved marketing and a rise in the Thompson ISI ranking of some journal titles have also contributed to the increased number of submissions. Subsequently, the total number of manuscripts accepted and rejected has also increased (Fig. 8).

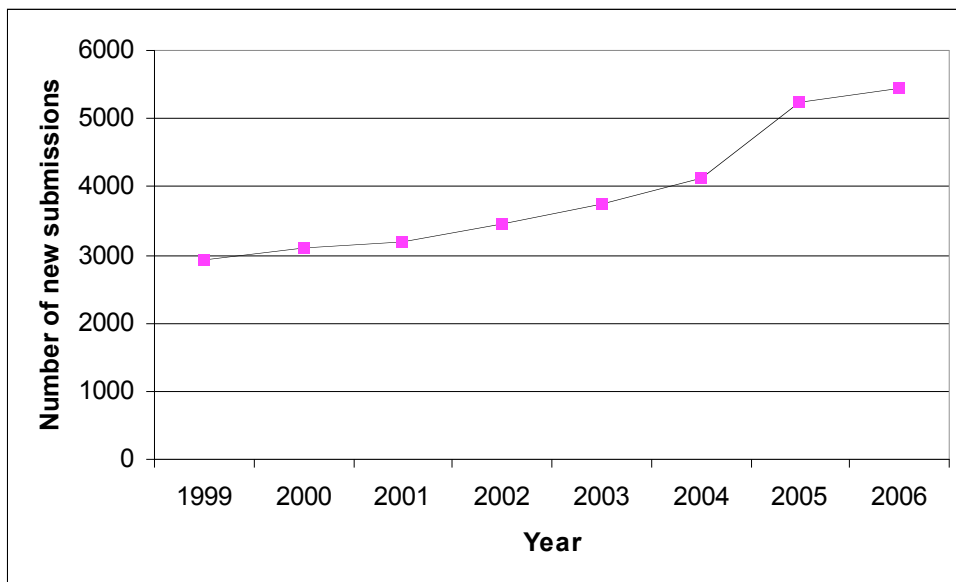


Figure 7: Total number of manuscripts received at NRC Research Press

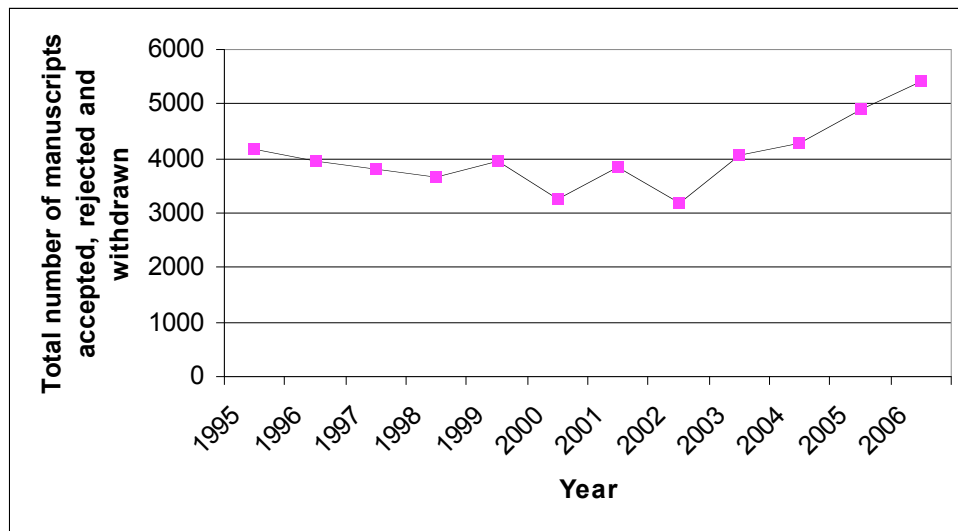


Figure 8: Total number of manuscripts accepted, rejected and withdrawn at NRC Research Press

Thanks to the use of an online system, there have been many changes in the way that authors, reviewers and editors work. Editorial staff now check PDF files for completeness and provide support to authors and reviewers using the system. Authors now receive an automatic acknowledgement of receipt of their submission immediately upon completion, and the paper is in the hands of the Editor faster. All correspondence is saved within the database for easy access. Prior to using an online system, acknowledgments would be sent by mail or e-mail and manuscripts would have to be delivered to the Editor by courier or expedited mail.

Reviewers are now sent an invitation to review a manuscript, as well as automatic reminders to submit their reviews. In the past, the manuscript would be sent to the reviewer by mail, courier or e-mail and a separate system would have to be in place to track the e-mails and any follow-up required. Reviewers can access manuscripts immediately and submit their reviews electronically, allowing Editors to have access to the reviews faster. Reviews and editorial decisions are tagged in XML and can be displayed in HTML or plain text format.

Editors now have access to all data required to process a manuscript from their desktop, regardless of their geographic location. They can now search for reviewers by expertise taxonomy keyword. Each journal has its own set of keywords, assigned to reviewers based on the reviewers' area of expertise. This functionality allows reviewers to be found quickly and their reviewer history, availability and performance, is available immediately.

One of the downsides of an electronic system is the learning curve. Editorial staff had to learn to work differently; instead of having stacks of mail on their desks, they now have an inbox full of e-mail. For convenience and backup, a paper record is still maintained in some cases, however; OSPREY is the primary repository for all manuscript data and correspondence.

3.2 Usability

One of our key design objectives was ease of use. We have received mixed reviews from users; some find the system very intuitive and others have difficulty uploading files during submission and finding links to files for download. Some users have also suggested that the number of clicks required to perform a function should be minimized. The usability issues will be addressed in the future.

3.3 Troubleshooting

For the architectural reasons presented previously, it was the right decision to make OSPREY distributed, accepting that the more distributed the system, the more complex the troubleshooting. OSPREY contains many components and troubleshooting can be very time-consuming. A recent investigation into a problem identified that certain types of corrupt source files could crash the conversion server. In this particular instance, the source file was not of a typical file type. Our testing procedures for conversion now include a thorough test of valid and corrupt files of all accepted file types.

4 Discussion

We recognized that there will be continuous maintenance and product enhancements when developing an in-house system. However, by building in-house we control the product lifecycle, features, priorities and release schedules. Maintenance includes such issues as supporting new versions of content-creation software (e.g., Microsoft Word), new web browsers, and updating the underlying software infrastructure (e.g., Apache web server), while product improvement enhances the application with new features and functions.

Online submission and peer review systems on the market today have been increasingly adding new functionality over the past few years. In other systems, editorial workflows are also customizable, parsing of references and linking to PubMed and CrossRef is now available, and some vendors are offering complete services from peer review to publication. Integrated database searching is also available in some products.[8]

4.1 Future Work

CISTI is currently implementing a workflow management system to manage the XML publishing process. Once this key component is in place, accepted manuscripts and metadata will flow seamlessly from OSPREY to the publishing system and will appear in the appropriate staff work areas automatically. Manuscript metadata and management data will be captured, and manuscripts will be forwarded to pre-editing. Upon publication, OSPREY will be updated with appropriate data (volume, issue, page number, date of publication).

The development and enhancement of OSPREY has opened the door to new opportunities. NRC Research Press currently has service agreements with 5 journals and plans to continue marketing OSPREY through its publishing services programme. A usability study is currently under way and will include interviews and an online survey of the user community. We will focus on ease of use and new functionality.

In addition, we are considering exploring the following functionality:

- parsing and tagging references to provide a link to abstract databases or full text (e.g. PubMed, CrossRef, and user organization link resolver)
- interfaces to allow authors to purchase paper and electronic reprints
- option for authors to identify papers for open access and supply payment or funding
- approving page proofs of accepted papers
- ability to access tasks directly from e-mail
- integration with external databases
- troubleshooting of common graphic file problems

OSPREY is a component-based system which offers us the flexibility to expand and enhance its functionality by changing components. Our current plans include replacing the existing conversion component. The next implementation will make use of web services instead of Java RMI.

5 Conclusion

There are many considerations when moving from a paper-based manual process to an online automated peer review and manuscript submission system or from one online system to another. When purchasing a commercial solution, or developing an in-house system, the impact on authors, reviewers and editorial staff must be considered and managed. Resources required to maintain a system are considerable, not only for software development but for upgrading hardware and software. Systems must be robust and flexible in their design to accommodate new requirements. Adequate training and user support must be put in place early in the project.

References

- [1] SCHUH, P. Integrating Agile Development in the Real World. Charles River Media, Inc of Hingham, Massachusetts, 2003.
- [2] KURNIAWAN, B. How Servlet Containers Work [Online]. May 23, 2003. [Cited April 10, 2007]. Available from the World Wide Web at http://www.onjava.com/pub/a/onjava/2003/05/14/java_webserver.html

- [3] Core J2EE Patterns - Data Access Object. [Online]. Sun Microsystems, Inc. [Cited April 10, 2007]. Available from the World Wide Web at <http://java.sun.com/blueprints/corej2eepatterns/Patterns/DataAccessObject.html>.
- [4] HUNTER, J.; CRAWFORD, W. Java Servlet Programming. O'Reilly & Associates, Inc, 1998. pp. 207-231.
- [5] An Overview of RMI Applications. [Online]. Sun Microsystems, Inc. [Cited April 10, 2007]. Available from the World Wide Web at <http://java.sun.com/docs/books/tutorial/rmi/overview.html>.
- [6] Miktex 2.5. Release August, 2006. Available from the World Wide Web at <http://www.miktex.org/>.
- [7] Adlib Express Server 3.8. Release February 15, 2007. Available from the World Wide Web at <http://www.adlibsoftware.com/ExpressServer.aspx>
- [8] WARE, M. Online submission and peer review systems. A review of currently available systems and the experiences of authors, referees, editors and publishers. United Kingdom: Association of Learned and Professional Society Publishers, 2005.

A Bachelor and Master Theses Portal: Specific Needs and Business Opportunities for the DoKS Repository Tool

Rudi Baccarne

Central Library, Katholieke Hogeschool Kempen, Geel, Belgium
e-mail: rudi.baccarne@khk.be

Abstract

A few years ago a portal for bachelor and master theses from Flemish university colleges was established by means of the open source repository software DoKS. At present approximately 3500 theses from Flemish university colleges are available online. The growing use of the portal has led to a new communication stream that requires supervision and maintenance. Social software components amongst others are or will be integrated in the portal to give users a platform to perform tasks such as communicate, annotate and advertise. Although different local DoKS repositories and the concept of the DoKS application are similar to repositories and tools within the scientific community, the scope and the aim of a theses repository for university colleges are different. The main part of the database consists of applied research and the majority theses comprise trainee reports. Thus, in addition to students and instructors, the portal is attractive to key players in industry, non-profit institutions and private users with a particular interest in a theses subject. This paper examines the different opportunities and specific needs of a bachelor and master theses portal, illustrated by real life examples. Social software components can breath new life into former static text documents. Users can add comments, create blogs, add tables, illustrations and suchlike. Content sensitive advertisements enhance the content and usage of our theses records and create revenues that can be used to make new improvements. In addition we will discuss the need for new and strict procedures with regard to content control, copyright issues and embargos when a bulk collection of industry related theses are published online.

Keywords: print on demand; content sensitive advertising; Electronic Theses and Dissertations (ETD); social software

1 Introduction

In 2003 the library of the Katholieke Hogeschool Kempen (KHK) launched a portal [1] for electronic theses and curricula vitae of graduating students at Flemish university colleges. In order to create this portal a new software DoKS (Document and Knowledge Sharing) was built. The project is funded by the Institute for the Promotion of Innovation by Science and Technology in Flanders [2], private industry partners and non-profit organizations. One of the main reasons for developing a new software was the need for a system that could be highly customized by users and tailored to needs specific to Flemish university colleges. The need for a flexible way to add local metadata (awards, trainee posts, credit points, etc.) in addition to commonly agreed sets such as Dublin Core or ETD-MS was in particular a high priority for the different colleges that were interviewed during preparatory meetings.

The DoKS portals from different colleges are decentrally stored yet at the same time available through one interface via the OAI-PMH protocol. Apart from searching harvested metadata via the central OAI-harvester, the user can search the full text of Electronic theses and dissertations (ETDs) from different institutes by means of the Google custom search engine [3], the latter allowing specification of the websites to be searched.

The KHK portal [4] receives 1500 daily visitors and offers an almost daily feedback and thus provides an indication of its usefulness for the labour market, graduates and the industry amongst others. Although the concept is similar to ETD-repositories and tools from the scientific community, a theses repository of a university college has other target users in mind. From the start private industry partners and other organizations supported the development of the DoKS theses portal, their reasons for supporting the DoKS project varying in accordance with their core business or particular interest. (IT, recruitment, valorisation of knowledge, screening of publications on business related content, electronic publishing etc.). The business plan carried out in the framework of the DoKS project and feedback from users raised new ideas and commercial opportunities for consideration. In addition to the publication of ETDs, several new and sometimes secured services (curricula vitae, ratings, etc.) were added to the free service of rendering theses from Flemish university colleges available

worldwide. The DoKS software is available by means of an open source license at Sourceforge (<http://sourceforge.net/projects/doksproject>). Reports and manuals are available via a wiki (<http://doks.khk.be/wiki/>) and the project website (<http://doks.khk.be>).

2 Specific Needs and Services for University Colleges

At international level, the focus of ETD-projects is on research theses from academic institutions. The main target audience for electronic theses projects is the research community. By making research theses more broadly available by means of open online repositories, researchers and research become more visible and as a consequence more widely cited. Advantages and value added services are generated for the researcher, his work and his place of work. [5, 6] Although university colleges are less focused on scientific research they show a growing interest in publishing electronic theses. From the beginning the DoKS repository tool was developed in accordance with the guidelines of the ETD community. As a result the portal is interoperable and combinable to a larger extent. On the other hand, the increasing use of the site and the feedback gained has shown that a theses-portal with mainly bachelor and master theses has different needs and offers other opportunities to exploit.

2.1 Curricula vitae

In the framework of the DoKS project a business plan was carried out. Part of the business plan was a quantitative and a qualitative research of different target users of an ETD portal. One important conclusion drawn from the business plan and based on the findings of the user surveys was that there seemed to be considerable interest from the private sector in the use of the portal as a recruitment tool. As a consequence curricula vitae can be filed in a standardized way together with the ETD records. This renders the portal a simple and cheap alternative as a starting point in the search for new employees. The force of this system lies in the accuracy of the data it contains and specific search facilities to retrieve this data (see Figure 1). In addition it is possible to raise an alert when new CVs matching specific criteria are added. Entries to the system are made by students and/or institutes and the system is therefore more complete and unique than any other built *ad hoc* and by a third party. All companies interviewed were prepared to pay for such a service by means of a subscription or registration service.

Furthermore the business plan created for the DoKS project outlined the opportunities that could render the portal self maintaining. Once optimized and well positioned on the market the CV module should create enough revenues for the employment of a 50% employee for maintenance and administration of the system. At the moment different steps are being taken to convert this potential into reality. Different approaches to commercialize the CV module include partnerships, pay on demand, subscriptions and sponsorship to name but a few. It is clear that the graduating students also benefit from the CV module. The DoKS system automatically creates a CV for all graduating students based on data from student administration files. The student can complete this by adding extra data. The result is a CV that can be converted to a Europass CV by one mouse click. Statistics on the number of updated CVs confirm the students' interest. At the moment 1578 CV records are available belonging to KHK students who have graduated over the last two academic years. About 25 % of CV records were completed by students one month after the automatic creation of their CV and more than 70 % are updated before the end of an academic year. We believe that this enthusiasm is related to the fact that the CVs are already formatted, half filled out and exportable to print and other formats (see Figure 1). In other words, the student does not have to make much effort to produce an attractive and easy-to-maintain CV. The idea of automatically generating many of the relevant metadata behind the scenes to avoid filling out lengthy electronic forms is relevant for different web environments. In the field of educational multimedia, for example, this is strongly expressed by Erik Duval in the slogan 'electronic forms must die' [7]. While going through the self-archiving wizard for their electronic thesis students must choose to what extent their CV will be available to employers. A majority choose the option 'CV fully available' thus allowing employers to contact the student directly. 'CV available without personal data (contact through DoKS website)' and 'CV not available' are two other options. Graduated students can access and update their CVs via a dedicated account for a limited period of two years.

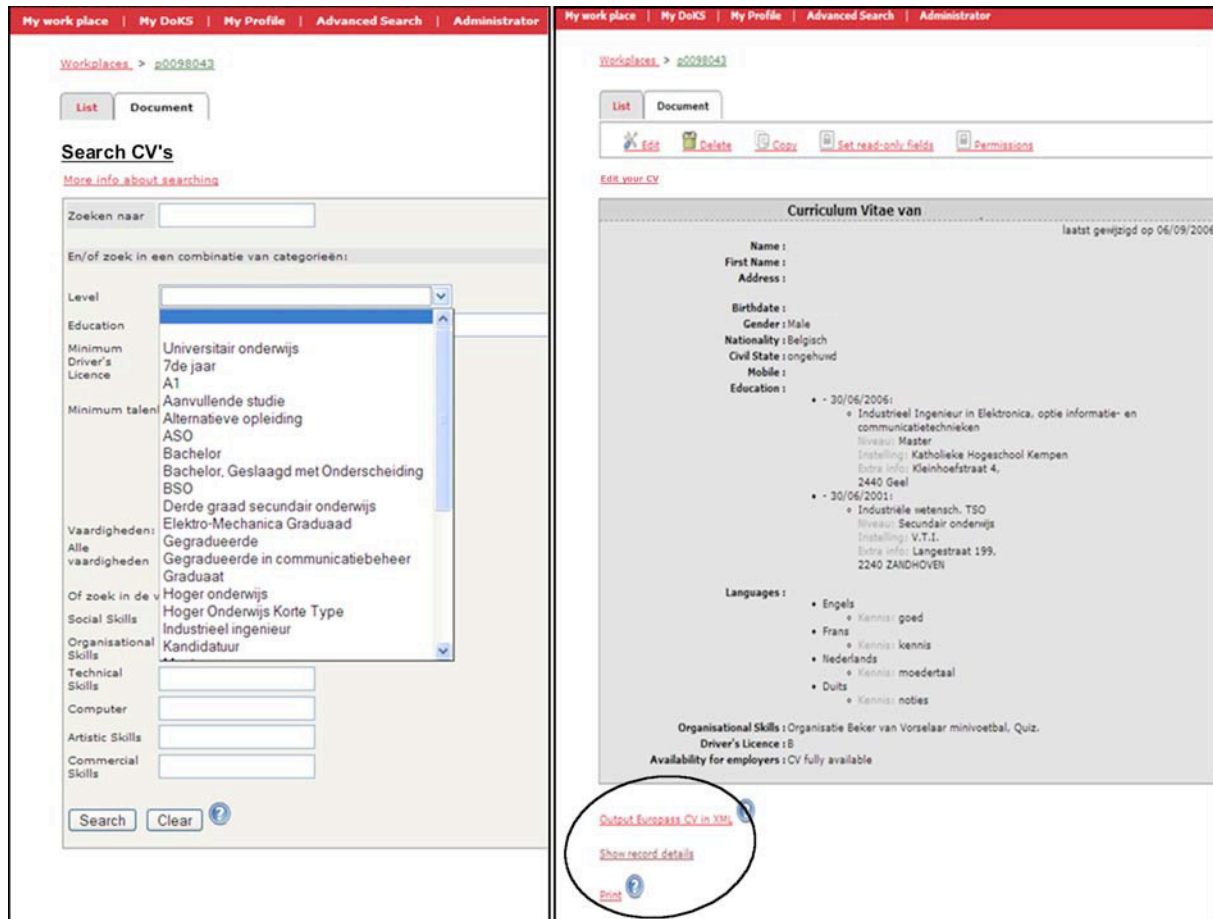


Figure 1: Search interface for CVs and basic example CV

In addition to the commercialization of the CV module, a good working DoKS portal can be exploited in several other ways. At the KHK the theses portal is linked to the Google AdSense program and creates revenues that are sufficient for replacement of hardware, new improvements, etc. Several other opportunities and partnerships with industry partners or non-profit organizations emerge once the portal is known to the different stakeholders. At the KHK this has led to partnerships with innovation agents, non profit partners and private industry partners that have a variety of reasons for supporting the maintenance of the portal [8, 9].

2.2 ETD-MS

ETD-MS is an interoperability metadata standard for electronic theses and dissertations [10]. The standard adds one element to the Dublin Core metadata elements, namely thesis.degree. This element has 4 qualifiers: thesis.degree.name, thesis.degree.level, thesis.degree.discipline, and thesis.degree.grantor. This standard is a result of the work established by the Networked Digital Library of Theses and Dissertations (NDLTD) which tries to coordinate the different worldwide initiatives. The aim of the standard is interoperability and is described as such on the NDLTD website [11] 'i.e., to make it possible to share information about ETDs. This will allow us to improve existing federated searches, create union databases, and provide greater consistency for researchers searching for theses and dissertations at different institutions.

To integrate bachelor and master theses in ETD union catalogues and repositories for scholarly communication we believe the level of education must be transparent and clearly distinguishable. This will help the end user to place the work in his context so he can judge it appropriately. The PKP-harvester software [12] we use for harvesting metadata from different local DoKS repositories did not however support ETD-MS. Therefore we recently created an ETD-MS plug-in [13] for this harvester. By using the plug-in users can perform searches on degrees and level of education to find graduates, their profiles and their learning outcomes.

Figure 2: ETD-MS plug-in for PKPHarvester2

When other ETD programs consistently use the same standards, users can search records in a similar manner across records from different countries. The end user has an immediate knowledge of the type of the work, the level of the work, the educational program in which it was produced, the related degree and so on without necessarily knowing the language in which it was written.

2.3 Content (applied research, less academic, ...)

At the KHK and by extension at most similar university colleges in Flanders nearly all theses comprise reports on work a student has done at a trainee post. As a consequence a thesis might contain confidential information. The industry partner where the trainee is placed has the option of requesting an embargo by means of strict procedures and dedicated forms. Although we expected to see a drastically increasing *a priori* demand for embargos once we started to publish ETDs, this seems not to be the case. Nevertheless the student must inform the trainee post about the online publication a clear demand for new embargos is seen once a thesis is online for a period of time. In some cases an embargo is requested because confidential information is published, but there seems to be several other reasons an industry partner does not want to see a trainee report published. This is often related to the high search engine ranking of our theses records - DoKS theses records are ranked higher than the web pages of the trainee posts - , old or false information on products is still available on the web, etc. In the future there are plans to give users a communication platform on which annotations can be made. In this way it is hoped that embargos can be avoided where the need for them goes beyond the publication of confidential information. On the other hand we see in literature [14] and also in practice a shift towards more transparency in domains (pharmaceutical industry, innovative IT companies) that were at first more resistant to the online publication of research data and material.

The power of the portal to serve the needs of innovation agents and intermediary organizations has resulted in the following collaborations:

Flemish Chamber of Engineers (VIK)

The Flemish Chamber of Engineers is developing an award program based on the DoKS repositories to stimulate entrepreneurship. The aim of this project is to filter theses with a high commercial or innovative character, especially those that have the potential to develop into enterprises. The idea stemmed from the fact that the annual number of new start-up businesses of an innovative nature in Belgium is very low compared with other countries [15]. Furthermore over the years the Chamber has kept files of new businesses that emerged from the basis of an innovative idea in a thesis;

'Innovatiecentrum West-Vlaanderen' (West Flanders regional innovation centre)

A study [16] carried out by the regional innovation centre of West Flanders pointed out that of all theses established in the context of a trainee post only a low percentage resulted in an economic surplus value for the firms involved. An analysis of the study however showed that the economic valorisation could be increased by taking measures such as recruitment of the student after graduation or extra guidance by the college. To achieve this the regional innovation centre allocates awards for students and valorisation budgets for the firms. The innovation centre urges University colleges from the region to set up a DoKS portal in order to improve and accelerate selection procedures for theses that would be considered for a valorisation trajectory;

Indiegroup

Indiegroup is an organization that develops software for the innovation market. Integrating innovative content from the theses of university colleges can create a surplus for their software 'Cognistreamer'. Cognistreamer is a platform for open innovation concepts. By means of RSS and XML crosswalks innovative theses projects from DoKS could be integrated and selectively disseminated via Cognistreamer to organizations that are working on related subjects.

2.4 Statistics

2.4.1 Daily Visitors

The statistics in Figure 3 are based on figures from Google Analytics and cover the last full 12 months the KHK DoKS portal was online. The extremely sudden peak on the 11th of January has a logical explanation. At that time it was noted that a majority of visitors downloaded the full text of a DoKS thesis directly via a Google result list bypassing the DoKS website. For several reasons we have now decided to use a URL rewrite mechanism so that users are always transferred to a DoKS thesis record from where they can download the full text of the document. First of all the figures for the use of our portal were seriously underestimated. First and foremost, users were downloading documents from the site without knowing they were reading a thesis document from a bachelor or master student at a Flemish university college.

In addition it is clear that there is stable use of the website which at the moment has an average of 1500 daily visitors. The trend is downwards during the weekend and holidays and use increases use during the periods the students are working on a thesis and need the portal to submit data and full text. The statistics from Figure 6 indicate that the use of the site is strongly related to the revenues created by Google Ads with the same steep increase from January 2007 onwards.

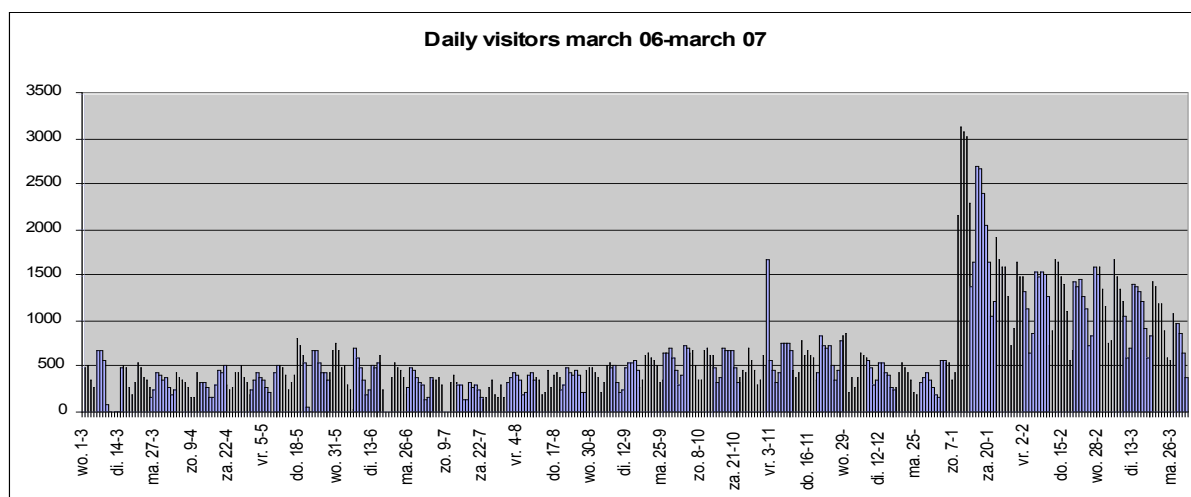


Figure 3: Daily visitors DoKS@KHK

2.4.2 Downloads

As shown in the graph below showing the number of these downloads from the KHK-portal a ‘long tail’ curve emerges. This illustrates the wide and varied interest in theses content. The usage shown by the curve seems to be typical for e-business websites and indicates new economic mechanisms that are related to the internet. Theses that perhaps never came to light when they were stored physically at the library are, once online, consulted more than the most borrowed hardcopy theses from that same library. These new models and mechanisms are described by Chris Anderson with regard to e-business sites from the amusement industry (Amazon, iTunes, etc.) [17].

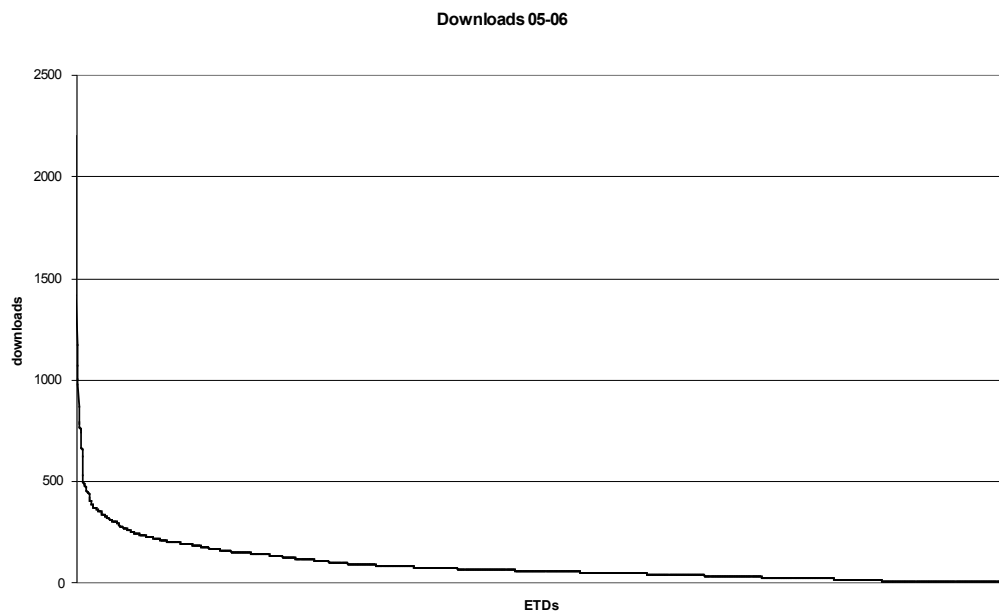


Figure 4: KHK-ETD downloads 05-06

2.5 User Feedback (categories)

When filled with a significant amount of content the DoKS repositories receive a high search engine ranking. As a result the number of daily visitors is significant. Among a variety of users, the feedback received in Flanders introduced several new business opportunities and interests. The feedback received can be categorized as follows:

- Job offers and offers for trainee posts
- Knowledge sharing
 - Collaboration proposals
 - Questions on thesis subject
 - Demand for annotations. Users want to comment on the content of a thesis and students want to add new views, opinions, corrections, etc.
- Editorial boards of journals
- Embargo requests from industry partners of the KHK
- Hardcopy requests (see section 3.3.5)
- Reporting on violations of the law with regards to:
 - professional confidentiality
 - copyright
 - privacy

2.6 Social Networking and Business Opportunities

The use and feedback on the portal clearly indicated the need to add social networking tools. Plans are being made for the future to integrate the features of the KNOSOS [18] platform in DoKS. Users who want to add, blog, annotate or tag to name but a few will be allocated to the collaborative working space provided by

KNOSOS. In preparation of a structured approach to overcome different needs, the first experiments have been set up. The following paragraphs describe the way in which we have already addressed some user demands.

2.6.1 A New Splash Page

Students are nowadays familiar with new technologies (Internet, multimedia, publishing, web 2.0, etc.) but are not supported in the use of them in a traditional hardcopy print environment. By following new internet trends, DoKS is able to keep track of the way young undergraduate students use the internet. We believe this is a necessary condition to conserve the enthusiasm of our most important supplier of information, the students themselves. As a result the record splash page (Figure 5) has been drastically changed in favour of a more user-friendly interface.

Use this url to link to this item: <http://hdl.handle.net/2161/etd.1090>

Nutrient deficiencies in maize and its relation with soil type and land use in West Africa
2003
Van Houdt, Steven
Industrieel Ingenieur in Landbouw en Biotechnologie

Trefwoorden: [History of agriculture](#), [Phytotechny](#), [horticulture](#), [crop protection](#), [phytopathology](#), [Soil science](#), [agricultural hydrology](#).

Abstract :
Continued land use intensification and the change from low analysis fertilizers to high-analysis fertilizers in West Africa may lead to increased deficiencies in elements other than the main el farmers shift increasingly from Single Super Phosphate, which contains considerable quantities of sulfur, to Triple Super Phosphate that is low in sulfur. Sulfur deficiencies have recently been rep A nutrient omission trial was conducted in the Northern Guinea Savanna of Nigeria, and in the central region of Togo. The nutrient omission trial compared 8 treatments which were randomly field the 8 treatments encompassed (i) a farmer's practice control plot (referred to as 'P0'), (ii) a full-nutrition control treatment (referred to as 'P40'), (iii) a treatment 'P20' with full nutrition bu and (iv) five treatments ('P0', 'K0', 'S0', 'Zn0' and 'B0') that received the full nutrition as in 'P40' except for one nutrient that was omitted (one at a time). Fields were sampled for soil and ear le Yields were much lower in Nigeria, due to acidifying effect of the fertilizers on the roots of the cornplants, treatmenteffects were subordinate to this effect. In Togo, yields were higher and yield K0 and the S0 treatments. Calculation of P DRIS indices showed a deficiency of P in the Togolese fields and yields responded to a TSP fertilization.

Full text:

File	Size	Type	Checksum	
thesis2003241.pdf	1 MB	PDF	MDS	Open file

Dit document werd 137 keer bekeken en 33 keer gedownload.

0 diggs

digg it [BOOK:996K](#)

bookmark this on del.icio.us saved by [other person](#)

[Show record details](#)
[Show ETD - Dublin Core](#)

Fertilizer
Great free information on Fertilizer
[FertilizerDirectory.info](#)

Organic Fertilizers
Complete fertilizers for optimal crop yields
[www.fermofeed.com](#)

Dust Control
Products and Services, Engineering and Compliance
[syntechproducts.com](#)

Bio fertilizer & manure
supply highly effective fertilizer, natural dry bacteria for the humus
[www.biozms.com/maize2.html](#)

Google AdSense : context sensitive advertisements

Figure 5: A thesis record splash page

In the following paragraphs the benefits of the major adjustments, namely the integration of social bookmarking tools and context sensitive advertisements will be discussed.

2.6.2 Social Bookmarking

In the light of our current subject classification which is deemed insufficient [19], a new opportunity is presented by the use of folksonomies or a tagging system. Furthermore, an interactive way of supplying keywords or tags perfectly matches the broader aims of the DoKS project, namely, knowledge sharing and community building. At the moment social bookmarking tools are provided on the theses records. By means of the 'Delicious Tagometer' (see Figure 5) it is easy to find out which other people have tagged a particular thesis record. This will lead you to the bookmark pages of people with common interests. It also allows you to see how what resources other users have tagged on the same subject. A desired feature that until now has not been available is a way of aggregating tags from different users of our portal in a tag cloud. Once such a feature is available we can provide this aggregation of tags to our users.

2.6.3 Instant Messaging

Although a part of the metadata (author, department, degree title, address, etc.) is automatically imported from files received by the library from the Institute's general administration department, another part (title, abstract, language, volumes, contact details, number of desired copies, instructor, trainee post, trainee supervisor) must be submitted by the students via the DoKS repository software. In addition the full text must be submitted by a self-archiving approach. Given that on an annual basis as many as 800 students submit their thesis data, they have to follow strict procedures and guidelines. By giving students the opportunity to ask questions whilst submitting data, the administrators can assist students directly should they encounter problems. In DoKS this communication is provided by a Meebo [20] widget and is similar to Instant Messaging systems the use of which

is very familiar to the students. Another attractive feature that comes with the integration of the Meebo widget is the ability to keep track of the number of concurrent users of your site.

2.6.4 Google AdSense

Once installed and filled with content a DoKS repository creates revenues via Google AdSense that can cover maintenance costs (upgrading hardware backup tapes) and/or new improvements. This is already the case for the repository of the Katholieke Hogeschool Kempen. The idea to start experimenting with Google Ads was based on a very practical mail question received from a user. The user had downloaded a thesis about laying out a private swimming pond. In the thesis prices for products were given which appeared to be much lower than the ones experienced by the user. Instead of contacting the student behind the particular theses we thought it might be easier to provide a direct path to suppliers of goods related to a thesis subject. Via Google AdSense relevant context sensitive ads are displayed on the pages containing theses records. The ads are related to the visitors' search and thus create a way to both monetize and enhance the theses records. We have currently been experimenting for almost a year with the system and evaluated that both benefits seems to be fulfilled. The ads are in most cases relevant and enhance our content.

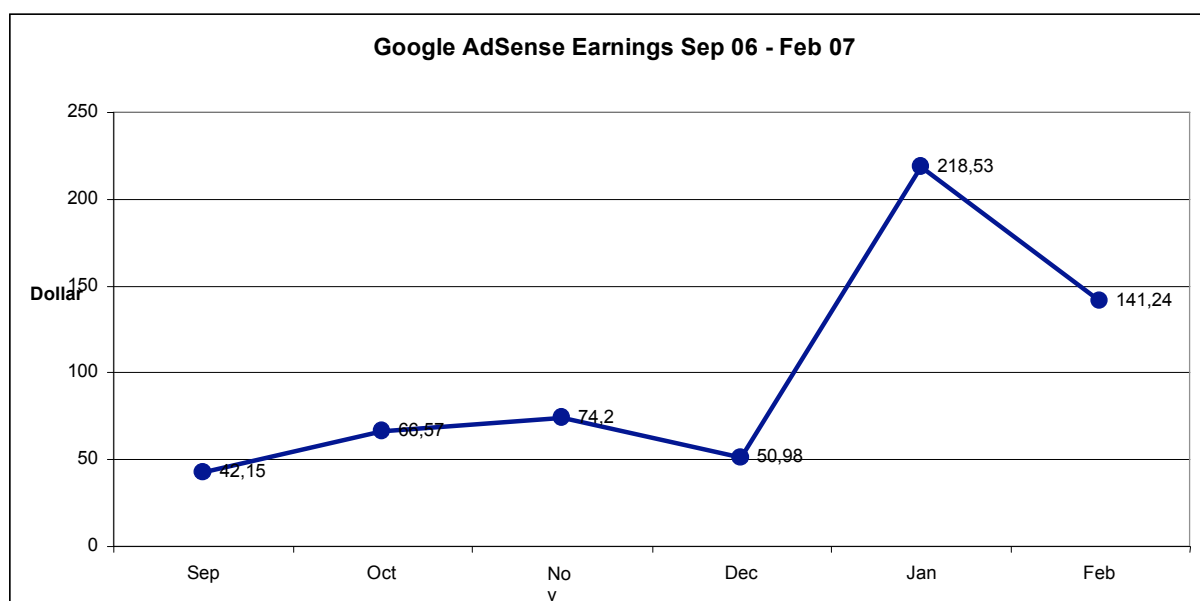


Figure 6: Google AdSense earnings Sep 06 – Feb 07

2.6.5 Print On Demand

With great surprise we noted a significant demand from our users to obtain a hardcopy version of the theses we published electronically. At first the intention was to deny these requests because it was thought that they would occur very occasionally and there were not the resources to give an appropriate answer. However more requests for hardcopies arose and by coincidence the DoKS portal caught the attention of a new player on the market of print on demand and self publishing, i.e. WWAOW (world wide association of writers) [21] This resulted very recently in a new collaboration and the first theses from different university colleges in Flanders are available via WWAOW (see Figure 7). Students are informed about this opportunity during the self-archiving procedure where they can choose whether they want to make their thesis available by means of the WWAOW website.

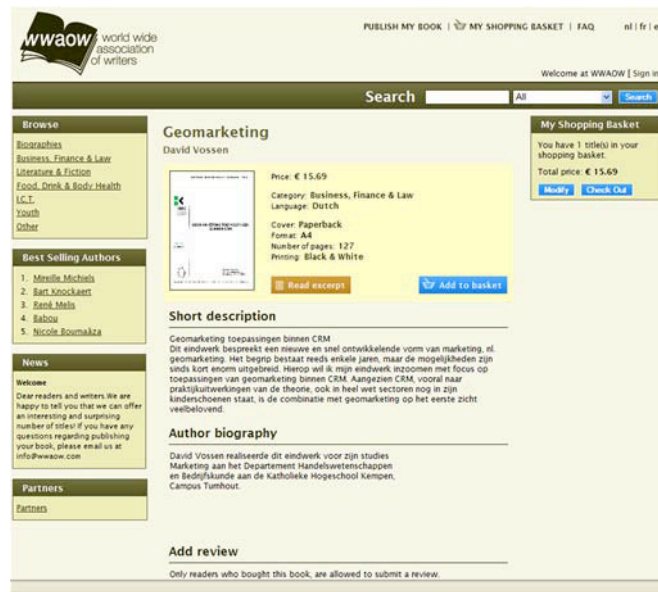


Figure 7: Print on demand via WWAOW

2.6.6 Alerting via Persistent Query Mechanisms and RSS

All authenticated users have a personal profile page in which they can store keywords (My Topics). By means of a persistent query mechanism search queries can be saved and can be executed again. This technique is used to provide an alert system. Once logged in a personal homepage is displayed (MyDoKS). On this homepage there appears a list showing documents which are new since the last time the user logged in.

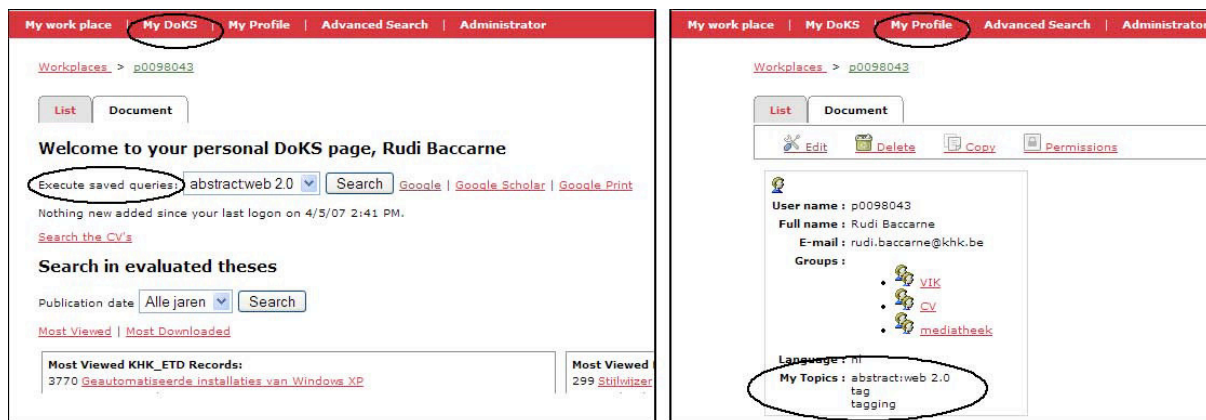


Figure 8: DoKS personal homepage and profile page

The built in RSS functionality can be used as an alert system as well. It is possible for example to subscribe to search queries via RSS, with the result that whenever a new item is published that reflects your query you will be alerted by you RSS reader. This RSS functionality is also a means to publish automatically updated lists. For example you can subscribe to a list of theses that are available by the print on demand system (see Figure 9), and new CVs to name but a few.

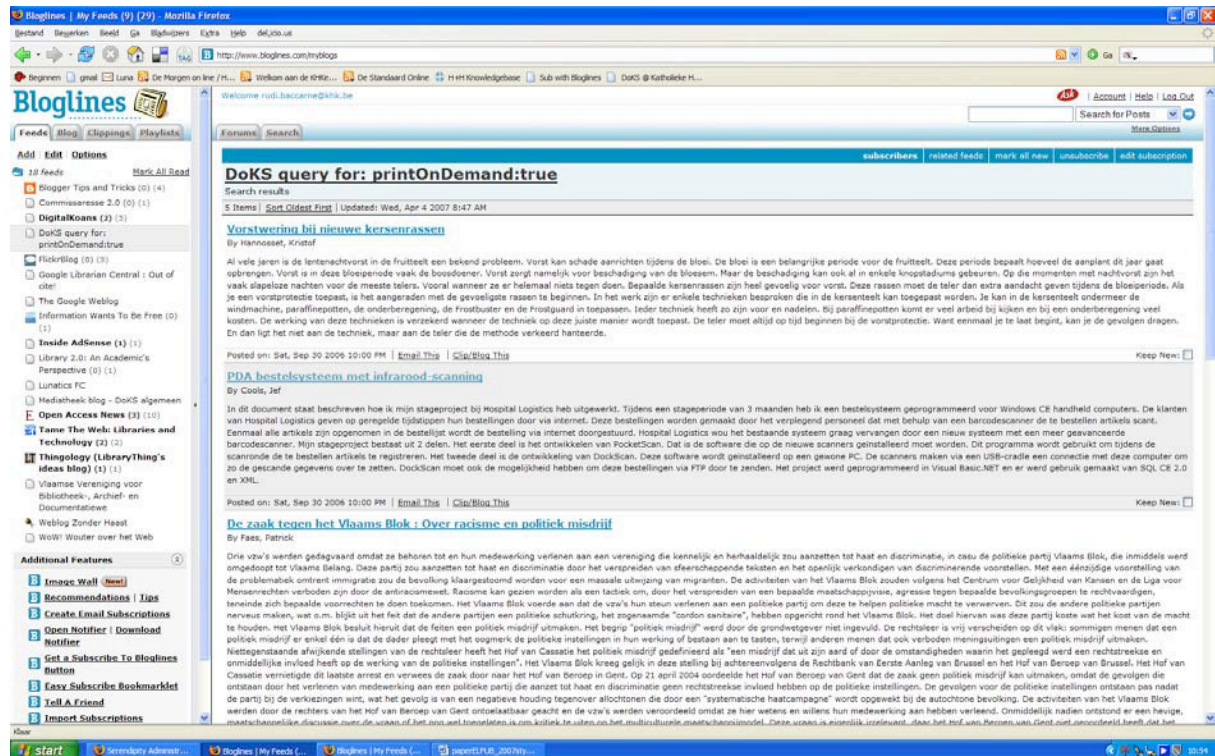


Figure 9: Example of a DoKS RSS search query subscription via Bloglines

4 Conclusion

Apart from being more visible to the scholarly community as well as to the labour market, students and university colleges will profit in the long term by contributing to the portal in several ways. The submitted work will have to meet certain conditions before being accepted for publication. Students learn about digital publishing and structured authoring. They have to deal with choices between different file and image formats, reducing file size and structured authoring, to name but a few. The wider availability of the work creates a mentality change amongst students and lecturers towards different phases in the electronic publication chain (citing, references, copyright, technical implications of electronic publishing, etc.). In the long term this will lead to better quality in electronic documents whilst at the same time students with a rather resistant attitude to computers, internet and the like will be introduced to the internet and electronic publishing.

When filled with a significant amount of content the DoKS repositories receive a high search engine ranking. As a result the number of daily visitors is significant. Among a variety of users, the feedback received at the KHK, has introduced several new business opportunities and interests. In this sense a successful repository can be seen as a powerful public relations tool. Because DoKS supplies a java and JavaScript-like scripting engine (Beanshell) for task automation, complex work flows, specialized import/export, etc. formerly manual processes such as MARC-export for the library, collection of abstracts and titles, publishing and so on are automated. Furthermore at the KHK DoKS is used to support services such as the employment agency and the research department. As a result the tool is highly appreciated by the users of the institution. Other benefits of integrating student scholarship in institutional repositories are discussed in many blogs and publications. A collection of similar and other arguments from which the quotation below is extracted is listed in the 'Law Librarian Blog' from Carol A. Parker [22]. 'The students' scholarship would attain visibility on a scale never before seen, and the students would enjoy the benefit of informing the subsequent work of others. Plagiarism should not be an issue because most institutional repositories are indexed at the full-text level, meaning that a simple Google search would quickly identify an existing paper that was later used without proper attribution. ...

Digital collections of student work can also be used for publicity and outreach, especially with alumni. Many schools already inform alumni of recent faculty publications; alumni could also be informed of student scholarship published in repositories. Making student scholarship available in digital collections provides students with a connection to their schools after graduation.

Law schools would also be sending the message that they take student scholarship seriously. Knowing that their work will also be subject to scrutiny beyond the four walls of their professors' offices would give law students added incentive to produce better scholarship.'

References

- [1] ETD portal for Flemish University colleges. See <http://www.doks.be>
- [2] <http://www.iwt.be>
- [3] See <http://www.google.com/coop/>
- [4] <http://doks2.khk.be/eindwerk>
- [5] The Open Citation Project - Reference Linking and Citation Analysis for Open Archives. *The effect of open access and downloads ('hits') on citation impact: a bibliography of studies*. Retrieved 5 April 2007 <http://opcit.eprints.org/oacitation-biblio.html>
- [6] PIWOWAR, H. A.; DAY, R. S.; FRIDSMA, D. B., (2007). *Sharing Detailed Research Data Is Associated with Increased Citation Rate*, PLoS ONE, March 21, 2007 Retrieved 5 April 2007 <http://www.plosone.org/article/fetchArticle.action?articleURI=info:doi/10.1371/journal.pone.0000308>
- [7] DUVAL, E. (2004). We're on the road to..., in ED-MEDIA 2004. Lugano, Switzerland. Retrieved 6 april 2007 <http://www.cs.kuleuven.ac.be/~hmdb/publications/files/pdfversion/41316.pdf>
- [8] For an overview of DoKS partners see: <http://www.doks.be/partners.htm>
- [9] BACCARNE, R (2005). *From central administration of hardcopy Bachelor- and Master theses towards a decentralized ETD-system with value added services* / Conference Paper, ETD2005, the 8th International Electronic Theses and Dissertations Symposium. The Scientia, University of New South Wales, Sydney, Australia, 28-30 september 2005. Retrieved 5 April 2007 <http://adt.caul.edu.au/etd2005/papers/045Baccarne.pdf>
- [10] For a full description of ETD-MS see: <http://www.ndltd.org/standards/metadata/current.html>
- [11] <http://www.ndltd.org>
- [12] http://pkp.sfu.ca/harvester_download
- [13] The plug-in can be downloaded at:
<http://doks.khk.be/do/record/Get?dispatch=view&recordId=SDoc413ebf1711bbc5e40111bc0f15560001>
- [14] TAPSCOTT, D; WILLIAMS, A. D. , *The new Science of Sharing* In: Business Week, March 2, 2007 Retrieved 29 march 2007 http://www.businessweek.com/innovate/content/mar2007/id20070302_219704.htm
- [15] DE CLERCQ, D.; MANIGART, S.; CLARYSSE, B.; CRIJNS, H.; DE SUTTER, M.; VERZELE, F. *Global Entrepreneurship Monitor: Executive report for Belgium and Wallonia*, Vlerick Leuven Gent Management School, 2003, 95 p. Retrieved 5 april, 2007 <http://www.gemconsortium.org/document.asp?id=265>
- [16] VANNESTE, P.; BLOMME, E.; DESAEGER, A.; GRYMOPREZ, P. (red.), *Kennisvalorisatie als opstap naar innovatiebij KMO's en kleine non-profit organisaties*, VZW Kortrijks Ondernemerscentrum, Kortrijk, 2006, 53p.
- [17] ANDERSON, C. *The Long Tail: Why the Future of Business is Selling Less of More*. New York : Hyperion, 2006.
- [18] <http://www.knosos.be>
- [19] For an essay on the insufficiency of traditional categorization methods for the electronic world, See: Clay Shirky. *Ontology is overrated: Categories, links, and tags*, 2005. Retrieved 10 April 2007 [http://www.shirky.com/writings/ontology overrated.html](http://www.shirky.com/writings/ontology%20overrated.html).
- [20] <http://www.meebome.com/>
- [21] <http://www.wwaow.com>
- [22] PARKER, C. A.[blog]
http://lawprofessors.typepad.com/law_librarian_blog/2007/02/institutional_r.html

The FAO Open Archive: Enhancing Access to FAO Publications Using International Standards and Exchange Protocols

Claudia Nicolai; Imma Subirats; Stephen Katz

Food and Agriculture Organization of the United Nations
Viale delle Terme di Caracalla 1, 00153 Rome, Italy
e-mail: Claudia.Nicolai@fao.org; Imma.Subirats@fao.org; Stephen.Katz@fao.org

Abstract

Since 1998, the Food and Agriculture Organization of the United Nations (FAO) has been publishing its electronic publications in the FAO Corporate Document Repository (CDR). The electronic publishing workflow is maintained by the Electronic Information Management System (EIMS). The EIMS-CDR holds more than 38 500 documents and is the gateway to FAO's publications. The EIMS-CDR coexists with the FAODOC – the online catalogue for documents produced by FAO. FAODOC catalogues and indexes both electronic and printed documents while the EIMS-CDR manages full text documents and a minimal set of metadata. This paper discusses the merger of the EIMS-CDR and the FAODOC into a unique FAO Open Archive based on the integration of the electronic publishing and the bibliographic cataloguing requirements. The FAO Open Archive will be the foundation for the collection, management, maintenance and timely dissemination of material published by FAO. To improve the effectiveness of the proposed repository, it is necessary to streamline the current electronic publishing workflow. The merger of the EIMS-CDR and the FAODOC will strengthen FAO's role as a knowledge dissemination organization. Especially, as one of the principal tasks of the FAO is to efficiently collect and disseminate information regarding food, nutrition, agriculture, fisheries and forestry.

Keywords: open access; open archive initiative; interoperability; digital repositories; data content standards

1 Introduction

The Food and Agriculture Organization of the United Nations (FAO) has more than 50 years of experience in the production and the dissemination of information, both through its headquarters-based regular programme and through field projects. The collection, analysis, interpretation and dissemination of information relating to nutrition, food and agriculture are FAO's main functions [1]. The World Wide Web has proven to be a powerful means for FAO to disseminate multilingual information.

In this context, FAO was an early implementer of:

1. an online catalogue for documents produced by FAO (FAODOC, Figure 1), a multilingual online catalogue which contains bibliographic metadata of FAO electronic and printed documents [2];
2. the Electronic Information Management System (EIMS), a workflow management tool and database which manages the publication of electronic documents and multimedia resources on FAO's Web sites [3]; and
3. the Corporate Document Repository (CDR, Figure 2), a corporate output interface for FAO full text electronic publications stored in the EIMS [4, 5].

The FAODOC is a multilingual, online catalogue of documents and publications produced by FAO since 1945. The system uses UNESCO's CDS/ISIS software [6]. More than 160 000 documents have currently been catalogued. Since its inception, the FAODOC has focused on the production of high quality bibliographic records.

The FAO Web site was released in 1995 and the first electronic publishing workflow (through EIMS) was initiated in 1998. Currently, more than 38 550 resources (full text documents and multimedia items) are managed by the EIMS (Table 1). Photos, videos and audio are accessible through different systems on the FAO Web site. The CDR was conceived as the online digital library of FAO electronic documents and publications, as well as selected non-FAO material. At present, more than 23 000 full text documents are available through the CDR.

Resource type	Number of Records
full text documents	23 000
photos	8 500
videos	6 300
audio	750
Total	38 550

Table 1: Resources at FAO (as at 10 April 2007)

For each system described above, the objectives are different. The FAODOC focuses on the cataloguing of FAO documents. The EIMS deals with electronic publishing, especially the management at the full text level (rather than the description of documents). The CDR focuses on the dissemination of FAO documents archived through the EIMS. In 2003, a link between both databases was created, linking the FAODOC records to the full text documents archived in EIMS-CDR.

Figure 1: FAODOC user interface

This paper describes the process of merging the EIMS-CDR and the FAODOC and the creation of the FAO Open Archive. The result will be one unique sustainable digital repository offering a solid foundation for the collection, management, maintenance and timely dissemination of material published by FAO. To improve the effectiveness of the proposed repository, it will be necessary to streamline the existing electronic publishing

workflow and to integrate the current functions into new modules. The FAO Open Archive is based on three key elements:

1. a metadata set based on international description guidelines and format;
2. a workflow procedure that guarantees the processing of all documents published by FAO; and
3. a system architecture based on cataloguing and electronic publishing.

This paper is divided into the following sections: Section 2 presents the current situation for the EIMS-CDR and the FAODOC; Section 3 details the objectives of the FAO Open Archive; Section 4 describes the workflow procedures, the new architecture, the compliance to International Standards for Bibliographical Description (ISBD) [7] and metadata sharing with other systems; and Section 5 is the conclusion and the next steps in implementing the FAO Open Archive.

FAO CORPORATE DOCUMENT REPOSITORY

Home FAO Home WAICENT Portal Ask FAO العربية 中文 English Français Español

Advanced search About Us Contacts Help

NEW RELEASES

Report of the Regional Workshop on Rehabilitation of Agriculture in Tsunami affected Areas
 In the wake of the devastation caused by the tsunami of 26 December 2004, FAO initiated several projects to assess the damage to agricultural land, plan interventions, and support agricultural workshops at the country level in Indonesia and Sri Lanka ...details

Comprehensive Africa Agriculture Development Programme
 The New Partnership for Africa's Development (NEPAD) programme for agriculture, the Comprehensive Africa Agriculture Development Programme (CAADP), focuses on investment into three mutually reinforcing 'pillars' that could make the earliest difference to Africa...details

Trade Reforms and Food Security
 Between 1999 and 2002 FAO undertook a series of 23 country case studies to evaluate the impact of the WTO Agreement on Agriculture (AoA) on agricultural trade and food security in developing countries ...details

FOCUS on ... Manuals

- Addressing HIV/AIDS through Agriculture and Natural Resource Sectors: A Guide for Extension Workers
- Urban food supply and distribution in developing countries and countries in transition: A guide for planners
- Building on gender, agrobiodiversity and local knowledge: a training manual

The State of...

- ...Food and Agriculture
- ...World Fisheries and Aquaculture
- ...World's Forest
- ...Food Insecurity in the World
- ...Agricultural Commodity Markets

A word about PDF files

Read about how to download FAO publications using Acrobat PDF. Choose which files to download, according to your connection speed ... more information

Comments? Please write to the Webmaster Publication Index | © FAO 2007

Figure 2: CDR user interface

2 Objectives

The objective of the FAO Open Archive is to create a unique sustainable digital repository for the dissemination of FAO publications and simultaneously, enhance interoperability with other information systems. The FAO Open Archive will guarantee efficient electronic publishing and metadata management, the effective dissemination of FAO information resources and the preservation of the Organization's institutional memory.

3 Current Situation for EIMS-CDR and FAODOC

FAODOC has been managing all bibliographic information for FAO documents and publications for over 30 years (since 1976). Since 1998, FAO established a workflow to manage the electronic publishing and dissemination of FAO full text documents through the EIMS-CDR [8]. The EIMS-CDR and the FAODOC workflows, actors and content are described below.

3.1 EIMS-CDR, the Electronic Publishing and Digital Repository

There are four different user profiles in the EIMS-CDR workflow:

- originator – the person within the FAO unit responsible for providing the source files and/or the printed copy of the publication;
- data owner – the FAO unit responsible for the content of the publication;
- focal point – the person responsible in EIMS-CDR for managing requests from FAO units [9]; and
- liaison officer – the person within a FAO unit who ensures that publications are made available online. The liaison officer is the link between the originator and the focal point.

Detailed guidelines of the EIMS-CDR workflow are available to all FAO users and EIMS-CDR administrators. Following is a brief description of standard workflow steps:

1. The originator provides source files to the external printing unit. When the publication is printed, the external printing unit provides the focal point with the source files, the PDF version and the hard copy. In some cases files are provided by the originator;
2. The data owner creates and locates a record in EIMS;
3. The data owner notifies the focal point of the record and the uploaded files;
4. The focal point completes the record. Conversion to HTML or PDF is handled by focal points or outsourced to an external company. When conversion is completed, the focal point notifies the data owner of the test URL for reviewing the publication;
5. The data owner reviews the publication and either approves it or requests changes, by notifying the focal point;
6. The focal point reviews the final publication, publishes it and notifies the data owner of the public URL. If no conversion is required, the focal point prepares an HTML table of contents that links to the low-resolution PDF files and notifies the data owner of the public URL (in some cases only PDF files are published without the associated HTML pages).

Publications are made available in various electronic formats:

- Full HTML version; HTML loads quickly and is easier to read on-screen. ~14 000 records;
- Full PDF version; PDF is better for printing and downloading a local copy. ~2 200 records;
- Full HTML version and PDF version. ~6 500 records; and
- HTML table of contents linked to Full PDF version. ~500 records

3.2 FAODOC, the Online Catalogue

The FAODOC cataloguing process involves various actors:

- originator – the person within the FAO unit responsible for delivering to FAODOC the hard copy of the publications and/or the full text documents to be published in EIMS-CDR;
- EIMS-CDR focal point – the person who notifies the FAODOC cataloguer of a new record in EIMS-CDR, so they link the FAODOC record to the EIMS-CDR full text document; and
- cataloguer – the person who selects and catalogues the publications (hard copies and full text documents from EIMS-CDR).

The FAODOC manages the cataloguing of document and the dissemination of bibliographic information through an Online Public Access Catalogue (OPAC). There are procedures for the exchange of information between the FAODOC and the document producers, but there is no specific electronic tool to manage the reception of documents, as exists in the EIMS-CDR workflow. The lack of any workflow management system makes it difficult to control the reception and cataloguing of documents.

3.3 Main Differences between EIMS-CDR and FAODOC

The process of merging the two existing databases is a challenging task, as each has a different structure and workflow procedure. The first step towards the FAO Open Archive was to determine the similarities and differences between the EIMS-CDR and the FAODOC.

3.3.1 Software Overview

The EIMS-CDR was developed by FAO to manage the electronic publishing workflow. The CDR and the EIMS both run on a Microsoft Windows platform with an Oracle 9 database server. The software uses Microsoft's ASP programming language (Active Server Pages), with some ad hoc modules and functionalities developed in ASP.Net (the successor to ASP). The EIMS architecture results from the interaction of several modules that manage different aspects of the overall workflow. All modules interact with a single database that stores the records' descriptive metadata and detailed workflow information.

The FAODOC uses CDS/ISIS, a software package for information storage and retrieval – developed, maintained and disseminated by UNESCO. It is freely available for non-commercial purposes. The customization of data input and output interfaces occurred in Poland at the Institute for Computer and Information Engineering and at FAO.

3.3.2 Metadata Structure

CDS/ISIS manages a database whose main content is text, while the EIMS-CDR uses a relational Oracle database. The structure and logic of the two databases are completely different. However, these differences are not a barrier for the merger into a new single relational database.

Both systems use a very similar set of metadata fields to describe documents. The FAODOC contains detailed document information, while the EIMS-CDR provides fewer details on the actual document, but stores much information related to the actors, workflow and full text management. The mapping of the EIMS-CDR and the FAODOC databases has already occurred. It was not a complicated procedure, as both systems use a similar metadata field set. The compliance of both databases to the Dublin Core metadata standard and the AGRIS AP [10] at export level, facilitated the mapping. Only those fields required for the EIMS-CDR workflow have been added to those that already exist in the FAODOC.

3.3.3 Database Content

The EIMS-CDR and the FAODOC currently use FAO cataloguing guidelines. The decision to adopt international cataloguing standards was taken to guarantee interoperability with other digital repositories.

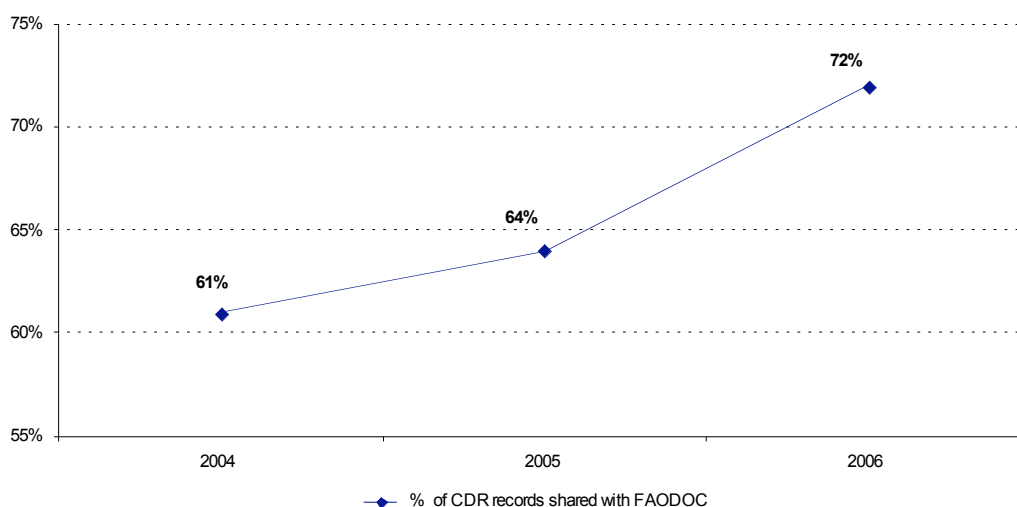


Figure 3: Percentage of the EIMS-CDR records catalogued in the FAODOC

In the EIMS-CDR, each record corresponds to one document (e.g., a book or a meeting report). The FAODOC catalogues documents and their analytics (e.g., a document is considered a book and the analytics are its chapters). Therefore, a book can have more than one record. The one-to-many relationship of records will be taken into consideration when merging data from the two databases.

The content of the two databases partially overlap, resulting in duplicate bibliographic records. The percentage of the EIMS-CDR full text documents linked from the FAODOC has increased over time (Figure 3): 72 percent of all records created in 2006 in the EIMS-CDR have been linked to from the FAODOC. This implies a duplication of effort (at metadata management level) and jeopardizes the dissemination and the maintenance of the FAO's institutional memory.

4 The Approach to Create the FAO Open Archive

The FAO Open Archive is based on the integration of the electronic publishing and the bibliographic cataloguing requirements. This merger requires the analysis of current workflows to detect similar procedures and reorganise them into a single coherent workflow. This process should focus on:

1. system architecture;
2. workflow procedure;
3. compliance with international data content standards; and
4. exposing metadata in a standardized way.

4.1 The New System Architecture

The architecture of the FAO Open Archive should integrate all features that are currently managed through the EIMS-CDR and the FAODOC. The FAODOC only manages the cataloguing process, but the FAO Open Archive must include the facility to deal with the reception of documents workflow, and improve the cataloguing module. The electronic publishing system is structured as a modular system where each module deals with a specific aspect of document publication. This approach will remain in the new architecture, integrated with new functionalities.

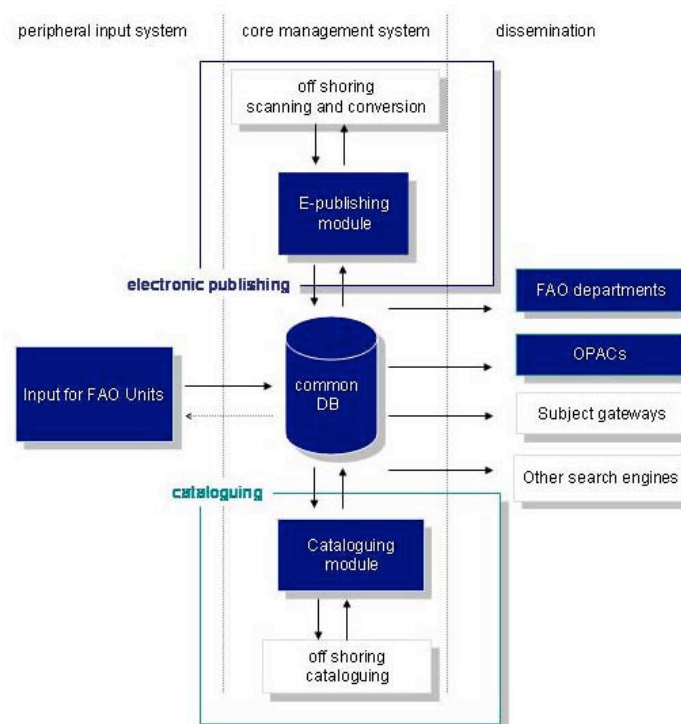


Figure 4: FAO Open Archive architecture

The FAO Open Archive architecture is detailed in Figure 4. The following elements define the architecture of the system:

1. integrated workflow; from left to right, the flow of information starts from the peripheral input system elements, passes through the core of the management system and to the dissemination interfaces;
2. common database; and
3. management of the two main functions of the FAO Open Archive; electronic publishing and cataloguing.

The objective of the system architecture is to manage all aspects of the electronic document life cycle. Electronic publishing and cataloguing will be managed through the same system and share the same database, e.g., from the document's creation, to its cataloguing, indexing and conversion to a suitable electronic format, to its dissemination on the Web.

Input for FAO units. This module will be used for data input and will be developed based on the current EIMS. FAO units now have individually customized EIMS interfaces. Each customization involves a basic internal workflow that can vary from one-step to multiple-step approval. FAO units are responsible for the introduction (and minimal description of documents) into the electronic publishing workflow. In the FAO Open Archive, FAO units will continue to provide data through a user-friendly system describing the document with a minimal set of metadata. With the FAO Open Archive, electronic publishing and cataloguing will share a common data entry point. The records that the FAO Open Archive will manage includes documents and multimedia files (photos, videos and audio) and non-FAO material (publications written in collaboration with FAO, yet FAO does not hold the copyright).

Electronic publishing. FAO will continue to publish documents online in electronic format. They will be managed through two modules:

- core module for electronic publishing – this module will be used to review the information from FAO units, based on EIMS, and to manage the conversion of full text documents into electronic formats (HTML, PDF, etc.); and
- scanning requests managing module – this module will be directly connected to the core module for electronic publishing and will be used to keep track of the work assigned to internal resources or of the work orders sent for scanning and/or conversion to external companies.

Cataloguing. FAO will offshore the cataloguing, using the minimal set of metadata and the full text provided by the FAO units. FAO cataloguers will check and validate the offshored records in order to guarantee the quality of the bibliographic description for the full text documents. Cataloguing will also be managed through two modules:

- core module for cataloguing – this module will be used to select records to be offshored for cataloguing and indexing and to check metadata quality. It will be used exclusively by cataloguers to manage the information to be released into the Open Archive; and
- cataloguing offshoring module – this module will be directly connected to the core module for cataloguing and will be used to manage the XML exports of data to be catalogued by external companies and to manage import and validation of offshored records.

4.2 Workflow Procedures

As well as the architecture, the workflow of the FAO Open Archive must integrate two main activities that so far have been conducted separately: electronic publishing and cataloguing. Figure 5 shows a top-down representation of the new workflow:

1. FAO units initiate a record by inserting a minimal set of metadata into the data input module. Only minimal information is requested to initiate a record: author, title, year and job number (a FAO unique identifier). The system verifies whether the job number exists in the database. A simple validation workflow within the peripheral input system will ensure that the records inserted are eligible for publication in the FAO Open Archive.

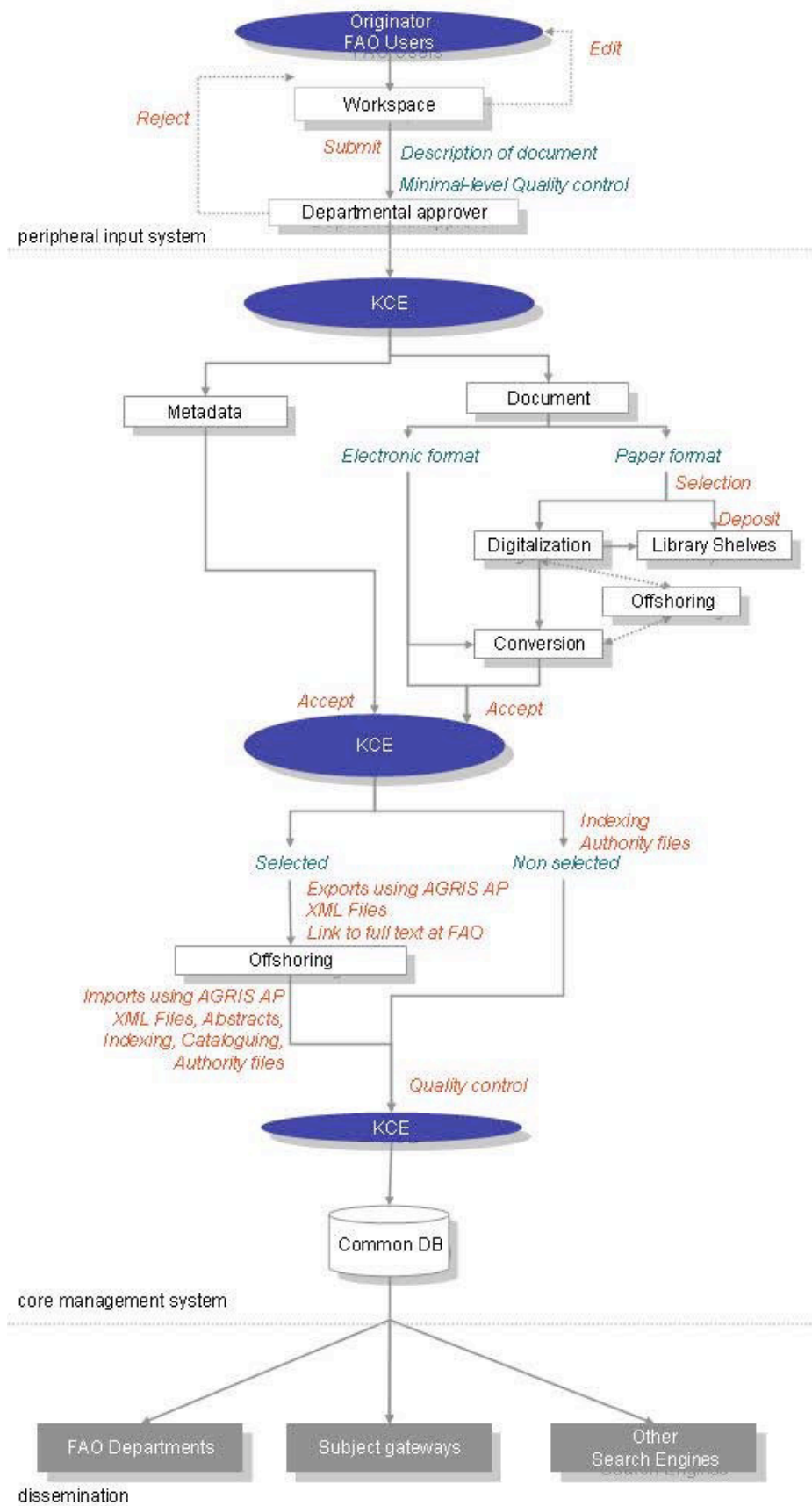


Figure 5: FAO Open Archive Workflow

2. The electronic publishing administration and the cataloguing administration are notified of the addition of a new record. They can take action simultaneously on the full text and the metadata of the records.
 - 2.1. If the document received is already in electronic format it requires validation and conversion to the most suitable format. This task can be carried out in-house or can be offshored. If the document needs digitalization then it is offshored for scanning.
 - 2.2. Using the minimal set of metadata in the system and the link to the full texts, the documents are catalogued and indexed by FAO and/or external cataloguers. The records that are selected for offshoring are exported using XML. When exported records are received from the external company they are imported into the system, checked and validated.
3. Validated records are disseminated through FAO Web sites. Moreover, search engines, services providers and digital libraries will harvest the records' metadata enhancing access to FAO documents.

4.3 Compliance with International Data Content Standards, ISBD

During the past few years, ISBD [11] has been identified as the standard most suitable for FAO. In April 2006, a study of the impact of changing FAO cataloguing rules recommended the adoption of ISBD rules:

"... recommend that FAO adopt the ISBD rules and build a system that will send and accept queries according to the OpenURL standard. In this way, FAO will build a system that will work with (interoperate with) other catalogues, while making FAO documents far more accessible to users. FAO, OCLC and other databases can create OpenURLs based on records that follow international guidelines and in this way, create an interoperable system [12]".

ISBD rules are rigorous and exact. ISBD is based on the principles of adequate identification, searchability and consistency so that:

1. no two different documents can be confused with each other; and
2. the many details comprising a description, are presented in a uniform manner so that they can be interpreted without unnecessary ambiguity [13].

By applying the ISBD rules, FAO will not only enhance the international exchange of FAO records, but will also assist in the interpretation of records across languages, because ISBD records can be interpreted on a first level (identification of elements) by users of every language. This is because of the fixed order of ISBD records. Finally, ISBD is independent of any metadata format. In conclusion, ISBD rules are simple, exact, widely used and supported by the International Federation of Library Associations and Institutes (IFLA). ISBD will facilitate the interoperability with other institutions and/or services providers, as it is an international standard followed by many of the world's major libraries and bibliographic institutions.

One of the biggest challenges will be the handling of the legacy data; old records require re-cataloguing, e.g., titles need to be transcribed according to ISBD rules. A possible solution could be to import bibliographic records from databases that have already catalogued FAO documents, ignoring fields that are not relevant to FAO's needs and adding specific information already existing in FAO records, e.g., AGROVOC Thesaurus [14] descriptors. However, the legacy data can be updated, prioritizing those records which have the full text available and/or are accessed on a regular basis. The introduction of an additional code to distinguish old from new ISBD records is required.

The FAO units will introduce a minimal-level description based on ISBD and the offshored and FAO cataloguers could then bring the records to full ISBD level.

4.4 Exposing Metadata in a Standardised Way

This is a very important issue, and it has been addressed successfully by the Open Archives Initiative (OAI). Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) is a simple protocol that allows data providers to expose their metadata for harvesting to services providers. The FAO Open Archive will be OAI compliant, so the FAO metadata can be harvested by any services providers and/or digital libraries.

The concept of OAI-PMH can be applied to a wide range of digital materials, e.g. images, audio or videos. It is mandatory to expose metadata as Dublin Core. It is important to note that the protocol enables multiple metadata

formats. These alternative forms of metadata can be as rich as is necessary to describe content. During the last few years, FAO has made an intensive effort to promote the exchange of high-quality metadata within the AGRIS Network, an international initiative based on a collaborative network of institutions in agriculture and related subjects. The AGRIS AP is a metadata format that facilitates sharing of metadata across different information systems. It is a metadata schema which uses elements from metadata standards such as Dublin Core (DC), Australian Government Locator Service Metadata (AGLS) [15] and Agricultural Metadata Element Set (AgMES) [16] namespaces. The standard enhances the quality of the description of agricultural information resources, enabling greater processing possibilities by service providers. The AGRIS AP has proved to be a successful initiative, and as a result, the FAO Open Archive will be fully compliant with the AGRIS AP at export level.

In conclusion, exposing metadata will:

1. improve the retrieval of FAO documents from a large number of sources (e.g., portals, aggregators and services providers);
2. allow aggregators to detect FAO documents and thereby help to disseminate them; and
3. enhance the visibility and awareness of FAO's available resources.

5 Conclusions and Next Steps

This paper illustrates the first phase for the creation of the FAO Open Archive, focussing on finding a strategy to solve:

1. the duplication of efforts in creating and managing metadata; and
2. the lack of integration of electronic publishing and cataloguing.

The relevant findings from this first phase are:

- The FAODOC and the EIMS-CDR will use a common database and a workflow supported by a workflow management system. FAO will supply FAO bibliographic metadata together with the full text.
- The conversion of the FAODOC and the EIMS-CDR to the FAO Open Archive will facilitate the data input and maintenance of information. The FAO units will continue to be involved in the metadata creation process.
- The use of ISBD rules will simplify the creation of metadata. The legacy data will be updated to ISBD standards, prioritizing those records, which a) are accessed on a regular basis, and b) have the full text available to improve the effectiveness of the OpenURL protocol.
- The visibility and dissemination of FAO documents will be maximized by exposing content through OAI-PMH. The FAO Open Archive should have the ability to transfer and use information in a uniform and efficient manner across multiple organisations and information technology systems.

The creation of the FAO Open Archive will strengthen FAO's role as a knowledge dissemination organization. The following phase is related to the software implementation. The integration of open source software into FAO Open archive is still under evaluation.

Acknowledgements

We would like to thank Anne Aubert, Johannes Keizer, Giorgio Lanzarone, Romolo Tassone and Jim Weinheimer for their valuable contributions.

Notes and References

- [1] FAO Constitution, Article I. <http://www.fao.org/docrep/x1800e/x1800e01.htm#1> Last accessed in April 2007.
- [2] Catalogue for Documents produced by FAO (FAODOC) <http://www4.fao.org/faobib/index.html> Last accessed in April 2007.

- [3] Electronic Information Management Services (EIMS). <http://www.fao.org/eims/> Last accessed in April 2007.
- [4] Corporate Document Repository (CDR) <http://www.fao.org/documents/> Last accessed in April 2007.
- [5] The Knowledge Exchange & Capacity Building Division (KCE) of FAO is the responsible for all the above mentioned systems.
- [6] AGRIS/CARIS Centre of Information Management for international agricultural research <http://www.fao.org/Agris/> Last accessed in April 2007.
- [7] International Standards for Bibliographic Description (ISBDs <http://www.ifla.org/VI/3/nd1/isbdlist.htm> Last accessed in April 2007.
- [8] SALOKHE, G.; PASTORE, A.; RICHARDS, B.; WEATHERLEY, S.; AUBERT, A.; KEIZER, J.; NADEAU, A.; KATZ, S.; RUDGARD, S.; MANGSTL; ANTON. *FAO's role in Information Management and Dissemination – Challenges, Innovation, Success, Lessons Learned*. 2005. <ftp://ftp.fao.org/docrep/fao/008/af238e/af238e00.pdf> Last accessed in April 2007.
- [9] This task involves the scanning and conversion of documents, corrections, modifications and the publication of HTML/PDF files.
- [10] The AGRIS Application Profile for the International Information System on Agricultural Sciences and Technology Guidelines on Best Practices for Information Object Description <http://www.fao.org/docrep/008/ae909e/ae909e00.htm> Last accessed in April 2007.
- [11] In 1969 the International Federation of Library Associations and Institutes (IFLA) created a general framework for the creation of standards to regularize the form and content of bibliographic descriptions (Byrum, J.D., "The Birth and Re-birth of the ISBDs: Process and Procedures for Creating and Revising the International Standard Bibliographic Descriptions". *IFLA journal*, Vol. 27, No. 1, 2001). The work resulted in the ISBD rules which specify the requirements for the description and identification of the most common types of resources that are likely to appear in library collections.
- [12] WEINHEIMER, J. (2006). *Consequences of changing FAO cataloguing rules & format with ISBD/AACR2/MARC21: a report for the Food and Agriculture Organization of the United Nations*. Internal report.
- [13] COETZEE, H. (2005). *Do we still need bibliographic standards in computer systems?* http://www.liasa.org.za/interest_groups/igbis/papers/IGBIS_WSJul04_Bib_Stds_Helena_Coetzee.doc Last accessed in April 2007.
- [14] AGROVOC is a multilingual structured and controlled vocabulary designed to cover the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains. http://www.fao.org/aims/ag_intro.htm
- [15] AGLS Metadata Standard http://www.naa.gov.au/recordkeeping/gov_online/agls/summary.html Last accessed in April 2007.
- [16] Agricultural Metadata Element Set (AgMES) http://www.fao.org/aims/intro_meta.jsp Last accessed in April 2007.

Five Years on – The Impact of the Budapest Open Access Initiative

Melissa R. Hagemann

Information Program, Open Society Institute
400 West 59th Street, New York, NY 10019, USA
e-mail: mhagemann@sorosny.org

Abstract

Open Access was first defined by the Budapest Open Access Initiative following a meeting organized by the Open Society Institute/Soros foundations. The subsequent Open Access movement has had a large impact on the scholarly communications system. This is seen in the growing number of Open Access journals and institutional and subject-based repositories which have developed over the past five years. The reaction of publishers to the movement has been mixed with individual publishers (both commercial and non-profit) experimenting with the model while large publishers' associations have generally shown resistance. However, the movement continues to gain strength as research funding agencies adopt Open Access mandates to the research they support.

Keywords: open access; Budapest open access initiative; Open Society Institute

1 Introduction

February 14, 2007 marked the fifth anniversary of the release of the Budapest Open Access Initiative (BOAI), which offered the first definition of Open Access [1]. This paper examines the impact of the BOAI over the past five years. Background information on the role of the Open Society Institute (OSI)/Soros foundations will be provided, followed by an examination of key objective measurements for analyzing the impact of the BOAI.

In 2001 OSI's Information Program began to follow the developments of several projects which shared the ultimate goal of making peer-reviewed scholarly content freely available online. Among these projects were arXiv.org, the preprints archive for Physics, Mathematics, Computer Science and Quantitative Biology and the Public Library of Science's petition, which called on researchers not to submit their articles to any publisher which did not allow articles to be freely available after six months. OSI organized a meeting in Budapest in December 2001 which brought together a group of leaders who were exploring alternative publishing models. During the meeting it was decided to link the blossoming repository (or self-archiving) movement with Open Access journal publishing. Thus the BOAI defined these as two complementary strategies for achieving Open Access. The simultaneous promotion of the two strategies has proven to be highly productive. Ultimately to succeed, both strategies rely on mandating Open Access to publicly funded research.

Following the release of the BOAI, OSI's Information Program pledged \$3 million to support Open Access initiatives. While OSI initially intended to spend these funds over a three year period, we realized that the transition to Open Access will require a longer time commitment on the part of OSI and more funding than initially pledged. This paper documents both the impact of OSI's direct funding of the principles outlined in the BOAI, as well as broader policy and funding discussions which followed the release of the BOAI.

2 Methodology

Key objective measurements for evaluating the impact of the BOAI include:

- a review of meetings which have followed the BOAI;
- the number of Open Access journals and institutional and subject-based repositories which have developed in the past five years;
- the number of sites which link to the BOAI as well as to some of the Open Access projects which OSI has funded;
- a review of the response of publishers to the Open Access movement;
- an examination of the major declarations and funders' policies regarding Open Access which have followed the BOAI.

3 The Development of a Movement

Having defined Open Access, the BOAI inspired lively debates among publishers, academics, librarians, and funders (both governmental and private) regarding the future of scholarly communication. Much of OSI's funding in the past five years has been dedicated to meetings, conferences and workshops which introduce the concept of Open Access. As of January 2007, OSI has provided \$441,300 in funding to support over 40 meetings to introduce and promote Open Access throughout the world.

In addition to supporting meetings on Open Access, OSI has funded projects which directly advocate for Open Access. Examples of these are the Open Access News blog, which is written by Peter Suber. Open Access News has come to be regarded as the main source for information on the Open Access movement and this can be seen in the over 5,400 sites which link to it. OSI also supports some of SPARC's (the Scholarly Publishing and Academic Resources Coalition) work to advocate for Open Access. SPARC has developed the Alliance for Taxpayer Access, an organization representing taxpayers, patients, physicians, researchers and institutions that support Open Access to taxpayer-funded research.

Seeing the need to facilitate the discovery and use of Open Access journals and repositories, OSI funded the development of the Directory of Open Access Journals (www.doaj.org) and the Directory of Open Access Repositories (www.opendoar.org). The DOAJ was developed by Lund University Libraries and as of April 2007 lists 2,622 Open Access journals, an increase of over 2,300 since its launch in 2003.

To complement the DOAJ, OSI brought together a group of funders to support the development of the Directory of Open Access Repositories by the University of Nottingham and Lund University Libraries. Currently OpenDOAR lists 853 institutional repositories and 15,400 sites link to the OpenDOAR. As of April 2007, 522 sites link to the BOAI. In particular, organizations often link to the BOAI in reference to defining Open Access.

Beyond the Open Access meetings and projects which OSI has funded, the discussion regarding Open Access has been broadened since 2002 to include national, international and institutional funders. In 2003, the Howard Hughes Medical Institute (HHMI) and the Max Planck Society both held meetings which addressed Open Access from a funder's perspective. The HHMI meeting produced the Bethesda Statement [2] (the meeting was held at HHMI's headquarters in Bethesda, Maryland) and the Max Planck conference developed the Berlin Declaration [3]. Both the Bethesda Statement and the Berlin Declaration provide definitions of Open Access which focus on the role of funders. Thus adding the Budapest definition to this mix, many refer to the "BBB" definition of Open Access.

4 Publishers' Reaction to Open Access

The BOAI received stiff criticism from publishers' associations when it was announced in February 2002. Sally Morris of the Association of Learned and Professional Society Publishers (ALPSP) said: "We are convinced all of our scholarly communities will be ill-served by an initiative which promotes systematic institutional archiving of journal content without having in place a viable alternative model to fund the publication of that content. This can only serve to undermine the formal publishing process which these communities value. She warned against those who would 'give it all away first and then start worrying later' [4].

However by the fall of 2002, ALPSP and OSI held a joint workshop in London which described the Open Access publishing model. This was the first in a series of three ALPSP/OSI workshops. By the third workshop, Martin Richardson of Oxford University Press (OUP) described how OUP was experimenting with the hybrid model of Open Access. Through the hybrid model publishers offer authors the choice of paying the article processing fee and having their article made freely available online, or they can elect not to pay and then only journal subscribers will have access to the article. This model seems attractive to authors, as by electing to have their article made freely available through Open Access, it has the potential to reach a larger audience. When OUP adopted the hybrid model for their *Journal of Nucleic Acids*, they found that a high percentage of authors elected to pay the article processing fee. Based upon this response, OUP converted the journal to full Open Access [5]. The hybrid model offers publishers of traditional subscription-based journals a way to experiment with Open Access and allow the pace of change to be dictated by the authors themselves. Jan Velterop, former publisher of BioMed Central and currently the Director of Open Access at Springer, described how the hybrid model can work for publishers wishing to experiment with Open Access in his *Guide to Open Access Publishing and Scholarly Societies* [6] commissioned by OSI. Within Springer, Velterop leads the Springer Open Choice Program which allows authors who submit their articles to all Springer journals to choose the hybrid model of

Open Access. Through Springer Open Choice, authors are allowed to retrain their copyright. Springer has adopted the Creative Commons Attribution License 2.0 as the Springer Open Choice License [7].

In addition to subscription-based journals which are converting to Open Access, there are many new Open Access journals which have been developed. Today the largest commercial Open Access publisher is BioMed Central which publishes over 175 titles. SciELO (the Scientific Electronic Library Online), based in Brazil, publishes over 200 Open Access titles and is supported by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo), in partnership with BIREME (the Latin American and Caribbean Center on Health Sciences Information). Hindawi Publishing, based in Cairo, publishes over 60 titles among a wide range of fields including Engineering, Life Sciences, Mathematics, and Physical Sciences. Most importantly, Hindawi Publishing has shown that the article processing fee business model is sustainable. As Paul Peters, Head of Business Development at Hindawi explained: “Based on our experience as a publisher of both subscription-based journals and author-pays open access journals, I would not only argue that the author-pays publishing model is sustainable, but also that it has many economic advantages over the subscription model. Even though our open access journal collection is only a few years old, we have already achieved profitability for the collection as a whole. Moreover, using a business model based on publication charges has enabled us to expand our publishing program in a much more sustainable way than we were able to using a subscription model” [8].

The Public Library of Science (PLoS), launched by Nobel Laureate Harold Varmus, Mike Eisen, and Pat Brown, has demonstrated that Open Access journals can compete with the top subscription-based journals in terms of producing high quality journals. *PLoS Biology* is ranked as the most highly cited general biology journal with an impact factor of 14.7 [9]. And PLoS is pushing the boundaries of the traditional concept of a journal with their new PLoS ONE which represents cutting edge innovation which could fundamentally change how research is communicated.

While individual publishers are experimenting with Open Access, some of the publishers’ associations continue to strongly oppose it. This was highlighted in January 2007 when *Nature* revealed that the Association of American Publishers (AAP) had hired a high profile public relations firm, Dezenhall Resources headed by Eric Dezenhall, to attack the Open Access movement [10]. Dezenhall’s corporate clients have reportedly included Enron and Exxon Mobil. “Dezenhall told the association’s professional and scholarly publishing division, he could help – in part by simplifying the industry’s message to a few key phrases that even a busy senator could grasp. Phrases like: ‘public access equals government censorship,’ and ‘government is seeking to nationalize science and be a publisher.’ The publishers liked what they heard” [11].

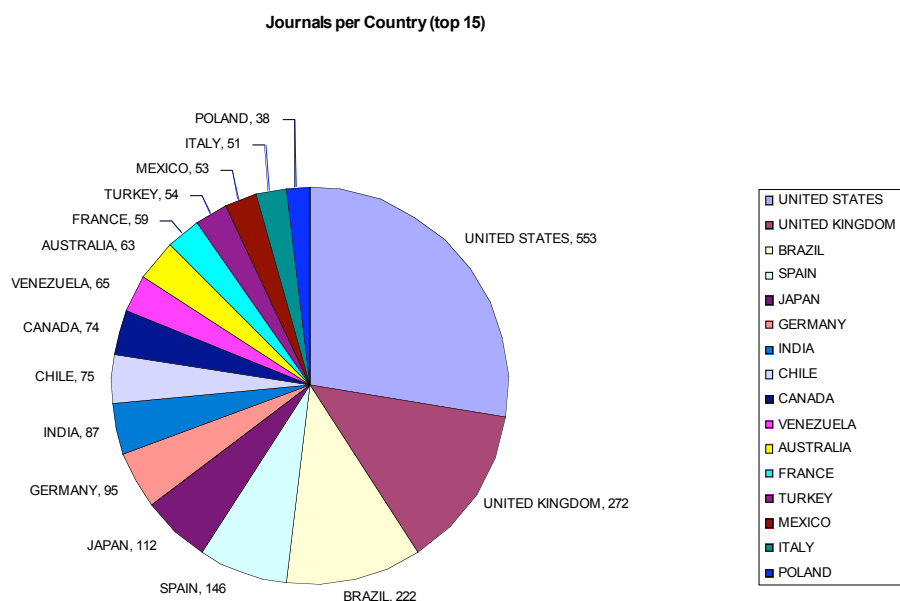


Figure 1: Journals per Country in DOAJ (top 15)

At the grassroots level, the reaction of publishers as well as users of research material to the Open Access movement can be seen by looking at statistics from the DOAJ. As previously mentioned 2,300 titles have been added to the DOAJ since its launch in 2003, thus 2,300 journals have either converted to or been launched as Open Access journals during this time. The importance of the DOAJ can be seen by the fact that 17,800 sites link to it. The DOAJ also receive over 5 million visits per month (although this figure does include robots).

By examining the countries where journals in the DOAJ are based (see Figure 1) it is clear that while many journals are based in the United States and the United Kingdom, Open Access has been adopted by publishers throughout the world, including those based in many developing countries.

	February 2007	December 2006
1	1582079: 37.32%: .com (Commercial)	1234542: 37.58%: .com (Commercial)
2	381159: 18.10%: [unresolved numerical addresses]	365680: 23.31%: [unresolved numerical addresses]
3	145734: 6.36%: .net (Networks)	189425: 9.08%: .net (Networks)
4	227688: 6.21%: .za (South Africa)	160363: 5.77%: .za (South Africa)
5	73271: 3.85%: .edu (USA Higher Education)	50301: 2.39%: .br (Brazil)
6	93426: 3.61%: .org (Non Profits)	42332: 2.03%: .edu (USA Higher Education)
7	8678: 3.28%: .bg (Bulgaria)	16593: 1.89%: .org (Non Profits)
8	77620: 3.15%: .ch (Switzerland)	44513: 1.88%: .ch (Switzerland)
9	55465: 1.75%: .br (Brazil)	34474: 1.58%: .uk (United Kingdom)
10	44309: 1.68%: .de (Germany)	15913: 1.44%: .de (Germany)
11	38913: 1.38%: .uk (United Kingdom)	24092: 0.84%: .fr (France)
12	22442: 0.80%: .ca (Canada)	15505: 0.71%: .it (Italy)
13	23282: 0.73%: .fr (France)	9933: 0.61%: .in (India)
14	18516: 0.73%: .my (Malaysia)	8485: 0.59%: .jp (Japan)
15	15142: 0.61%: .se (Sweden)	10227: 0.54%: .tr (Turkey)
16	11221: 0.61%: .in (India)	9997: 0.50%: .ca (Canada)
17	14596: 0.57%: .mx (Mexico)	8553: 0.47%: .pl (Poland)
18	15897: 0.54%: .it (Italy)	9999: 0.45%: .se (Sweden)
19	7332: 0.45%: .jp (Japan)	10889: 0.44%: .gr (Greece)
20	10317: 0.43%: .nl (Netherlands)	8329: 0.41%: .mx (Mexico)
21	10485: 0.41%: .tr (Turkey)	8584: 0.40%: .be (Belgium)
22	4753: 0.38%: .dk (Denmark)	6726: 0.40%: .es (Spain)
23	9422: 0.38%: .au (Australia)	7104: 0.39%: .nl (Netherlands)
24	7384: 0.36%: .es (Spain)	7727: 0.36%: .pt (Portugal)
25	7944: 0.36%: .pl (Poland)	6114: 0.33%: .fi (Finland)

Figure 2: Hits to DOAJ based upon country (top25).

The high global appeal of Open Access journals is also seen by examining the hits to the DOAJ based upon country (see Figure 2). While the wealthy research countries are represented, many developing countries also make the top 25, thus demonstrating that Open Access journals have been promoted widely and deely to the global research community. The high percentage of hits coming from unresolved domains could be due to the fact that many users in developing countries access the DOAJ through internet cafes and other third party access points.

	2004 Feb	2004 Nov	2005 Feb	2005 Nov	2006 Feb	2006 Nov	2007 Feb
Successful requests:	264,931	1,318,720	1,225,736	1,945,841	2,632,710	2,607,935	3,062,684
Redirected requests	57,660	513,306	395,886	328,585	525,862	1,745,736	2,318,193
Distinct files requested:	33,016	171,181	272,397	280,800	487,478	738,879	776,702

Distinct hosts served:	33,107	120,320	81,189	171,378	138,900	231,663	175,055
Data transferred MB:	1,570	12,960	11,440	20,420	27,330	23,900	25,990
Link to journal						750,677	836,151
Explanation:							
Successful requests	Each time a user prompts the server to show a file it is a request						
Redirected requests	A redirected request can be either a redirection within DOAJ (i.e. from a bibliographic record to an abstract) or from DOAJ to an external server (i.e. from an abstract in DOAJ to the full-text on a publisher's site).						
Distinct files requested:	Indicates how many different files in the DOAJ have been requested during one month.						
Distinct hosts served:	Indicates how many different registered IP-addresses have consulted the DOAJ during one month.						
Data transferred MB:	Indicates how much data has been transferred (downloaded) from DOAJ during one month. Take in consideration that one metadata record is only a very small number of bytes - 1000 Megabytes= 1 Gigabyte.						
Link to journal	Indicates how many times users have used the DOAJ to go to an abstract or full-text on the publishers sites during one month.						

Figure 3: DOAJ requests

And finally, the high use of the Open Access journals in the DOAJ (see Figure 3) is seen in the growing number of requests which the DOAJ has received over the past three years.

5 Mandating Open Access – The Role of the Funding Agencies

The role of the research funders within the Open Access movement is extremely important. By 2003 the Open Access movement was advocating for Open Access to research supported by both governmental and private research funders. The research funders have begun to adopt mandates for Open Access (or Public Access in the case of government-supported research as this research is supported by tax dollars). The message that research funders (and taxpayers) are essentially paying twice for the same information has resonated with funders. In the case of government funded research, the public supports the research itself through grants from the federal research agencies and then the public (through libraries, hospitals, etc.) must purchase the journals in which the publicly funded research is published to access the research results.

In 2003 the Wellcome Trust published an economic analysis of scientific publishing [12]. Based upon this report, the Trust decided to pursue an Open Access policy for the research which it funds. This ultimately led to the Trust becoming the first funder to mandate Open Access to all of the research it funds in September 2006 [13].

The Science and Technology Committee of the House of Commons launched an Inquiry into the state of Scientific Publishing in 2004. Its final report concluded that “the current model of scientific publishing is unsatisfactory” and “recommends that the Research Councils and other Government funders mandate their funded researchers to deposit a copy of all articles in repositories.” [14]. Although the report was released in 2004 it took some time for the Research Councils in the UK to adopt policies mandating Open Access. Today five out of the seven Research Councils [15] mandate Open Access to the research which they fund. Of particular significance, among the five which mandate Open Access is the Medical Research Council. This coupled with the mandate from the Wellcome Trust insures that the bulk of medical research funded in the UK will be available through Open Access.

A 2006 study by the European Commission on the Economic and Technical Evolution of the Scientific Publication Markets of Europe recommended public access to publicly-funded results [16]. This study was discussed at a meeting organized by the Commission on Scientific Publishing in the European Research Area in February 2007 in Brussels. As a result of the meeting, the Commission will now include the costs of Open

Access publishing as an eligible cost in Community funded projects and will begin discussions with the European Parliament and the Council regarding mandating Open Access [17].

In the U.S., the Federal Research Public Access Act (FRPAA) will be re-introduced this spring. FRPAA would mandate Public Access to research funded by the eleven largest government departments and agencies (i.e. National Institute of Health, National Science Foundation, Department of Energy, etc.). FRPAA would require that every federal agency with an annual research budget of \$100 million or more implement a public access policy which would require researchers who receive full or partial support from the agency to deposit a copy of their article in a stable digital repository maintained by that agency or in another suitable repository that permits free public access, interoperability, and long-term preservation no later than six months after the article has been published in a peer-reviewed journal. This would be a huge improvement over the current NIH Public Access Policy which “requests” NIH funded authors to deposit a copy of their article in PubMed Central and has seen a very low compliance rate on the part of the authors [18].

Funding agencies in developing and transition countries are also considering mandating Open Access to the research which they fund. In Ukraine, a Parliamentary Inquiry on Harmonization of Governmental Educational Policies was launched in December 2005 and concluded that the Ministry of Education and Science should encourage the development of Open Access resources in science, technology and education with Open Access a condition of state funded research. Subsequently, an Open Access Working Group was formed in Ukraine with representatives of the Parliamentary Committee on Science and Education, the State Fund for Fundamental Research, the Scientific and Publishing Council of the National Academy of Science of Ukraine, the Ministry of Science and Education, the National Library of Ukraine, the State Department of Intellectual Property, the Kyiv public administration, and the International Renaissance Foundation (Soros Foundation–Ukraine) [19]. In South Africa, the South African National Research Foundation has pledged to support all costs associated with their grantees publishing in Open Access journals. And the Library of the Chinese Academy of Sciences held the first Open Access meeting in China in June 2005 and is working with other government funding bodies to support Open Access.

6 Lessons Learned

As mentioned earlier, OSI initially pledged \$3 million to support the Open Access movement when the BOAI was launched in 2002. Since then OSI has seen that the transition to Open Access will require a longer time commitment on our part and more funding than initially pledged. In 2002 it was hoped that other foundations would join in supporting Open Access. With the exception of the Gordon and Betty Moore Foundation and the Sandler Family Supporting Foundation which have provided generous support to PLoS, other foundations have not embraced Open Access, although some of the leading American foundations provide substantial support to other open content issues such as Intellectual Property Rights reform and the development of open source software. More philanthropic support directed at advocating for the adoption of Open Access mandates by government and research funding institutions would be extremely helpful in countering the lobbying efforts of the large publishers’ associations.

From OSI’s experience with the BOAI it is clear that it was important to first define Open Access and develop specific strategies for achieving it. This allowed the key stakeholders to develop communities and subsequently a movement to support Open Access. This could serve as an example for the development of other movements around open content issues, such as open educational resources.

7 Directions for the Future

While the developments over the past five years are encouraging, much still remains to be done for Open Access to meet its full potential. Among the top priorities for the movement are:

1. Mandates from governments/funding agencies: Europe appears to be leading the way in terms of adopting significant mandates with the leadership of the Wellcome Trust and the five Research Council in the UK which have adopted mandates. In the U.S., while the FRPAA will be re-introduced this year in the Senate, strong opposition to it, led by AAP, poses a real obstacle to its adoption and increased support for public access advocacy will be needed;
2. Mandates from universities for deposit of material in repositories: In addition to developing repositories, more universities must adopt mandates for the deposit of all research written by

those affiliated with the university in the institutional repositories. This will require continued advocacy at many levels of the university administration and faculty,

3. The development of more Open Access journals: Some estimate that there are 24,000 peer-reviewed journals, thus this would mean that just over 10% are Open Access if one considers that the DOAJ lists 2,622 Open Access titles. More Open Access journals must be developed so that authors can have a choice to publish in an Open Access journal as opposed to a subscription-based journal. In addition to the numbers, it is important that the quality of the Open Access journals is high so that authors will elect to publish in them,
4. Continued unity of the Open Access movement: The Open Access movement (the Open Access publishing and the self-archiving/repositories communities) must remain united behind the common goal of making peer-reviewed content freely available and not allow differing mandates directed at journals or repositories to divide the movement.

8 Conclusion

The impact of the BOAI is clearly seen when one considers that before the meeting in Budapest, there was not even a term or definition for Open Access. Now Open Access is being debated by governments and publishers and mandated by funding bodies and universities. Much still remains to be achieved, but it is clear that Open Access has permanently changed the field of scholarly communication.

Acknowledgements

I would like to thank Leslie Chan, Program Supervisor for the Joint Program in New Media Studies and the International Studies Program at the University of Toronto at Scarborough for his guidance in the development of this paper as well as Lars Bjørnshauge, Director of Lund University Libraries, for providing a wealth of data on the DOAJ.

Notes and References

- [1] The BOAI defines Open Access as the free availability of peer-reviewed literature on the public internet, permitting any user to read, download, copy, distribute, print, search, or link to the full texts of the articles. See <http://www.soros.org/openaccess/>.
- [2] Bethesda Statement on Open Access Publishing: <http://www.earlham.edu/~peters/fos/bethesda.htm>.
- [3] Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities: <http://oa.mpg.de/openaccess-berlin/berlindeclaration.html>.
- [4] DODD, D. Access to research should be open to all, say many in the scientific community. Try telling that to publishers, *Information World Review*, April 15, 2002, p. 9.
- [5] Oxford University Press Release. Oxford Journals takes bold step towards free access to research. 26 June, 2004. http://www.oxfordjournals.org/our_journals/nar/narpressjun04.pdf.
- [6] VELTEROP, J.M. *Open Access Publishing and Scholarly Societies*. July 2005. http://www.soros.org/openaccess/scholarly_guide.shtml.
- [7] Springer Open Choice License: http://www.springer.com/dal/home/open+choice?SGWID=1-40359-12-161193-0&teaserId=55557&CENTER_ID=115382.
- [8] See Paul Peters comments in the *Nature Newsblog*, June 21, 2006: http://blogs.nature.com/news/blog/2006/06/openaccess_journal_hits_rocky.html.
- [9] See overview of PLoS Journals: <http://www.plos.org/journals/index.html>.
- [10] GILES, J. PR's 'pit bull' takes on open access. *Nature*, Vol. 445/25 January 2007, p. 347.
- [11] WEISS, R. Publishing Group Hires 'Pit Bull of PR'. *Washington Post*, January 26, 2007.
- [12] Economic Analysis of Scientific Publishing. Commissioned by the Wellcome Trust, January 2003. <http://www.wellcome.ac.uk/assets/wtd003182.pdf>.
- [13] Wellcome Trust position statement in support of open and unrestricted access to published research. Last updated 14 March 2007. http://www.wellcome.ac.uk/doc_WTD002766.html.

- [14] Scientific Publications: Free for all? *Select Committee on Science and Technology, Tenth Report*, 7 July 2004. <http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/39914.htm>.
- [15] See SHERPA's Juliet site on research funders' open access policies. www.sherpa.ac.uk/juliet.
- [16] DEVROEY, J.P.; DUJARDIN, M.; VANDOOREN, F. *Study on the economic and technical evolution of the scientific publications markets in Europe*. Commissioned by DG-Research, European Commission, January 2006. http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf.
- [17] Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee on scientific information in the digital age: access, dissemination and preservation. COM(2007) 56 final, 14 February 2007. http://ec.europa.eu/research/science-society/document_library/pdf_06/communication-022007_en.pdf.
- [18] See The Alliance for Taxpayer Access site: FRPAA: <http://www.taxpayeraccess.org/frpaa/index.html#issue>
- [19] See Access to Knowledge, Ukraine site. Open Access Working Group formed in Ukraine: <http://www.a2k.org.ua/news.php?id=1172&lng=en>