Sharing the Know-how of a Latin American Open Access only e-journal: The Case of the Electronic Journal of Biotechnology

Graciela Muñoz[1]*; Atilio Bustos-González*[2]*; Alejandra Muñoz-Cornejo*[2]

[1] Instituto de Biología, Facultad de Ciencias Básicas y Matemáticas, Pontificia Universidad Católica de Valparaíso, Av. Brasil 2950, Valparaíso, Chile
e-mail: gmunoz@ucv.cl
[2] Sistema de Biblioteca, Pontificia Universidad Católica de Valparaíso
Av. Brasil 2950, Valparaíso, Chile
e-mail: {abustos; biotec}@ucv.cl

Abstract

Scientific communication is essential for the advancement of science and in generating benefits for the general society. Also it is fundamental in strengthening the knowledge society with a positive effect on innovation and economic growth. The Open Access journals have demonstrated to be important in providing a reliable and a more accessible mean in communicating science. An example as such, is that they are evaluated by the ISI Thomson Scientific –recognized as an authority for evaluating journals- following the same rigorous selection process as journals in print media. The measurement of the impact factors in the electronic publications demonstrates that these receive a smaller citation level than print journals, ranking in general in the lower half of journals in their subject category. Moreover, the low usage of the electronic media demonstrates a lack of confidence of authors in this new mean of communication. In this sense, editors have to provide answers to some unsolved issues regarding e-publications in order to make these journals more reliable and confident to the scholarly community. The journals edited in Latin America with international visibility represent 0.63% of the total number of publications covered by the ISI Web of Science. In the year 2005, there were 44 Latin American journals covered by the Science Citation Index, of which 89% of these are considered Open Access publications. In that same year, these publications reached an impact factor average of 0.447 in comparison with the impact factor average of 1.588 for all the journals of Open Access at world-wide level. The Electronic Journal of Biotechnology is the only Open Access Latin American journal edited exclusively in the electronic format which is covered by the ISI Science Citation Index. The experience of this journal shows that with commitment to international diversity, quality, academic rigor of the peer review process, transparency, responsibility to scientists, innovation and international cooperation, a high level of visibility and accessibility can be obtained, as demonstrated by an average of more than 83,000 readers during year 2006 and an impact factor of 0.725, which is over the mean value of Latin American journals, offering an unique opportunity to fulfill the ever increasing public demand for science information.

Keywords: ISI Web of Science; electronic journal; scientific communication; impact factor; open access

1 Introduction

Science and its communication are essential in science advancement and in building the knowledge society with a positive effect on innovation and economic growth. Its dissemination, accessibility and understanding play key roles in its impact on research funding policies and in the benefits for a knowledge-based economy. The success of modern science depends on social acceptance of new scientific results and requires a permanent dialogue with an informed civil society where an open communication system, accessible and visible, is of primary importance for the benefit of society [1].

Although the introduction of the web based technology has raised a continuous debate on science communication, the journal system has remained stable and scientists value journal articles as a recognized mean of communicating original, peer-reviewed and edited information. The key problems regarding the use of this journal system continue to be the high costs of subscriptions, technical barriers, and the specialized language used in the scientific articles which leads to non-equity and discrimination across the international science community [2, 3].

The increasingly pervasive impact of science and technology is reaching every aspect of human welfare and is therefore urgent to make information more accessible and more usable by offering electronic journals a unique opportunity to reach these goals through the instant access to and dissemination of scientific information [4]. An important recent trend has been the development of the Open Access movement, which promotes free online access to full text research articles in every academic field [5, 6].

2 Are Scientists and Society Living in the Plato's "Myth of the Cave"?

The changes in information technology and communication that made possible the rise of the knowledge society generated a new type of illiterates. They are citizens and scientists, as authors and publishers, have not sufficiently incorporated the new expertise and possibilities of communication that are based on the use of information technology. This technology not only changed the form of communicating, but also reinvented the strategies used to recover, to analyze, and to diffuse scientific information. In recent years, societal changes have rapidly progressed and caused the previous guidelines of scientific communication to be insufficient for today's society.

An important decision made by the members of the international scientific community was to adopt new guidelines of communication for their results. Also, as evidenced in this article, a large part of the world's researchers continue to resist the new forms of communicating science based in web technology. The consolidation of new communication channels as they are: exclusively electronic journals, the repositories of preprints and postprints, and the institutional repositories. The enhancement of procedures of quality control of science, such as public and open peer review. Furthermore, the opportunities created by multimedia and hypertext that can exist exclusively in an electronic format. New alerting services and new indicators of scientific production, such as the index of Hirsh or the citation tracker, constitute sufficiently forceful changes that have moved for always the guidelines of scientific communication.

All these changes that characterize the knowledge society can lead one to recall Plato's "myth of the cave". It is possible to imagine a parallel link between the scientists who do not incorporate these new guidelines of scientific communication and the inhabitants of the mythical Greek cave. The cave dwellers were convinced that the shadow projected inside the cave was the reality. They were wrong. The reality is that of scientists that live outside of the cave and dominate the useful tools of communication that the world of today offers them. However, the one that remains inside the cave sees a only piece of the present reality and thus runs the serious risk of thinking that what he or she sees is the whole truth. Also, publishers who maintain printed journals and those that impose high subscription costs, have the most part of society living within the cave, thus making difficult the equitable access to high quality scientific data and the possibilities for science to benefit all of society.

3 The Traditional System of Academic Journals and the Challenges of the Digital Environment

Key changes must occur in order to make scientific knowledge more accessible, visible, and usable. More editors and publishers should commit themselves to the requirements of the overall society, which claims for innovations that depend in the scientific information.

Moreover, governments should assume a more proactive and strategic role in addressing key international issues regarding the importance of science for society, supporting an efficient communication system. Although the digital era offers a unique opportunity to cope with these goals, the number of Open Access journals indexed in the ISI Web of Science is still low, representing less than 3% of the total number of journals published by this database.

In spite of the well known and unique advantages provided by the electronic journal format in comparison with the print version, such as increased visibility, accessibility to all issues, lower costs of edition, use of hypermedia, the adoption of only electronic journals still poses a challenge to the editor [7, 8].

Some unsolved issues regarding e-publications, for example electronic archiving and uncertainty about future access, generate significant concerns, skepticism, and distrust in the scholarly community. It has taken some time for only e-journals to become integrated into scientific information systems, indexed by major services, appear in library catalogs, and cited by other researchers in main stream journals.

As a result only e-journals covered by the ISI Thomson Scientific database have low impact factors affecting the prestige of these journals. Also it is worth noting that authors tend to stick to traditional formats and do not make use of the advantages offered by electronic media when writing manuscripts for e-journals. The publication of videos, audio, and three-dimensional images between others are all examples of such advantages. In summary, the distrust of the scientific community in this new media directly affects not only the prestige of these journals but also the possibility to have an accessible communication system that satisfies the needs of the scientific community.

4 The ISI Thomson Scientific and Impact Factors

The Institute for Scientific Information, ISI, was founded by Eugene Garfield in 1960. Then, in 1992, it was acquired by the Thomson Scientific & Healthcare, thus changing the name to Thomson ISI. It is now a sector of the Thomson Corporation referred to as Thomson Scientific [9].

Recognized by the widespread scientific community as an authority for evaluating journals, it covers the world's leading journals of science and technology. Thomson Scientific, or ISI, offers bibliographic database services, covering thousands of academic journals in all scientific disciplines, social sciences, and arts and humanities that consistently achieve and maintain high quality standards in their editorial processes. The ISI Web of Science includes the Science Citation Index (SCI) with 6,623 journals [10], the Social Sciences Citation Index (SSCI) with 1,962 journals [11], and the Arts and Humanities Citation Index (AHCI) with 1,158 journals [12], all of which are available online through the Web of Science database, a part of the Web of Knowledge database collection.

While the evaluation process is independent of the journal’s business model, it depends exclusively on quality standards that are independent of the journal’s format, whether it be print or electronic [13].

The ISI Thomson Scientific writes: “E-Journals undergo the same rigorous selection as journals in print media. Publishing Standards, Editorial Content, International Diversity, and Citation Analysis are all considered”. This gives clear evidence that both paper and electronic formats are equally reliable and genuinely able to communicate science.

Thomson Scientific also publishes an annual Journal Citation Reports, which lists an impact factor for each of the journals of the SCI and SSCI. This is a quantitative tool, which measures the frequency of citation of an “average article” from a journal in other publications covered by a citation index within a two year period previous to its publication. The impact factor is calculated based on a three-year period, and can be considered to be the average number of times published papers are cited up to two years after publication [14]. For example, the 2007 impact factor for a journal A, which is known in the following year, is calculated as follows:

X: 2007 cites in ISI journals to articles published in 2006-2005 by journal A
Y: total number of articles published in 2006-2005 by journal A

Impact Factor 2007 = $\mathbf{\frac{X}{Y}}$

Although traditional journals have attained high impact factors, 49.794 being the highest record in 2006, electronic journals in general rank in the lower half of journals in their subject category. Table 1 shows the highest impact factors of e-only journals ranking among the top 12%.

Journal	Impact factor	Open Acess	ISI subject category	Highest IF of the category	Lowest IF of the category	No. journals of the category
PLos Biology	14.672	yes	- Biochemistry & Molecular Biology	33.456	0.097	261
PLos Medicine	8.389	yes	- Medicine General & Nternal	44.106	0.067	105
Genome Biology	9.712	no	- Biotechnology & Applied Microbiology	22.738	0.024	139
BMC Developmental Biology	5.41	yes	- Developmental Biology	23.69	0.66	33
BMC Structural Biology	5.00	yes	- Biophysics	16.175	0.169	65

BMC Bioinformatics	4.96	yes	- Biochemical Research Methods - Biotechnology & Applied Microbiology	9.876 22.738	0.404 0.024	53 139
Physiological Genomics	4.636	no	- Biochemistry & Molecular Biology - Cell Biology - Physiology	33.456 29.852 28.721	0.097 0.207 0.082	261 153 75
BMC Molecular Biology	4.49	yes	- Biochemistry & Molecular Biology	33.456	0.097	261
BMC Evolutionary Biology	4.45	yes	- Evolutionary Biology - Genetics & Heredity	14.864 25.797	0.675 0.08	33 124
Pediatrics2	4.272	no	- Pediatrics	4.272	0.208	73

Table 1: Ranking of impact factors of the top 12% only e-journals [15, 16]

The information provided in the table is self explanatory. An effort must be done by editors and publishers of only e-journals in order to make this media more reliable and useful to communicate science and therefore to achieve higher impact factors and to locate journals in the upper half of their subject category.

5 The Latin American Context

The journals of Latin America have a low representation in the international databases as in the ISI Web of Science, where 44 journals are covered by the Science Citation Index. These represent a 0.43% of the total of journals on a worldwide basis included in this database. The ranking the impact factors of the Latin American journals is between 0.078 and 3.234, with an average value of 0.442.

It is worth to mention as Table 2 shows, that a high percentage of these publications are Open Access.

	Number	**Percentage %**
Open Access	39	89
Non Open Access	5	11
Total	44	100

Table 2: Comparison between Open Access and non Open Access Latin American journals covered by the Science Citation Index

The ISI criterion to identify Open Access journals is that they are available in full text for the data bases DOAJ , J-Stage and SciELO.

The SciELO (Scientific Electronic Library Online) project is an initiative by FASESP (Foundation of Support to the Research in the State of Sao Paulo) and by BIREME (Latin American and Caribbean Centre with Information in Health Sciences) who is headquartered in Brazil. It includes a selected collection of scientific articles in full-text from Latin American scientific publications. Thanks to this project, the selected Latin-American journals are publishing their articles in the electronic format, remaining freely available in the SciELO website and thus acquiring the character of Open Access.

6 The Case of the Electronic Journal of Biotechnology

The Electronic Journal of Biotechnology is an Open Access, scientific international peer-reviewed journal which has gained a position in the international scene as the only Latin American journal edited exclusively in the electronic format that belongs to the 1% core of only e-journals covered by the ISI Web of Science. It has an impact factor of 0.725, over the average of the impact factors of journals in Latin America and is positioned number 6 in ranking of impact factors in the 44 Latin American journals covered by the ISI Science Citation Index.

It was created in 1998 by the Pontificia Universidad Católica de Valparaíso, Chile with the declared purpose of servicing the international scientific community to make information more accessible, searchable, relevant, and usable. It supports the principles of equal opportunities and freedom of access to scientific information, making the full contents of all articles permanently accessible and searchable for anyone. Therefore, it satisfies the demands of Open Access initiatives. Also, no charge is required for publication and articles are published under the Creative Commons Public License, where no restrictions apply on subsequent redistribution, allowing unlimited use, distribution, and reproduction in any medium, provided the original work is properly cited. Moreover, the provision of CD-ROMs with the Electronic Journal of Biotechnology website to UNESCO and the subsequent distribution to least developing countries allows for a shortening of the digital divide between countries with and without internet facilities, as the CDs also contains the browser internet explorer.

We have an outstanding international academic editorial board, composed of 72 members from 21 countries with Dr James D. Watson (Nobel Prize Laureate) as the Honorary Member of the board.

The journal covers a broad scope of topics in biotechnology, from molecular biology and the chemistry of biological processes, to policy, educational, and ethical issues related directly to this topic. It publishes review and research original articles, short communications and technical notes after submission to full and strict peer review, engaging a geographically broad group of well-recognized scientists as evaluators. Manuscripts are handled electronically, which drastically reduces the time of publication and accepted articles are published in HTML and PDF formats.

In order to maximize its visibility, the journal is located on two servers, one in the Northern hemisphere (http://www.ejbiotechnology.info) and the other in the Southern hemisphere (http://www.ejbiotechnology.cl) receiving in March 2007 more than 110,000 visits and over 1 million hits per month. Also, the use of CrossRef, a citation-linking network, allows the connection of cited references with full text papers while enhancing visibility and accessibility. The knowledge and skills developed during our 10 years of publication can be summarized in seven commitments:

6.1 Commitment to Internationality

The editorial board is international, conformed by 72 members, 34% from North America, 33% from Latin America, 28% from Western Europe, 3% from Near East, 1% from Pacific, and 1% from Asia [17] (see Fig. 1).

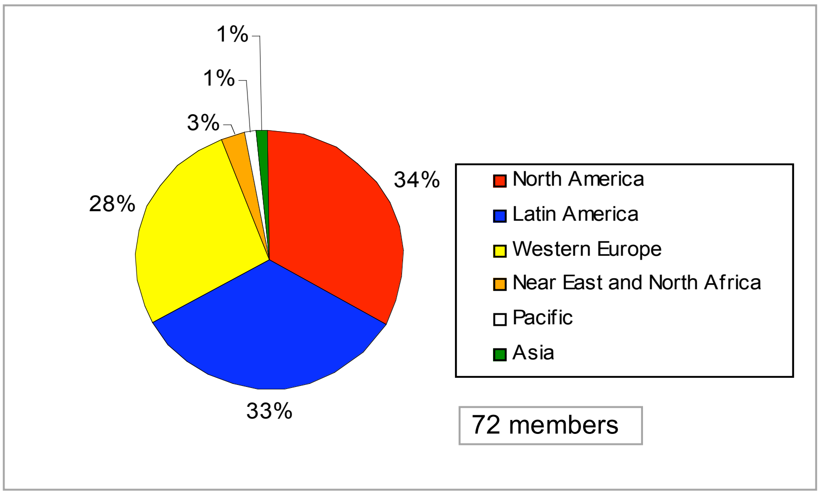


Figure 1: Editorial board internationality

Also, the internationality applies to authors (see Fig. 2) and reviewers, as they come from nearly each region in the world.

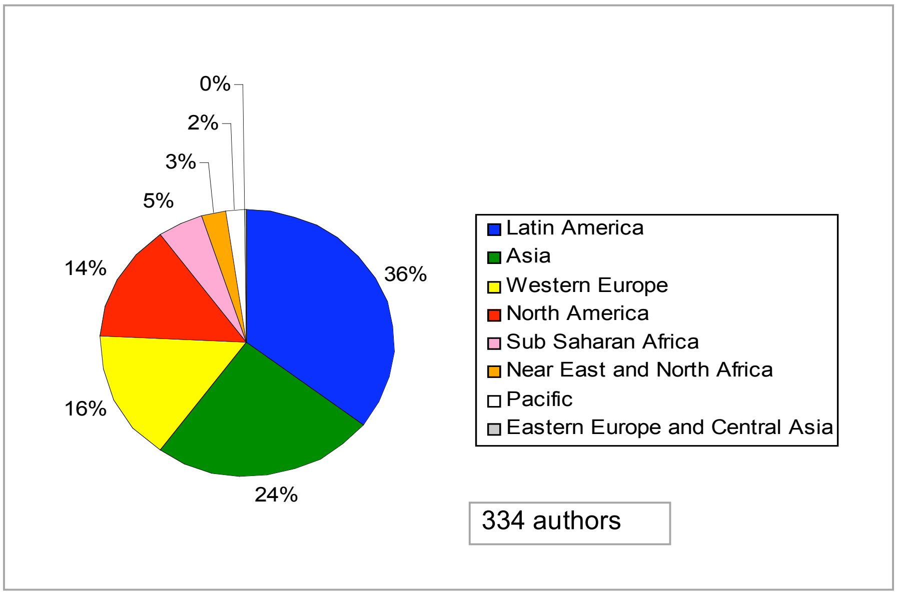


Figure 2: Corresponding authors internationality

A statistics software included in the server shows that readers also hail from different regions, with the USA ranking first as the most active country with visitors on both servers. India, UK, Malaysia, Singapore, Canada, Italy, Germany, Australia and Chile follow for the server located in the Northern hemisphere. The activity of the website located in the Southern hemisphere shows that visitors are mainly from Mexico, Chile, Colombia, Spain, Argentina, Brazil, Peru, Germany and India.

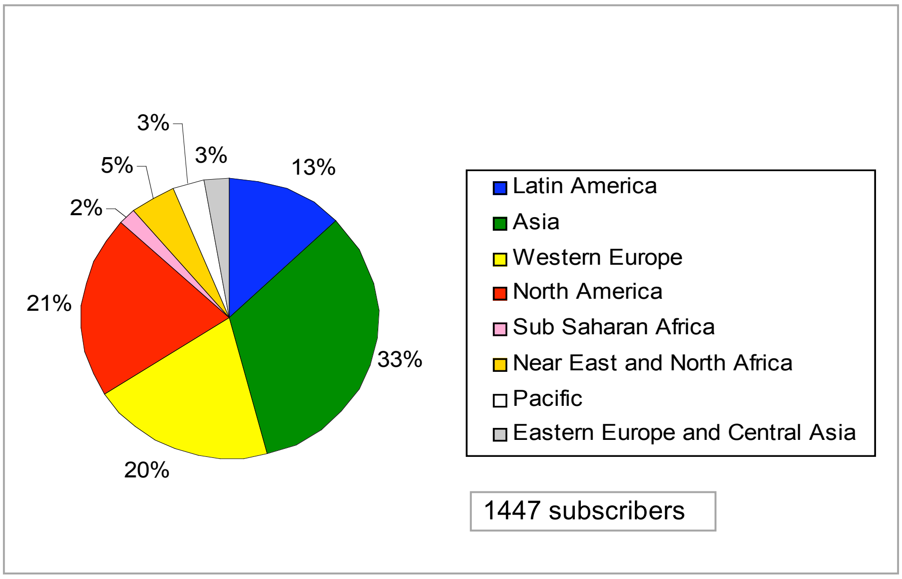


Figure 3: Subscribers internationality

Figure 3 shows the international diversity of the subscribers to an email alerting service of the Electronic Journal of Biotechnology.

6.2 Commitment to Quality

The journal follows the high standards of scientific publications recommended by the ISI Thomson Scientific. Editorial board members are selected by their publishing records taking into account where their articles have been published and if the manuscripts have been cited. A prominent Honorary Member and a well-recognized editorial board are among the best guarantees for the scientific quality of published articles and are indicative of a reliable media of communication.

The Electronic Journal of Biotechnology follows international editorial conventions, for example, the informative journal title, abstracts, full address information for every author and keywords between others. Also the journal is strictly published according to its stated frequency, 4 times a year, in order to comply with guidelines of publication, which is an important standard criteria for quality.

Complete bibliographic information for all cited references is essential and authors are required that at least 75% of the cited bibliography must be from the last decade while at the same time from ISI indexed journals.

6.3 Commitment to Academic Rigor in the Peer Review Process

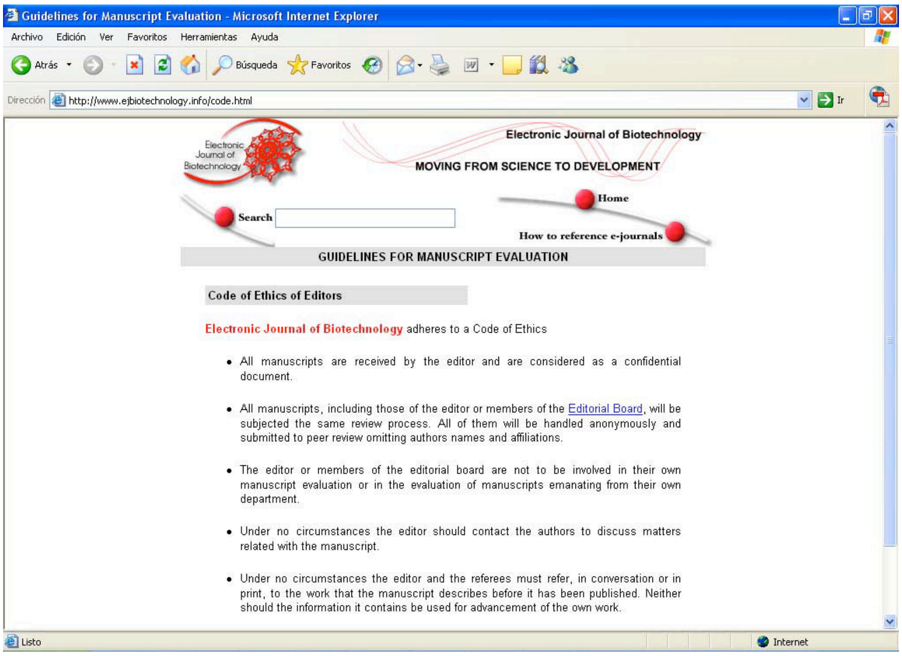
We follow an independent, international and blind peer review process. Evaluators are selected by their expertise from international bibliographical databases and the success of this system is demonstrated not only by the high quality of revision performed on each manuscript, but also because several reviewers have subsequently submitted their manuscripts to the Electronic Journal of Biotechnology in order to be considered for publication. It is worth mentioning, that the refusal of manuscripts has been increasing with time, reaching at present over 70% of rejection (see Table 3).

Published articles	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**TOTAL**	**Percentage %**
Research	3	7	3	13	18	13	21	26	69	23	196	59
Review	17	7	11	3	2	5	5	2	3	1	56	17
Short communications			5	3	5	6	2	4	7	5	37	11
Educational Resources	0	0	0	0	0	0	0	0	0	1	1	0
Biotechnology issues Developing Countries				8	8	5	4		3	1	29	9
Issues in Biotechnology Teaching					5	1	2	3	0		11	3
Letter to editor					1		2		1		4	1
Total published	**20**	**14**	**19**	**27**	**39**	**30**	**36**	**35**	**83**	**31**	**334**	**100**
Rejected articles												
Research	4	2	3	9	9	15	53	74	79	80	328	72
Review		1	4	1	1	12	4	7	7	1	38	8
Short communications			2	1	9	9	3	6	16	33	79	17
Educational Resources									1	1	2	
Biotechnology issues Developing Countries						1	3	3	1	1	9	2
Issues in Biotechnology Teaching										0		0
Minireview								1	0	0	1	0
Total rejected	**4**	**3**	**9**	**11**	**19**	**37**	**63**	**91**	**104**	**116**	**457**	**100**

Table 3: Comparison of received, published and rejected articles

6.4 Commitment to Transparency

Instructions to authors, the composition of the editorial board and the statistics of the website are all easily visible and accessible from the homepage of the journal. The items considered in the evaluation process are also transparent to the authors, the originality of the work being of utmost importance. Furthermore, a code of ethics is also visible to the visitors (see Fig. 4).


Guidelines for Manuscript Evaluation - Microsoft Internet Explorer
Archivo Edición Ver Favoritos Herramientas Ayuda
Atrás
Búsqueda
Favoritos
Dirección http://www.ejbiotechnology.info/code.html
Ir
Electronic Journal of Biotechnology
Electronic Journal of Biotechnology
MOVING FROM SCIENCE TO DEVELOPMENT
Home
Search
How to reference e-journals
Listo
Internet

GUIDELINES FOR MANUSCRIPT EVALUATION

Code of Ethics of Editors

Electronic Journal of Biotechnology adheres to a Code of Ethics

- All manuscripts are received by the editor and are considered as a confidential document.
- All manuscripts, including those of the editor or members of the Editorial Board, will be subjected the same review process. All of them will be handled anonymously and submitted to peer review omitting authors names and affiliations.
- The editor or members of the editorial board are not to be involved in their own manuscript evaluation or in the evaluation of manuscripts emanating from their own department.
- Under no circumstances the editor should contact the authors to discuss matters related with the manuscript.
- Under no circumstances the editor and the referees must refer, in conversation or in print, to the work that the manuscript describes before it has been published. Neither should the information it contains be used for advancement of the own work.

Figure 4: Code of ethics of Electronic Journal of Biotechnology

6.5 Commitment to Scientists

The editor is obliged to scientists, and must acknowledge within three working days the reception of a manuscript. Also, the editor has to respond to the requirements of every author and reader, independent of their academic position and geographic location. The commitment to the scientific community is also demonstrated by the support for an Open Access journal with Open Access licenses that clearly facilitate the retrieval of manuscripts.

6.6 Commitment to Innovation

Electronic Journal of Biotechnology provides a good graphical user interface which enhances the usability of the website. This is based on the scientists' requirements of speed and efficiency which are necessary for the identification and retrieval of articles and documents of interest. Also, the use of searchable descriptive metadata greatly increases the accessibility of the journal to search engines. As for example, if the term "journal biotechnology" is searched in Google, one the first documents to be retrieved is the Electronic Journal of Biotechnology.

Also we have adopted the DOI system [18], which provides a persistent and unequivocal identification of each article. It allows the use of CrossRef, a citation linking system that permits a researcher to click on a cited reference and link directly to that reference on the publisher's platform, subject to the publisher's requirements regarding the access to information [19].

6.7 Commitment to Cooperation

Electronic Journal of Biotechnology welcomes cooperation with any group interested in communicating scientific results in the area of biotechnology. In this way, we have interacted with UNESCO, Bioline International, REDBIO/FAO Co-operation Network on Plant Biotechnology for Latin America and the Caribbean.

In summary, Open Access electronic journals offers a unique opportunity to fulfil the increasingly public demand for making scientific information more accessible, visible and usable. Scientific knowledge must be made public, as it is a right of education and essential to human development. The problem of the distrust in electronic communication must be overcome by the inclusion of more e-journals in international scientific information systems. The ISI Thomson Scientific database publishing company has ensured that both paper and electronic formats are equally trustworthy and legitimate to communicate science.

Acknowledgements

The authors wish to acknowledge the contribution of the team of Electronic Journal of Biotechnology. Also the authors are grateful to Mrs. Marcela Aguirre, Departamento de Información, Comisión Nacional de Investigación Científica y Tecnológica (CONICYT) Chile.

References

[1] 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. Brussels, COM(2007) 56 final. {SEC(2007)181}. February 14, 2007. Available from <http://ec.europa.eu/research/science-society/document_library/pdf_06/communication-022007_en.pdf>.[cited March 30 2007].

[2] TENOPIR, C. Lessons for the future of journals. *Nature***,** October 2001, vol. 413, no. 6857, p. 672-674.

[3] TENOPIR, C.; KING, DW. The use and value of scientific journals: past, present and future. *Serials*, 2001, vol. 14, p. 113-120.

[4] ICSU <http://www.icsu.org/index.php> [cited March 30, 2007].

[5] Budapest Open Access Initiative <http://www.soros.org/openaccess/read.shtml> [cited March 30, 2007].

[6] LIESEGANG, TJ.; SCHACHAT, AP.; ALBERT, DM. The Open Access initiative in scientific and biomedical publishing: fourth in the series on editorship. *American Journal of Ophthalmology*, January 2005, vol. 139, no. 1, p. 156-167.

[7] TENOPIR, C.; KING, DW. Reading behaviour and electronic journals. *Learned Publishing*, 2002, vol. 15, p. 259-265.

[8] ROWAN, L. Editorial Electronic paperless scientific communication, are we ready? *Electronic Journal of Biotechnology*, April 2003. [cited March 30 2007]. Available from <http://www.ejbiotechnology.info/content/vol6/issue1/editorial.html>.

[9] Thomson Scientific. <http://scientific.thomson.com/index.html> [cited March 30, 2007].

[10] Science Citation Index Expanded(™) (*Web of Science*) <http://www.thomsonscientific.com/cgi-bin/jrnlst/jloptions.cgi?PC=D> [cited March 30, 2007].

[11] Social Sciences Citation Index® *(Web of Science)* <http://www.thomsonscientific.com/cgi-bin/jrnlst/jloptions.cgi?PC=J> [cited March 30, 2007].

[12] Arts & Humanities Citation Index® (*Web of Science*) <http://www.thomsonscientific.com/cgi-bin/jrnlst/jloptions.cgi?PC=H> [cited March 30, 2007].

[13] The Thomson Scientific Journal Selection Process. <http://scientific.thomson.com/free/essays/selectionofmaterial/journalselection/> [cited March 30, 2007].

[14] The ISI impact factor. <http://scientific.thomson.com/free/essays/journalcitationreports/impactfactor/> [cited March 30, 2007].

[15] Journal Citation Reports 2005 Thomson Scientific.

[16] Directory of Open Access Journals <http://www.doaj.org> [cited March 30, 2007].

[17] Classification of regions according to Science and Engineering indicators 2002. National Science Foundation

[18] The DOI System <http://www.doi.org/> [cited March 30, 2007].

[19] CrossRef <http://www.crossref.org/> [cited March 30, 2007].

Open Access Journals: A Pathway to Scientific Information in Iran

Alireza Noruzi

University of Tehran, Dep. of Library and Information Science, Tehran, Iran
e-mail: nouruzi@gmail.com

Abstract

This paper reviews the movement of open access (OA) journals in Iran, investigates and compares the influence of Iranian journals in terms of citation ranking, using the Citation Indexes of Thomson-ISI. There has been growth in the number of open access journals in Iran. The advantages of open access for Iranian researchers are: (i) provides access to other research done in their research fields; (ii) speeds up scholarly communication and scientific dialog between researchers; (iii) provides greater visibility and possibly greater impact, although only if open access to the full text is provided. Authors' experiences and motivations have a vital and key role to play in open access. This study indicates that for linguistic reasons, Iranian (Persian-language) journals may not receive and attract the attention that they deserve from the international scientific community. Since there has been little or no discussion in the literature on the impact that the increasing use of OA journals has on scientific production and academic institutions in developing countries, this case study of Iranian experience should be useful for developing countries.

Keywords: open access; scientific journals; Persian-language; Iran

1 Introduction

The traditional model of scholarly publishing (i.e., publication through peer-reviewed journals) and the new information and communication technologies (i.e., the Internet and the Web) have converged to publish scientific open access (OA) journals, which are freely available to those who want to read, download and print them. Open access has removed many access barriers to the scholarly literature, sharing the knowledge of developed countries with developing countries and vice versa, accelerating research and enriching education. In this new strategy, researchers generally publish the results of their research in scholarly open access journals without payment. Open access can increase the internationality, readership, visibility and Impact Factor (IF) of a journal.

Open access means making the full text of an article available online to all users free of charge, immediately and permanently. It has been defined as "free availability of [scholarly literature] on the public Internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of 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 [1]. Stevan Harnad [2] argues that "open access is free, immediate, permanent online access to the full text of research articles for anyone, webwide." However, there remains one constraint linked to copyright: authors control the integrity of their work and must be properly acknowledged and cited. An open access journal is defined here as full text available toll-free on the Web.

The open access movement is a global effort to provide electronic free access to scholarly literature, especially peer-reviewed journals. Open access to the scholarly literature means removing access barriers and limits (e.g., subscription fees, limited access, and so on) to scientific work. There are two parallel "routes" towards open access: open access journals and self-archiving. “Scholarly articles can be made *freely* available to potential readers in one of two main ways – by being published in an open access journal or by being deposited in an electronic repository which is searchable from remote locations without restrictions on access” [3].

OA journals make published articles immediately freely available on their web site, a model mainly funded by charges paid by the author (usually through a research grant). The alternative for a researcher is “self-archiving” (i.e., to publish in a traditional journal, where only subscribers have immediate access, but to make the article available on their personal and/or institutional web sites [including so-called repositories or archives]), which is a practice allowed by many scholarly journals [4].

Open access journals allow, potentially, a greater number of people to access materials compared with subscription-based journals only and in turn helps to solve the research access/impact problem - where restricted access results in the loss of potential research impact [5, 6]. MacCallum and Parthasarathy argue that papers freely available in an open access journal will be more often read and cited than those behind a subscription barrier [7]. A study by Eysenbach [4] reveals that self-archived articles are also cited less often than OA articles from the same journal.

In Iran, open access journals are coming of age, and in a relatively short time have become a mature medium for reporting the results of research. The convenience of open access journals makes them an attractive information resource for Iranian readers and they are increasingly becoming accepted as credible sources of scholarly information.

Iranian open access journals began to proliferate as the Web emerged. For example, 20 journals published in either English or Persian by *Tehran University of Medical Sciences* are open access. Iran is making an electronic version of its English and Persian journals, originally published in print format, freely available to the world. Providing open access to journals is consistent with their strategic goal of enhancing the global visibility of their research output through increasing readership, attracting more contributors, and expanding the influence of Iranian authors in general. The use of open access journals in other fields such as education, science, health, culture, art and development is also maximizing research access through publishing peer-reviewed articles. The Iranian community can now gain access to the results of research by participating in an open access model of research dissemination and individual researchers achieve increased impact typically measured by the number of times a paper is cited and Iranian science gains recognition.

Ensuring that the results of research supported by public funds are made accessible and available for consultation by the research community and others is an integral part of the research process. It involves a partnership between all players involved (universities and other employers of researchers, funders, publishers, libraries, as well as researchers themselves). Ideas and knowledge derived from publicly-funded research are made available and accessible for public use, interrogation, and scrutiny, as widely, rapidly and effectively as practicable [8].

In Iran, open access has already improved the productivity, quality and effectiveness of scientific output, facilitating scholarly communication between Iranian researchers and their foreign colleagues as well as increasing the impact factor (IF) of scientific work based on their citations. Iranian scholarly literature is vital to national productivity and well-being. Publicly-funded research undertaken in Iranian universities and research centers lies at the heart of a productive economy, as well as supporting the social, cultural and physical health of the Iranian nation. Therefore, open access is a valuable vehicle to promote the scientific productivity of Iran. As Brody says: "increased *access* generates the increased *impact*" [9]. The purpose of this paper is to examine the state of open access scholarly journals in Iran and to analyze their visibility through citations to Iranian OA journals in Thomson-ISI Citation Indexes.

2 Literature Review

Previous studies demonstrate that open access articles are more immediately recognized and cited than non open access articles, although it depends on the field of science. The effect of OA publishing may be even higher in fields where journals are not widely available on the Web and where articles from the control group remain *toll-access* [4]. Open access increases impact factor [4, 7, 10, 11], that is, authors who make their peer-reviewed articles open access are cited more than those whose full texts are available only on a subscription-basis from the same refereed venue. It is expected that the growth and use of OA will increase as awareness spreads among authors that OA increases visibility, resulting in more citations and therefore leading to greater impact [6].

Eysenbach [4] shows that articles published as an immediate OA article on the journal *(PNAS: Proceedings of the National Academy of Sciences)* site have higher impact than self-archived or otherwise openly accessible OA articles. It is also shown that OA authors are cited more often per paper. He found strong evidence that, even in a journal that is widely available in research libraries, OA articles are more immediately recognized and cited by peers than non-OA articles published in the same journal. He deduces that OA is likely to benefit science by accelerating dissemination and uptake of research findings and suggests that OA journals facilitate knowledge dissemination to a greater degree than self-archiving, presumably because few scientists search on Google for articles if they have encountered an access problem on the journal web site.

Any scientific open access journal's success depends on authors choosing to submit their research to it for publication. Authors publish research in order for the value of their findings to be recognized. The kudos granted by a solid publication record is crucial for a scientific career. If a journal had a reputation for publishing poor science, it would not receive submissions. Thus the system is inherently self-correcting [12]. However, Ghane reports that a large proportion of randomly selected faculty members, as authors, are not familiar with the concept of open access. Thus, the attitudes and experiences of authors, as owners of the copyright of articles, who have published work in open access journals, play an important role in promoting the idea of open access [13].

Antelman studies the impact of freely available articles in different disciplines (philosophy, political science, electrical / electronic engineering and mathematics). The data of the study show a significant difference in the mean citation rates of open access articles and those that are not freely available online in all four disciplines. The relative increase in citations for open access articles ranged from a low of 45 percent in philosophy to 51 percent in electrical and electronic engineering, 86 percent in political science, and 91 percent in mathematics [14].

Thomson-ISI recently conducted a study of the overall performance of OA journals, using a selection of OA journals in the field of natural sciences and focusing on determining whether OA journals perform differently from other journals in their respective fields. The study's initial findings indicate that there was no discernible difference in terms of citation impact or frequency with which the (open access) journal is cited [15]. On the other hand, Lawrence, investigating the impact of free online articles citation rates in the field of computer science, reported that there is a clear correlation between the number of times an article is cited and the probability that the article is online. More highly cited articles, and more recent articles, are significantly more likely to be open access [11].

The impact factor of journals continues to attract a lot of attention, especially from journal editors, publishers, authors and librarians. Librarians may use the ISI impact factor as one element in selection and de-selection procedures; scientists may be interested in journals with high impact factors in order to reach the highest possible visibility for their published results; funding agencies may consider the impact factors of the journals in which researchers given a grant publish funded research; and university research councils may use journal impact factors as indices in local evaluation studies [16].

3 Research Questions

This research seeks to answer the following questions:

- What constitutes a successful open access journal and how can we ascertain and measure such success?
- What is the role of the authors?
- How is certification of an open access journal related to success?
- What incentives and assistance are needed?

4 Materials and Methods

The approach used in this study includes the following steps:

- First, we conducted a search on Google and Iranian directories of scholarly journals to find open access journals, see *Iranian Directory of Open Access Journals* [17];
- Second, to determine citation rates and Citation Impact[1][18], *Web of Science* (Thomson-ISI citation index) was searched on April 10, 2007, for all Iranian open access journals.

5 Results

It is noteworthy that there are 960 Iranian (either Persian or English language) print-based journals and magazines out of which 247 journals (i.e., 28 English and 175 Persian) are accredited by the Iranian *Ministry of Science, Research and Technology* (MSRT, [19]), and 113 journals (i.e., 23 English and 90 Persian) in the fields of medicine, health, nursing, dentistry, pharmacy, podiatry, and biomedicine are accredited by the *Ministry of Health and Medical Education* (MOHME, [20]). Almost all of the English-language journals accredited by *MSRT* and *MOHME* are now open access or *back access* (back-issue or back-volume open access) (see *Iranian*

[1] The Citation Impact is the ratio of the total number of citations received to the total of citable items published in a journal. Citation Impact can be used as a measure of the *impact* an article has had within its particular field [18].

Directory of Open Access Journals). It should be noted that *Thompson-ISI* citation indexes index only 15 English-language journals from Iran. The current study includes only OA journals published in English.

Table 1 shows the total number of citations to Iranian English-language OA journals, either the ministries accredited or not. The total number of citations (with or without self-citations) is a reliable indicator of scholarly impact and influence [21].

Journal title	Total No. of Citations in WoS	No. of Citations since OA began
Iranian Polymer Journal	304	304
Iranian Journal of Chemistry & Chemical Engineering	183	29
Iranian Journal of Public Health	152	24
Journal of Sciences (Islamic Republic of Iran)	150	31
Iranian Journal of Medical Sciences	119	77
Archives of Iranian Medicine	80	80
Acta Medica Iranica	76	10
Iranian Journal of Pharmaceutical Research	47	42
DARU	40	32
Journal of the Earth and Space Physics	38	0
Journal of the Iranian Chemical Society	35	35
Iranian Biomedical Journal	24	21
Journal of Agricultural Science and Technology	21	21
Iranian International Journal of Science	14	14
Journal of the Iranian Statistical Society	13	0
Iranian Journal of Biotechnology	11	11
International Journal of Endocrinology and Metabolism	10	10
Iranian Heart Journal	10	2
Iranian Journal of Pharmacology and Therapeutics	9	9
Webology	8	8
International Journal of Environment Science and Technology	7	7
Shiraz E-Medical Journal	6	6
Journal of Research in Medical Sciences	6	4
Iranian Journal of Allergy, Asthma and Immunology	6	3
Iranian Journal of Radiation Research	5	5
Iranian Journal of Immunology	4	4
Iranian Journal of Pharmaceutical Sciences	4	4
Iranian Journal of Reproductive Medicine	4	4
Iranian Journal of Veterinary Research	4	0
Journal of Medical Education	4	4
Iranian Journal of Pediatrics	3	1
Iranian Journal of Clinical Infectious Diseases	2	2
Iranian Journal of Radiology	2	2
Iranian Journal of Mathematical Sciences and Informatics	1	1
Journal of Dentistry of Tehran University of Medical Sciences	1	1
Advanced Research Yields across Atherosclerosis	0	0
Caspian Journal of Environmental Sciences	0	0
Hepatitis Monthly	0	0
International Journal of Hematology- Oncology and Bone Marrow Transplantation	0	0
Iranian Journal of Environmental Health Science & Engineering	0	0
Iranian Journal of Parasitology	0	0
Iranian Journal of Pathology	0	0
Iranian Rehabilitation Journal	0	0
Journal of Tehran Heart Center	0	0
Journal of Respiratory Disease, Thoracic Surgery, Intensive Care and Tuberculosis	0	0
Urology Journal	0	0

Table 1: Total Number of Citations to Iranian English-Language OA Journals

Table 1 is a ranked list of the English-language OA journals included in the study, although ranking by total citations obviously favors older and more famous journals. The last column shows the number of citations since OA began. It should be noted that Iranian English-language journals, published by well-known universities, are still in their infancy and need more time to be recognized by their peers and the international scientific community. It seems that one of the main reasons why Iranian journals are not widely cited is that they are not indexed and circulated by foreign databases, especially American and British databases (e.g., Medline, CAB, EBSCO, Proquest, ERIC, Web of Science, WorldCat, LISA, INSPEC, Agris, COMPENDEX, etc.). Therefore, not only open access but also wide circulation is important for a journal's acceptance and reputation.

Table 2 comprises a sample of Persian-language OA journals (including English-language abstracts), nationally well-known, for comparison with the English-language journals.

Journal title	Total No. of Citations in WoS	No. of Citations since OA began
Iranian Journal of Diabetes & Lipid Disorders	6	6
Iranian Journal of Nuclear Medicine	2	1
Audiology	0	0
HAYAT	0	0
Journal of Dental Medicine	0	0
Scientific Journal of School of Public Health and Institute of Public Health Research	0	0
Tehran University Medical Journal	0	0

Table 2: Total Number of Citations to Persian-Language OA Journals

The comparison between non-English-language and English-language open access journals from Iran shows that English-language journals are more cited. Examination of citations to Persian-language OA journals from English-language journals shows that they are infrequent and only cited by Persian-speaking authors. Therefore, it can be concluded that English-speaking authors do not cite Persian-language journals. It should also be noted that Thomson-ISI citation indexes have a bias towards the English-language, indexing few non-English-language journals.

6 Suggestions for Improvement

Iranian institutions should implement a policy:

- to encourage their researchers to publish their research papers in open access journals where a suitable journal exists (and provide the support to enable that to happen);
- to launch new open access journals, where necessary, to serve individual communities, and should support existing journals who want to make the transition to open access;
- to establish open access repositories for English and Persian papers written by Iranians;
- to require Iranian researchers to deposit a copy of all their published papers in a national open access repository; and
- to submit OA journals for inclusion in a large number of indexing and abstracting databases to be widely circulated and widely read.

Creating and developing digital repositories that will assist Iranian academic organizations in the ongoing process of curating –identifying, selecting, acquiring, managing, describing, and providing access to– their scientific collections is vital if the community is to successfully ensure the preservation and continuing access of electronic resources. It is recommended that the Iranian government and authors consider the following suggestions:

- The Iranian government should provide funds for all universities to launch open access institutional repositories, appointing a central body to launch a national repository for preservation and to coordinate the implementation of a network of institutional repositories;
- Iranian universities should launch and support open access journals, and encourage faculty members to take action in support of open access;
- All Iranian universities should establish institutional repositories as an important first step toward more radical change;

- Authors of articles based on government funded research should deposit articles in their institutional repositories, after publication;
- Authors should self-archive (deposit electronic articles into electronic archives);
- Iranian universities should call on all university faculties to self-archive a digital copy of every article accepted by a peer-reviewed journal into the institutional repository.

7 Discussion and Conclusion

To sum up, 'open access' to Iranian scholarly literature is the key element for Iran, improving and accelerating the scientific activities. The Internet makes it possible for Iranian research papers to be read more easily and therefore probably get cited more, because of free, unrestricted access to open access journals. Research institutions that support open access will benefit greatly in terms of impact and influence, due to the greater accessibility and visibility of their research. Iranian researchers should absolutely have the right to see the results of the research that their taxes have paid for.

Some Iranian journals (English or Persian language) currently offer delayed free access, or *back access*, making issues of journals free six months or a year after journal publication. It is worth noting that in fast-moving topics, such information may be out of date when the readers gain access, thus providing *back access* rather than open access. The overall costs of providing open access to scholarly journals are far lower than the costs of traditional print journals, therefore we suggest that Iranian journals, especially international English-language journals become OA, because it is not possible for a print journal to be circulated throughout the world.

Ensuring that the main outputs of research –knowledge and ideas- are disseminated widely is vitally important. Iranian universities should support moves by the research community and scholarly journal publishers to develop new publishing models that are based on the principle that research outcomes should be freely accessed and disseminated as widely as possible via the Internet. It should be noted that OA by itself does not guarantee greater impact and influence for an OA journal, except if the journal publicizes and circulates its contents as widely as possible via international discussion groups, listservs and databases.

Briefly, the advantages of open access for Iranian researchers are: (i) provides access to other research done in their research fields; (ii) speeds up scholarly communication and scientific dialog between researchers; (iii) provides greater visibility and possibly greater impact, although only if open access to the full text is provided.

Acknowledgments

The author wishes to thank Mrs. Marjorie Sweetko for her helpful comments in producing the paper.

References

[1] CHAN, L., et al. Budapest Open Access Initiative, (2002, February 14). Available at: http://www.soros.org/openaccess/read.shtml

[2] HARNAD, S. *American-Scientist-Open-Access-Forum*. Re Proposed update of BOAI definition of OA Immediate and Permanent, 2005. Available at: http://www.ecs.soton.ac.uk/~harnad/Hypermail/Amsci/4420.html

[3] SWAN, A.; BROWN, S. *JISC/Open Society Institute Journal Authors Survey*, 2004. Available at: http://www.jisc.ac.uk/uploaded_documents/JISCOAreport1.pdf

[4] EYSENBACH, G. Citation Advantage of Open Access Articles. *PLoS Biology*, 4(5) (2006), e157. Available at: http://biology.plosjournals.org

[5] HARNAD, S. *et al.* The green and gold roads to open access. *Nature*, 2004. Available at: http://www.nature.com/nature/focus/accessdebate/21.html

[6] HARDY, R.; OPPENHEIM, C.; BRODY, T.; HITCHCOCK, S. Open Access Citation Information. Final Report – Extended Version, JISC Scholarly Communications Group, September 2005. Available at: http://eprints.ecs.soton.ac.uk/11536/

[7] MACCALLUM, C.J., & Parthasarathy, H. Open Access Increases Citation Rate. *PLoS Biology*, 4(5) (2006), e176. Available at: http://biology.plosjournals.org

[8] Research Councils UK. Rcuk Position Statement on Access to Research Outputs, 2005, June. Available at: http://www.rcuk.ac.uk/access/statement.pdf

[9] BRODY, T. Citation Analysis in the Open Access World. Author eprint, 2004. Available at: http://eprints.ecs.soton.ac.uk/10000/01/tim_oa.pdf

[10] HITCHCOCK, S. The effect of open access and downloads ('hits') on citation impact: A bibliography of studies, 2005. Available at: http://opcit.eprints.org/oacitation-biblio.html

[11] Lawrence, S. Online or Invisible? *Nature*, 411(6837) (2001), p. 521.

[12] WEITZMAN, J.B. (Mis)Leading Open Access Myths, 2006. Available at: http://www.biomedcentral.com/openaccess/inquiry/myths/

[13] GHANE, M. A Survey of Open Access Barriers to Scientific Information: Providing an Appropriate Pattern for Scientific Communication in Iran. *The Grey Journal: An International Journal on Grey Literature*, 2(1), 2006.

[14] ANTELMAN, K. Do open-access articles have a greater research impact? *College & Research Libraries*, 65(5) (2004), 372-382. Available at: http://eprints.rclis.org/archive/00002309/

[15] Thomson-ISI. *The Impact of Open Access Journals: A Citation Study from Thomson ISI*, 2004.Available at: http://www.isinet.com/media/presentrep/acropdf/impact-oa-journals.pdf

[16] ROUSSEAU, R. Impact of African Journals in ISI Databases. *LIBRES: Library and Information Science Research Electronic Journal*, 15(2), (2002). Available at: http://libres.curtin.edu.au/libres15n2/index.htm

[17] Iranian Directory of Open Access Journals, 2007. Available at: http://nouruzi.googlepages.com/IDOAJ.doc

[18] BRODY, T. *et al.* The effect of open access on citation impact. *National Policies on Open Access (OA) Provision for University Research Output: An International meeting*. Southampton University, Southampton, UK, 19 February 2004. Available at: http://opcit.eprints.org/feb19oa/brody-impact.pdf

[19] MSRT. Ministry of Science, Research and Technology, 2006. Available at: http://www.msrt.gov.ir/

[20] MOHME. Ministry of Health and Medical Education, 2006. Available at: http://www.net.hbi.ir/new/dynamic/journals/journal-index.php

[21] CRONIN, B.; MEHO, L. Using the *h*-index to rank influential information scientists. *Journal of the American Society for Information Science and Technology*, 57(9) (2006), 1275–1278.

Automatic Sentiment Analysis in On-line Text

Erik Boiy; Pieter Hens; Koen Deschacht; Marie-Francine Moens

Katholieke Universiteit Leuven, Tiensestraat 41 B-3000 Leuven, Belgium
e-mail: erik.boiy@law.kuleuven.be; pieter.hens@econ.kuleuven.be
koen.deschacht@law.kuleuven.be; marie-france.moens@law.kuleuven.be

Abstract

The growing stream of content placed on the Web provides a huge collection of textual resources. People share their experiences on-line, ventilate their opinions (and frustrations), or simply talk just about anything. The large amount of available data creates opportunities for automatic mining and analysis. The information we are interested in this paper, is how people feel about certain topics. We consider it as a classification task: their feelings can be positive, negative or neutral. A sentiment isn't always stated in a clear way in the text; it is often represented in subtle, complex ways. Besides direct expression of the user's feelings towards a certain topic, he or she can use a diverse range of other techniques to express his or her emotions. On top of that, authors may mix objective and subjective information about a topic, or write down thoughts about other topics than the one we are investigating. Lastly, the data gathered from the World Wide Web often contains a lot of noise. All of this makes the task of automatic recognition of the sentiment in on-line text more difficult. We will give an overview of various techniques used to tackle the problems in the domain of sentiment analysis, and add some of our own results.

Keywords: sentiment analysis; document classification; artificial intelligence

1 Introduction

Automatic sentiment analysis is a topic within information extraction that only recently received interest from the academic community. In the previous decade, a handful of articles have been published on this subject. It's only in the last five years that we've seen a small explosion of publications. The idea of automatic sentiment analysis is important for marketing research, where companies wish to find out what the world thinks of their product; for monitoring newsgroups and forums, where fast and automatic detection of flaming is necessary; for analysis of customer feedback; or as informative augmentation for search engines.

The automatic analysis of sentiments on data found on the Web is useful for any company or institution caring about quality control. For the moment, getting user feedback means bothering him or her with surveys on every aspect the company is interested in. The problems with this approach are making a survey for each product or feature; the format, distribution and timing of the survey (asking to send a form right after purchase might not be very informative); and the reliance on the goodwill of people to take the survey. This method can be made obsolete by gathering such information automatically from the World Wide Web, where the large amount of available data creates the opportunity to do so. One of the sources are blogs (short for "web logs"), a medium through which the blog owner makes commentaries about a certain subject or talks about his or her personal experiences, inviting readers to provide their own comments. Another source are the electronic discussion boards, where people can discuss all kinds of topics, or ask for other people's opinions. We define a topic as the subject matter of a conversation or discussion, e.g. an event in the media or a new model of car, towards which the writer can express his or her views.

There are several additional advantages to this approach. First, the people who share their views usually have more pronounced opinions than average, which are additionally influencing others reading them, leading to so-called word-of-mouth marketing. Extracting these opinions is thus extra valuable. Second, opinions are extracted in real-time, allowing for quicker response times to market changes and for detailed time-based statistics that make it possible to plot trends over time.

This paper is organized as follows: In section 2 we will go over the concepts of emotions in written text. Section 3 gives an overview of various methods that can be used to analyse the sentiment of a text, making a distinction between symbolic techniques and machine learning approaches. In section 4 we describe some challenges in the field that need to be overcome. Section 5 provides a comparison of results from the literature using the aforementioned techniques, to which we add some of our own results1. In section 6 we shortly discuss those results, before coming to conclusions in section 7.

2 Concepts of Emotions in Written Text

2.1 Concept of Emotions

Before attempting to classify sentiments, we must ask the question what sentiments are. In general we can state that sentiments are either emotions, or they are judgements or ideas prompted or coloured by emotions[2]. An emotion consists of a set of stages, namely: appraisal, neural and chemical changes and action readiness. We will give a quick overview of each of these states.

An emotion is usually caused by a person consciously or unconsciously evaluating an event, which is denoted *appraisal* in psychology. Appraisal does not only denote the evaluation whether something is positive or negative, but it also denotes other measurements such as the significance of an event, the personal control or the involvement of the own ego. In general, the same appraisal gives rise to the same emotion. Appraisal causes *mental and bodily changes*, that make up the actual experience of an emotion. Emotions urge for actions and prompt for plans: an emotion gives priority for one or a few kind of *actions* to which it gives a sense of urgency. We use the term "action" to denote all mental or physical actions (that are the result of an emotion). This includes actions such as moving away from a negative event, mental processes, such as worrying about the event, and other effects that are direct result of the emotion, such as crying or going pale.

2.2 Emotions in Written Text

The study of emotions in text can be conducted from two points of view. Firstly, one can investigate how emotions influence a writer of a text in choosing certain words and/or other linguistic elements. Secondly, one can investigate how a reader interprets the emotion in a text, and what linguistic clues are used to infer the emotion of the writer. In this text, we'll take the second point of view. We are interested in the way people infer emotions, so we can mimic this process in a computer program. In the remainder of this section we will investigate how linguistic elements describing appraisal and action-readiness are used in texts to convey the emotion of the author, as they comprise the majority of clues to infer emotion from text.

Appraisal

A lot of linguistic scholars agree on the three dimensions of Osgood and al. [1], who investigated how the meaning of words can be mapped in a semantic space. Factor analysis extracted 3 major dimensions: (1) positive or negative evaluation (2) a power, control or potency dimension and (3) an activity, arousal or intensity dimension. Although these dimensions are originally proposed as the dimensions of a semantic space, they can also be used to organize linguistic categories of emotion or for the automatic detection of emotions. Most research is devoted towards the appraisal component of emotions, and we will look into it a bit deeper by briefly going over Osgood's dimensions, giving some examples along the way.

(1) Evaluation (positive/negative)
The evaluation dimension is fairly straightforward; it contains all choices of words, parts of speech, word organization patterns, conversational techniques, and discourse strategies that express the orientation of the writer to the current topic. Evaluation is often expressed by using adjectives.
e.g. “It was an *amazing* show.”

(2) Potency (powerful/unpowerful)
This dimension contains all elements that general express whether the writer identifies and commits himself towards the meaning of the sentence or whether he dissociates himself. From a psychological standpoint these phenomena are related to approach and avoidance behaviour. This dimension consists of 3 sub-dimensions: proximity, specificity and certainty.

(2.1) Proximity (near/far)

[1] This research is sponsored by the company Attentio (Brussels, Belgium): attentiocontact@attentio.com. We are very grateful to Per Siljubergsåsen and Simon Mc Dermott for the funding and the relevant remarks. We thank Kalina Lipinska for the annotation of one of the corpora we use in our experiments. We thank Roxana Angheluta for her valuable help when integrating our work in Attentio’s system architecture.

[2] Adapted from the Merriam-Webster On-line Search dictionary.

This category contains all linguistic elements that indicate the 'distance' between the writer and the topic. The proximity from the writer to the current topic expresses whether the writer identifies himself with the topic or distance himself from it.
e.g. “I'd like you to meet John.” versus “I'd like you to meet Mr. Adams.” (social proximity)

(2.2) Specificity (clear/vague)
Specificity is the extent to which a conceptualized object is referred to by name in a direct, clear way; or is only implied, suggested, alluded to, generalized, or otherwise hinted at.
e.g. “I left *my* / *a* book in your office.” (particular vs general reference)

(2.3) Certainty (confident/doubtful)
This dimension expresses the certainty of the writer towards the expressed content. A stronger certainty indicates that the writer is entirely convinced about the truth of his writings and possibly indicates a stronger emotion.
e.g. “It *supposedly* is a great movie.” versus “It *definitely* is a great movie.”

(3) Intensifiers (more/less)
When expressing emotions, a lot of the emotional words used do not express an emotion, but modify the strength of the expressed emotion. These words, the intensifiers, can be used to strengthen or weaken both positive and negative emotions.
e.g. “This is *simply* the best movie.” (adverb)
“He had cuts *all* over.” (quantifier)
“Where *the hell* have you been?” (swearing)

Direct Expressions

The most direct way to express an emotion is of course to express it directly, without making a detour by using appraisal or action readiness. This can be done among others by using verbs and adjectives [2, 3]. A typical way to express an emotion directly seems to be a pattern similar to "I am/feel/seem [adjective describing emotion]"
e.g. I *ache for* a cigarette.
I *am delighted* of the final results.

Elements of Action

Excellent examples of actions indicating emotions are of course crying and laughing, but more subtle signs that denote emotion in certain circumstances can be considered as well. An example is looking at your watch when watching a movie, which is most probably a result of boredom and a lack of interest.
e.g. I was *grinning* the whole way through it and *laughing out loud* more than once.

Remarks

There are additional ways of expressing emotions that don't strictly fall into above categories, like the use of figurative language and irony. It must also be noted that most techniques in sentiment classification focus on terms that do actually not really denote emotions, but denote evaluation, appreciation or judgement. Of course this is not surprising, because most techniques focus on reviews of movies, products, cars, etc., and basically in a review the reviewer evaluates the object under discussion. The sentiment of the reviewer is often not discussed, although of course, it is often easy to infer his emotions. Recognizing the fact that classifying a review is in essence classifying it according to appraisal, doesn't only improve understanding but can also lead to the discovery of new techniques.

3 Methodology

In the previous section we discussed the indicators of sentiment in text. In this section we will see methods of identifying this information in a written text. There are two main techniques for sentiment classification: symbolic techniques and machine learning techniques. The symbolic approach uses manually crafted rules and lexicons, where the machine learning approach uses unsupervised, weakly supervised or fully supervised learning to construct a model from a large training corpus.

3.1 Symbolic Techniques

3.1.1 Lexicon Based Techniques

The simplest representation of a text is the bag-of-words approach. Hereby, we simply consider the document as a collection of words without considering any of the relations between the individual words. Next, we determine the sentiment of every word and combine these values with some aggregation function (such as average or sum). We will discuss a selection of methods to determine the sentiment of a single word.

3.1.1.1 Using a Web Search

It was already indicated by Hatzivassiloglou and Wiebe [4] that adjectives are good indicators of subjective, evaluative sentences. Turney [5] recognizes that, although an isolated adjective may indicate subjectivity, there may be insufficient context to determine semantic orientation. For example, the adjective "unpredictable" may have a negative orientation in an automotive review, in a phrase such as "unpredictable steering", but it could have a positive orientation in a movie review, in a phrase such as “unpredictable plot". Therefore he used tuples consisting of adjectives combined with nouns and of adverbs combined with verbs.

The tuples are extracted from the reviews and the semantic orientation of a review is calculated as the average semantic orientation of the tuples taken from that review. To calculate the semantic orientation for a tuple (such as "unpredictable steering"), Turney uses the search engine Altavista. For every combination, he issues two queries: one query that returns the number of documents that contain the tuple close (defined as “within 10 words distance”) to the word "excellent" and one query that returns the number of documents that contain the tuple close to the word "poor". If the combination is found more often in the same context as "excellent" than in the same context as "poor", the combination is considered to indicate a positive orientation, and otherwise to indicate a negative orientation.

3.1.1.2 Using WordNet

Kamps and Marx use WordNet [6] to determine the orientation of a word. In fact, they go beyond the simple positive/negative orientation, and use the dimension of appraisal that gives a more fine-grained description of the emotional content of a word. Kamps and Marx developed an automatic method [7] using the lexical database WordNet to determine the emotional content of a word along Osgood et al.'s dimensions. In essence, the WordNet database consists of nodes (the words) connected by edges (synonym relations). Kamps and Marx define a distance metric between the words in WordNet, called minimum path-length (MPL). This distance metric counts the number of edges of the shortest path between the two nodes that represent the words. For example, the words "good" and "big" have a MPL of 3. The shortest path from the word "good" to the word "big" is the sequence <good, sound, heavy, big>.

To estimate the magnitude of a dimension of appraisal for a particular word, they compare the MPL of that word towards the positive and towards the negative end of that dimension. Both ends of a dimension are represented by prototype-words. The positive end of the evaluative dimension is represented by the word "good" and the negative end is represented by the word "bad". The prototypes for the potency dimension are respectively "strong" and "weak" and for the activity dimension "active" and "passive".

Only a subset of the words in WordNet can be evaluated using this techniques, because not all words are connected to one of the prototype words. After examination, it showed that the subset of words connected to either "good" or "bad" was composed of 5410 words. Interestingly, the subset of words connected to either "strong" or "weak" consisted of exactly the same 5410 words, and so did the subset connected to "active" or "passive". It seems that all important words expressing emotive or affective meaning are included in this one set.

3.1.2 Sentiment of Sentences

So far, we've seen different methods that determine the sentiment of a single word and assumed a simple approach to combine the sentiments of words within a single sentence. The bag-of-words approach has some important drawbacks. As already briefly indicated in section 3.1.1.1, it can often be advantageous to consider some relations between the words in a sentence. There are several approaches in this field; we mention here briefly Mulder and al.'s article [8], which discusses the successful use of an affective grammar. They note that simply detecting emotion words can tell whether a sentence is positive or negative oriented, but does not explain towards what topic this sentiment is directed. In other words, what is lacking in the research towards affect is the relation between attitude and object. Mulder and al. have studied how this relation between attitude and object can be formalized. They combined a lexical and grammatical approach: (1) lexical, because they believe that affect is primarily expressed through affect words, and (2) grammatical, because affective meaning is intensified and propagated towards a target through function-words and grammatical constructs.

3.2 Machine Learning Techniques

In this section a description and comparison of state-of-the-art machine learning techniques used for sentiment classification are discussed. First a description is given of a selection of different features that are commonly used to represent a document for the classification task, followed by an overview of machine learning algorithms.

3.2.1 Feature Selection

The most important decision to make when classifying documents, is the choice of the feature set. Several features are commonly used, like unigrams or part-of-speech (the linguistic category of a word, further shortened to “POS”) data. Features and their values are commonly stored in a feature vector.

Unigrams
This is the classic approach to feature selection, in which each document is represented as a feature vector, where the elements indicate the presence (or frequency) of a word in the document. In other words, the document is represented by its keywords.

N-grams
A word N-gram is a subsequence of N words from a given sequence (e.g. a sentence). This means that the features in the document representation are not single words, but pairs (bigrams), triples (trigrams) or even bigger tuples of words. For example, “easy" followed by “to" becomes “easy to" in a bigram. Other examples of positive oriented bigrams are: “the best”, “I love”, “the great”, ... and negative oriented: “not worth”, “back to”, “returned it”, ... [9]. With the use of N-grams it is possible to capture more context. N-grams are for example effective features for word sense disambiguation [10]. When using N-grams, the feature vector could take on enormous proportions (in turn increasing sparsity the of the feature vectors). Limiting the feature vector size can be done by setting a threshold for the frequency of the N-grams, or by defining rule sets (e.g. only incorporate N-grams that satisfy a certain pattern like *Adjective Noun* or *Adverb Verb*).

Lemmas
Instead of using the words as they literally occur in the text, the lemmas of these words can be used as features for the document. This means that for each word its lemma, being its basic dictionary form, is identified. Examples are:

writes -> write *was -> be* *better -> good*
written -> write *cars -> car* *best -> good*

The advantage with lemmatisation is that the features are generalized and it will be easier to classify new documents, but this is not always true: you still have to look out for overgeneralization. Dave et al. [9] report a decrease in accuracy of sentiment classification when the words in the documents are conflated to their dictionary form. Lemmatisation comes with loss of detail in the language. For example, Dave notes that negative reviews tend to occur more in the past tense, which cannot be detected after lemmatisation.

Negation
Another extension of the unigram approach is the use of negation. When you only consider the words in a sentence and someone writes “*I don't like this movie”,* a program can think that this person loved the movie, when it looks at the word ”like". A solution for this is to tag each word after the negation until the first punctuation (with for example NOT_). The previous sentence will then become: “*I don't NOT_like NOT_this NOT_movie”*. This was done by [11]. In this experiment, the negation tagging gives a slight decrease in performance. Dave et al. [9] note that simple substrings (N-grams) work better at capturing negation phrases.

Opinion Words
Opinion words are words that people use to express a positive or negative opinion [12]. Opinion words are obtained from several POS classes: adjectives, adverbs, verbs and nouns [13, 14]. These opinion words can be

incorporated into the feature vector, where they represent the presence or absence of such a word. Two techniques can be used to define opinion words:

- Use a predefined lexicon; Wiebe and Riloff [14] constructed such an opinion word-list. This approach combines the lexicon based method described above with the machine learning methods.
- Identify the words (mostly adjectives; see below) that describe a certain feature of a product in a text [12]. e.g. After nearly 800 pictures I have found that this camera takes *incredible* pictures.

Adjectives

Wiebe noted in [15] that adjectives are good indicators for subjectivity in a document. According to these findings you can assume that documents only represented by their adjectives should do well in sentiment classification. Experiments where only adjective features are used, were done in [11, 16]. The results showed that you get better results when using all POS data. This doesn't mean that adjectives are bad sentiment classifiers, as adjectives only represent on average 7.5% of the text in a document.

Salvetti used WordNet to enrich the only-adjective feature vectors. He translated the adjectives into synsets of adjectives and used hypernym generalization on them (both synsets and hypernyms can be found using WordNet). Using this procedure he found a decrease in the accuracy of the sentiment classification, which was due to the loss of information produced by the generalization.

3.2.2 Machine Learning Techniques

Supervised Methods

In order to train a classifier for sentiment recognition in text, classic supervised learning techniques (e.g. Support Vector Machines, naive Bayes Multinomial, Maximum Entropy) can be used. A supervised approach entails the use of a labelled training corpus to learn a certain classification function. The method that in the literature often yields the highest accuracy regards a Support Vector Machine classifier [11]. In the following section we discuss a selection of classification algorithms. They are the ones we used in our experiments described below.

(1) Support Vector Machines (SVM)

Support Vector Machines operate by constructing a hyperplane with maximal Euclidean distance to the closest training examples. This can be seen as the distance between the separating hyperplane and two parallel hyperplanes at each side, representing the boundary of the examples of one class in the feature space. It is assumed that the best generalization of the classifier is obtained when this distance is maximal. If the data is not separable, a hyperplane will be chosen that splits the data with the least error possible.

(2) Naive Bayes Multinomial (NBM)

A naive Bayes classifier uses Bayes rule (which states how to update or revise believes in the light of new evidence) as its main equation, under the naive assumption of conditional independence: each individual feature is assumed to be an indication of the assigned class, independent of each other. A multinomial naive Bayes classifier constructs a model by fitting a distribution of the number of occurrences of each feature for all the documents.

(3) Maximum Entropy (Maxent)

The approach tries to preserve as much uncertainty as possible. A number of models are computed, where each feature corresponds to a constraint on the model. The model with the maximum entropy over all models that satisfy these constraints is selected for classification. This way no assumptions are made that are not justified by the empirical evidence available.

Unsupervised and Weakly-supervised Methods

The above techniques all require a labelled corpus to learn the classifiers. This is not always available, and it takes time to label a corpus of significant size. Unsupervised methods can label a corpus, that is later used for supervised learning (especially semantic orientation is helpful for this [17]). Turney's technique using AltaVista (see section 3.1.1.1) can be viewed as a form of weakly supervised learning, where a set of seed terms is expanded to a collection of words. We mention two more methods for determining the sentiment of single words based on weakly-supervised methods. Hatzivassiloglou and McKeown[18] presented a method for determining the sentiment of adjectives by clustering documents into same-oriented parts, and manually label the clusters positive or negative. OPINE [19] is a system that uses term clustering for determining the semantic orientation of an opinion word in combination with other words in a sentence. The idea behind this approach comes from the fact that the orientation of a word can change with respect to the feature or sentence the word is associated (e.g. The word *hot* in the pair: *hot water* has a positive sentiment, but in the pair *hot room* it has a negative sentiment).

4 Challenges

With the techniques described above, pretty good results can be obtained already (see section 5), but nevertheless, there are some challenges that need to be overcome.

4.1 Topic-Sentiment Relation

Our goal is to determine sentiments towards a certain topic. It often happens that a person expresses his opinion towards several topics within the same text or sentence. For example, in a movie review he may state he dislikes the special effects and some of the acting, but likes the movie nonetheless. His opinion about these topics is in contradiction with his thoughts about the movie in general. When a sentence contains a lot of negative subjectivity, but all expressed toward a different topic than the one we are investigating, the sentence is still classified as negative. Therefore, it is useful to investigate the relation of the sentiment to the topic. One way of doing this is by looking into the sentence parse tree (i.e. a syntactic analysis of the sentence according to the language's grammar) to derive better features.

Related to this problem is the classification of whole texts. Until now we have only looked at the classification of sentences, in which topic and terms indicative for the sentiment are assumed to appear together. This is however not a realistic assumption. For the detection of the topic-sentiment relation in texts, coreference resolution needs to be applied across sentences. Even when there is only one topic in the text, it is also advantageous to use a more advanced metric to combine the predictions for the sentences than a simple sum of the sentiments found in the individual sentences. Taboada and Grieve [20] state that opinions expressed in a text tend to be found in the middle and the end of that text. Therefore, they weigh the semantical orientation of a sentence based on its position in the text, giving improved results.

4.2 Neutral Text

A first question is what to do with neutral text, as not all text is either positively or negatively oriented. It is often useful to determine whether a piece of text expresses subjective or objective content. Subjective sentences are used to communicate the speaker's evaluations, opinions, emotions and speculations, while objective sentences are used to convey objective, factual information [21]. Both kinds often appear in the same text, for instance in movie reviews, where the writer can express his attitude toward the movie (which is the semantic orientation of the document), but can also describe, within the same review, objective statements about the movie itself (e.g. a summary of the plot). Most subjectivity classifiers use machine learning techniques (see [22]) and classify between subjective and objective sentences or between positive, negative and objective sentences. To our best knowledge, there has been only one attempt to use a symbolic technique that classifies subjective sentences, done by Wiebe [23, 24].

4.3 Cross-domain Classification

Other research in the sentiment classification field regards cross-domain classification. How can we learn classifiers on one domain and use them on another domain (e.g. *books* and *movies*)? A reason why cross-domain classification might be necessary is because there is not always enough training data available to train a classifier for a specific domain. The classifier should then be trained with data from another domain. Tests are done by Aue et al. [25] and by Finn et al. [26]. Overall, they show that sentiment analysis is a very domain-specific problem, and it is hard to create a domain independent classifier. One possible approach is to train the classifier on a domain-mixed set of data instead of training it on one specific domain.

4.4 Text Quality

A last important issue is the quality of the text to be evaluated. When text is automatically gathered from the World Wide Web, one can expect a fair amount of junk to be returned (e.g. adds, web site menus, links, ...). This junk may be mixed with other information we are interested in, making it more difficult to filter it out. Also the language used by the writers may be of poor quality, containing lots of Internet slang and misspellings. Both issues have a negative influence on the classification for both types of methods discussed. Especially the junk may confuse a machine learner by providing it with a lot of irrelevant features. This also means extensive manual filtering of the text in order to acquire a good training corpus, and makes it harder to perform deeper

NLP techniques like parsing. An example of dirty input text (the topic is the movie “A Good Year”) is the following:

Nothing but a French kiss-off Search Recent Archives Web for (rm) else &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; ONLINE EXTRAS SITE SERVICES Movie Listings Friday Nov 10 2006 Posted on Fri Nov. 10 2006 MOVIE REVIEW A Good Year a flat bouquet Nothing but a French kiss-off Gladiator collaborators seem defeated by light-weight love story.By ROBERT W.

Needless to say that using a sentence parser (that detects the syntactic structure of the sentence) on this example will have little success.

5. Results

5.1 Evaluation Measures

As a first evaluation measure we simply take the classification accuracy, meaning the percentage of examples classified correctly. This measure is not sufficient when we classify individual sentences and include the neutral class as a third option next to the positive and negative ones. Neutral examples are a majority in texts and their correct detection largely influences simple accuracy results. With this in mind, it makes more sense for us to use precision for positive and negative examples as the evaluation measure. When generating sentiment statistics a high recall is just as desirable as a high precision. Other evaluation metrics that influence the performance are also considered: the speed of the classification method, the feature vector size and the available resources.

5.2 Symbolic Techniques

In the previous section we have seen a selection of two symbolic techniques, for which we give concrete results. Turney reports accuracies ranging between 65.83% on a collection of movie reviews, to 84.0% on a collection of automobile reviews when applying his method using a web search engine. Kamps and Marx achieved an accuracy of 68.19% on the manually constructed list of the General Inquirer for classification along Osgood's evaluative dimension, when applying their approach using WordNet. The accuracy rose to 76.72% when increasing the interval for which words are considered neutral.

An interesting experiment done by Pang et al. [11] shows the difficulty to construct a lexicon (or another knowledge-resource) that has a (close to) complete coverage of the target domain. A lot of information is often not captured in the hand-built model and lost. In the experiment they compared the ability of humans in selecting appropriate words for an emotion lexicon, with automatic methods. Although the lists of words created by humans seemed intuitively valid, they resulted in poorer performance: the best human created list resulted in 64% accuracy (with 39% ties), while a simple automatic method (a count of the frequencies of words in positive and negative reviews) resulted in a list with 69% accuracy and only 16% ties. Interestingly, some words that have no significant emotional orientation were quite good indexes. For example, the word “still", was found to be a good indicator of positive orientation, because it appeared in sentences such as “Still, though, it was worth seeing".

Given the above results, we did not perform any experiments with symbolic techniques, instead we focused on machine learning techniques of which the results are given below.

5.3 Machine Learning Techniques

5.3.1 Corpora

A corpus is a large, electronically stored set of texts. Corpora are used by machine learning approaches both for training and testing (and just for testing in the case of symbolic approaches). Evaluation will often be performed by using cross-validation. This means that over several iterations, in each iteration part of the corpus will be used for training, and the other part for testing. After all iterations, each example from the corpus will have been used for testing once, resulting in a full evaluation of the corpus. In order to compare results of different approaches, they need to be compared on the same corpus, as some corpora can be considerably easier to work with than others. We performed tests on two corpora to obtain the results presented in this paper:

- Pang and Lee's[3] movie review corpus, consisting of 1000 positive and 1000 negative reviews, is often used to evaluate sentiment analysis approaches in the literature. These movie reviews seem hard to classify. A possible explanation of this phenomenon is the mix of words that describe the storyline and words that describe the evaluation of that particular movie.
- A corpus gathered from blogs, discussion boards and other websites containing 759 positive, 205 negative, 1965 neutral and 1562 junk examples, annotated with a sentiment towards the topic under evaluation. The latter two categories were considered as one for our test purposes. The topics include various movie titles and car brands. The examples are of poor quality, displaying the problems described in section 4.4 (the example given there was taken from this corpus). As the number of examples in each category is very unbalanced, corrective measures were taken by adding additional examples from the Customer Review Datasets corpus by Hu and Liu[4]. In total, 550 negative sentences from the customer reviews were added to the corpus, and 222 extra positive sentences were used for training only.

5.3.2 Our Experiments

In Table 1 we show some of our results on the movie review corpus, indicating the features that perform well in the literature (discussed above), optional processing and the machine learning methods used. For both the support vector machine (SVM) and naive Bayes multinomial (NBM) methods the Weka[5] implementation was used, the Maxent[6] package from OpenNLP was used as implementation of the maximum entropy classifier. For our tests using SVM's, an error tolerance of 0.05 was set for training, the other parameters (e.g. linear kernel) were kept default for all methods. We used QTAG as POS tagger for obtaining the adjectives. It achieves a rather low accuracy[7], but it is fast and easy to incorporate into software. “Subjectivity analysis” stands for a simple subjectivity analysis using a NBM classifier, trained on the subjectivity dataset introduced in Pang and Lee [22], which removes all objective sentences from the examples. A cut-off of four was used for the bigram feature, meaning that only bigrams occurring at least four times were included in the feature vector. Frequencies of the features were used in the feature vector for SVM and NBM, while binary feature presence was used for Maxent.

Features	*SVM*	*NBM*	*Maxent*
Unigrams	85.45%	81.45%	84.80%
Unigrams & subjectivity analysis	86.35%	83.95%	87.40%
Bigrams	85.35%	83.15%	85.40%
Adjectives	75.85%	82.00%	80.30%

Table 1: Results in terms of accuracy on the movie review corpus for different machine learning methods using a selection of features (and processing)

Table 2 shows our results on the second corpus that realistically represents blogs found on the World Wide Web. The corpus was extended with 550 negative review sentences, which are included in the results. In the first column are the baseline results on the corpus. The baseline uses the approach that gives the best results for the movie corpus (see Table 1), i.e., an approach comparable to the literature and with a low novelty factor. In the second column are our latest results. A total of 84 examples were beyond the reach of our current methods and are excluded from the results. In order to include those examples, we could consider them as neutral; resulting in a slight decrease in the total accuracy and in the recall for positive and negative, compared to the results shown in the second column, while still being much better than the baseline results. The methods, features and processing used to arrive to these results may not be disclosed by us. For more information on the methods used, the reader may contact Attentio, the company that sponsors our research.

[3] Available at http://www.cs.cornell.edu/people/pabo/movie-review-data.

[4] Available at http://www.cs.uic.edu/~liub/FBS/FBS.html.

[5] See http://www.cs.waikato.ac.nz/~ml/weka/.

[6] See http://maxent.sourceforge.net/.

[7] Our own experiments indicate an accuracy of about 86%, while current state of the art POS tagging achieve ca. 96% accuracy.

	Baseline NBM	***Our latest approach***
accuracy %	84.25	90.25
precision/recall % for positive	64.52/49.93	74.39/75.62
precision/recall % for negative	88.48/72.96	87.43/82.70

Table 2: Results on the blog corpus, comparing the results of the baseline approach (cf. Table 1) and those of our latest methods

6 Discussion

Although we have not done any experiments using symbolic techniques ourselves, we deemed machine learning approaches more promising after reviewing methods from both categories, and conducted our research in that direction. Judging from the good results we have achieved, this seems like it has been the right choice.

The results in Table 1 show that there is rather little difference in accuracy between the experiments using different features (except for the adjectives). With this in mind, it becomes interesting to look at other factors influencing the choice of which features and processing to use. The advantages of unigrams and bigrams over the other features are that they are faster to extract, and require no extra resources to use, while e.g. adjectives require a POS tagger to be run on the data first, and subjectivity analysis requires an additional classifier to be used. A downside is the feature vector size, which is substantially (over 5 times for unigrams) larger e.g. than when only adjectives are included. For the machine learning method we see a more substantial difference between NBM and both SVM and Maxent. It might however still be advantageous to use NBM, as it is considerably faster. The results of the state of the art techniques for sentiment classification on the movie review corpus shown in Table 1 are comparable with the ones found in the literature that use this corpus.

The results from Table 2 need some more explanation. The blog corpus used in the experiments of Table 2 is considerably more difficult to work with, and is annotated in three classes (including neutral), where the movie review corpus (results in Table 1) only had two. However, compared to the baseline (current state-of-the-art) approach, our latest method performs significantly better. The lower precision and recall for the positive class compared to the negative one, are due to the added negative examples from the easier Customer Review Datasets corpus, and due to the higher correlation of positive examples with neutral ones, making misclassifications between those classes more common. The results we obtained are encouraging, and show that it is possible to overcome the difficulties explained in section 4.

7 Conclusion

In this paper we have indicated the usefulness of sentiment classification, and have given an overview of the various methods used for this task. While many of the methods show encouraging results, there are still challenges to be overcome when applying them to data gathered from the World Wide Web, especially from blogs. We have demonstrated that in these circumstances improvements over state of art methods for sentiment recognition in texts are possible.

References

[1] OSGOOD, C. E.; SUCI, G. J; TANNENBAUM, P. H. *The Measurement of Meaning*. University of Illinois Press, 1971 [1957].

[2] BIBER, D; FINEGAN, E. Styles of stance in english: *Lexical and grammatical marketing of evidentiality and affect.* Text 9, 1989, pp. 93-124.

[3] WALLACE, A. F. C.; CARSON, M. T., *Sharing and diversity in emotion terminology*. Ethos 1 (1), 1973, pp. 1-29.

[4] HATZIVASSILOGLOU, V.; WIEBE, J., *Effects of adjective orientation and gradability on sentence subjectivity*, Proceedings of the 18th International Conference on Computational Linguistics, ACL, New Brunswick, NJ, 2000.

[5] TURNEY, P., *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*, Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 417–424.

[6] FELLBAUM, C. (ed.), *Wordnet: An electronic lexical database*, Language, Speech, and Communication Series, MIT Press, Cambridge, 1998.

[7] KAMPS, J.; MARX, M.; MOKKEN, R. J.; DE RIJKE, M., *Using WordNet to measure semantic orientation of adjectives.* LREC 2004, volume IV, pp. 1115—1118.

[8] MULDER, M.; NIJHOLT, A.; DEN UYL, M.; TERPSTRA, P., *A lexical grammaticaimplementation of affect*, Proceedings of TSD-04, the 7th International Conference Text, Speech and Dialogue, Lecture Notes in Computer Science, vol. 3206, Springer-Verlag, Brno, CZ, 2004, pp. 171–178.

[9] DAVE, K.; LAWRENCE, S.; PENNOCK, D. M. *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews.* In Proceedings of WWW-03, 12th International Conference on the World Wide Web, ACM Press, Budapest, HU, 2003, pp. 519–528.

[10] PEDERSEN, T. *A decision tree of bigrams is an accurate predictor of word sense.* In Proceedings of the Second Annual Meeting of the North American Chapter of the Association for Computational Linguistics, 2001, pp. 79–86.

[11] PANG, B.; LEE, L.; VAITHYANATHAN, S. *Thumbs up? Sentiment classification using machine learning techniques.* In Proceedings of EMNLP-02, the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Philadelphia, US, 2002, pp. 79–86.

[12] HU, M.; LIU, B. *Mining opinion features in customer reviews.* In Proceedings of AAAI-04, the 19th National Conference on Artificial Intellgience, San Jose, US, 2004.

[13] BETHARD, S.; YU, H.; THORNTON, A.; HATZIVASSILOGLOU, V.; JURAFSKY, D. *Automatic extraction of opinion propositions and their holders.* In James G. Shanahan, Janyce Wiebe, and Yan Qu, editors, Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, Stanford, US, 2004.

[14] RILOFF, E.; WIEBE, J.; WILSON, T. *Learning subjective nouns using extraction pattern bootstrapping.* In Walter Daelemans and Miles Osborne, editors, Proceedings of CONLL-03, 7th Conference on Natural Language Learning, Edmonton, CA, 2003, pp. 25–32.

[15] WIEBE, J. *Learning subjective adjectives from corpora.* In Proceedings of AAAI-00, 17th Conference of the American Association for Artificial Intelligence, AAAI Press / The MIT Press, Austin, US, 2000, pp. 735–740.

[16] SALVETTI, F.; LEWIS, S.; REICHENBACH, C. *Impact of lexical filtering on overall opinion polarity identification.* In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, Stanford, US, 2004.

[17] BEINEKE, P.; HASTIE, T.; VAITHYANATHAN, S. *The sentimental factor: Improving review classification via human-provided information.* In Proceedings of ACL-04, the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, ES, 2004, pp. 263–270.

[18] HATZIVASSILOGLOU, V.; MCKEOWN, K. R. *Predicting the semantic orientation of adjectives.* In Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Madrid, ES, 1997, pp. 174–181.

[19] POPESCU, A.; ETZIONI, O. *Extracting product features and opinions from reviews.* In Proceedings of HLT-EMNLP-05, the Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing, Vancouver, CA, 2005.

[20] TABOADA, M.; GRIEVE, J. *Analyzing appraisal automatically.* In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, Stanford, US, 2004, pp. 158–161.

[21] WIEBE, J.; BRUCE, R. F.; O'HARA, T. P. *Development and use of a gold-standard data set for subjectivity classifications.* In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics, College Park,US, 1999, pp. 246–253.

[22] PANG, B.; LEE, L. *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.* In Proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Barcelona, ES, 2004, pp. 271–278.

[23] WIEBE, J. *Recognizing Subjective Sentences: A Computational Investigation of Narrative Text.* Technical report, SUNY Buffalo Dept. Of Computer Science, Buffalo, NY, 1990.

[24] WIEBE, J. *Tracking point of view in narrative.* Computational Linguistics, 20 (2), 1994, pp. 233–287.

[25] AUE, A.; GAMON, M. *Customizing sentiment classifiers to new domains: a case study.* In Submitted to RANLP-05, the International Conference on Recent Advances in Natural Language Processing, Borovets, BG, 2005.

[26] FINN, A.; KUSHMERICK, N. *Learning to classify documents according to genre.* J. American Society for Information Science and Technology, Special issue on Computational Analysis of Style, 57(9), 2006.

A Comparison of the Blogging Practices of UK and US Bloggers

Sarah Pedersen

Department of Communication and Media, The Aberdeen Business School
The Robert Gordon University, Aberdeen, UK
e-mail: s.pedersen@rgu.ac.uk

Abstract

This paper describes the results of an investigation into the differences and similarities between the blogging techniques of UK and US bloggers undertaken in the winter and spring of 2006-7 and funded by the UK Arts and Humanities Research Council. Blogging started in the US, while British bloggers are relative latecomers to the blogosphere. How has this late arrival impacted on the ways in which Britons blog in comparison to US bloggers? A survey was administered to 60 UK and 60 US bloggers and data was also collected directly from their blogs and by means of online tools. A blog was also set up in order to discuss the findings of the research within the blogosphere. Since blogging started in the US, the majority of research into blogging so far has focused on the US and it is suggested that this focus has resulted in all bloggers being defined through the US experience. The findings of this project suggest that bloggers outside the US may have different approaches to blogging and find different satisfactions. It also suggests a new financial motivation for blogging, which had not previously been identified, and which may be an indication of the way in which the blogosphere is evolving.

Keywords: blogs; weblogs; comparitive analysis

1 Introduction

This paper reports on the first attempt to compare and contrast the blogging practices of US and UK bloggers. While there is a growing body of academic research into this new form of computer-mediated communication, the vast majority of such research has so far focused on North American bloggers. Blogging started in the US while British bloggers are relative latecomers to the blogosphere. The paper asks how this later arrival has impacted on the ways in which Britons blog in comparison to US bloggers.

Blogging has joined e-mail and home pages as a mass use of the internet. Blogs are usually defined, following Blood as 'frequently updated, reverse-chronological entries on a single webpage'.[1] The original blogs were filter-type web pages, directing the reader to other blogs and websites on the Internet and offering commentary and often the opportunity for readers' discussion. Blogging took off as a publishing platform, at first mainly in the US and predominantly amongst students or recent graduates (see Schiano et al [2]; Herring et al [3]). The expansion of the blogosphere that we see today, when the blog-tracking website Technorati claims to be tracking over 70 million blogs, occurred after the introduction of cheap and easy-to-use build-your-own-blog software such as Blogger, Pitas and Groksoup in 1999. Unlike early bloggers, who needed advanced programming skills to construct their blogs, it is now possible for anyone with access to the Internet to set up their own weblog. Filter blogs have been joined by so-called 'journal' blogs, which tend to have fewer links, fewer readers and are more like online, public diaries. However, both types of blog conform to the pattern of frequently updated posts arranged in a chronological order, the majority offering the opportunity for readers to post comments. The ability to simply publish a weblog online is ever increasing with community websites such as MySpace adding blogging to their services.

Although the first blogs did not appear until 1997, there has been a remarkably swift growth in academic research into this new form of computer-mediated communication. Early research focused on the categorisation of blogs or bloggers. For example, Krishnamurthy proposed the classification of blogs into four basic types along two dimensions: personal versus topical and individual versus community.[4] Another focus for scholarly research has been the on-going debate about the role of blogging as a form of journalism (for examples of the research on this topic see Singer; Matheson; Kahn and Kellner; Wall; Pedersen and Chivers)[5]. At the other end of the blogosphere, research teams lead by Herring, Schiano and Nardi have investigated journal bloggers, and in particular pointed out that, despite the media's focus on blogs written by white, educated, US males, over 50% of journal blogs are actually written by women and young people.[6] Other research has investigated the dynamics of different communities of bloggers; for example, Huffaker and Calvert have surveyed teenage bloggers[7] while

Mortensen and Walker discussed the way in which academics can use blogs as a research tool.[8] As Thelwall points out, 'It is difficult to summarise the findings of the extremely diverse body of blog research except by pointing to the wide variety of uses of blogs and the fact that blogs do sometimes create genuine online communities'.[9] One thing that does immediately stand out when reading through early research into blogging is the focus on North American bloggers. There has been a very limited amount of research into the second wave of blogging that occurred outside the US, although this is now being addressed, for example, Trammell et al's recent examination of the state of the Polish blogosphere, Tricas-Garcia & Merelo-Guervos' work on the Spanish blogosphere and Abold's discussion of the use of blogs in the 2005 German election campaign.[10] There has been a limited amount of research into the UK blogosphere. Discussion of gender issues within the UK blogosphere has been undertaken by Pedersen & Macafee while Auty has investigated the blogs of UK politicians and Thelwall undertook a descriptive analysis of blog postings around the London bomb attacks of July 2005.[11]

2 Methodology

Following a pilot stage of the project, a survey was administered to 60 UK and 60 US bloggers. The bloggers were selected randomly through two blog directories: Britblog and Globe of Blogs. Both directories offered the possibility of selecting blogs by state or county and so it was possible to ensure that all regions of the UK and US was covered by the survey. The survey was distributed to equal amounts of male and female bloggers. The criteria for selection were that the blogger had to have posted on their blog within a month of the start of the selection process, that the blog was written in English, that the blogger was resident in either the UK or the US and was over the age of 18. Teenage blogging is acknowledged by most researchers to be a very different type of computer-mediated communication from that of adult bloggers, associated generally with use of community sites such as Bebo and MySpace, and therefore it was decided to focus only on bloggers over the age of 18, which also avoided many ethical issues. Data was also collected directly from the survey respondents' blogs and by means of online tools (Technorati, The Truth Laid Bear and SurfWax). Areas investigated included average time spent blogging and when that blogging occurs; the promotion of blogs; attitudes to blogging and issues of privacy and openness related to the use of photographs and other personal material. A measure of success was devised (based on traffic, links and directory rankings). In addition, a blog related to the research was established. This gave the researcher first-hand experience of the challenges of blogging and also offered the opportunity for further data collection since the surveyed bloggers were invited to comment on the research as it was ongoing, an opportunity which they took up with enthusiasm.

3 Results

Demographics

It has already been stated that efforts were made to send the survey to an equal number of men and women. In the final analysis, the respondents were as follows: 32 UK males, 30 UK females, 32 US males and 28 US females. It should be noted that, during the period of research, two of the UK females actually moved to North America. Out of this random group, one of the UK women identified herself as a lesbian, one of the US males as gay and one of the UK men as a transvestite.

Age ranged from 18 to 73. 40% of the UK respondents were under 30, in comparison to 26% of the US respondents. 4% of UK respondents were over 56, with the US figure being 18%. Thus, for this random sample, the US bloggers were on average older than the UK bloggers. Differences between the two countries were also found in terms of educational attainment, with US respondents having achieved higher educational attainment. 47% of US respondents were educated to bachelors degree level, compared to 32% of the UK respondents, and 35% of US respondents held a postgraduate degree, in comparison to 18% of UK respondents. 28% of the UK respondents reported that their highest level of educational attainment was as a school leaver compared to only 10% of US respondents. To a certain extent, this must be linked to the youth of the UK respondents, although it should be noted that 11 UK respondents were currently undertaking education compared to 13 US respondents.

Previous studies into the blogosphere have characterised bloggers as usually educated to graduate level or beyond. However, these studies were based in the US and investigated the first wave of bloggers. Some of these studies even focused on university bloggers through their selection of survey participants. It may be, therefore, that the second wave of blogging outside the US is attracting a different type of person to the blogosphere. Is blogging in the UK more associated with youth culture?

As far as employment is concerned, 105 of the survey respondents stated that they were employed, with most of those not employed being either retired or looking after dependents in the home. The number of those employed on a part-time basis was evenly split across the two countries. It should be noted that the number of women working part time was much greater than the number of men in part-time employment: 40% (21 out of 52 women who stated that they were employed) in comparison to 11% of all employed men (6 out of 53). As will be seen later, this may be related to the higher number of women who were attempting to gain financial reward from their blogging.

Blogging practices

Respondents were asked where they did the majority of their blogging. Only 14% of all respondents indicated that they blogged at a workplace outside the house while 43% blogged at home. Interestingly, another 30% stated that they blogged at home, which was also their place of employment. It is possibly not surprising that few bloggers choose to blog at a workplace outside their home considering the number of high-profile cases of bloggers being sacked or reprimanded for blogging at or about their place of employment. However, it is noteworthy that such a high number of respondents *worked* at home, either fully or partially, and blogged from there. When this was discussed on the blog set up in association with the research, several bloggers suggested that it was only because they worked at home that they had the time to blog:

> *Speaking as a blogger who works from home, I sometimes wonder how a person who doesn't work from home would find significant time to blog. I have the luxury of an extra couple of hours which would otherwise be used as commute time to compose my thoughts, and blog. I also have intervals during the day where I can break from my client projects to respond to reader comments.*

A respondent to the survey agreed that blogging was easier because she worked as a freelance: 'I am self-employed so sometimes it [blogging] will be immediately after something happens in work' although another confessed that easy access to the Internet brought its own problems: 'Being self-employed I can almost do as I please (as long as I get the work done) and there have been times in the past when I've been obsessed with blogging and work would suffer as result.'

Thus respondents were more likely to blog at home than outside the home. This finding is also related to *when* they blogged, with the most popular choices being evenings and breaks in the day. US bloggers were somewhat more inclined to blog in the mornings before they went to work, with 35% of US respondents admitting to this in comparison with 17% of UK bloggers. Again, this was discussed on the blog with one US blogger suggesting: 'Americans are familiar with a fast-paced, over-time-heavy schedule, which means that getting up in the morning to blog is a convenient way to blog each day but still get to work and/or started on the home-based work before the 8am rush.' A report in *The Guardian* of 29 November 2006 of a survey conducted by the European Interactive Advertising Association also supports the finding that Europeans access the Internet later in the day:

> *The survey also looked at when people access the internet. From 6am to 10am the majority of European internet users prefer to listen to the radio or read a paper. But that picture inverts dramatically as the day wears on. From 5.30pm to 9pm, three-quarters of web users are watching TV but almost as many are accessing the internet.*[12]

68% of all respondents admitted blogging for up to 5 hours a week, with another 18% blogging for between 5 and 10 hours a week. One or two respondents were blogging for up to 35 hours a week, although it should be noted that some respondents were what might be called professional bloggers, either setting up blogs for others or using their blogs as part of newspaper columns. A surprising 52 respondents, spread evenly across the countries, admitted to writing more than one blog. The reason usually given for this was to focus on different subjects, although some bloggers kept one blog private, accessible only through the use of a password, while the other was public access. For example, one female blogger kept a blog about her pregnancy private while publicly blogging about food and cooking.

Content of blogs

The respondents' blogs were analysed for content. Unfortunately, only 112 of the blogs were able to be analysed in this way because during the eight-month period of research eight of the blogs were abandoned and removed from the Web. On the basis of the last ten postings made on the blog they were placed in one of the following categories: personal, creative work, criticism, politics and opinion, IT, business and work, religion, chance discoveries and food. It should be noted that it was sometimes difficult to distinguish between the IT and business and work categories since those who blogged about IT were usually also working in IT. Therefore the IT category should be seen as a subset of the business and work category.

51 out of the 112 blogs (46%) were categorised as 'personal'. It should be noted that far more female blogs were categorised as personal than male blogs (15 men and 36 women) and, in particular, only four US male blogs were categorised as personal. In contrast, five male blogs were characterised as being about religion, but none of the female blogs. 12 male blogs were characterised as opinion and politics, but only three female blogs. If work and business and IT are seen as one category, 21 blogs are found here: 5 of these were female blogs and the other 16 were male. 3 blogs (all female) were characterised as being about food.

Blog promotion

Success in the blogosphere is linked to popularity. The more links to a blog, the greater its success rating. Although it must be conceded that not all bloggers are interested in the type of success that comes with membership of the 'A list' (Technorati's list of the 100 most linked blogs), the survey did investigate how far bloggers promoted their blogs to other bloggers in order to encourage a higher readership and more incoming links. As far as promotion of their blog was concerned, the most popular methods used by respondents were to submit their blog to a blog directory or blog search engine (97 respondents) or to post on other blogs (66 respondents). Male US respondents were the most likely to submit an RSS feed to a blog directory or search engine, rather than merely submitting their URL, which perhaps suggests a higher level of technical ability in blogging.

A form of promotion particularly popular amongst the UK bloggers was blogrings. Blogrings connect a circle of blogs with a common theme or purpose. A link to the blogring is displayed on a blog and clicking on that link takes the reader to the blogrings page, where the other members of the blogring are listed. Alternatively, clicking on the link takes the reader directly to the next blog in the ring. UK respondents were more likely to state that they used blogrings to promote their blog. 26 UK respondents (11 men and 14 women) admitted to using blogrings in comparison to 15 US respondents, only five of whom were male. Analysis of respondents' blogs showed a large selection of blogrings with few being named by more than one or two bloggers. The more popular blogrings were either those which linked bloggers of the same sex, such as 'Blogs by Women' or 'Crazy/Hip Blog Mamas', or those which linked geographically similar bloggers. 24 bloggers linked to blogrings related to location, such as 'Blogging Brits', 'Scots Bloggers' or 'Expat Bloggers'. Male bloggers were more likely to belong to a blogring which promoted an interest or hobby, such as blogrings for birdwatchers, Methodists or transvestites, which reflects the male bloggers' preference for issue-based blogging, while female bloggers were more likely to belong to blogrings that celebrated their femininity (16 female bloggers belonged to female-only blogrings), which again reflects the female proclivity for more journal blogging with a focus on themselves.

Survey respondents were asked their opinion of membership of blogrings. While most acknowledged that they could provide more traffic, in terms of readers, to a blog, the opinion of many was that they were not worth joining any more, having been replaced in usefulness by blog directories. One respondent commented: 'I think blog rings can be a little random. I'd rather have a focused directory that points specifically to my site'. Others were concerned that it would be assumed that they would have identical opinions with others in the blogring. Several respondents explained that they had joined blogrings at the start of their blogging, but would not join any more now. The relative popularity of blogrings amongst the British bloggers – and the high number of blogrings related to the UK or regions of the country – is noteworthy in comparison with the lower interest from US bloggers, in particularly US male bloggers, and may point to a desire to mark themselves out as different, or a need to group together, in the face of the much more numerous US bloggers.

Concerns about privacy

56 respondents – just under half – had concerns about privacy. These were divided equally between the two countries. Privacy concerns tended to be about two areas of the bloggers' lives: their family and their work. Respondents reported that they tried not to mention their family on their blog or to make their address identifiable. Those that worried about colleagues or management at work identifying them, which might lead to trouble at work, also mentioned the worry that potential future employers might search for them online in order to assess their suitability for employment.

Are bloggers identifiable through the information they give on their blog? US female respondents were the least likely to state their full name on their blog, with only 14% of surveyed blogs giving this information in comparison to the rest of the surveyed blogs where the figure was around 50%. On average 70% of the blogs did not show an identifiable photograph. However, the US males again seem slightly different to the others. 54% of the US male blogs analysed did show an identifiable photo. From the anecdotal evidence given in the survey responses it seems that bloggers are right to be concerned about being identified through photographs. Several respondents told stories of being identified from the photos on their blog, including one man who was accosted by a complete stranger while walking through the departures lounge of an airport.

Opinions on blogging

Respondents to the survey were asked a number of questions about the way in which they perceived blogging in order to ascertain any differences in attitude between the two countries. Firstly, they were asked whether their blogging had replaced any sort of paper documentation. The most frequently reported replacement by a blog was a diary. 28 respondents agreed that they had replaced their diary by blogging. 17 stated that project journals had been replaced by a blog and 15 that a travel diary had been replaced by blogging.

There were some differences between the sexes to be discerned here. 19 of those who had replaced diaries were women, as opposed to 9 men, whereas 10 men had replaced travel diaries in comparison to 5 women. While this might conform to gender stereotyping to a certain extent, it is interesting to note that 12 women stated that blogging had replaced project journals as opposed to only 5 men.

Respondents were also asked whether they blogged mainly for their own records. Interestingly, UK female respondents answered this question very differently to all other respondents, with 50% declaring that they did blog mainly for their own records. Other respondents were far less likely to respond positively, for example only 4 (13%) of US males agreed that they blogged mainly for their own records.

Respondents were asked whether, in general, they considered blogging to be a form of publishing, journalism, creative writing, diary keeping or other. Overall there was uniformity in many responses, with many respondents selecting all four of the named choices. However, it should be noted that US males were particularly unlikely to see blogging as a form of diary keeping, with only 12 selecting diary keeping in comparison to 21 for UK males and females and 23 for US females.

Respondents were asked how they saw their own blogging activity. They were asked to select as many as necessary from a selection of statements. It is obvious from the results that blogging is seen very much as a leisure activity by respondents on both sides of the pond. 74 selected ‘Leisure time activity’ and another 60 selected ‘A welcome distraction’. The small amount of students amongst the respondents was again demonstrated by the low number who saw blogging as a quick break from, or an adjunct to, studying – only 7. Men were slightly more likely to see blogging as a quick break from work (23 to 10), and this is probably related to the higher numbers of male than female respondents in full-time employment.

What was particularly interesting in the response to this question was the number of respondents who indicated another way of looking at blogging: as a form of income generation. The work of teams led by Schiano and Nardi on the motivations of bloggers suggests that there are five main reasons for blogging. These are: documenting the author’s life; providing commentary and opinions; expressing deeply felt emotions; working out ideas through writing; and forming and maintaining communities and forums. They note that such motivations for blogging are not mutually exclusive. Pedersen’s work on the motivations of women bloggers suggests that another motivation may be the women’s need for validation of their thoughts and actions.[13] However, this survey has brought a further motivation to light: that of financial reward. Among many responses to the question of why respondents blogged along the lines of the motivations outlined by Schiano et al was the introduction of a financial motive. Preliminary findings from this research have suggested that the financial motivation is particularly strong amongst women bloggers, who may be looking for ways in which to generate income as an alternative to full-time employment outside the home. Of the 31 respondents who mentioned a financial motivation in their written responses to the survey, 21 were women, and their responses showed very clearly that they were hoping that their blogging would lead to some sort of financial gain. As one female respondent stated: ‘I hope to eventually make enough money from my blog to support my family, I see it as the beginnings of an online business.’

The ways in which bloggers hoped to make money through their blogging differed. Some bloggers used their blog as a marketing tool for themselves or for their businesses. 24 respondents agreed that their blogging brought custom for their business. For example, one UK female blogger stated: ‘I started the blog as a way of promoting my online business, enhancing online word-of-mouth marketing for my business and developing my brand’. Another, who blogs about parenthood, stated that her blogging had started as a leisure activity but was now opening up serious work opportunities. One respondent, who worked as a children’s book illustrator, reported that she showcased her work and sold associated greetings cards through her blog. Another respondent, who described herself as a courtesan, explained that her blog helped attract suitable clients.

Blogging might also offer direct financial reward as a profession – one UK respondent worked as a freelance blogger, setting up blogs for West End shows and individual actors. An American blogger reported that her blogging ‘started out as a leisure time activity and has become my work. The postings on my blog are the same as the reviews that now appear in my syndicated column of movie reviews, which appear in various newspapers across the Northeast, thanks to a deal made with a company that saw the work on my blog and hired me to be their critic.’ Interestingly enough, her blog was one those ranked as least popular by this research (see below), indicating that unpopularity online by no means translates into lack of success elsewhere.

Blogs can also make money through carrying advertising or requesting subscriptions. One of the most famous bloggers on the web is Heather B Armstrong, the author of the blog Dooce.com, who reportedly supports her entire family through the advertising that her blog carries. While none of the respondents to this survey mentioned such large financial earnings, a UK male respondent’s blog carries a section offering the possibility of running a banner advertisement at the top of his blog for a month with the guarantee that no other advertising will be accepted during this time. He charges £200 for this privilege. As well as carrying advertisements on their blogs, bloggers might also earn money through ‘pay-per-post’ advertising where bloggers write about certain products or services in their blogs in return for payment. Bloggers might even hope for income through the paper publication of their entire blog. Blogs which have been successfully published as books include *Belle de Jour: Intimate Adventures of a London Call Girl* or Tom Reynolds’ *Blood, Sweat and Tea: Real Life Adventures in an inner-city ambulance* (taken from his blog ‘Random Acts of Reality’). Recent press coverage in the UK has focused on the £70,000 book deal given to ex-*Sunday Times* education correspondent Judith O’Reilly for her blog *Wife in the North*. Several respondents to the survey mentioned hopes that their blogging would attract potential publishers: ‘I have aspirations to write a book about the food industry and I believe that writing the blog is a tool to (1) exercise my writing muscles and developing a voice; (2) distinguishing or creating a unique vice; (3) offer me opportunities for credibility and to be viewed as a subject matter expert.’ In fact, one male respondent from the UK reported that his blogging had helped clinch a book publishing contract for a book on his subject specialism.

In contrast, a few respondents reported that blogging had actually lost them money. One US male felt that the strongly held views of the government policy he discussed on his blog had lead to loss of work from the defence industry. He also considered that blogging was a threat to his career as a journalist: ‘Sadly I feel my work abilities (writer, reporter, photojournalist) are going to go the way of the dinosaur. Note the rise in cheap digital stock imagery, blogging, "citizen" journalist submissions to network and cable TV, websites, publications, etc. (all for free, mind you).’

There were some differences between UK and US bloggers when discussing the gains – financial or otherwise – to be found in blogging. US bloggers were far more willing to acknowledge that they found blogging ‘useful’. 31 US bloggers stated that they found blogging useful because it widenened the audience for their intellectual work, in comparison with 14 UK respondents, and 44 US respondents felt that it widened the audience for their creative work, in comparison with 26 respondents. UK respondents were far more likely to respond that blogging had no use. One stated: ‘It’s not the pretentious thing you seem to think it is. It’s sharing. It’s putting yourself out there. Not for recognition or to “help” people, although that might happen on occasion. It’s not there for me to make people like or respect me. It’s just me, warts and all. No other agenda.’

Blogrolls

A blogroll is a collection of links to other blogs and is seen as a list of recommended reading. The majority (82) of the blogs surveyed for this project offered their readers access to a blogroll, although interestingly more than 82 survey respondents answered the questions about their blogroll. Respondents were asked what they had in common with the contacts on their blogroll. The most popular choice here was ‘Interests’ (92 respondents). 59 respondents, just under half, also chose ‘A sense of humour’. The least popular choice was ‘Economic or domestic circumstances’, with only 8 respondents. Interestingly, bearing in mind the popularity of blogrings which linked bloggers located in the same geographic region, ‘Part of the world’ was also an unpopular choice with only 19 respondents. Bearing this in mind, an analysis of the blogrolls of all respondents was undertaken in order to ascertain how willing bloggers were to link to blogs from outside their own country.

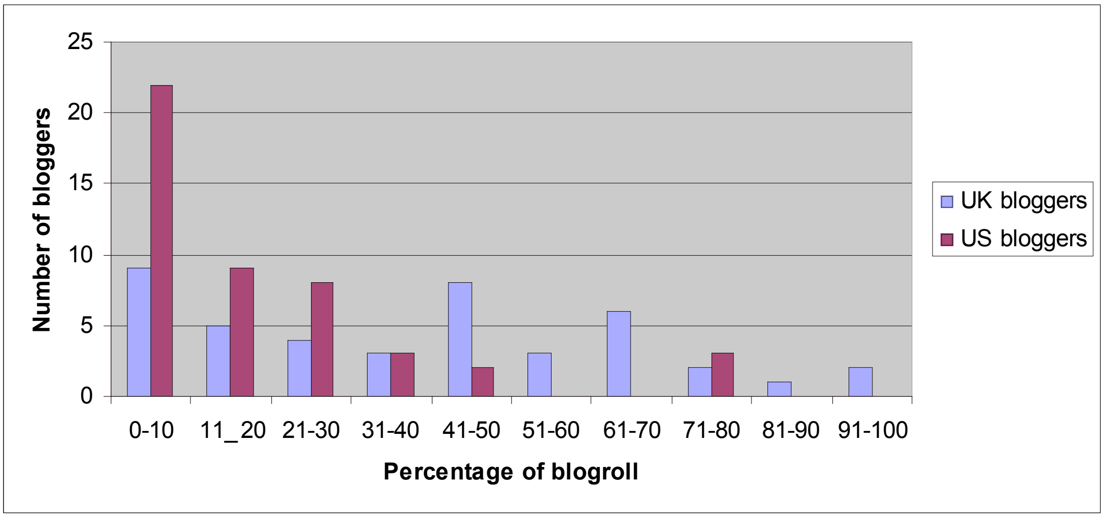


Figure 1: Percentage of blogrolls containing links to foreign blogs

Of the 47 US blogs which carried a blogroll, 31 (66%) had less than 20% of their blogroll links to blogs from outside the US. 15 of these bloggers had no links at all to blogs from outside the US. The male blogger with the most links to outside the US was actually a German expat living in the US who wrote a blog on international affairs and culture, primarily German. The female blogger with the most links to blogs located outside the US wrote a blog about the English author Jane Austen. Only three bloggers had more than 50% of links in their blogrolls to blogs outside the US.

Of the UK blogs, 14 out of the 43 which featured blogrolls had less than 20% of their blogroll linked to blogs from outside the UK. Of these, 7 had no links to any blogs located outside the UK. 14 bloggers had more than 50% of their links to blogs outside the UK. It is not that surprising that UK bloggers link more to blogs outside the UK since there ARE more blogs outside the UK. Riley estimated in July 2005 that there were 2.5 million British bloggers, compared to up to 30 million US bloggers, although there are difficulties in enumerating specifically British blogs, because of what Riley calls ‘the Anglosphere problem’, i.e. the existence of a common body of service providers and readership across the English-speaking internet.[14] However, the limited amount of linking that the average US blogger does to blogs outside the US should be noted.

While bloggers might prefer to link to other bloggers in their own country, there was less evidence that they preferred to link to bloggers within their own state or town. 64% of all bloggers had less than 10% of their blogroll devoted to links to others in their area. Only 8% had more than 50% of the links in their blogroll devoted to local bloggers. However, all four of the bloggers whose links were 100% local were from the United States and only two UK bloggers had more than 50% of their webroll devoted to local links.

Ranking in terms of popularity

Using data gathered from the blog-monitoring sites Technorati and The Truth Laid Bear and information concerning the number of links made to a blog’s front page from Surfwax, the 120 survey respondents’ blogs were ranked in terms of popularity. The Truth Laid Bear (http://truthlaidbear.com/) and Technorati (http://www.technorati.com) are websites that use links from other blogs as the measure of the relative worth of a blog. Surfwax is a metasearch engine whose Site Snaps function offers a quick abstract of any web page, including the number of links made to that page. Since popularity, as demonstrated by number of links, is used as the main criteria for success in the blogosphere, the surveyed blogs were ranked using the data collected and in the top and bottom twenty in the listing were analysed to discover common characteristics.

The top 20 blogs were as follows: 12 US respondents (10 males and 2 females) and 8 UK respondents (4 males and 4 females). The bottom 20 blogs were 8 US respondents (1 male and 7 female) and 12 UK respondents (6 male and 6 female). What is suggested by this exercise is that the survey’s US male respondents are on average more successful in the blogosphere than the other three groups. This finding corresponds to the general tenor of research findings about gender in blogging. (For references to the extensive online debate, see Pollard; Ratcliff; Garfunkel [15]). Ratcliff has recently produced evidence that men’s postings receive more comments than women’s.[16] Meanwhile, Henning suggests that women’s blogs make up only 15% of all blogrolls [17]. It has also been claimed, in the North American context, that a greater amount of attention is given in the media to male bloggers (Herring et al[18]).

Surfwax data was also used to investigate the bloggers in terms of number of links, number of images used and number of words used in their blogs. In terms of number of links, again the US males dominated with six in the top ten. They included a birding enthusiast, an evangelical Christian, an expert in global current affairs, the expert in German culture and an expert in betting on American football. The two UK males included another Christian, this time a minister, and a blogger with a long blogroll relating to mental illness. The female bloggers were both promoting their businesses through the Internet, one as a children's book illustrator and the other as a sex therapist. If we are therefore seeing a high number of links in the blog as evidence of success, again we have more successful American male bloggers, but it is also obvious that bloggers who focus on one particular subject, which may or may not be a career or source of income for them, are the most active in terms of links. Out of the ten most successful bloggers, five were blogging about some aspect of their career.

In terms of the number of images used on the blogs, the top ten bloggers included seven US males, two UK females and one US female. Four of the top bloggers here were also in the top links list above: the birding expert, the international affairs expert (who writes for a variety of magazines and journals on the subject), the evangelical Christian and another blogger whose blog focuses on funny and strange things to be found on the Internet. One UK female blogger uses her blog as part of her online shop which sells objects for the home and therefore illustrations and photos are very necessary. It appears that US bloggers are happier to use photos on their blogs than UK bloggers.

The top ten blogs with the most amount of words were those belonging to six US males, two UK males and one US female and one UK female. Again the blogs with the most words are dominated by those with a theme or focus. Of the two female bloggers, one discussed right-wing politics while the other reviewed crime novels. One of the two UK males wrote about military affairs, having been a soldier, while the other was a policeman writing anonymously about policing in the UK. Of the five US males, two were religious bloggers, one wrote about international affairs, one was the expert on the subject of American sports and betting, one was a solider writing about military affairs and one was the German expat blogging about international culture.

As can be ascertained from the above descriptions, many of the bloggers who were in the top ten for amount of words were also in the top ten blogs for either use of images and links. The two bloggers who were in the top ten for everything were a young, evangelical Christian male (US) and a blogger who wrote about international politics, with an emphasis on technology (US male). Bloggers in the top ten for at least two out of three: were the US male birding enthusiast; the German expat living in the US and writing about cultural issues; the group blog on American sports and betting (US male); a US male minister writing from a Christian viewpoint. Thus all the particular dominant bloggers in the survey, according to Surfwax data, were US males.

4 Conclusions

This project set out to compare and contrast the blogging techniques of UK and US bloggers. However, what it has discovered is noticeable differences between US males and the rest of the blogosphere. The US male bloggers surveyed dominated the rankings as far as links, use of images, amount of words and overall popularity (as defined by Technorati and The Truth Laid Bear) are concerned. The content of the US male blogs was also more likely to focus on an interest, business or hobby and less likely to be categorised as personal. They were least likely to write a blog purely for their own records or to see blogging as a form of diary keeping. In other words, US male bloggers were less likely to write 'journal' blogs, which confirms the findings of Herring et al in their analysis of such blogs. In terms of the debate about the dominance of male bloggers in the US blogosphere, it has been suggested that men are more likely to blog about external events, rather than personal ones, and are therefore more likely to be found by prospective readers when using a search engine and thus more likely to be linked to, raising their popularity ranking. The US male bloggers also seemed less concerned about privacy, giving their full name and showing identifiable photographs of themselves more frequently. Thus the main finding of the project is that the dominance of male bloggers in the US, as identified by many commentators in the last few years, also translates into a dominance of the international, anglophone blogosphere.

Some statements can be made about differences between the UK and US bloggers surveyed. The US bloggers were on average older than the UK bloggers and differences between the two countries were also found in terms of educational attainment, suggesting that the picture of bloggers as, on average, educated to graduate level gained from earlier US-based studies needs to be questioned by more research into the blogosphere outside the US.

Bloggers were much more likely to blog at home than at work, although US bloggers were more likely to blog in the morning than UK bloggers. This finding concurs with other research into Europeans' use of the Internet.

In terms of the promotion of their blog, UK respondents were more likely to use blogrings to promote their blog. Whilst US respondents tended to dismiss blogrings as of less use than blog directories, UK bloggers were happier to use them, in particular those that identified the blogger as part of the UK or its regions. Given the very different sizes of the US and UK blogosphere, this may well be in order for UK bloggers to identify each other and to maintain a sense of a UK identity against the overwhelming US group. UK bloggers were also more ready to make links to overseas blogs in their blogrolls, while US bloggers as a group were less ready. More US bloggers also had blogrolls which contained only local links. Obviously, a great part of the explanation for this is the size of the US blogosphere compared to the rest of the world. It will be interesting to see if this US-centric approach changes in the future as the blogosphere continues to expand.

US bloggers were more likely to see blogging as a useful activity, attracting readers for the intellectual or creative work. However, an equal number of bloggers in both countries identified financial gain as a motivation for blogging. This was particularly true of female bloggers and can probably be linked to the higher number of women bloggers who worked part time. Blogging is now being seen as a viable income generator for those who need a flexible approach to employment.

Overall, the project suggests that further research needs to be undertaken into the blogosphere outside the US. Since blogging started in the US, the majority of research into blogging so far has focused on the US and it is suggested that this focus has resulted in all bloggers being defined through the US experience. The findings of this project suggest that bloggers outside the US may have different approaches to blogging and find different satisfactions. It also suggests a new financial motivation for blogging, which had not previously been identified, and which may be an indication of the way in which the blogosphere is evolving.

References

[1] BLOOD, R., Weblogs: a history and perspective. *Rebecca's Pocket*, 7 September 2000**.** http://www.rebeccablood.net/essays/weblog_history.html (accessed November 2006).

[2] SCHIANO, D., et al, Blogging by the rest of us. Conference on Human Factors in Computing Systems, 24-29 April 2004, Vienna. Published in CHI '04 extended abstracts on Human factors in computing systems, ACM Press, New York. 1143-1146.

[3] HERRING, S. C.; SCHEIDT, L. A.; BONUS, S.; WRIGHT, E., Bridging the gap: A genre analysis of weblogs. Proceedings of the 37th Hawaii International Conference on System Sciences, 5-8 January 2004, Big Island, Hawaii. Los Alamitos: IEEE Press.

[4] KRISHNAMURTHY, S., The multidimensionality of blog conversations: The virtual enactment of September 11. AOIR Internet Research 3.0: Net/Work/Theory. Maastricht. October 13-16 2002.

[5] SINGER, J. B., The Political J-Blogger: Normalising a new media form to fit old norms and practices, *Journalism*, 6(2), 2005, 173-198; Matheson, D., Weblogs and the epistemology of the news: some trends in online journalism, *New Media and Society*, 6(4), 2004, 443-468; Kahn, R. and Kellner, D., New media and internet activism: from the 'Battle of Seattle' to blogging, *New Media & Society* 6(1), 2004, 87-95; Wall, M., Blogs of War, *Journalism*, 6(2), 2005, 153-172; Pedersen, S. and Chivers, A., 'Empowering citizens to join the debate? What draws readers to news blogs?', submitted to *The International Journal of Technology, Knowledge and Society*, 2007.

[6] HERRING, S. C. et al, Conversations in the blogosphere: an analysis "from the bottom up". *Proceedings of the thirty-eighth Hawai'i International Conference on System Sciences (HICSS-38)*, 2005, Los Alamitos: IEEE Press; Schiano, D. J., Nardi, B. A., Gumbrecht, M. and Swartz, L., Blogging by the rest of us. *CHI 2004, April 24-29 2004, Vienna, Austria*, http://home.comcast.net/~diane.schiano/CHI04.Blog.pdf (accessed on 3rd June 2004); Nardi, B. A., Schiano, D. J. and Gumbrecht, M., Blogging as a social activity, or, would you let 900 million people read your diary? *Proceedings of computer supported cooperative work 2004*, http://home.comcast.net/%7Ediane.schiano/CSCW04.Blog.pdf (accessed 23 February 2006).

[7] HUFFAKER, D.; CALVERT, S., Gender, identity, and language use in teenage blogs, *Journal of Computer-Mediated Communication*. 10.2, 2005.

[8] MORTENSEN, T.; WALKER, J., Blogging Thoughts: Personal Publication as an Online Research Tool. In A. Morrison (Ed.), Researching ICTs in Context. InterMedia, University of Oslo, 2002, 249-279.

[9] THELWALL, M., Bloggers during the London attacks: Top information sources and topics. 15th International World Wide Web Conference, Edinburgh, Scotland, May 23-26, 2006, 2.

[10] TRAMMELL, K. D.; TARKOWSKI, A.; HOFMOKL, J.; SAPP, A. M., Rzeczpospolita blogów [Republic of blog]: Examining Polish bloggers through content analysis, *Journal of Computer-Mediated Communication, 11(3),* 2006; Abold, Roland, 1000 Little Election Campaigns: Utilisation and Acceptance of Weblogs in the Run-up to the German General Election 2005, 2006 ECPR Joint Sessions of Workshops, 25-30 April, 2006. Nicosia/Cyprus; Tricas-García, F., and Merelo-Guervos, J. J., The Spanish-Speaking Blogosphere: Towards the Powerlaw? IADIS International Conference WWW/Web Based Communities, Lisbon. 24-26 March, 2004.

[11] PEDERSEN, S.; MACAFEE, C.,The Practices and Popularity of British Bloggers. 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 Dobreva. pp. 155-164 http://elpub.scix.net/cgi-bin/works/Show?213_elpub2006; Auty, Caroline, UK elected representatives and their weblogs: first impressions. Aslib Proceedings: New Information Perspectives, 57.4 (August 2005). 338-355; Thelwall, M., Bloggers during the London Attacks.

[12] WRAY, R, Surfers glued to web for 11 hours a week. *The Guardian*, 29 November 2006. http://technology.guardian.co.uk/news/story/0,,1959548,00.html (accessed 29 November 2006).

[13] PEDERSEN, S., Women users' motivation for establishing and interacting with blogs (web logs). *International Journal of the Book 3(2),* 2005, 85-90.

[14] RILEY, D., Blog count for July: 70 million blogs. The Blog Herald. b5media, 2005. http://www.blogherald.com/2005/07/19/blog-count-for-july-70-million-blogs/.

[15] POLLARD, D., Is the blogosphere sexist? How to save the world, 30 October 2003. http://blogs.salon.com/0002007/2003/10/30.html (accessed 9 March 2006); Ratliff, C., Whose voices get heard? Gender politics in the blogosphere. Culture Cat, 25 March 2004. http://culturecat.net/node/303 (accessed 12 September 2005); Ratliff, ., *The* link portal on gender in the blogosphere. Culture Cat, 21 December 2004. http://culturecat.net/node/637 (accessed 12 September 2005); Garfunkel, J., Promoting women bloggers: a timeline of relevant discussions. Civilities media structures research, 15 March 2005. http://civilities.net/PromotingWomenBloggersTimeline (accessed 23 February 2006).

[16] RATLIFF, C., WATW by the numbers. Culture Cat, 18 February 2006. http://culturecat.net/node/1030.

[17] HENNING, J., The blogging iceberg. Perseus, 4 October 2003. http://www.perseus.com/blogsurvey/iceberg.html (accessed 5 October 2005).

[18] HERRING, S. C.; KOUPER, I.; SCHEIDT, L. A.; WRIGHT, E. L., Women and children last: the discourse construction of weblogs. *In*: Laura J. Gurak et al., eds. *Into the blogosphere: rhetoric, community, and culture of weblogs*. Minneapolis: University of Minnesota, 2004. http://blog.lib.umn.edu/blogosphere/women_and_children.html (accessed 7 March 2006).

Enhancing Traditional Media Services Utilising Lessons Learnt from Successful Social Media Applications – Case Studies and Framework

Asta Bäck ; Sari Vainikainen

VTT, Media and Internet, P.O.Box 1000, FI-02044 VTT, Finland
e-mail: asta.back@vtt.fi; sari.vainikainen@vtt.fi

Abstract

The paper presents a framework for describing electronic media services. The framework was created by utilising earlier models and case studies of successful social media applications. Wikipedia, YouTube and MySpace were analysed because they are among the most popular sites in the world and they highlight different aspects of social media applications. The proposed model consists of two main parts: Concept and system, and Content and user. Both of them were further divided into four subgroups. With the help of a radar view, various applications can be described and compared and their further development opportunities identified. A prototype application, StorySlotMachine, is used as a case example, where the framework is used.

Keywords: social media; YouTube; Wikipedia; MySpace

1 Introduction

The initial vision by Tim Bernes-Lee [1] was that internet would be a platform for interactive information sharing. During the recent years we have seen development that has made this vision true. The tools needed for digital content production have become easy to use and cheap enough for a large number of people. More and more people have broadband Internet access, which makes Internet the natural way to share digital media. Terms like Web 2.0, social media and user-generated content are being used to describe the services that have been created on top of this development during the last years.

This development has raised many questions among media companies. The traditional media approach has been product centred. The aim is to offer a product to as large audiences as possible. The role of a media company has been to organise and filter content and to package it into a marketable whole. Marketing is often done at two directions: the product is sold to readers or viewers and this audience is then sold to advertisers.

Media services can be roughly categorised into two main groups: information or entertainment. In practice, most traditional publications try to address both of these with the main emphasis on one or the other. We can also claim that many publications promote some sense of community among their readers or viewers by giving people common topics to talk about. Community building is also obvious in publications with relation to some religious, political or some other idealistic movement. But also in these cases, the role of individual end users has mostly been invisible and they have been seen as a target group.

Social media applications give people new ways to find information and entertainment and also to build communities, which has meant that the role of traditional media has become weaker. Traditional media companies do not host the most popular or most quickly growing sites on the net.

Several terms are being used in connection to these new services. The word participatory media has been used to emphasise the nature of user participation in media creation. Other frequently used terms are social media, social software and social networking. These terms bring out the social aspect – users not only act as content creators but social interaction between users has become possible and visible in the applications.

From the content or media object point of view, the most important change is the lengthened life cycle of media objects. In the traditional publishing models, the content selection at media object level was made by professionals who compiled the aggregations that then were offered to consumers as packages. As content has become digital and it has become easy to refer to and discuss single media objects, the life of a media object may be lengthened considerably (Fig 1). Media is also more and more consumed as smaller fragments - video clips instead of whole shows, single songs instead of whole albums.

This development raises the question of what traditional media companies could learn from the successful social media applications to revitalise their own business. The aim was to create a framework that would help in characterising social media applications and would be usable as a tool in finding new opportunities to developing new services to traditional media companies.

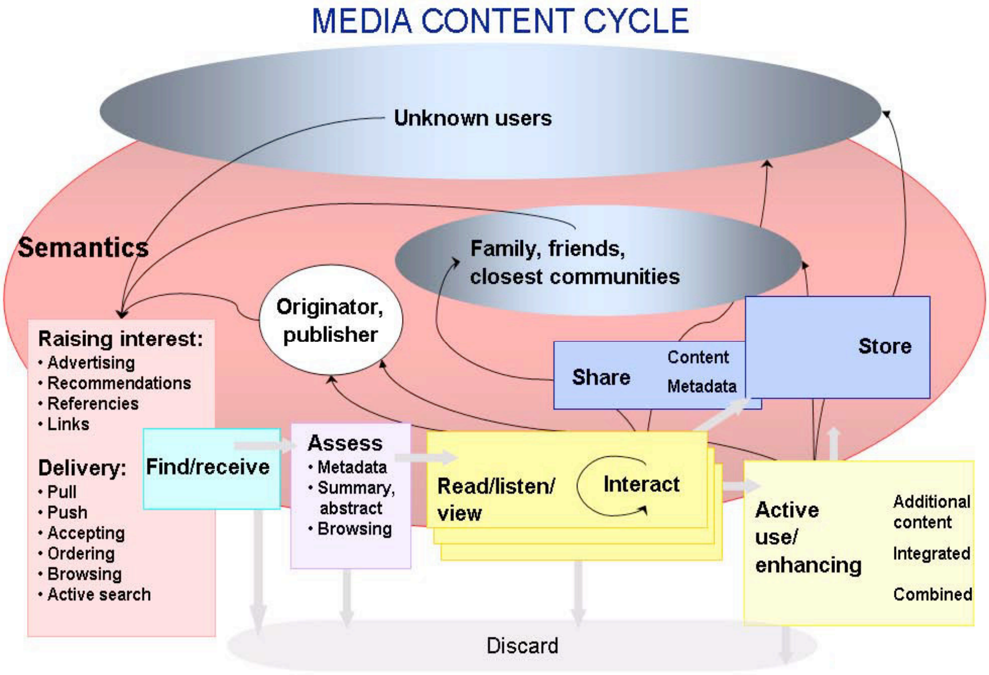


Figure 1: The life content cycle is prolonged for the electronic media

2 Methodology

The chosen research method was to make case studies of successful social media applications. In order to make the case studies in a systematic way, the main characterising features of social media applications were identified. Here, also the traditional media processes were taken into consideration in order to find and pinpoint the areas were the differences are greatest.

Case studies were made by going through three hugely popular applications, Wikipedia, YouTube and MySpace, and analysing their features. These three are among the most popular sites in the world, and they highlight different aspects of social media applications. Additional information and research findings were searched from the literature. No user studies were made, so the analysis is based on what can be seen at the website and what research results and other information is available.

3 Results

3.1 Characterisation of Social Media Applications

We can look for characteristics for describing social media applications from two directions: from general IT application adoption point of view and in comparison to traditional media. Technology acceptance model (TAM) has been applied successfully as a theoretical framework to explain the adoption of IT applications [2]. It was originally developed relating to office applications, but is has been found to apply to other IT applications as well.

The model is simple. It explains the acceptance with the help of two factors – ease of use and usefulness. People must perceive benefits from using an IT application and the application must be easy to use. These are important criteria for social media applications, since the user must be able to learn to use the service by him or her self, and the value of these applications increases as the number of users increases. The first seconds a first time visitor spends with the application are the most important ones, because if the user is confused and does not know right away how to use the service and what benefit there are available, he or she may never return. The traditional media process consists at top level of three main steps:

1. Concept and longer term product development: what kind of content products will be offered and to which target groups;
2. Marketing to customers and advertisers;
3. Production – creating and selecting the content for individual issues. Here, depending on the type of media, content is either made based on commissions or offered for publishing;
4. Feedback from customers is received two ways: direct feedback on single articles but mostly in the form of increasing or decreasing sales and subscriptions.

Since social media applications are IT applications, we need to pay attention to the ease of use aspect, which has not been relevant in traditional media products. The usefulness or value proposition is important in any service or product, since other features are irrelevant if there is not perceived value in using an application. In social media, the aspects are particularly worth addressing users and content. Based on these features, the following characteristics were chosen to be used in the analysis:

1. What is the main service concept, and which additional needs people may have in connection to media products (Concept and value proposition);
2. What kind of content is being created and shared and how users participate in content creation and management, what is required from users to be able to participate in content creation and management (Content and user participation);
3. How visible the users are in the service, does the service support identity and community creation (User identity and networking);
4. When and how the service is being used and marketed (Use and marketing).

3.2 Wikipedia

Wikipedia[1] has become an important collection of knowledge and it can be regarded as the open content counterpart to the open source development.

Concept, value proposition

The Wikipedia product concept – a collaboratively created encyclopaedia, was defined by its founders. The key promise was to make a free and open knowledge source and this way participating in Wikipedia entailed participating in a big common goods endeavour. The five pillars [3] depicting the key Wikipedia principles are as follows:

- Wikipedia is an encyclopaedia
- Wikipedia has a neutral point of view
- Wikipedia is free content
- Wikipedia has a code of conduct
- Wikipedia does not have firm rules

Larry Sanger, one of the Wikipedia founders, sees the following causes behind the Wikipedia success [4]:

1. Open content licence
2. Focus on encyclopaedia
3. Openness – anyone can contribute
4. Ease of editing
5. Collaborate radically, anybody can edit anybody's article
6. Offer unedited, unapproved content for further development
7. Neutrality
8. Start with a core of good people
9. Enjoy the Google effect

[1] http://en.wikipedia.org

Content and user participation in content creation

There are two main parts in the Wikipedia:

- the platform with many features supporting collaborative and unmanaged content creation;
- the content creation and management process that lets any user act as a content creator or editor.

The Wikipedia platform supports collaborative editing by storing all versions of the content together with the user name or IP address. All versions of the article are available and comparisons may be made between any of them. If one of the earlier versions is considered better than the current one, it is possible to revert to it. This can also be used to fight spam.

Anyone is able to create a new entry on any subject they see worth writing about. Existing articles can be used as examples. Most active users act as editors and, for example, mark articles that need further refinement. New articles are being created by making an internal Wikipedia reference to a topic that does not yet exist. This means that the person who creates or names a page, need not know or write about it but only to be of the opinion that this is a topic that should be included in the Wikipedia. There is no process for selecting which articles to write. However, users may create what is called a portal within Wikipedia to promote and support creating articles relating to a topic. Such portals have been created for multiple topics, for example for various sports or countries.

In order to promote high-quality articles, articles may be requested a peer-review, and an article may get evaluated as a good article or a featured article. The focal point of the users' work is creating a coherent and balanced article about a topic. Self-organising is the supported way to get things done. Active users may become selected into managerial tasks within the Wikipedia community.

User networking

Mediawiki[2], the platform used for creating Wikipedia, supports creating pages or articles. Information about single users may be presented in the same way: A user may write an article about him- or herself and give whatever information he or she wants to share with others - or remain unknown and participate anonymously. There is no direct support to find information about user connections, but users must explicitly create any such information.

Use and marketing

Users may access articles either by making a search, clicking the Random page link or following the links in the articles. Users are encouraged to embed internal Wikipedia links within the articles, and group the external links at the end of the article, if such links exist.

There is not direct information available about how much the pages have been viewed. There is a Statistics page that gives some information about the total number of pages, as well as lists the most popular pages. There is no direct support to invite other people to visit Wikipedia.

Currently search engines are an important driver to Wikipedia articles. Wikipedia articles are ranked highly, and some search engines support targeting searches directly to Wikipedia. Also, the free access to Wikipedia content has given it visibility, as various open initiatives, such as Semapedia[3], have utilised the Wikipedia content. Free content brings with it the opportunity to get free visibility.

3.3 YouTube

Concept, value proposition

YouTube[4] is technically a platform for sharing online videos. The service was created to support this particular media format at the point when people had started creating more videos, but there was no easy way to share them.

We can say that the value proposition has two levels: the immediate and practical value and potential value. The immediate value is being able to share content over the internet with family and friends, and the potential value is in being able to reach new people and contacts, and even international fame.

[2] http://www.mediawiki.org/wiki/MediaWiki

[3] http://www.semapedia.org/

[4] http://www.youtube.com

Content and user participation

The service relies completely on users to upload video clips. The platform is open for any person who wants to upload a video. At uploading, users give some basic metadata about their video (category, description, tags).

User participation is not restricted to bringing in new content, but users may participate in assessing and evaluating content. There are several opportunities to that such as rating, discussing, and adding to favourites. Also the mere act of viewing a video is utilised as a metadata, because number of times each video is viewed is shown and the most viewed videos can easily be found. Users' may collect videos into playlists and these lists may be made public, which supports finding related items. A user may also create a channel that others may subscribe. A collaborative way of aggregation videos is to set up a group with a special theme.

User networking

Each user gets a home page that gathers information about that user and his or her activities in the community. Users may easily communicate with each other and also create a permanent link between themselves by connecting as a friend. User favourites, subscriptions and subscribers are visible there. User networks support both finding other users with similar interests and also finding content.

Use and marketing

The YouTube website shows about each video how many times it has been viewed. Also information is given on which websites have links to each video, as well as honours, if the video has received such. Once a video has been viewed the system suggests additional videos for viewing in order to make people stay longer at the website. User networks and the ability to embed any YouTube video in other websites have played a key role in making YouTube known and popular. YouTube fits well to the online communication where users post links to each other.

Jawed Karim, one of the YouTube founders considers the ease of linking to the videos as well as the opportunity to embed a YouTube video by copying a piece of HTML code to ones web page as factors contributing to the rapid growth that the site experienced [5]. Additional key feature was the ease of use: videos could be viewed immediately without downloading a viewer or codec and discussions could be made around videos, so that viewing a video was not any longer a single, detached event, but became a social event with possible many steps.

3.4 MySpace

Concept, value proposition

MySpace[5] is a social networking site. It main value proposition is to offer a public web space to present oneself and to connect to other people with similar interests. Also here, the opportunity to become famous, particularly in the music scene, is an important part of the value proposition.

Content and user participation

The content in MySpace is gathered around profile pages. People add media that they find interesting and supportive to their profile and personality. Also various commercial products like cars or films may have a profile page that people may connect to, if they find the product valuable to them. Users may, for example, rate videos, but these ratings are not published as directly and openly as for example in YouTube.

User marketing. The break through however was made when small offline communities between 100 and 1000 members were attracted to the users of the site. This contributed to launching the network effect.

3.5 Lessons from the Social Media Applications

networking

User networking is the key functionality in the service.

Use and marketing

According to danah boyd [6] there are several reasons why MySpace became so popular, particularly among 14-24 people. This age group has a bigger need for creating a public profile than other age groups. Youth also needs their own public space, and MySpace has offered this in virtual space.

[5] http://www.myspace.com

Also, the communication opportunities offered by MySpace fit the existing communication patterns: the youth communicates via instant messaging for immediate communication needs, and MySpace is the complementing asynchronous communication channel. MySpace has become part of their daily routine, which means that when people have their computer on, one of the sites they have open, is MySpace.

Gabbay in his case study of MySpace [7] claims as one of the decisive factors contributing to the MySpace success the freedom that the platform and its developers offered to the users. Photos and music were according to him the most important content elements. As to the initial marketing of the site, various kinds of marketing actions were utilised including contests and email
Each of these three successful applications has a clear focus or value proposition, which is complemented with additional values.

Wikipedia is the most idealistic application where a common goal to create an important knowledge mass motivates people. Here a single user may remain anonymous, if he or she wishes so. And also, there is least support for making and showing user connections. Wikipedia differs from the other two also in that a common media object, an encyclopaedia article is being created in collaborative fashion. The model is disruptive in the sense that users are encouraged to modify and alter other users' texts. Also, the self-organising way of creating content can be regarded as disruptive. The critical point in Wikipedia is user motivation and maintaining it. There are little opportunities to external rewards for Wikipedia content creators and editors.

YouTube offers both immediate value (video sharing) and potential value as a platform in seeking feedback, popularity and fame. User information and profile are visible and even though the content is shared freely among users, each user may utilise the platform for making him or herself more known. There are several ways for users to connect with each other. Users may also create groups that concentrate on some special interest. The role of users in organising the content is also important: the video related metadata becomes richer as people rate and comment videos and add them to their favourites. Finding videos would be practically impossible without this additional user generated metadata.

MySpace turns the relation between users and media objects into reversed order. Here, the users or user profiles are in the focus of attention and media objects are used to complement the user profiles. MySpace lets users play with their creativity and use the system in many, also in unforeseen ways.

In traditional media, products like an issue of a magazine or book, are being created by utilising well-defined processes with input and output. Web-based social media applications have a very different life cycle: a platform is offered to users and as the service gets used, its starts to take shape and may get new form.

3.6 Framework

Based on the initial characteristics used in describing the social media applications and the findings that were made, the characteristics were further elaborated into a framework. The framework supports getting a quick overview of the features a service and pinpoints to areas where there could be additional opportunities for further development by taking into account the features that have been successfully utilised in social media applications. The features were grouped into two groups: Concept and system, and Content and user. There were further divided into four subgroups each.

Concept and system (Fig 2) is divided into the following subgroups:

1. Main attention in the service
   - Content
   - Users
   - Other

2. Value proposition timeframe
   - Immediate
   - Long term, cumulative

3. Value proposition type
   - Social
   - Emotional
   - Rational, practical

Social and emotional values are indicated as separate, even though they are somewhat overlapping. We wanted, however, to give the opportunity to separate them. For example, a movie typically offers emotional value without strong social aspect.

4. Usability and improvements
    - Ease of use
    - Evolutionary development

Content and use (Fig 2) is divided into four subgroups:

1. Content enhancements
    - Aggregating content (e.g. playlists)
    - Modifying content
    - Opinion expression (ratings, metadata) and its visibility

Aggregating content refers to the opportunities of combining available content into aggregations or combinations that a user finds meaningful or practical to him- or herself, for example making a playlist of videos or combining bookmarks into a readlist. Modifying content refers to changing the content or features within a media object, for example editing the text or picking a part of a video and combining it with some other video. Opinion expression refers to user generated metadata, which may be utilised to make recommendations, and/or shown explicitly to other users.

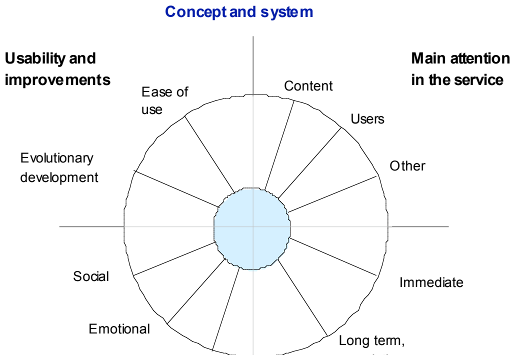


Figure 2: The framework for describing a media service features. The more a certain feature or functionality is supported a position closer to the outer circle should be chosen

2. Content type and sharability
    - User-generated
    - Commercial, professional
    - Exportability (embedding, APIs)

3. Creation opportunities
    - Alone
    - Small group
    - Community

Here, the term 'Small group' refers to the opportunity of organising groups with special focus or aim, or with one's friends and family. The 'Community' refers to larger scale collaborative work and sense of community within the service.

4. User visibility and networking
    - Identity building
    - User networking
    - Inviting and attracting new users

By going through the whole framework, it is possible to get an overview as to where the main opportunities are for further development. Also more detailed analysis may be carried out by picking two of the subgroups and looking at their interrelations as a quadrangle. For example, the value proposition time frame and type could be analysed as a quadrangle, or Content enhancement opportunities and Content type and sharability, or Content enhancement opportunities and Creation opportunities. The framework does not imply some ideal value, even though the general interpretation is that the larger a figure comes out when evaluating the features of a service, the more opportunities there are for user interaction and value creation. It may not, however, be sensible or even possible to combine all the features in one service.

Figure 3 shows two of the analysed services described with the help of the framework. We can see that YouTube has more support on the user and user connectivity side. The clear difference in the approach to content enhancement is that YouTube supports aggregating and playing with video clips where as the main focus of Wikipedia in the modifying and working on the actual content, mainly the articles. Regarding concept and use, we can see that YouTube includes many more of the identified features than Wikipedia.

In Figure 4, an application prototype, called StorySlotMachine, is described with the help of the framework. StorySlotMachine [8] was built in one of our earlier research projects. The project aimed at exploring the opportunities of utilising semantic metadata and user-created content together with commercial media content. The application lets users explore and play with media content. Content may come from various sources: user's own content may be complemented with that from other users and what a commercial media company offers. The users are encouraged to add some information about where their photos were taken, and what is shown in the photos. Based on these clues, additional related content is offered. Users may also create their own collections out of the existing material, either travelling plans, guides or reports.

The prototype deals with travelling related content and focus. Many media companies have extensive archives that are not effectively utilised as end user services. The StorySlotMachine is an example of how their content could be offered in a more interesting ways than as mere searches and search result lists.

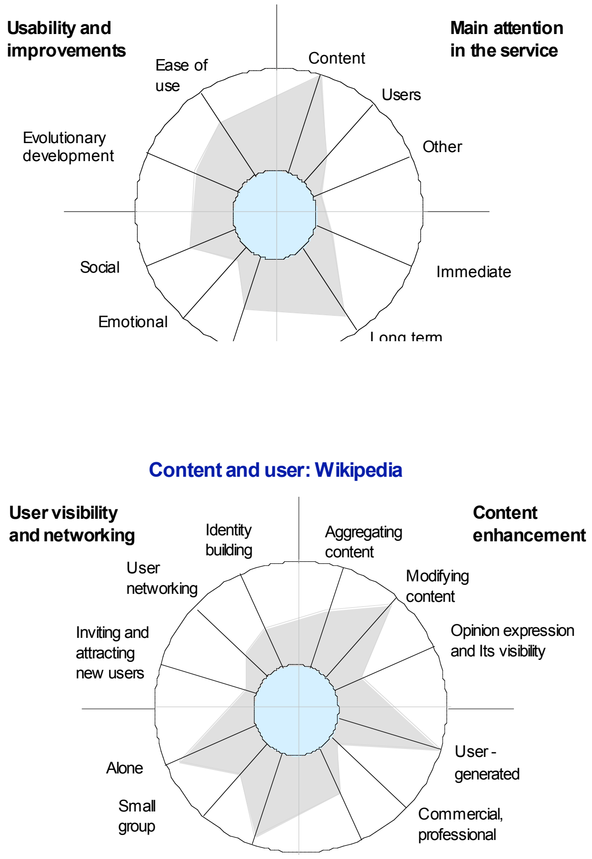


Figure 3a-b: Two social media applications, Wikipedia and YouTube characteristics depicted in the proposed framework. The differences in the emphasis of the applications come clearly out

The prototype development was started three years ago. The idea of combining user-generated content with professionally created commercial content was emphasised as well as the aspect of being able to easily play with content components. The social aspects of such an application were not taken so strongly into the focus in system development. In user tests, users liked the idea to get information as ready made stories and to combine their own content with commercial content, and were interested in features that let them view and utilise aggregations that other users had created. We clearly see the opportunities in further development by adding more visibility as to what other people are doing in the service, and also being able to connect to other users.

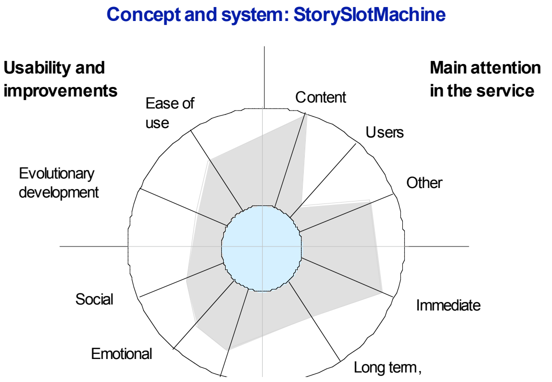


Figure 4: A prototype application, StorySlotMachine described with the framework. We can see that the development opportunities exist particularly in adding support for user visibility and social aspects of content creation and sharing

4 Discussion

The framework gives a quick way of capturing the features of a media service and comparing them with the op–portunities and the lessons learned from successful social media applications. The framework does not address the business side. Currently advertising is the key business model, and people rarely pay for the access to content [9].

Traditional media companies may benefit from the web experiments that other companies have made, but there are also many challenges. The most successful applications have originated from new companies, and not from traditional media companies. New entrants have not had any existing business and this had given them the opportunity to adopt very new and even disruptive models. During the first Internet wave in late 1990'ies, the belief was that the key was to attract as many users as possible, and that the business would come along with it. This belief is still valid in many cases. Also, creating new services on the web for new customers is in many cases easier and more sensible than trying to offer web-based services to users who are more accustomed and happy with their traditional media products.

As to the framework, we see that it should be tested further. More services should be described with it in order to validate it more. Also comparisons should be made with not so successful applications – does this framework help in differentiating these from the more successful ones. Also, wider application of the model as a tool in exploring the development opportunities for existing media services should be done.

The focus is utilising the framework should not be in debating where the exact position on each axis is but to discuss, whether these different aspects should be supported in a service, and if so, what would be the best way to do it.

5 Conclusions

Social media applications have brought the users into focus in media applications. The traditional model where users were regarded as a target group remains also valid but the growth opportunities have been in activating and giving tools to users. New opportunities have also been found in taking something else than content as a starting point: in MySpace the users are the centre piece, and in our StorySlotMachine the travelling sights.

The framework gives a good starting point for exploring the development opportunities based on success stories. But, we must pay attention to new opportunities that may emerge but which have not been explored yet. Adding the mobile dimension and utilising context can be mentioned as two areas where to look for new opportunities.

Involving users creates the opportunity to grow the sense of community and ownership of the service. The most critical point is the start – how to get the community to attract and active users. Also, connectivity to other users and content on the net is needed and helps in creating successful applications.

Notes and References

[1] BERNERS-LEE, T., Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web. Harper San Francisco; 1st edition (September 22, 1999).

[2] DAVIS, F. D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quartely:13/1989, pp. 319–339.

[3] Wikipedia: Five pillars. http://en.wikipedia.org/wiki/Wikipedia:Five_pillars. Accessed April 10, 2007.

[4] ANON. The Early History of Nupedia and Wikipedia, Part II. http://features.slashdot.org/article.pl?sid=05/04/19/1746205&tid=95. Accessed April 10, 2007.

[5] KARIM, J., YouTube: From Concept to Hyper-growth. http://www.youtube.com/watch?v=nssfmTo7SZg. Accessed April 10, 2007.

[6] BOYD, D. "Identity Production in a Networked Culture: Why Youth Heart MySpace." *American Association for the Advancement of Science*, St. Louis, MO. February 19. 2006. Available at http://www.danah.org/papers/AAAS2006.html. Accessed April 10, 2007.

[7] GABBAY, N. MySpace Case Study: Not a purely viral start. September 2006. http://www.startup-review.com/blog/myspace-case-study-not-a-purely-viral-start.php. Accessed April 10, 2007.

[8] BÄCK, A.; Vainikainen, S.; Näkki, P.; Reti, T.; Sarvas, R.; Seppälä, L.; Hietanen, H.; Turpeinen, M. Semantically supported media services with user participation. Report on the RISE-project. VTT, Espoo. VTT Publications: 612. 2006.

[9] KANGAS, P.; TOIVONEN, S.; BÄCK, A.; (ed.). Google advertisements and other social media business models (in Finnish, Googlen mainokset ja muita sosiaalisen median liiketoimintamalleja) VTT, Espoo. VTT Tiedotteita - Research Notes: 2369. 2007.

The Fight against Spam - A Machine Learning Approach

Karel Jezek; Jiri Hynek

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

Abstract

The paper presents a brief survey of the fight between spammers and antispam software developers, and also describes new approaches to spam filtering. In the first two sections we present a survey of the currently existing spam types. Some well-mapped spammer tricks are also described, although the imagination of spam distributors is endless, and therefore only the most common tricks are covered. We present some up-to-date spam blocking techniques currently integrated into today's spam filters. In the Methodology and Results sections we describe our implementation of Itemsets-based, Naïve Bayes and LSI classifiers for classifying email messages into spam and non-spam (ham) categories.

Keywords: unsolicited mail; spam filter; machine learning; latent semantic indexing; classification

1 Introduction

The term "electronic publishing" commonly refers to the distribution of e-books and periodicals, as well as websites, blogs, etc. E-mail is just another means of information dissemination. It thereby demonstrates the features of electronic publishing. If used properly, it perfectly serves for information exchange among individuals, but when used maliciously (which is more often the case), it serves for broadcasting of (mis)information to the general public. What we have in mind is, of course, spam, also known as Unsolicited Bulk Mail (UBM), Excessive Multi-Posting (EMP), Unsolicited Commercial Email (UCE), spam mail, junk mail, or bulk email, as opposed to the term "ham" used for legitimate mail.

The very first spam was distributed in 1978 via ARPANET, notifying all network users of the newly developed DEC-20 computer. The first unsolicited mail that was actually labeled "spam" for the first time in history was distributed in 1993 by Richard Depew, who mistakenly distributed 200 messages to newsgroups administered by him. He apologized immediately, using the word "spam".

The antispam industry is constantly developing new techniques to fight sophisticated tricks used by spammers. On January 24, 2004, Microsoft chairman Bill Gates presumptuously announced that "spam will be solved by 2006". However, neither Microsoft, nor any other company, has yet found a solution. The spam-filter-review statistics of 2006 (see http://spam-filter-review.toptenreviews.com/spam-statistics.html) show the following data: spam constitutes 40% of all email messages, there are 12.4 billion spam emails distributed per day, and 2,200 spam messages are received per person per year. The most common categories of spam are product advertisements (25%), financial (20%), adult (19%), scams (9%), and health (7%).

Spam messages pose a serious problem due to multi-billion dollar costs. The MSSP Survey of 2006 claims that unsolicited emails now consume approx. 819 terabytes of bandwidth every day, representing 85% of global mail traffic. Fortunately for email users, antispam software has become increasingly effective in recent years

Many computer users have been given hope by antispam laws, such as the US federal Can-Spam Act [1] of 2003, but only 26 states had implemented any antispam legislature by 2006. This required spam senders to allow recipients to opt out of receiving future messages. It also prescribed imprisonment for violators. According to the US Federal Trade Commission, the volume of spam declined in the first eight months of 2005, but the decline was short-lived. At the beginning of 2006, spam was again out of control.

An example of another useful initiative is the OECD's "Task Force on Spam" [2]. The OECD (Organization for Economic Co-operation and Development) has launched its Anti-Spam "Toolkit" as the first step in a broader initiative to help policy makers, regulators and industry players orient their policies relating to spam solutions and restore trust in the Internet and email.

2 State of the Art

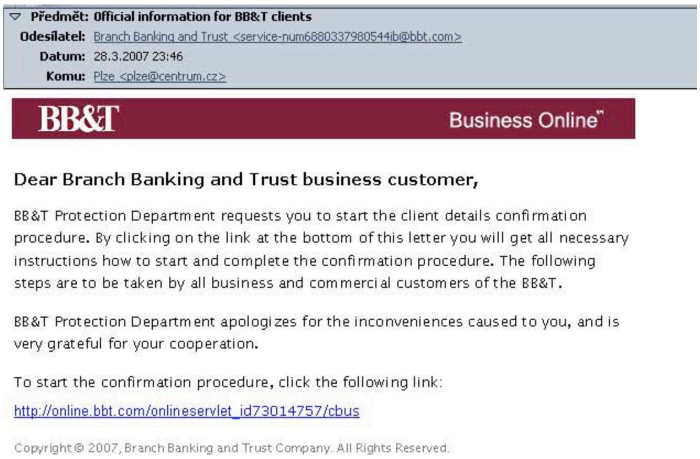
2.1 A Survey of Currently Existing Spam Types

2.1.1 Stock Spam, „Pump and Dump“

The term “pump and dump” on the Internet represents unsolicited mail offers of very inexpensive goods (typically below $1), urging mail recipients to quick purchase. This evokes massive demand for goods which have already been sold in most cases. Nonetheless, the price of the goods is gradually increased (“pumped”). This type of unsolicited mail often includes links to small or non-existing companies, as it is almost impossible to track any information on the company making the attractive deal. In some cases, “pump and dump” spam is designed to hurt the good name of an existing company, as the consequences of illegal business deals are borne by the actual company, not the spammers.

2.1.2 Phishing

Phishing (see Figure 1) is used for messages designed to elicit personal data (such as bank account numbers, credit card numbers, passwords, etc.) from email recipients. The term is derived from “fishing”, which is exactly what spammers do – distribute “bait” and wait to see what happens. Spammers commonly use exploits such as using the company’s image, inserting links to the real company site, or using email that appears to be from the spoofed company.



Předmět: Official information for BB&T clients
Odesílatel: Branch Banking and Trust <service-num6880337980544ib@bbt.com>
Datum: 28.3.2007 23:46
Komu: Plze <plze@centrum.cz>

Dear Branch Banking and Trust business customer,

BB&T Protection Department requests you to start the client details confirmation procedure. By clicking on the link at the bottom of this letter you will get all necessary instructions how to start and complete the confirmation procedure. The following steps are to be taken by all business and commercial customers of the BB&T.

BB&T Protection Department apologizes for the inconveniences caused to you, and is very grateful for your cooperation.

To start the confirmation procedure, click the following link:

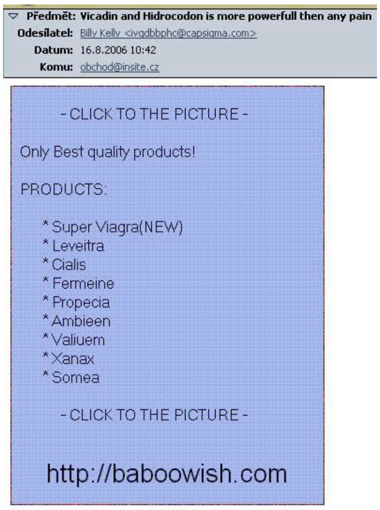
http://online.bbt.com/onlineservlet_id73014757/cbus

Copyright © 2007, Branch Banking and Trust Company. All Rights Reserved.

Figure 1: Example of a phishing spam

2.1.3 Image-Based Spam

Tricks used to distribute unsolicited mail get more and more sophisticated. The best way to get around statistical text filters is to use images instead of text (see Figure 2). Image handling is quite difficult for antispam software, regardless of the actual image form – plain text converted into an image, various interference items on the background, use of animations, etc. Although use of images for spamming is not a new concept, it is definitely gaining popularity. According to various studies, approximately one-third of all unsolicited mail was represented by image-based spam at the end of 2006. It seems that spammers are quite content with the hit rate of their messages, and keep converting all their text-based mails into images.



Předmět: Vicadin and Hidrocodon is more powerfull then any pain
Odesílatel: Billy Kelly <ivgdbbphc@capsigma.com>
Datum: 16.8.2006 10:42
Komu: obchod@insite.cz

- CLICK TO THE PICTURE -

Only Best quality products!

PRODUCTS:

- Super Viagra(NEW)
- Leveitra
- Cialis
- Fermeine
- Propecia
- Ambieen
- Valiuem
- Xanax
- Somea

- CLICK TO THE PICTURE -

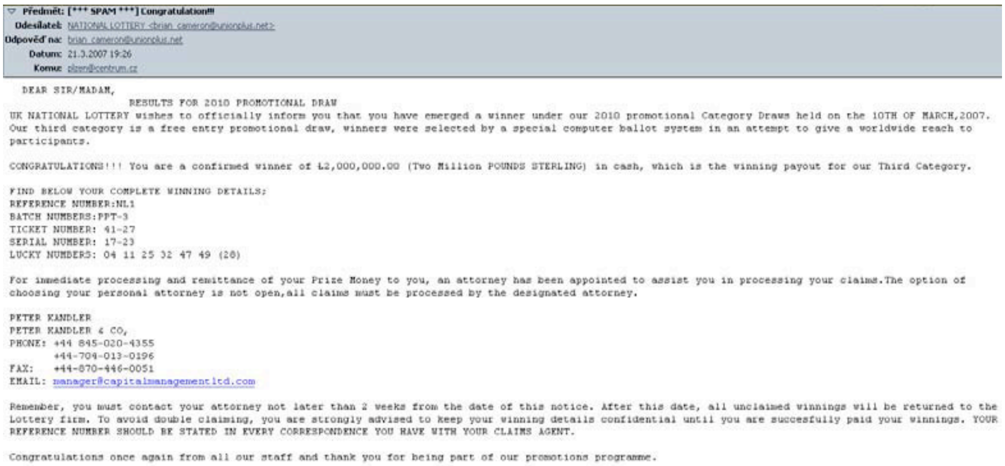
http://baboowish.com

Figure 2: Example of a clickable image-based spam

2.1.4 Text Spam

Text spam is just unsolicited commercial mail distributed in textual form (see Figure 3). Typical features of the text spam are listed below (please note that the majority of these features are language-independent):

- HTML text contained in message body,
- High proportion of capital letters (usually more than 30%),
- Exclamation mark(s) in the message subject,
- Instructions on how to unregister from the distribution list,
- Instruction to click on a link,
- Text lines longer than 200 characters,
- High priority assigned to the message,
- Nonsense date of sending (such as 1st January 1970),
- Disclosed message sender,
- More (or disclosed) message recipients.



Předmět: [*** SPAM ***] Congratulation!!!
Odesílatel: NATIONAL LOTTERY <brian_cameron@unionplus.net>
Odpověď na: brian_cameron@unionplus.net
Datum: 21.3.2007 19:26
Komu: plzen@centrum.cz

```
  DEAR SIR/MADAM,
                    RESULTS FOR 2010 PROMOTIONAL DRAW
UK NATIONAL LOTTERY wishes to officially inform you that you have emerged a winner under our 2010 promotional Category Draws held on the 10TH OF MARCH,2007.
Our third category is a free entry promotional draw, winners were selected by a special computer ballot system in an attempt to give a worldwide reach to
participants.

CONGRATULATIONS!!! You are a confirmed winner of £2,000,000.00 (Two Million POUNDS STERLING) in cash, which is the winning payout for our Third Category.

FIND BELOW YOUR COMPLETE WINNING DETAILS;
REFERENCE NUMBER:NL1
BATCH NUMBERS:PPT-3
TICKET NUMBER: 41-27
SERIAL NUMBER: 17-23
LUCKY NUMBERS: 04 11 25 32 47 49 (28)

For immediate processing and remittance of your Prize Money to you, an attorney has been appointed to assist you in processing your claims.The option of
choosing your personal attorney is not open,all claims must be processed by the designated attorney.

PETER KANDLER
PETER KANDLER & CO,
PHONE: +44 845-020-4355
       +44-704-013-0196
FAX:   +44-870-446-0051
EMAIL: manager@capitalmanagementltd.com

Remember, you must contact your attorney not later than 2 weeks from the date of this notice. After this date, all unclaimed winnings will be returned to the
Lottery firm. To avoid double claiming, you are strongly advised to keep your winning details confidential until you are succesfully paid your winnings. YOUR
REFERENCE NUMBER SHOULD BE STATED IN EVERY CORRESPONDENCE YOU HAVE WITH YOUR CLAIMS AGENT.

Congratulations once again from all our staff and thank you for being part of our promotions programme.
```

Figure 3: Example of a text spam

2.2 Common Spammer Tricks

2.2.1 How to Get More Victims: Email Address Harvesting

Having enough email addresses to distribute spam to is the basic prerequisite for the success of any "advertising campaign" on the Internet. Spammers must therefore adopt various high-tech tricks to identify as many email recipients as possible, often looking for publicly available emails posted on the web. According to the CAN-SPAM Act [1], advertisers are prohibited from "harvesting" email addresses from web sites in the first place. However, such activities are difficult to monitor or penalize.

An “Internet bot” is a standalone Internet application designed to perform predefined tasks. The largest group of bots is represented by web spiders, whose task is to collect information from web pages. “Positive spiders” present no problem, as they focus namely on page indexing for Internet search engines. On the other hand, “spambots” are designed to search pages and look for email addresses only. A set of robots, i.e. computers hosting the same bot, represents a “botnet”, i.e. a network of spambots, that can be utilized for coordinated attacks, namely for high-volume spam distribution. Spammers thus misuse the computers of other people on the Internet to commit their illegal activities. They are, of course, immune to blacklists. Networks consisting of thousands of computers are available on the Internet to be leased for distributing unsolicited mail.

2.2.2 Traditional Tricks Used by Spammers to Fool Spam Filters

Over time, spammers have adopted many more or less sophisticated tricks to fool spam filters, namely those that are based on statistical parameters of spam messages. Here are some examples:

- Avoidance of keywords (such as *stock*, *Viagra*, etc.),
- Frequent change in sender’s address,
- Message encoding (such as base64, commonly used for secure message transfer),
- Hashing (e.g. insertion of HTML tags into messages),
- Use of images instead of plain text (namely GIF, JPEG, and PNG).

2.2.3 New Spammer Tricks

In the following paragraphs you will find a sample survey of new tricks used by authors of “new generation” spam. Unfortunately, the list is far from exhaustive, as new approaches are constantly being developed to obfuscate spam filters.

Character hashing in words

Spammers use this trick to make typical spam keywords illegible for a filter, although they present no problem for a human brain. Should the user label such a message as spam manually, a few new keywords are added to the keyword database used by the antispam software, with no effect until re-training the filter.

Example of a message with character hashing:

```
I finlaly was able to lsoe the wieght I have been sturggling to lose for
years! And I couldn't bileeve how simple it was! Amizang pacth makes you
shed the ponuds! It's Guanarteed to work or your menoy back!
```

HTML code interleaving

HTML code is inserted into the middle of words. This presents no problem for email clients with HTML code support, as the message is kept in perfectly legible form. However, it is difficult for the filter to detect keywords split by HTML code. On the other hand, this HTML code interleaving trick is quickly losing popularity among spammers. Here is an example of an email encoded in HTML table:

```
<table cellSpacing=0 cellPadding=0 align=center border=0>
      <tr vAlign=bottom>
      <td rowSpan=2>Inc</td>
      <td rowSpan=2>reas</td>
      <td rowSpan=2>e&nbsp;S</td>
      <td rowSpan=2>exual&nbsp;Desi</td>
      ...
```

Commercial attachments in the form of Microsoft Office documents

This is a way to avoid contents analysis by spam filters altogether; the message is passed as long as it can stand an antivirus check. On the other hand, the user must be curious enough to open the attachment, which is rarely the case, as the message comes from an unknown sender, and usually contains neither body text nor subject line in order to pass the spam filter.

Keyword masking by repeating characters

Spammers try to obfuscate keywords by repeating some characters. The message remains legible for humans, but makes detection by statistical filters difficult.
Here is an example: `Buuuyyyy cheeeeaaap viaaagraaa.`

Word obfuscation by replacing characters by punctuation marks, spaces or images

Statistical spam filters typically look for certain keywords such as *Viagra*, *tablets*, or *watches*, so spammers have adopted techniques to obfuscate them by using spaces and various punctuation marks, while preserving the legibility of their messages for humans. However, heuristics (i.e. sophisticated lexical analysis) can be integrated into text-based spam filters to fight this technique.

Examples of word obfuscations:

```
\/laGr@
Need a{} Dpiloma?
sh1pp1ng //orldwide
S0ft T4bs
Ci@li$
repl1ca w4tches from r0lex
```

Use of CSS styles for color setting and/or visibility of letters

The widespread application of CSS styles for web page formatting gives spammers a new opportunity to use the same technique to format their messages and circumvent spam filters based on statistical parameters.
Example – Insertion of CSS styles into HTML tags to "encode" the word *Cialis*:

```
<span style="display: yes; display: none">g</span>C
<span style="display: yes; display: none">l</span>I
<span style="display: yes; display: none">o</span>A
<span style="display: yes; display: none">c</span>L
<span style="display: yes; display: none">s</span>I
<span style="display: yes; display: none">z</span>S
```

The only word that is actually displayed upon opening the message is "CIALIS" – a term that is notoriously known to all spam filters.

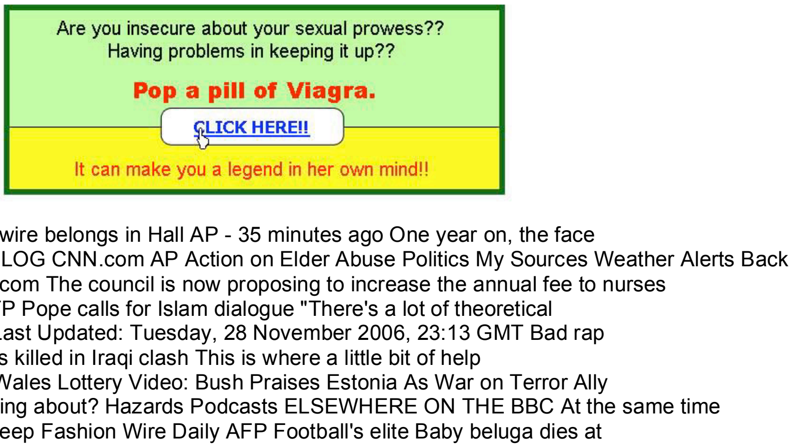
ASCII art

Spammers sometimes rely on some good old tricks, believing that they have already been forgotten. ASCII art is a good example dating back to the era of DOS systems. This is yet another way to go around the filter and push through a commercial message perfectly legible for humans. Statistical filters have very little chance in this case, as keywords can only be found in the subject line.
Example of ASCII art (quite non-commercial, but you get the idea):

```
     \|||||/
    ( o   o )
-ooO--(_)--Ooo---------------------------------------
```

Good word attacks

Spammers attack statistical spam filters by inserting "good" words into their messages. Such words can be chosen from a dictionary (*a dictionary attack*). There is a more sophisticated approach to utilize words that appear most frequently in legitimate mail, such as Reuters news, or USENET messages (such English corpora are freely available). In Figure 4 below you can see a typical spam embellished with a few pieces of news to fool statistical spam filters.


Are you insecure about your sexual prowess??
Having problems in keeping it up??
Pop a pill of Viagra.
CLICK HERE!!
It can make you a legend in her own mind!!

Russa says McGwire belongs in Hall AP - 35 minutes ago One year on, the face
live! EDITORS' BLOG CNN.com AP Action on Elder Abuse Politics My Sources Weather Alerts Back
Security SPACE.com The council is now proposing to increase the annual fee to nurses
Freeman dies AFP Pope calls for Islam dialogue "There's a lot of theoretical
CSMonitor.com Last Updated: Tuesday, 28 November 2006, 23:13 GMT Bad rap
to top ^^ Five girls killed in Iraqi clash This is where a little bit of help
28, 6:33 AM ET Wales Lottery Video: Bush Praises Estonia As War on Terror Ally
ANALYSIS Mucking about? Hazards Podcasts ELSEWHERE ON THE BBC At the same time
Victims Were Asleep Fashion Wire Daily AFP Football's elite Baby beluga dies at

hands-on situation." 'My mother was assaulted' Entertainment Search World Radio 2 Google together Mr Litvinenko's movements on 1 November, the day he fell...

Figure 4: Example of spam with "good words" inserted

2.3 Today's Spam-Blocking Techniques

2.3.1 Protecting Web Pages from Email Harvesting

Authors of web pages use various techniques to protect email addresses presented on the Internet, thus making email harvesting by robots more difficult, if not impossible. Protecting email addresses from appearing in spammers' lists is by far the best prevention.

JavaScript

JavaScripts run on the client's side and can be used to display (or change the format of) an email address upon page load when the onLoad event occurs.

Replacement of @ character by an image or another string

The @ character can be replaced by an icon representing the same, which makes email detection by robot impossible, as robots can "see" just plain text, not images. E-mail is therefore undetected.

String reverted by CSS3 cascading style sheets

Thanks to CSS3 (technology not yet supported by all web browsers), text strings (such as emails) can be reverted upon page load. For example, the original reverted string such as <<ten.niamod@eman>>, which is actually stored in the web page, is reverted to <<name@domain.net>> and displayed to the user. Cascading style sheets must be enabled in order to present the address correctly, which is the main disadvantage of this trick.

2.3.2 Blacklist Filter

A simple technique blocking unwanted email by filtering messages coming from a specific list of senders. The blacklist is usually defined by users, systems administrators, or third parties (see, for example, [3] or [4]). Blacklists include email or IP addresses. Blacklist filters check whether the address of a new message is on the blacklist; if it is, the message is rejected. Spammers routinely switch IP and email addresses to cover their tracks; therefore, the blacklist goes out of date quickly. Spammers have also overcome this strategy by infecting computers of credible users, who (unaware) downloaded viruses sending out spam in large numbers.

2.3.3 Whitelist Filter

Contrary to the above, the whitelist filter blocks out junk mail by specifying which senders to accept. Legitimate addresses are placed in a list of trustworthy senders. This method suffers from the same drawbacks as the blacklist, in addition to disabling messages from new legitimate senders.

2.3.4 Greylist Filter

It takes advantage of the fact that many spammers attempt to distribute a spam batch only once. The receiving mail server firstly rejects the message from an unknown sender and generates a failure message to the sender's server. If the message is re-sent, the greylist filter assumes the message is not a spam and puts it in the inbox, while adding the sender's address to the list of legitimate senders. Unfortunately, the greylist filter delays time-sensitive messages.

2.3.5 Fighting Image-Based Spam

Conversion into text - Optical Character Recognition (OCR)

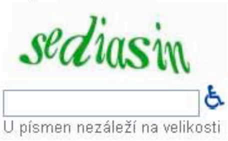
Spammers have recognized that intentional distortion of words or putting the text inside an image can easily outwit word filtering. Pre-processing of documents is therefore necessary, involving scanning of email images using character recognition techniques, applying a sophisticated text filtering method in the second phase (see below). Image filters must be trained similarly to text-based filters. OCR is applied to detect text contained in images and convert the message into a standard ASCII document. However, spammers have adopted obfuscation techniques, such as replacement of letters with numbers or other similar symbols, use of similar words, etc. Spammers are enhancing their messages by adding various noise items (such as randomly placed dots, lines or waves) on the background. Such emails remain legible for humans, but become hard to handle for OCR methods. Some OCR algorithms are language-dependent, which is a great disadvantage in the context of spam filtering.

Recurrent Pattern Detection (RPD)

Pattern detection is a typical machine learning approach based on comparing new patterns with those already detected. It can be applied in detecting image-based spam. In order to achieve a sufficient reliability level, spam must be "tracked" within the first minutes after being released, and it must be isolated regardless of the language. RPD technology is not based on image analysis, text mining or searching for keywords, but rather on comparing image patterns with those detected in unsolicited messages. Millions of messages are handled each day and stored in a so-called Signatures Repository. Client applications make queries to the central server, comparing (in real time) new emails with repository patterns. The quality of detection is gradually improved by machine learning. Language-independence is one of the best advantages of this approach.

CAPTCHA

The CAPTCHA tool (Completely Automated Public Turing test to Tell Computers and Humans Apart) [5] is frequently utilized on the web (namely in newsgroups) to tell apart pieces submitted by people from those submitted by robots, in order to prevent software applications from inserting commercial texts on the web. CAPTCHA is based on the Turing test. It presents an image containing more or less misshaped text that is usually easily readable for humans (see Figure 5), but not quite so for web spiders utilizing OCR technology. The user must repeat the character sequence to pass the test in order to make his or her submission to a blog or newsgroups, for example.



U písmen nezáleží na velikosti

Figure 5: Example of a CAPTCHA text used for opening a new Gmail account

Adaptive Image Filtering (AIF) - wavelet transform

AIF technology has been adopted to block image spam by means of the wavelet transform. This is a process that transforms a graphical image into a mathematical formula representing the original message. According to the Tumbleweed company, which authored this technology, the method can capture even those spam messages that were deliberately embellished by randomly inserted graphical elements to prevent spam filtering.

2.3.6 Analysis of Text-Based Spam

Antispam software developers fought successfully, for a time, with the help of various filtering strategies. Antispam programs scan emails and analyze keywords contained in these emails. Web sites referenced from these emails are analyzed as well. Filtering strategy is based on the use of statistical techniques. The filter must determine which words are more likely to be a part of a legitimate message rather than spam.

For example, the text spam message shown in Figure 3 above was checked by a spam filter (SpamAssassin 3.1.0.). Needed to say, the authors of this "lottery winning message" did not apply any exploits to fool the filter. Here is the extract from the spam-filter report:

```
X-Spam-Level: ***
X-Spam-Status: Yes, score=3.0 required=3.0 tests=ADVANCE_FEE_1,ADVANCE_FEE_2,
       ADVANCE_FEE_3,ALL_TRUSTED,DEAR_SOMETHING,PLING_PLING,UPPERCASE_25_50
       autolearn=no version=3.1.0
X-Spam-Report:
       * -1.4 ALL_TRUSTED Passed through trusted hosts only via SMTP
       *  1.6 DEAR_SOMETHING BODY: Contains 'Dear (something)'
       *  0.0 UPPERCASE_25_50 message body is 25-50% uppercase
       *  1.8 ADVANCE_FEE_3 Appears to be advance fee fraud (Nigerian 419)
       *  0.5 PLING_PLING Subject has lots of exclamation marks
       *  0.0 ADVANCE_FEE_1 Appears to be advance fee fraud (Nigerian 419)
       *  0.6 ADVANCE_FEE_2 Appears to be advance fee fraud (Nigerian 419)
```

Note that the filter detected several keywords frequently occurring in spam messages, in addition to excessive use of uppercase characters (more than 25%), and multiple exclamation marks in the Subject line.

3 Methodology

Content-based filters apply various techniques, from a simple handmade list of words frequently used in spam messages up to sophisticated machine learning methods. As mail filtering is actually a classification task, all classification methods can be involved. In this section we describe the techniques we have implemented to fight spam.

Our primary goal was to examine the antispam abilities of the methods we have partly designed and partly modified for this application area. These are namely our Itemsets Method, originally designed for document categorization, the LSI Method modified by us for spam-filtering purposes, and another traditional method, the Naïve Bayes classifier. All these methods must be trained initially using a collection of messages, a priori labeled as either spam or legitimate. All these methods can be trained individually on a per user basis, in addition to being adaptable in run-time (i.e. they have the ability to learn).

3.1 Spam Collections for Spam-Filter Testing

SpamAssassin public mail corpus [6] is a selection of mail messages suitable for testing spam filtering systems. It contains slightly more than six thousand messages (legitimate messages posted to public forums), with about a 31% spam ratio.

PU123A [7] are four public corpora based on private mailboxes. These are relatively small collections of spam messages and legitimate emails (encoded).

Ling-spam [8] is a mixture of 481 spam messages and 2412 messages sent via the Linguist list, a moderated (hence, spam-free) list about the profession and science of linguistics.

3.2 The Naïve Bayes Filter

The Naïve Bayes filter examines a set of known spam emails and a set of emails known to be legitimate. After teaching itself the vocabulary used by spammers from this known list, it will use Bayesian probabilities to calculate whether a message is spam.

This filter is based on the Bayes theorem. Applied to spam, it states that the probability of an email being spam is equal to the probability of finding the same words in this email and spam, times the probability that any email is spam, divided by the probability of finding those words in an arbitrary email. Expressed in a conditional probability formula:

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \times \Pr(A)}{\Pr(B)}$$

Pr(A|B) is the probability that a message is spam should it contain the word B.
Pr(B|A) is the probability of the word B in spam. This value is computable from the training collection.
Pr(A) is the probability that the email is spam (i.e. the number of spam messages divided by the number of all emails in the training collection). No information on B is used.
Pr(B) is the probability of word B in the collection.
Each word in the email contributes to the e-mail's spam probability. This probability is computed across all words in the email. Should the total exceed a certain threshold, the message is blocked out.

3.3 The Itemsets Filter

The Itemsets method is our original categorization method for short documents developed in 1999. Application of this method for spam filtering was presented at ELPUB [9]. We have suggested potential application of itemsets for categorization in 2000 (see [10]).

In the training phase we search for sets of characteristic terms (words or word sets) for each category (categories being spam and ham). The itemset $\prod_j$ is characteristic for class T_i if its weight w_{ij} is sufficiently high. Let us denote $D\prod_j$ the set of messages containing the itemset $\prod_j$ and DT_i the set of documents in class T_i, where $i \in$ {spam, ham}. From the different approaches taken, the best results were achieved using the following formula for computing itemset weights (j-th itemset for the spam/ham category):

$$w_{ij} = \frac{\left|D\Pi_j \cap DT_i\right|}{\left|DT_i\right| \times \left[1 + \left|D\Pi_j\right| - \left|D\Pi_j \cap DT_i\right|\right]} \qquad i = 1, 2$$

The terms with the highest weights for class T_i form the set of C_i's characteristic terms. In the classification phase, a document is assigned to class T_i, for which the following sum is the highest:

$$SumT_i = \sum_{j=1}^{|C_i|} w_{ij}$$

3.4 The LSI Filter

Latent semantic indexing (LSI) has been used in information retrieval (IR) applications since the beginning of the 1990s. Compared to other traditional IR methods, this approach can guarantee higher recall, with detrimental impact on precision. In general, LSI proves efficient for collections of heterogeneous documents that use different terms to represent the same concept. On the other hand, this technique is not suitable for homogeneous document collections (as far as terms are concerned), as it introduces additional noise to the collection.

LSI is (as with Itemsets) based on space reduction. LSI is an application of the SVD (singular value decomposition) mathematical theory in the area of information retrieval. In this method we decompose the "term by document" matrix A (i.e. matrix of words × emails) into three matrices, say T, S, D.

$$A_{t \times d} = T_{t \times n} S_{n \times n} (D_{d \times n})^T,$$

where n = min(t, d), with item $a_{t\,d}$ representing the frequency of the term t in document d.

T and D are orthonormal, and S is a diagonal matrix containing singular values in descending order. We can choose some k < n and approximate A by A' in the reduced k-dimensional space (i.e. we constrain T, S, D in only the first k-columns, thereby obtaining T', S', D'). It has been proven that approximation of A by A' is the optimal projection of terms and documents into the new reduced space.

In the training phase we decompose matrix A and evaluate the matrix $B = S'D'^T$. Classification of a message consists of a correlation evaluation $C = (T'^T m)^T B$, where m is the vector of terms (words) of the message being classified. Consequently, we find the global maximum, i.e. the document demonstrating the highest semantic similarity.

LSI in brief

The training phase:

- Compute singular value decomposition (SVD) on matrix A (documents × terms),
- Compute matrix B (using S and D matrices representing the reduced space) and save B and T matrices (representing the reduced space).

The classification phase:

- Construct the query q (i.e. prepare the e-mail to be classified),
- Compute correlation coefficient using the original documents C = (TTq)TB, looking for the global maximum, i.e. the document whose semantic similarity is the highest.

4 Results

We have tested the above methods on the PU1 email collection [7]. It contains 481 spam messages and 618 legitimate emails, in total including 849,977 term positions (24,745 unique terms). Lemmatization and stop-list application techniques were utilized if they were found useful. The collection was split into ten parts. Nine were used for training and one for testing. Our spam classifier returns a text string, which is inserted into the message header. The string includes detailed information to decide whether the message should be moved to the spam folder (see below):

```
X-SPAM: ********** (3/3)
>> Itemsets: ********** (100.0%)
>> LSI: ***** (50.39196180000235%)
>> SVM: ******** (78.4891665%)
>> Pattern matching: ********** (100.0%)
>> Black&White: ***** (50.0%)
```

This means that according to the Itemsets filter, the message is certainly a spam (100%). The sender was found in neither black nor white lists, therefore, we have insufficient information to decide based on this criterion (thus 50%). Certainty level in percent is also converted to star signs (*), which is utilized for filter personalization.

The results of our practical testing are shown in the tables below. Please note that substantially better results can be achieved in real-life filter application by applying additional heuristic techniques. In the tables below, FPI means False Positive Identification, and FNI stands for False Negative Identification.

FPI = (#ham as spam) / #ham, i.e. the proportion of legitimate messages deleted by mistake.
FNI = (#spam as ham) / #spam, i.e. the proportion of spam passing through the filter.

	dim = 50	dim = 100	dim = 150	dim = 200
FPI [%]	10.32	9.78	11.96	11.41
FNI [%]	11.72	10.34	8.27	8.27

Table 1: LSI-based spam filter results

Table 1 above shows LSI-based classifier results. We observed the impact of reduced-space dimension on the classification accuracy and effectiveness. According to Table 1, the best results were achieved when reducing the space to the dimension 50 - 100.

	1-itemsets							
	100	200	**300**	400	500	700	1000	1500
FPI [%]	0.49	0.49	**0.52**	2.21	2.19	2.74	2.17	2.17
FNI [%]	11.05	9.66	**4.17**	4.17	2.78	2.78	2.08	2.08

Table 2: Itemsets-based spam filter results

Table 2 above shows Itemsets-based classifier results. We observed filter accuracy and effectiveness depending on the number of 1-itemsets used for classification. According to our experiments, a classification category is relatively well described by approx. 300 characteristic terms.

	NB	Itemsets	LSI
FPI [%]	1.08	0.52	9.24
FNI [%]	15.81	4.17	11.72

Table 3: Results of the spam filters implemented

Table 3 above shows the best results achieved by our implementation of the Naïve Bayes classifier, Itemsets classifier and LSI-based classifier. Crucial is the FPI rate (i.e. the proportion of legitimate mails deleted by mistake), where the results of the Itemsets classifier were relatively acceptable in this experimental setup (not in real life – hardly anyone would accept the deletion of a good message in every 200 received). It is necessary to note that even the worst results of the LSI-based classifier are relatively good – although it deletes approx. 10 % of legitimate mail, it also filters out 90% of spam messages.

5 Discussion

According to Symantec's Antispam Technology Brief [11], competitive spam filters are those with a false positive rate (i.e. legitimate messages deleted by mistake) of 1 in 100,000, i.e. accuracy of 99.999%. Accuracy for the best in class filter should be as high as 99.9999% (i.e. one false positive in 1 million messages).

Rates for effectiveness (i.e. proportion of spam messages detected) are not so strict, corresponding to 85% for competitive filters and over 95% for best in class filters.

Looking at the above ranking by Symantec, our spam filters are competitive in terms of effectiveness (especially in the case of the Itemsets-based filter), but far from competitive in terms of accuracy, as too many legitimate messages are deleted by mistake. Nonetheless, we have applied just a "plain" text classifier with no heuristics implemented. For example, we pay no attention to random character hashing, repeated characters, insertion of HTML tags, or replacement of letters by images.

In general, the efficiency of spam filters is also strongly influenced by "good word attacks" (see section 2.2.3 above). Please note that in the case of the popular Naïve Bayes filter, an attacker can get as much as 50% of currently blocked spam past the filter by adding 150 words or fewer [12].

The testing collection used for experiments also has a strong impact on classification results. Statistical filters that demonstrate exceptionally good results are often tested on single-topic collections, such as email collections harvested from newsgroups on the Internet. It is therefore easier to distinguish spam from legitimate messages, as all legitimate mails pertain to a relatively narrow topic, featuring characteristic words typical for this topic.

6 Conclusions

Additional information on spam filtering can be found at http://spam.abuse.net and http://spam.getnetwise.org. Various anti-spam filters are freely available on the Internet, e.g. http://spammotel.com, http://www.hms.com/spameater.asp, and http://www.mailwasher.net. A useful collection of links to various spam filters and other tools can be found at http://www.spamarchive.org. A summary of our work can be found at http://www.textmining.cz.

Our next investigation will focus on the use of compression algorithms for spam filtering. Although this novel approach may not prove effective for some categories of spam, we believe that taking this new road will be interesting. It appears that the compression-based technique may surpass some traditional machine learning systems [12, 13]. Fighting image-based spam is another field we want to concentrate on, as this spam category is gaining vast popularity and a lot of work is yet to be done. The fight against spam is not lost – as long as we remain one step ahead of its distributors.

Acknowledgements

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

Notes and References

[1] The CAN-SPAM Act: Requirements for Commercial Emailers, available at: http://www.ftc.gov/bcp/conline/pubs/buspubs/canspam.pdf

[2] OECD Task Force on Spam, available at: http://www.oecd-antispam.org/

[3] SpamCop Blocking List, available at: http://www.spamcop.net/bl.shtml

[4] Distributed Sender Blackhole List (DSBL), available at: http://dsbl.org/main

[5] The Captcha Project, available at http://www.captcha.net/

[6] SpamAssassin Public Mail Corpus, available at: http://spamassassin.apache.org/publiccorpus/

[7] PU123 Public Corpora, available at: http://www.aueb.gr/users/ion/data/

[8] Spam Corpora, available to download at: http://www.iit.demokritos.gr/skel/i-config/downloads/

[9] HYNEK, J.; JEŽEK, K. 2002. Use of Text Mining Methods in a Digital Library, pp. 276-286. In: *Proceedings of the Sixth International Conference on Electronic Publishing – elpub2002 Karlovy Vary, Czech Republic,* Joao A. Carvalho, Arved Hübler, Anna A. Baptista (Eds). Verlag für Wissenschaft und Forschung Berlin, Germany, ISBN 3-897-0035

[10] HYNEK, J.; JEŽEK, K. 2000. Document Classification Using Itemsets, pp. 97-102. In: *Proceedings of 34th Spring International Conference MOSIS 2000*, *Rožnov pod Radhoštěm, Czech Republic*, J. Zendulka (Ed.). MARQ, Czech Republic, ISBN 80-85988-45-3

[11] Antispam Technology Brief: "Filtering Technologies in Symantec Brightmail Antispam 6.0", available at: http://www.symantec.com

[12] LOWD D.; C. MEEK. Good word attacks on statistical spam filters. In: The Conference on Email and Anti-Spam (CEAS), 2005. Available at: http://www.ceas.cc

[13] BRATKO, A.; CORMACK, G.; FILIPIC, B.; LYNAM, T.; ZUPAN, B. Spam filtering using statistical data compression models. *Journal of MachineLearning Research 7* (Dec. 2006).

[14] GOODMAN, J.; CORMACK, G.; HECKERMAN, D. Spam and the Ongoing Battle for the Inbox. In: *Communications of the ACM*, February 2007, Vol. 50, No. 2

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

Irene Buonazia; M. Emilia Masci; Davide Merlitti

Laboratorio LARTTE, Scuola Normale Superiore di Pisa
Piazza dei Cavalieri 7, 56126, Pisa, Italy
e-mail: i.buonazia@sns.it; e.masci@sns.it; d.merlitti@sns.it

Abstract

In September 2004 the Italian Ministry of Cultural Heritage and Activities (MiBAC) committed to Scuola Normale Superiore di Pisa (SNS) the scientific and technical project for the Italian Culture Portal. The project was delivered during 2005, together with a prototype which had the function to verify and test the project's issues and has been provided as reference for the implementation. In 2006 MiBAC selected, through a public competition, the IT company Reply for developing the Portal and Electa Napoli for providing the editorial office and plan. The Portal is now under development and will be delivered during 2007. SNS is presently working as consultant of MiBAC to give support to the whole staff employed in the fulfilment of the Portal and to help in the difficult activity of the mapping of various resources to be harvested and published in the Portal. This paper illustrates the project of the Italian Culture Portal delivered by SNS, describing in particular the solutions adopted for guaranteeing the interoperability, accessibility and usability tasks. One of the main objectives of the Portal is to offer open access to information on the "Italian Culture", which is a wide, evolving concept comprehensive of tangible and un-tangible cultural patrimony. Resources pertaining to this vast and complex domain are therefore of very different kinds and formats, moreover, they are codified following different schemas. For guaranteeing the interoperability among such cultural resources, a Dublin Core Application Profile has been specifically designed for the Portal. An official publication of this AP is currently under development: it has been recently refined and improved on the basis of the first mapping experiences and is anticipated in this contribute in this updated form.

Keywords: open access; interoperability; metadata standards; application profile

1 Introduction

The scientific and technical project for the Italian Culture Portal was promoted by the Italian Ministry of Cultural Heritage and Activities (MiBAC) and delivered by Scuola Normale Superiore di Pisa (SNS) during 2005 [1]. At the moment SNS is working as a consultant for MiBAC to flank the company which is carrying out the Portal, which will be named "CulturaItalia".

The main mission of the Italian Culture Portal is to communicate to different kinds of users the whole ensemble of Italian culture, as a media conceived for the diffusion of knowledge, promotion and enhancement of cultural heritage. Thus, CulturaItalia will offer access to the existing resources on cultural contents and will give more exposure to the vast amount of websites pertaining to museums, libraries, archives, universities and other research institutions: users will access resources stored in various repositories browsing by subjects, places, people and time. It will be possible to visualise information from the resources and to further deepen the knowledge directly reaching the websites of each institution.

The Portal will harvest metadata from different repositories and will export metadata to other national and international Portals. It will also provide contents created and managed by an editorial office, to offer updated news on the main cultural events and to provide thematic itineraries for a guided navigation through the harvested contents.

Resources originating from various data-sources will remain under the control of institutions responsible for their creation, approval, management and maintenance: data will not be duplicated into the Portal's repository and will be retrievable through a unified and interoperable system.

In order to guarantee the interoperability of various kinds of cultural resources and to allow retrieval and indexing functions on their contents, a specific Dublin Core Application Profile has been designed on the basis of the complex domain of "Italian Culture". The PICO AP (so called from the Project's acronym) [2], which will be exposed in this paper, has been currently reviewed and improved according to the first mapping experiences made by SNS on some repositories, whose contents have been chosen to be harvested by CulturaItalia. The PICO AP will be soon published on a PURL (Persistent Uniform Resource Locator) [3].

2 Methodology

The project for CulturaItalia has been developed through the following steps:

- users and domain analysis
- definition of user scenarios and use cases
- overall architecture design
- content analysis
- analysis of the state of the art on descriptive metadata standards
- design of the metadata schema (PICO AP)
- design of the user interface
- project prototype

The identification of potential users of the Portal moved from the requirements issued by MiBAC, which pointed out that the Culture Portal should be distinguishable in its domain and functionalities both from the official web site of MiBAC [4], oriented to people in charge of management and preservation of Cultural Heritage, and from the Portal for Italian Tourism.

Moreover, potential addressees of a cultural portal have been identified with the analysis of some of the most important European and international portals (e.g. French www.culture.fr, British http://www.24hourmuseum.org.uk/), websites of cultural institutions such as museums, theatres, universities, etc.

Eight user scenarios have been written, describing eight different approaches to the Portal, by different kind of users. Scenarios described the following users and functionalities:

- Foreign tourist: language selection, access from the map, browsing and e-booking;
- General user: disambiguation of query results, use of contents suggested by the editorial staff and linked to results of user's query;
- Italian teacher with partial visual deficit: accessible set up, simple and advanced search, registration to the Portal, submission of a comment to the editorial staff;
- Foreign researcher: free and advanced search, access to the web site identified through the Portal;
- Journalist: search amongst cultural events, purchase of printable pictures, registration, download;
- Publishing house: search amongst images, contact for banner exchange;
- Tourists with motion deficit: browsing from place and events, visualization on the map, participation to forum;
- Italian high school student: simple search, print function and e-commerce tools.

Adopting UML (Unified Modelling Language), such descriptive scenarios have been transformed in use cases diagrams, which identified:

- Actors, human (different final users both of the front end and of the back office) and IT components;
- inter-actions between actors and the system from the first query to the final result.

UML has been useful also to improve cooperation in a staff composed by IT developers as well as cultural domain experts, overcoming the gap of different languages. On the basis of main functionalities identified by the user requirements, the core components of the System Architecture have been designed; as the project should be used as a -non mandatory- feasibility study for the final development, costs and benefits of some existing systems and components have been considered.

Moving from the analysis of the contents foreseen for the Portal, the best solution has been identified in an harvesting procedure. A study of the state of the art collected different metadata standards and categories for describing cultural resources, such as Dublin Core, VRA -Visual Resources Association, CDWA- Categories for the Description of Works of Art, CIMI core set, EAD- Encoded Archival Description, MARC- Machine-Readable Cataloguing format, CIDOC-CRM, etc. Moreover, most relevant thesauri (from UNESCO to ULAN and AAT) concerning cultural domain, have been taken into consideration. This analysis served to decide to adopt a specific DC application profile, which will be in depth described later. Finally the user interface has been designed, specially focussing on the functionalities of searching and browsing.

The Portal is currently under development. Reply S.p.A. is developing the technical system. The editorial staff, under MiBAC supervision, is preparing contents and identifying new providers. SNS is flanking MiBAC in testing functionalities and interfaces of the system, and works as consultant for identifying new content providers and data sources, analysing the data models adopted by each provider, defining mappings to the PICO AP, monitoring and improving results of harvesting procedures.

3 Analysis: Users' Identification, Mission and Domain

The project is based on the analysis and definition of the expected users' target, consequently on the identification of users' needs and requirements, of the mission of the Portal and of its domain, which necessarily corresponds to the domain of Italian Culture. The target of the Portal will be Italian and foreign users, such as:

- tourists and people interested in, and passionate of, culture
- business users (publishers, merchandising, etc.)
- young people, from primary to high school
- culture professionals such as scholars, museums curators, researchers, etc.

Special contents and services will be created for each kind of user. It is important to notice that each user can be a person with physical or cognitive disabilities: the Portal must be accessible also for those categories. The mission of CulturaItalia identifies the following goals:

- To promote Italian culture and heritage in Italy and abroad:
  - to integrate Italian culture in the international contest;
  - to attract web users toward cultural themes;
  - to give visibility to Italian cultural institutions;
  - to support activities and projects focused on culture;
  - to integrate cooperation between public and private institutions.

- To promote and integrate existing resources:
  - to offer an index of Italian cultural resources and heritage;
  - to create flexible and scalable relations between resources;
  - to identify existing digital resources, websites, databases, digital libraries;
  - to allow interoperable queries on indexed subjects, places, events, and people.

The Domain of "Italian Culture" is a wide concept, conceived in different ways. MiBAC is responsible for preservation, management, research and exploitation of the Italian cultural patrimony, which is composed by:

- Tangible heritage:
  - architectural and environmental objects;
  - artworks and collections;
  - manuscripts, edited books and the current literature;
  - archaeological and demo-ethno-anthropological objects;
  - contemporary art and architecture.

- Un-tangible heritage:
  - music;
  - dance and theatre, circuses and street performances;
  - cinema;
  - humanities;
  - scientific culture.

4 Harvesting of Contents

CulturaItalia will give integrated access for information pertaining to the domain of “Italian Culture”, as it has been defined in the previous chapter. Resources coming from various data-sources will not be duplicated into the Portal’s repository. On the contrary, it will offer an index of those contents by harvesting metadata pertaining to their data.

Before being harvested, metadata will be mapped into one metadata schema, which will permit the indexing, browse and query functions on the whole ensemble of harvested contents. Metadata will be harvested using OAI-PMH [5]. This protocol allows the metadata migration from content providers to one or more harvesters, adding services as indexing system or automatic classification. OAI-PMH uses HTTP protocol for data transfer and XML for data coding.

Each institution responsible for contents to be harvested will establish, together with MiBAC, which data will be accessible from the Portal, as some resources or part of them could contain confidential information that shouldn’t be published.

5 The PICO DC Application Profile

Contents coming from external data-sources will be imported in the Portal trough the harvesting of metadata and the mapping in one metadata schema. As the Portal will join different kinds of contents, it seamed unsuitable to use a data model with predefined entity types. For guaranteeing system’s scalability, a flexible solution has been preferred, which consists in the designing of a unique metadata schema: to respect world wide used standards, the Italian Culture Portal will adopt a metadata set based on DC (Dublin Core) standard [6].

This standard is very used because it consists in one scheme that can be applied to every kind of resource, distinguished by the element <dc:Type>. Anyway, it is not really efficient for cultural resources because, as the DC Element Set (the so called ‘Simple DC’) is very restricted, many different information must be grouped into one element [7]. For this reason, in the last years Dublin Core Metadata Initiative (DCMI) divulged the ‘Qualified DC’ schema, which refines DC Element Set using Element Refinements and supporting Encoding Schemes, to attribute to a given property the value selected from a controlled vocabulary, a thesaurus, or an ontology [8].

Thanks to Dumbing Down algorithms, now developed in XML, in the data sharing with a system that supports Simple DC, it is possible to reduce Qualified DC values into Simple DC values. With this process, there is a minimum loss of information and more possibility to obtain a significant retrieval. At the same time, interoperability between repositories based on Dublin Core is assured.

DCMI suggests to institutions and research groups to develop DC Application Profiles for specific applications and domains, designing schemas which can join:

- All, or a selection of, DC Elements and Refinements;
- Elements from one or more element sets;
- Elements from locally defined sets [9].

A DC Application Profile has been designed for the Portal of Italian Culture on the basis of recommendations, documents and samples published by DCMI, in order to define further extensions specially conceived to retrieve information pertaining to Italian culture. This application profile could be further expanded for harvesting eventually unexpected contents in the future, by adding Refinements and Encoding Schemes that could be necessary for data retrieval.

The PICO AP has been designed by I. Buonazia, M. E. Masci and D. Merlitti (SNS working group on metadata, supervised by U. Parrini). It has been recently improved on the basis of the first mappings performed on some data-models or metadata schemas related to contents to be harvested by CulturaItalia. An official publication is currently under development. It will be edited on a PURL, following the DC AP Guidelines [10]. This DC Application Profile joins in one metadata schema:

- All DC Elements;
- All DC Element Refinements and Encoding Schemes from the Qualified DC;
- other refinements and encoding schemes specifically conceived for the CulturaItalia domain.

Therefore, the following namespaces are included into this metadata schema: ‘dc:’, ‘dcterms:’, ‘pico:’. In the following sections the additional extensions and qualifiers of PICO AP to the Qualified DC are exposed in detail.

5.1 Extensions to DCMI Type Vocabulary

The resource’s type has been further extended with the PICO Type Vocabulary, which joins the types ‘Corporate Body’, ‘Physical Person’ and ‘Project’, to the types foreseen in the DCMI Type Vocabulary [11].

5.2 Qualifiers Added to the Qualified DC Element Refinements and Encoding Schemes in the PICO AP

In the following table are resumed the Element Refinements and Encoding schemes added in PICO AP to the Qualified DC: all DC qualifiers are implicitly included in the AP. In the right column qualifiers added by the PICO AP are specified for each DC Element, indicated in the left column.

DC ELEMENTS	**PICO AP QUALIFIERS**

dc:creator	label= Author
	definition= Any living or dead physical person, any corporate body and institution, responsible for the creation of a resource. It can be a writer, a painter, an architect, a musician, a photographer, a collectioner (as the author of the collection).
	comments= It is recommended to use Author instead of Creator when the creator of the resource can be mentioned with a proper name.
	type= element-refinement

	label= Commissioner
	definition= Any living or dead physical person, corporate body and institution, responsible for the commission, the order and/or the funding of the design of a resource.
	type= element-refinement

	label= ULAN - Union List of Artist Names
	definition = Controlled vocabulary by The Getty Research Institute. Reference at: http://www.getty.edu/research/conducting_research/vocabularies/ulan/.
	comments= It is recommended to use DCSV syntax for expressing ULAN values. For the name, indicate the ‘Preferred Name’. For the value, use the ID code assigned by ULAN. E.g. name=Cerquozzi, Michelangelo; value=500007713.
	type= encoding scheme

dc:subject	label= Theaurus PICO
	definition= Thesaurus composed by hierarchically structured keywords for indicating the topic of all the resources included into CulturaItalia. This ontology includes terms for assigning the resources to the index and to the themes menu of the Portal.
	type= encoding scheme

	label= UNESCO Thesaurus
	definition= Thesaurus for indicating the topic of resources on education, culture, natural, human and social sciences, communication and information. Multilingual: English, French, Spanish. Reference at: http://databases.unesco.org/thesaurus/.
	type= encoding scheme

	label= AAT (Art and Architecture Thesaurus)
	definition= Thesaurus defined by Getty Research Institute for indicating the topic of resources pertaining to art and architecture objects. Reference at: http://www.getty.edu/research/conducting_research/vocabularies/aat/.
	comments= It is recommended to use DCSV syntax for expressing AAT values. For the name, indicate the ‘Preferred Name’. For the value, use the ID code assigned by AAT. E.g. name=doric; value=300020111.
	type= encoding scheme

	label= ICONCLASS definition= Taxonomy of the iconographic subjects for the Western Art, from Medieval to the Contemporary Art. Multilingual: English, German, Italian, French, Finnish. Reference at: www.iconclass.nl. comments= It is recommended to use DCSV syntax for expressing ICONCLASS values. For

DC ELEMENTS	PICO AP QUALIFIERS
	the name, indicate the subject name, for the value, use the related code. E.g. name=angels fighting against other evil powers; value=11G34. type= encoding scheme

dc:description	label= Information
	definition= Information about the resource, as opening and closing ours.
	comments= It is generally used for resources with type: CorporateBody.
	type= element-refinement

	label= Contact
	definition= Information about contacts related to the resource.
	comments= Examples of Contact include: telephone number, fax, address, e-mail address, etc. It can't be used for indicating contacts of people which contribute to the resource.
	type= element-refinement

	label= Services
	definition= Services offered by the resource. E.g. cafeteria or restaurant services, services for unpaired people, laboratories and activities, extra.
	comments= It is generally used for resources with type: CorporateBody.
	type= element-refinement

dc:publisher	label= Distributor
	definition= Any living or dead physical person, any corporate body and institution, responsible for the distribution of an edited or published resource.
	comments= The usage of this term is recommended for resources as musical records and films.
	type= element-refinement

	label= Printer
	definition= Any living or dead physical person, any corporate body and institution, responsible for the print of an edited or published resource. This term comprehends both printers of physical (books, journals, images, etc.) and digital (CD, DVD, etc.) resources.
	type= element-refinement

dc:contributor	label= Editor
	definition= Any living or dead physical person, any corporate body and institution, responsible for the making, editing or organisation of the resource. E.g. the editor of a volume of proceedings or of an exhibition.
	comments= The usage of this term is recommended for resources with type: Text or Event.
	type= element-refinement

	label= Performer
	definition= Any living or dead physical person, which contributes to the execution of the resource by acting a performance, with reference to some entertaining events in particular. E.g. an actor, dancer, singer, musician, etc.
	type=element-refinement

	label= Producer
	definition= Any living or dead physical person, any corporate body and institution, responsible for the artistic and/or economic production of the resource. This term is used for producers of cinema, music, theatre, etc.
	type= element-refinement

	label= Responsible
	definition= Any living or dead physical person, any corporate body and institution, responsible for the management, organisation, administration, etc. of the resource or of a part of it. In some cases it coincides with the contact person, whose contacts are indicated for people who are looking for information about the resource. E.g. the responsible of a project or of one of its work packages, a museum director, the director of a university or of a department, etc.
	comments= For resources catalogued following ICCD (Central Institute for the Catalogue and the Documentation – Italy) schema, it indicates the cataloguing responsible.
	type= element-refinement

	label= Translator
	definition= Any living or dead physical person who made the translation of the resource
	type= element-refinement

DC ELEMENTS	PICO AP QUALIFIERS
	label= ULAN - Union List of Artist Names (see above, under dc:creator)

dc:type	label= PICO Type Vocabulary
	definition= Controlled vocabulary which includes some resource types specifically conceived for the Italian Culture Portal domain: Corporate Body, Physical Person, Project. Those types are not foreseen by the DCMI Type Vocabulary.
	type= encoding scheme

	label= CDType - Collection Description Type Vocabulary
	definedBy= http://purl.org/cld/terms/
	definition= A list of types that categorize a collection.
	comments= Reference at: http://www.ukoln.ac.uk/metadata/dcmi/collection-application-profile/#cldCLDT
	type= encoding scheme

dc:format	label= Material And Technique
	definition= The material of the object and of its support and the technique of execution of a resource with type: PhysicalObject
	type= element-refinement

dc:identifier	label= ISBN - International Standard Book Number
	definition= The International Standard Book Number is an uniform and persistent identifier for a given title or for the edition of a title pertaining to a given publisher. Reference at: http://www.isbn.it/.
	comments= It is generally used for resources with type: Text.
	type= encoding scheme

	label= ISSN - International Standard Serial Number
	definition= The International Standard Serial Number is the international identifier for serial publications such as printed or digital newspapers and periodicals. Reference at: http://www.issn.org/.
	comments= It is generally used for resources with type: Text.
	type= encoding scheme

dc:relation	label= Preview
	definition= Any form of abstract, reduction, image, video streaming used as anticipation of the resource.
	type= element-refinement

	label= Promotes
	definition= The described resource promotes and/or organizes the referenced resource.
	type= element-refinement

	label= is Promoted By
	definition= The described resource is promoted and/or organized by the referenced resource.
	type= element-refinement

	label= Manages
	definition= The described resource manages with different responsibilities (scientific, administrative, technical, etc.) the referenced resource.
	type= element-refinement

	label= Is Managed By
	definition= The described resource is managed with different responsibilities (scientific, administrative, technical, etc.) by the referenced resource.
	type= element-refinement

	label= Is Owner Of
	definition= The described resource owns the referenced resource.
	type= element-refinement

	label= Is Owned By
	definition= The described resource is owned by the referenced resource.
	type= element-refinement

	label= Produces definition= The described resource produces in its physical, or administrative, or any other issue, the referenced resource.

DC ELEMENTS	PICO AP QUALIFIERS
	comments= It is recommended to express value as URI. This relation should be used when it is possible to point at another resource, pertaining to the work produced by the described resource. type= element-refinement

	label= Is Produced By
	definition= The described resource is produced in its physical, or administrative, or any other issue, by the referenced resource.
	comments= It is recommended to express value as URI. This relation should be used when it is possible to point at another resource pertaining to who produced the described resource; otherwise it is recommended to use Producer.
	type= element-refinement

	label= Performs
	definition= The described resource performs, directly participating (e.g. as actor or musician) to, the referenced resource.
	comments= It is recommended to express value as URI. This relation should be used when it is possible to point at another resource pertaining to the work performed by the described resource.
	type= element-refinement

	label= Is Perfomed By
	definition= The described resource is performed by the referenced resource.
	comments= It is recommended to express value as URI. This relation should be used when it is possible to point to another resource, pertaining to who performs the described resource; otherwise it is recommended to use Performer.
	type= element-refinement

	label= Is Responsible For
	definition= The described resource is anyhow responsible for, or is the contact person of, the referenced resource.
	comments= It is recommended to express value as URI. This relation should be used when it is possible to point to another resource, the described resource is responsible for.
	type= element-refinement.

	label= Has As Responsible
	definition= The described resource has as responsible and/or contact person the referenced resource.
	comments= It is recommended to express value as URI. This relation should be used when it is possible to point at another resource, pertaining to who is responsible for the described resource. Otherwise it is recommended to use Responsible.
	type= element-refinement

	label= Contributes To
	definition= The described resource contributes anyhow to the referenced resource.
	comments= It is recommended to express value as URI. This relation should be used when it is possible to point at another resource, pertaining to something/someone that receives contributions from the described resource.
	type= element-refinement

	label= Has As Contributor
	definition= The described resource is produced, managed, organized with the contribution of the referenced resource.
	comments= It is recommended to express value as URI. This relation should be used when it is possible to point at another resource, pertaining to something/someone that is giving contributions to the described resource. Otherwise it is recommended to use Contributor.
	type= element-refinement

	label= Digitises
	definition= The described resource is responsible of the digitisation of the referenced resource.
	comments= It is recommended to express the value as URI. This relation is generally used for resources with type: Physical Person or Corporate Body.
	type= element-refinement

	label= Is Digitised By

DC ELEMENTS	PICO AP QUALIFIERS
	definition= The described resource is digitised by the referenced resource. type= element-refinement

	label= Anchor
	definition= Reference to the URL of the web-page publishing the resource described by the metadata record.
	comments= It is recommended to use DCSV syntax as follows: title= e.g. Website of the Scuola Normale Superiore of Pisa; URL=http://www.sns.it
	type= encoding scheme

dc:coverage	label=Date of Birth
	definition= Date of Birth pertaining to resources with type: Physical Person.
	type=element-refinement

	label= Date of Death
	definition= Date of Death pertaining to resources with type: Physical Person.
	type= element-refinement

	label= Place of Birth
	definition= Place of Birth pertaining to resources with type: Physical Person.
	type= element-refinement

	label= Place of Death
	definition= Place of Death pertaining to resources with type: Physical Person.
	type= element-refinement

	label= ISTAT Code
	definition= Code assigned by Istituto Nazionale di Statistica italiano (Italian National Institute for Statistics), which identifies inhabited places in the Italian territory. Reference at http://www.istat.it/strumenti/definizioni/comuni/
	comments= ISTAT code must be composed by 8 numbers: first 2 identify the Region; following 3 identify Province; final 3 identify the City (o smaller inhabited place) within the Province.
	type=encoding scheme

	label= Postal Address
	definition= Postal address of a resource with type: Physical Object or Corporate Body. It is expressed with the DCSV syntax as specified in the following example: PlaceType=Via /piazza / Largo, etc.; PlaceName=Dante; PlaceNumber=26; ZipCode - CAP=57124; City=Roma; Province=RM; Region=Lazio; Country=Italia.
	type=encoding scheme

6 User Interface

CulturaItalia will publish different kinds of contents:

- static contents: Head and logo, access to multilingual versions, credits, contact information, mission, site map, copyright;
- dynamic contents, from CMS: news, itineraries, focus, press release, forum, FAQ, newsletter, specific areas (e.g.: young users);
- dynamic contents, from harvesting: metadata harvested from external repositories;
- business logic contents, depending on the user session: search results, bookmarks, etc.;
- user inputs: layout personalization controls (font, contrast, colour), registration area to access in a private area to save bookmarks, annotate events in agenda, etc.

The interface will allow data retrieval on those contents trough different possibilities for searching and browsing. User will access contents through three kinds of searches:

- free search: user composes one or more words, using Boolean syntax;
- advanced search: user refines the query in the catalogue, selecting if the item to be retrieved is “place”, “person”, “event”, or “object”;
- geographic search, selecting a place on a list or on a map related to a GIS system.

It will be possible to browse the catalogue through the Main Menu or the Theme Menu. According to the 4 High Level Elements of DC Culture, defined by Aquarel project and approved by MINERVA project, the Main Menu of the catalogue is structured in:

- Who: people, institutions, administration offices, museums, archives, libraries, universities, etc.;
- What: art objects, monuments, documents, books, photos, movies, records, theatre and music productions, etc.;
- When: contents retrievable trough temporal periods;
- Where: browse by region, province, town, on a controlled list or directly on a GIS.

User will browse the catalogue using a 'facettes' system: he can start the query from one of the four elements and further refine the results range. A simplified alternative for browsing is the Themes menu. It groups the resources according to the following arguments: Archaeology, Architecture, Visual Arts, Environment and Landscape, Cinema and Media, Music, Entertainment, Traditions, Humanities, Scientific Culture, Education and Research, Libraries, Literature, Archives, Museums, Exhibitions.

The Portal will not publish only resources harvested from external repositories, but will produce also new contents: an editorial office will prepare and manage contents to provide interesting relations between resources and make the user discover them through links among different kinds of information. Those new contents will be tailored on different users' targets and will be distributed into the following sections of the Portal:

- Itineraries: articles focused on a theme, aimed at suggesting a virtual tour through some resources selected from the catalogue;
- Focuses: short monographs on a single argument;
- Events: information on cultural events (exhibitions, concerts, theatre, conferences);
- News: selected news on Italian culture.

Finally the project recommends that the Portal would provide the following services, to be eventually implemented in a later phase after the first realization:

- Multilingual versions;
- Newsletter;
- Forum;
- Young users area.

7 Recommendations for Usability and Accessibility

The project for the Italian Culture Portal deals with recommendations both for usability and for accessibility by impaired people. The Portal, which will be maintained by a public institution, must be usable by impaired people, e.g. by people with visual, auditive, motion and cognitive deficit, through the use of assistive devices and technologies.

In Italy such recommendation is ruled by the law n. 4 issued on 2004/01/09, "Disposizioni per favorire l'accesso dei soggetti disabili agli strumenti informatici" [12] (Recommendations for favouring access to IT tools by impaired people). This law imposes specific obligations both for the purchase of goods and for the providing of services, even with the possibility of making the service void.

The abovementioned law refers to the Italian Constitution (art. 3 "Every citizen has equal social dignity and is equal for the law, without distinctions of […] personal and social conditions. It is duty of the Republic to remove economic and social impediments which […] prevent the full development of the human being and the complete participation to the political, economical and social organization of the State.") and to norms issued by the Ministry of Public Administration (Ministero della Funzione Pubblica) and by the AIPA – Authority for IT in Public Administration, often ignored.

The law, together with recommendations for technical requirements for hardware (e.g. keyboards, devices for remote control) and software (e.g. user interface, maintenance of set up defined by users, textual information, buttons for accessing the assistive devices, etc.) rules also how web sites must be created (and tested) for guaranteeing accessibility.

Concerning web sites accessibility, the law is based on international recommendations as chapter 1194.21 of Section 508 of the USA Rehabilitation Act [13], and on guidelines provided by international bodies such as World Wide Web Consortium (W3C [14]) and, in particular, the recommendations of Web Accessibility Initiative (WAI [15]).

Moreover the project for the Italian Culture Portal adopts the guidelines proposed by MINERVA EU Project, concerning requirements and testing methodologies for the creation of "good quality web sites", to be not only technically accessible but also completely usable. The handbook for "Quality in cultural web sites[1]" recommends a technical test for accessibility to contents and a subjective test for the usability of information and services.

The minimum level of accessibility for all users (including people with complete or partial visual disabilities) takes into consideration what appears in the browser window, for technical aspects and for the contents, and imposes (amongst others) the following requirements:

- It is mandatory to adopt a DTD Strict and to use XHTML. Such recommendation implies to separate content from layout, and forbids to open new windows within the present one. Such an issue imposes some specific constraints specially when pieces of contents are imported from other web sites (e.g. through RSS feed);
- The use of frames must be avoided;
- Every non textual object must have an equivalent textual alternative. Therefore, images, audio and video streamings must be integrated with a text (from simple captions to a complete synchronized under-titling) in order to make assistive devices able to read all the objects of the page;
- Sensible maps must be client side or, if not possible, they must be linked to textual alternatives;
- It must be possible to easily distinguish main information from the background, both for graphic or audio components. This remarkably impacts on the use of colours and backgrounds;
- Layout and contents must be resizable, without overlapping or loss of information;
- Table-based layouts should be avoided; it is recommended to adopt a CSS based layout, using the element <div>
- Tables of data must be provided with information to be correctly interpreted by assistive devices, such as screen readers. Forms too must be created taking into consideration that they could cause problems when read by assistive devices;
- Pages must be usable even when scripts and applets are disabled or not supported;
- Links must be understandable even if read out of their context. User must be able to click them even through keyboard commands, technologies of keyboard emulation and pointing devices alternative to the mouse. This implies constraints of their position in the page, as they must not be too close each other.

Evaluation procedures are based both on automatic and semiautomatic validation systems and on the analysis carried on by an expert in web technologies and accessibility. The project suggests to evaluate the web site with the cooperation of a user panel, including impaired users, according to the following (subjective) quality criteria:

- perception
- comprehensiveness
- efficiency
- consistency
- safety
- security
- transparency
- easiness of learning system functionalities
- availability of helps and documentation
- tolerance to errors
- look and feel
- flexibility

Such criteria must be taken into consideration both during the design of the web site and in the development of the interface after this first evaluation.

[1] http://www.minervaeurope.org/publications/qualitycriteria.htm/

Notes and References

[1] Project responsibles for MiBAC: A. P. Recchia, R. Caffo. Project responsible for SNS: S. Settis. Coordinators: B. Benedetti, U. Parrini. Working group: P. Baccalario, I. Buonazia, M. Delcaldo, M. E. Masci, D. Merlitti. Consultants: G. Cresci, O. Signore, P. Valentino.

[2] PICO AP – Portal of Italian Culture Online - Application Profile.

[3] PURL - Persistent Uniform Resource Locator: http://purl.oclc.org/

[4] MiBAC website: http://www.beniculturali.it/

[5] OAI-PMH – Open Archive Iniziative – Protocol for Metadata Harvesting: http://www.openarchives.org/OAI/openarchivesprotocol.html/

[6] DCMI - Dublin Core Metadata Initiative: http://dublincore.org/

[7] For DCES – Dublin Core Element Set, see: http://dublincore.org/documents/dces/

[8] DC elements and terms are defined in: http://dublincore.org/documents/dcmi-terms/. See also *DC Metadata Registry*: http://dublincore.org/dcregistry/. For Qualified DC, see: *Using DC Qualifiers*: http://dublincore.org/documents/usageguide/qualifiers.shtml/; *Expressing Qualified DC in RDF/XML*: http://dublincore.org/documents/dcq-rdf-XML/

[9] Definition of 'Application Profile' from the DCMI Glossary: "In DCMI usage, an application profile is a declaration of the metadata terms an organization, information resource, application, or user community uses in its metadata. In a broader sense, it includes the set of metadata elements, policies, and guidelines defined for a particular application or implementation. The elements may be from one or more element sets, thus allowing a given application to meet its functional requirements by using metadata elements from several element sets including locally defined sets. For example, a given application might choose a specific subset of the Dublin Core elements that meets its needs, or may include elements from the Dublin Core, another element set, and several locally defined elements, all combined in a single schema. An application profile is not considered complete without documentation that defines the policies and best practices appropriate to the application". See: http://dublincore.org/documents/usageguide/glossary.shtml

[10] This document is downloadable at: ftp://ftp.cenorm.be/PUBLIC/CWAs/e-Europe/MMI-DC/cwa14855-00-2003-Nov.pdf.

[11] See: *DCMI Type Vocabulary*: http://dublincore.org/documents/dcmi-type-vocabulary/

[12] Disposizioni per favorire l'accesso dei soggetti disabili agli strumenti informatici - http://www.pubbliaccesso.gov.it/normative/legge_20040109_n4.htm

[13] See: http://www.pubbliaccesso.gov.it/normative/rehabilitation_act/index.htm

[14] W3C website: http://www.w3.org/

[15] WAI website: http://www.w3c.it/wai/

Building Bridges with Blocks: Assisting Digital Library and Virtual Learning Environment Integration through Reusable Middleware

Santiago Chumbe[1]*; Roddy MacLeod*[2]*; Marion Kennedy*[2]

[1] Institute for Computer Based Learning, School of Mathematical and Computer Sciences
Heriot Watt University, Edinburgh, EH14 4AS, UK
e-mail: santiago@macs.hw.ac.uk
[2] Heriot-Watt University Library
Heriot Watt University, Edinburgh, EH14 4AS, UK
e-mail: r.a.macleod@hw.ac.uk; m.l.kennedy@hw.ac.uk

Abstract

Various studies have reported that achieving effective use of increasingly heterogeneous scholarly objects within institutional learning and teaching frameworks is becoming critical to the performance of educational institutions. The integration of digital information environments, such as a University library, within a virtual learning environment (VLE) encapsulates this challenge. This paper presents reusable middleware to achieve effective digital library (DL) and VLE integration. The aim of the study is to demonstrate that the use of open standards and service-oriented architectures (SOA) to build "light" web-services-based middleware is a suitable alternative for embedding digital library information sources in learning and teaching frameworks. We argue that by using open-source and open-standards approaches rather than software and practices developed specifically for a particular VLE product, it is possible to obtain open reusable middleware that can simplify the DL-VLE integration and bridge the functionality of both environments. We hope that our methodology can provide a common foundation on which a variety of institutions may build their own customized middleware to integrate scholarly objects in VLEs. The study has assessed the impact of the VLE-library integration on academic users of both the library and the VLE. Performance issues of the proposed digital library-VLE integration are also discussed. A secondary but important finding of our study is that much more effort is required to open and standardize the closed, restricted and proprietary approach of digital publishers to the reuse of scholarly material. This approach can be a serious obstacle to effective digital library-VLE integration and can limit the publishers' ability to allow the discovery, integration and reuse of scholarly material. Current research in this area is analyzed and discussed.

Keywords: reusable middleware; SOA; SRU/SRW; federated search; VLE; open standards

1 Introduction

It is becoming clear from a number of perspectives that allowing effective discovery and use of scholarly objects within learning and teaching frameworks such as VLEs and institutional portals will be critical to the performance of educational institutions [1-7]. However, as Low B. [8] has noticed, resource discovery has been overlooked as a function of VLEs by vendors. We believe that this deficiency needs to be addressed urgently and with an "open standard" perspective. Digital libraries (DL) and VLEs both support learning and teaching in academic institutions. Institutions use library management system or digital libraries (DLs) to gain access to the content of scholarly objects from local databases such as institutional repositories or other collections of research papers, e-theses, technical reports, OPACs, image banks, etc., as well as from subscribed external content such as scientific papers provided by journal publishers or aggregators, and remote digital libraries, directories and online databases. On the other hand, VLEs are integrated environments of components (e.g. online discussions, course materials, e-mailing communication, submission of assignments, assessment, etc.) in which learners and tutors participate in "online" interactions of various kinds, involving online learning (VLEs are also known as Learning Management Systems (LMS) outside the UK.) However, despite the fact that both DLs and VLEs are oriented to support learning and teaching, previous studies have reported that the process of integrating DL and VLE systems can raise technical issues that require in depth investigation and complex solutions [9-11] (non-technical but important issues are beyond the scope of this study.) For example, systems run on different operating systems, use different data formats, have different authentication requirements and different web interfaces, etc. This paper sheds some light on a cost-effective methodology for overcoming such technical issues and confirms that a service-oriented approach combined with web services technology that makes use of standards or specifications for interoperability is a simple solution for achieving effective DL-VLE integration.

This paper first briefly describes the PerX toolkit, an open-source federated search software application produced by the *Pilot Engineering Repositories Xsearch* (PerX) Project [12]. It then presents the web services-based and open shareable SOA-compliant middleware used to embed Library functionality within the VLE. When describing the work done for encapsulating the middleware in the VLE used in this study, the paper mentions the commercial VLE *Blackboard* platform [13]. However, the work presented is not dependent on *Blackboard,* because it uses open standards and open source. Thus, our work can provide a common foundation on which a variety of institutions may build their own customized middleware to integrate their scholarly objects in their own VLEs. The only requirement is that the VLEs support XML-based retrieval via HTTP, preferably using standard web services communication. Further information and discussion on the middleware implementation and details of the "*Building Block*" encapsulation is presented in section three.

We also discuss the current status of digital publishing with respect to DL-VLE integration, finding that, within this context, most digital publishers have adopted a closed, restricted and non-standard approach. Publishers of scientific papers are one of the main sources of DL content and their lack of participation in sharing and reusing of scholarly metadata via open standard mechanisms can have a negative impact on DL-VLE integration success. Some recommendations for increasing interoperability and reuse in digital publishing are outlined at the end of section four.

The study has assessed the impact of the proposed VLE-library integration on academic users of the VLE and library services. Use case scenarios highlighting experiences gained and implications for stakeholders arising from the pilot are described in section five. The outcomes of these experiences are used as a basis for recommendations for future development of the pilot as well as for institutions planning to integrate their library with institutional VLEs.

After a discussion of the implications and some performance issues of the proposed digital library-VLE integration, the paper ends with conclusions obtained from the study.

2 The PerX Federated Search Toolkit

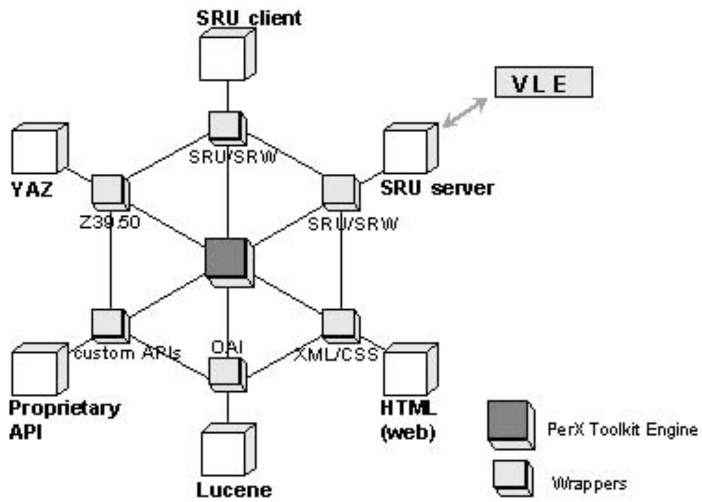


Figure 1: The PerX Toolkit Architecture

The core software component of this pilot is an open-source federated search toolkit produced by the *Pilot Engineering Repositories Xsearch* (PerX) Project, funded by the JISC Digital Repositories Programme. We chose this federated search software (referred to as the PerX toolkit) because it uses both an XML-based technology for system integration and a service-oriented architecture (SOA) [14] for achieving greater "loose" separation between its software components. In fact, the PerX toolkit is a reusable library of open source software applications integrated by a SOA model. It is a loosely coupled collection of proven, scalable and reusable software libraries and APIs (Application Program Interfaces) that have been combined via XML messages. Figure 1 represents the PerX toolkit architecture. Its main component is the *PerX Toolkit Engine*, which communicates with the rest of the software components via *wrappers*. The *wrappers* use XML-messaging for handling requests from/to the reusable APIs, which in turn deal with the database sources. The toolkit allows remote and local heterogeneous database sources to be cross-searched from one access point. It uses open standard technology for metadata exchange such as OAI-PMH (Open Archive Initiative-Protocol for Metadata Harvesting, [15]) and the search protocols SRU (Search/Retrieve via URL, [16]) and Z39.50 (International Information Retrieval Standard ISO 23950, [17].) Chumbe et al [18] presents a full description of searching databases with the PerX Toolkit as well as the processes involved in metadata exchange with data providers (harvesting, normalization, searching and rendering.)

3 Reusable Middleware Approach for Embedding DL Functionality within a VLE

Heriot-Watt University Library has recently collaborated with the Institute for Computer Based Learning (ICBL) on a *Blackboard* VLE - e-Library integration pilot. The core software component of this pilot is the PerX toolkit described in the previous section. However the key player, or broker, of the integration itself is the reusable web-services based and SOA compliant middleware used to embed the toolkit functionality within the VLE. An important condition for the pilot was that the middleware should know nothing about the hosting VLE environment and thus can potentially be reused within any VLE framework. Its only function was to provide a "live bridge" between the toolkit functionality and the VLE system.

Traditional client/server middleware has typically been deployed in a 2-tier, point-to-point architecture [19], which in our case would involve the installation of a proprietary API Client on the VLE server, and an API server on the PerX server machines. However, this is an expensive and inflexible model, because neither of the Client and Server APIs can be reused. A step forward in flexibility is an n-tier model, where an XML-based middleware API is installed between the client (VLE) and the server (DL) systems. The n-tier model offers the benefit of avoiding the development of two different APIs and the need to access source codes on both sides to enable interoperability. The XML-based middleware API is a kind of *wrapper* that hides the complexity of the native APIs of both server and client because it uses web services technology for exposing their services. This XML-based n-tier model has been used by the LEBONED Project to integrate the *eVerlage* Digital Library product into the *Blackboard* VLE [20]. However, the cost-effective factor is still unresolved by this approach, because such a specific *wrapper* will need to be written again for any other digital library system to be integrated into *Blackboard*. We present here a further step towards achieving inexpensive, reusable and flexible DL-VLE integration. Our approach is also based on a three-tier design pattern using an XML-based middleware API that sits between the *Blackboard* VLE (front end) and the PerX toolkit (back end) systems, but we do use the open standard SRU/SRW protocol for interoperability and XML message exchanging between the systems. The use of proven open standards effectively turns our XML-based middleware into a reusable *wrapper* or message broker. This approach would allow organizations to access virtually any SRU/SRW compliant system from within any SRU/SRW compliant DL system through a scalable service-oriented architecture. While simple, the middleware constitutes the basic infrastructure behind the DL-VLE integration, becoming a versatile alternative for integration. The basic deployment architecture of the proposed approach is shown in Figure 2.

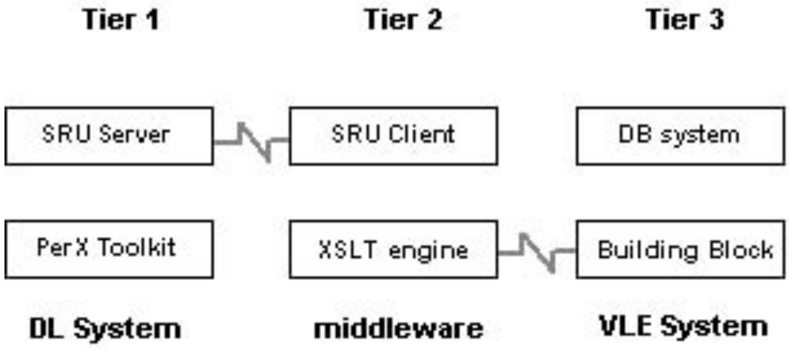


Figure 2: DL-VLE integration architecture using a three-tier design model

The approach described above offers a potentially rich system-level DL-VLE integration because it uses a standard specification for interoperability such as the Search/Retrieve via URL (SRU) protocol encapsulated in a reusable middleware. The SRU/SRW protocol is simple to implement because it is a standard REST-ful specification for providing Web Services functionality without the complexity of tightly coupled designs as found in remote procedure calls such as SOAP [21]. A REST-based protocol uses the HTTP mechanism to implement a client/server model using TCP/IP sockets [22]. The encapsulated middleware (the HTTP client) opens a connection to the PerX toolkit's SRU server (the HTTP server) and sends a request message consisting of a search query using the HTTP GET method. The HTTP server then returns a response XML message with the search results using the POST method and then closes the connection. The middleware then reformats the

XML message and puts the search results into the VLE database system, so they can be shareable in the VLE modules. The middleware is a kind of proxy or intermediary software that handles requests on behalf of the systems that it is bridging.

In order that the middleware be recognized by the VLE as one of its components, it needs to be encapsulated in a *building block* of the *Blackboard Learning System*. This is accomplished by issuing an XML configuration file (*manifest*) to identify the middleware as a "bridge type" *Blackboard* "*building block*," and by including *Blackboard* proprietary Java class Tag libraries to abstract user interface components [23]. The middleware implementation has followed as strictly as possible the current Java Servlet specifications for Web applications [24]. The *Blackboard* system includes a portal running on a *Tomcat servlet* [25] and in fact its "*building blocks*" are just local *portlets* that can be handled as web applications individually deployed on the local *servlet*. These *building blocks* do not adhere to the web services specifications for remote *portlets* (known as Web Services for Remote Portlets WSRP specification [26]), so they are not shareable from other portals or remote systems. However, this is not an issue for our implementation as we supply the share-ability via the REST-ful model described above. At its very core, the middleware is an SRU client that provides standards-based technology to achieve integrated behavior and performance at the system-level across diverse environments such as the federated search toolkit and the VLE system.

Unlike other open standards for interoperability, such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), implementations of the SRU/SRW protocol are poorly reported in the digital library literature. Equally, while OAI-PMH has attracted much attention in providing interoperability, despite reports of a number of important issues concerning OAI-PMH in the literature [18, 27-29], practical examples of SRU and its relevance in the digital library are seldom discussed. To the best of our knowledge the approach of combining SOA with SRU to embed DL functionality within a VLE for cross-searching remote databases and local repositories has not yet been reported fully, and practical implementations have not received much attention. An attempt to integrate VLEs and digital repository systems using the SRU protocol in the open source d+ toolkit has been reported by Low B. [8]. However, some issues with the software were uncovered when using d+ for interoperability with VLEs. Despite claims of adherence to the current service-oriented trend, it was found that, apart from the SRU functionality provided by licensed OCLC software, deployment and use of the d+ toolkit required hardcode configuration of the software components as well as of the digital repositories. Also, at the time of testing, d+ could only query one database at the time (a sequential searching approach in contrast to the desirable "simultaneous" cross-searching approach.) Performance issues were also noted. It seemed that the ability of d+ for searching Z39.50 targets was bound to the limitations of the JAFER toolkit [30], which is still not fully available for production. Other open source alternatives considered before PerX and found unsuitable for the work discussed here, were the MDC toolkit, MyLibrary and the software suite Greenstone. In the UK, the JISC – DiVLE research strand involved a number of projects looking at how library resources can be integrated into VLEs using open standards. Thus, for example, between 2002 and 2004, the OLIVE project has been focused on how the OpenURL standard can be used to link Reading lists and Learning objects from the VLE. It also explored the use of Web Services (SOAP.) However, little practical achievement was reported [31], and unfortunately most of the plug-in software developed by the project was dependent on the commercial platforms used for integration (*MetaLib*, *Blackboard*, *Aleph*, *Discover*, *SFX*, etc.) For example, the method for implementing OpenURL is tightly coupled to the search form in the *Building Block* and cannot be reused for other applications. Also, the approach of the OLIVE project of loading functionality on the *Building Block* for metadata management raised many interoperability issues, as the *Blackboard* metadata functionality proved to be unusable and inaccessible to other areas within *Blackboard*. In Australia, Richardson J. [32] also reported on a project at Griffith University to integrate library resources into the *Blackboard* system. She recognized the power of commercial products in this arena, such as Sentient *Discover*, which supports OpenURL and Z39.50, but also highlights the "cognitive disconnect" faced by users of *Blackboard* when are taken away to the *Discover* user-interface environment from the *Blackboard* user-interface environment.

On the other hand, SOA approaches in e-learning are being promoted as suitable alternatives by important organizations such as JISC (the United Kingdom's Joint Information Systems Committee), DEST (the Australian Department of Education Science and Training), ADL (the US Advanced Distributed Learning Initiatives), IMS (the Innovation Adoption Learning global learning consortium), NSDL (the US National Science Digital Library) and IC (Industry Canada). Need for stable and coherent technical frameworks or infrastructures where e-learning services can inter-operate harmoniously have been highlighted [33-36]. Our work is firmly in harmony with the above approach and recommendations, and it would be part of any standard e-learning framework where its functional components expose service behavior via loosely coupled interfaces. In this context, we follow with interest the work being carried out by related projects, such as the open source digital library architecture Fedora [37] and the NSDL Data Repository Architecture [38], as well as any research outcome from the JISC e-Framework for Education and Research Programme [39].

4 Issues of the Digital Publishing Model Regarding Reuse of Scholarly Material

An additional finding of our study is that integration of digital publishing is made difficult by the fact that publishers rarely use open standards to make their metadata available to third parties. Many publishers currently rely on large external aggregators in order to expose their scholarly contents to a wider audience. Frequently, digital libraries need to deal with these external aggregators in order to gain access to subscribed scholarly material using expensive commercial software tools, which in most cases do not use open standards. The consequence is that it is often difficult for institutions to get access to publishers' metadata and databases using suitable open standards and protocols for interoperability. The reality is that progress towards integration of scholarly digital information within VLEs is slowed down by commercial publishers and aggregators by not offering machine-to-machine access to their databases using open standards. Figure 3 illustrates a simplified view of a typical digital publishing model within the digital library context. Clearly noticeable is a need for a “consolidator” point for effective inter-operation between digital library and the rest of components of the model. Integration in a component-by-component basis would be unfeasible. Figure 4 sketches an alternative model where effective integration is enable by a suitable “middleware consolidator” created using technology presented in previous section. Advocacy for open standards is not about encouraging free access to resources but simply about providing effective ways to find (discover) and reuse resources. The PerX Project has produced a relevant report on the benefits for publishers of exposing their metadata via open standards [40]. Also other works [41, 42] advocate the use of best practices among data providers and argue that the business strategies of digital publishing in fact can benefit from the standards that are part of the digital library.

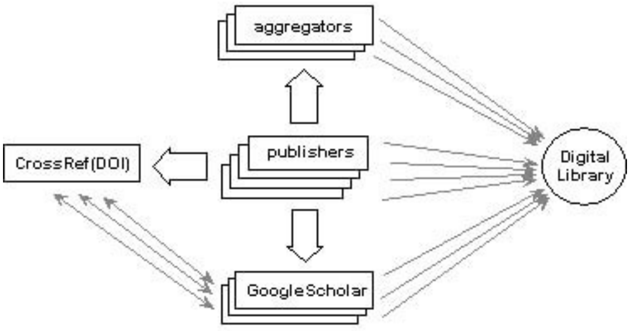


Figure 3 A typical digital publishing model.

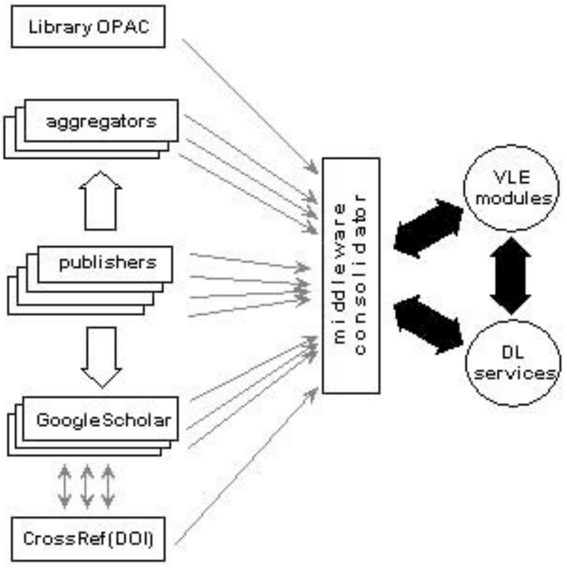


Figure 4: Publishing model with “middleware consolidator”

Currently there is a large movement towards openness in almost all aspects of digital publishing. Promising initiatives for solving important interoperability issues are not only coming from organizations that advocate open standards. Thus two technologies for enabling easier scholarly resource discovery have emerged, one from the publisher’s side (CrossRef) and another one, Google Scholar, from Google, the leading commercial search engine. (Microsoft Windows Live Academic Search and Scirus services could also be mentioned here.) CrossRef is being promoted by the publishing industry to make possible standard scalable linkage of scholarly material through Digital Object Identifiers (DOI) [43]. We have been investigating the feasibility of using the CrossRef OAI service to cross-search metadata for a selection of the 23 million records hosted by CrossRef as well as to provide openURL linkage via the CrossRef openURL resolver. Unfortunately access to the CrossRef OAI repository is not open to everyone, which again puts a limitation to the reuse of metadata. Also, the CrossRef OAI service uses a somewhat limited subject classification that makes subject-based implementations difficult. The usefulness of being able to only search on title, authors and citation of papers could be also challenged (CrossRef does not store abstract and keywords.) Google Scholar [44] via direct agreements with publishers is in fact crawling and cross-searching a very important and increasing portion of scholarly resources (e.g. peer-review articles, theses, preprints, technical reports, etc.) Google Scholar also works with CrossRef to use the DOI as the primary means to link to an article. Despite the facto that Google Scholar is still a very broad commercial oriented solution that includes any material that "looks scholarly" and that can come from unknown sources, it offers something more than other open access federated search services such as the DOAJ (Directory of Open Access Journals [45]) or Scirus [46]. It offers enhanced and fast search capabilities, cited references and links to subscribed resources through local link resolvers. On top of that, Google Scholar has the advantage of being quickly associated with the de facto ubiquitous discovery tool: Google is everywhere. As the value and usage of Google Scholar is significant, we have integrated Google Scholar in the DL-VLE prototype, via a custom API. Some researchers think that Google Scholar could be the possible solution to the cluttered access provided by traditional gateway and hosting providers used by the library. However, Google is still a commercial initiative, produced without open standards and developed in a way in which not everyone can participate, which raises concerns over obsolescence and dependency issues. For example, experience suggests that only file formats that use open standards can secure long-term preservation of scholarly material and avoid software dependencies.

In fact, the publisher with relatively cost-effective and simple solutions can produce chunk-able, reuse-able and embed-able metadata using open standards. The technical undertaking by publishers need not be large, though the potential benefits are. Publishers are already using complex and tailored mechanism to expose their data on demand basis. This approach can be ineffective and expensive. Let us consider, for example, the case of a publisher that wants its metadata to be included in CrossRef, and also be available (optionally the full-text, too) from various aggregators (MetaPress, ingentaConnect, Ovid, ProQuest, SwetsWise, etc.) and indexed by Google Scholar. Without standards, the publisher will need to set up and maintain different XML metadata files for Google, for CrossRef and for each aggregator. It will need to use an FTP-based mechanism for uploading data on the aggregators or allow them to crawl their servers hosting its data. All that could be avoided if the publishing industry agree on using a set of open standards, and better still if they work with librarians for enabling easy resource discovery, as both of them share the same goal: to make scholarly content available for the users that need it. The benefits of making online search a pleasing experience are for both the publishers and the digital libraries. We suggest that publishers start by implementing “light” open standards such as SRU/SRW, openURL, RSS feeds, and Dublin-Core (DC) metadata format. It is worth it for publishers to consider redirecting some of their IT resources to implementing open standards, automatic machine-to-machine access and simplified user interfaces. Diverse studies have already suggested that what online users want is fast and effortless access to the resources they need [47-50]. Users give little value to sophisticated user interfaces provided by publishers’ web sites. Publishers should take notice of the behavior of users. On the other hand, commitment to protocols or specifications that do not adhere in full to the open standard concept [51] should be avoided if possible, in case that “cutting edge” technology that is not backed by mature open standard bodies is abandoned. For example it can be instructive to follow the discussions on the reasons for the apparent decline in the use of the CORBA protocol [52], which has been providing interoperability for more than a decade. In summary, it would make a positive impact on interoperability in general, and possibly in their revenues too, if publishers implement open standards for enabling institutional and individual users to gain quick access to the content they need with almost no effort.

5 Study of the Impact of DL-VLE Integration on Library Users

A prototype working system demonstrating the VLE-Library system interoperability has been implemented and made available to stakeholders (students, academic and library staff) at the Heriot Watt University. It is being used to asses the impact of the VLE-library integration on academic and library users as well as a basis for gathering suggestions and recommendations for future developments to benefit institutional planning for library and institutional VLE integration. The prototype system, named as *PerX Building Block*, provides distributed searching of a sample of subscribed e-journals, the local library catalogue (OPAC) and the Google search engine. A facility for bibliographic export in RDF-based format is being added in the prototype. Testing is being carried out with a group of academic library users, and feedback is being gathered using a short questionnaire and informal interview. So far our study has confirmed the perception that in particular under-graduated students tend to ignore searching in databases subscribed by the library and prefer the ease of using Google [53]. Post-graduate researchers also feel attracted to Google capabilities. The current searchable web based interface of the library does not include links to Google or Google Scholar. If even lecturers and librarians use Google in their work, we expect that users will appreciate having Google Scholar embedded in the prototype. In fact Google Scholar can be used to drive users to the library web site and add value to the sometimes ignored library catalogues at not cost.

The reuse and sharing of DL content among the different VLE components is being explored with particular interest. We have had high interest in finding out how users rate the usefulness of cross searching from within the VLE and the convenience of onward use of search results in other VLE functions e.g. exporting, saving, emailing and posting them to discussion boards.

In parallel to our work, the University VLE Educational Support Team has been conducting consultation meetings with lecturers who are using the VLE in their courses to give them the opportunity to bring up any problems and provide feedback. In some of the meetings various issues were mentioned by lecturers that in fact would be solved by enabling machine-to-machine inter-operability between the VLE and the rest of University systems. This prompted the possibility of expanding the applicability of the reusable middleware for *bridging* systems such as the Students Registration System with the VLE.

Regarding possible performance issues of the proposed DL-VLE integration, we have noticed that SRU/SRW is not necessary relatively slow. We were expecting that the SOA-based prototype be significantly slower than fast, general purpose search engines because it uses XML-based messaging services, which typically consume more computing resources. However, after assessing the performance of the search services when searching various heterogeneous scholarly objects, users noticed that speed and performance were not issues in the prototype.

Finally, some recommendations for increasing the usability and the effectiveness of the prototype have been identified. In addition to more sophisticated retrieval and searching algorithms (e.g. full common Boolean support across heterogeneous databases), there are key operational enhancements that have been acknowledged as desirables. Enhancements include:

- Combining search results from multiple databases, which involves unified ranking;
- Comparing and consolidating search results (simplest case: removing duplicate search results; more complex case: fussy techniques for combining several databases´ results);
- Discovering inconsistencies and removing them in the search results (for example search results that seems to be different but in fact point to same resources).

6 Conclusions

By using open-source and open-standards approaches rather than products and practices developed specifically for an individual VLE product, we have obtained a reusable middleware that can provide a common foundation on which a variety of institutions may build their own customized middleware to integrate their scholarly objects in VLEs. Our study hopes to demonstrate that the use of service-oriented architectures (SOA) and REST-ful based (SRU) open source middleware is a cost effective, simple and open alternative for embedding digital library services within learning and teaching frameworks.

We have described relevant related works and software solutions. We have highlighted shortcomings and pros of those studies. Most of these studies have tended to produce solutions tied to commercial platforms or have given priority to questionable standard such as OAI-PMH, for achieving interoperability, as it was in the case of the BRICKS Project [54], which it seemed promising when presented web services based concepts for achieving integration. However the SOA factor and the ease of alternatives such as SRU/SRW were unnoticed by these projects.

Although SOA middleware reduces the need for system development and also management and maintenance burdens, the performance of SOA-based search services need to be monitored for large production services, because XML-based messaging services typically consume more computing resources and are slower than fast general purpose search engines. Early testes suggested that users have not found performance issues using the DL-VLE integration prototype system.

A key requirement for VLEs should be integration, and the tendency of using VLEs that do not support SOA, open standards and Web Services should be reversed. Main global and national organizations are working towards SOA e-Frameworks, where monolithic and centralized architectures are no longer taken into account for effective delivery of services. The ultimate aim of a VLE should be to provide a framework where service applications are embedded and integrated through agreed behaviors and interfaces using open web services technology to achieve interoperability.

The combination of open standards, "light" web services and SOA can produce powerful platforms that can help to develop information environments that are responsive to new generation of library users (the Net generation) that expect to find ubiquitous discovery tools, such as Google Scholar, in their learning environment systems.

Publishers and libraries share the ultimate goal of making scholarly content available for the users that need it. Both of them also face the same challenge produced by the movement towards openness in almost all aspects of e-learning. Clearly it has been demonstrated that both can benefit from open standards. Libraries and publishers no longer can expect that their users adapt to and learn about their existing closed, restricted and non-standard systems. It is them who need to provide open access to their assets for interoperability purposes.

Our study concludes that without open standards, any middleware used to integrated different systems is likely to become rather cumbersome and infeasible. The use of open standards reduces dependency and heterogeneity and it is a key facilitator for systems integration and for making reality service-oriented systems.

Acknowledgements

The study described here is based upon work supported by the PerX research project of the JISC-funded Digital Repositories Programme. The authors acknowledge the efforts and help of the Heriot Watt University Library staff as well as give special recognition to Dr. David A. Cole for his fundamental role in the installation of the middleware on the VLE described here.

Notes and References

[1] ROSENBERG, M. J. *e-Learning - Strategies for Delivering Knowledge in the Digital Age*. McGraw-Hill, 2001

[2] PINFIELD, S. *The changing role of subject librarians in academic libraries*. Journal of Librarianship and Information Science 33(1) pp. 32-38. 2001

[3] ALEXANDER, W. *Adaptive developments for learning in the hybrid library.* Ariadne Issue 24. 2000 [Available at http://www.ariadne.ac.uk/issue24/sellic/intro.html]

[4] MACCOLL, J. *Virtuous learning environments: the library and the VLE*. Program: electronic library & information systems, Volume 35, Number 3, pp. 227-239(13) 2001

[5] MCLEA, N.; LYNCH, C. *Interoperability between Library Information Services and Learning Environments - Bridging the Gaps*. A white paper on behalf of the IMS Global Learning Consortium [Available at http://www.imsglobal.org/digitalrepositories/CNIandIMS_2004.pdf]

[6] CARLSON, S. *The Deserted Library: As Students Work Online, Reading Rooms Empty Out Leading Some Campuses to Add Starbucks*. Chronicle of Higher Education Nov 16 2001 [Available at http://chronicle.com/free/v48/i12/12a03501.htm]

[7] FAULHABER, C. *Distance Learning and Digital Libraries: Two Sides of a Single Coin*. Journal of the American Society for Information Science, v47 n11 pp. 854-56 Nov 1996

[8] LOW, B.; MACCOLL, J. *Searching Heterogeneous e-Learning Resources*. Paper presented at the DELOS Workshop 2005, Digital Repositories: Interoperability and Common Services, May 2005

[9] BLACK, NANCY E. *Distance Library Services in Canada: Observations and Overview of Some of the Issues*. New Review of Libraries and Lifelong Learning 4, pp. 45-62. 2003

[10] FLECKER, D.; MCLEAN, N., *Digital Library Content and Course Management Systems: Issues of Interoperation.* Report of a study group funded by the Andrew W. Mellon Foundation. July 2004 [Available at http://www.diglib.org/pubs/dlf100/cmsdl0407.pdf]

[11] WHITING, J.; KARTUS, E.; RUNNER, E. *Challenges of Integrating Learning Resources within the Learning Management System at Deakin University*. Proceedings of the EDUCAUSE Australasia Conference, pp. 159-171. 2003.[Available at http://www.caudit.edu.au/educauseaustralasia/2003/EDUCAUSE/PDF/AUTHOR/ED030070.PDF]

[12] Pilot Engineering Repositories Xsearch (PerX) Project: http://www.icbl.hw.ac.uk/perx

[13] Blackboard VLE: http://www.blackboard.com/us/index.Bb

[14] HE, H. What Is Service-Oriented Architecture. Sept. 2003 [Available at http://www.xml.com/pub/a/ws/2003/09/30/soa.html]

[15] Open Archives Initiative: http://www.openarchives.org

[16] SRU/SRW Protocol: http://www.loc.gov/standards/sru

[17] Z39.50 Protocol: http://www.loc.gov/z3950/agency

[18] CHUMBE, S.; MACLEOD, R.; BARKER, P.; MOFFAT, M.; RIST, R. *Overcoming the obstacles of harvesting and searching digital repositories from federated searching toolkits, and embedding them in VLEs*. Proceedings of the 2nd International Conference on Computer Science and Information Systems, Greece. 2006

[19] ALONSO, G.; CASATI, F.; KUNO, H.; MACHIRAJU, V. *Web Services: Concepts, Architectures and Applications*. ISBN: 978-3-540-44008-6 Springer 2004

[20] OLDENETTEL, F.; MALACHINSKI, M.; REIL, D., *Integrating digital libraries into learning environments: the LEBONED approach.* Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries. 2003

[21] ZUR MUEHLEN, M.; NICKERSON, J.; SWENSON, K. *Developing Web Services Choreography Standards – The Case of REST vs. SOAP.* Decision Support Systems 37. Elsevier. 2004.

[22] FIELDING, R. *Architectural Styles and the Design of Network-based Software Architectures*. Doctoral Dissertation. University of California, Irvine, CA (180p.) 2000

[23] Blackboard Building Blocks. Developer Guide. Blackboard Learning System Release 7.1 2006

[24] JSR 154: JavaTM Servlet 2.4 Specification: http://jcp.org/en/jsr/detail?id=154

[25] Apache Tomcat Servlet: http://tomcat.apache.org/tomcat-6.0-doc/index.html

[26] POLGAR, J.; BRAM, R.; POLGAR, A. *Building and Managing Enterprise-wide Portals*. Idea Group Inc (IGI) 2006

[27] LAGOZE, C.; KRAFFT, D.; CORNWELL, T.; DUSHAY, N.; ECKSTROM, D.; SAYLOR, J. *Supporting education: Metadata aggregation and "automated digital libraries": a retrospective on the NSDL experience*. Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries JCDL '06, June 2006

[28] STER,N R.; MCELROY, M. *Virtual Collections: Challenges in Harvesting and Transforming Metadata from Harvard Catalogs for Topical Collections*. Proceedings of the DLF Fall 2006 Forum. Harvard University. Boston MA. 2006

[29] FOULONNEAU, M. at al. *Strand : Open Archives Protocol for Metadata Harvesting*. The Knowledge Exchange workshop on Institutional Repositories Report. Danish Library Agency in Copenhagen, Denmark. Feb. 2007 [Available at http://knowledge-exchange.net.dynamicweb.dk]

[30] JAFER Toolkit Project: http://www.jafer.org

[31] OLIVE Project Final Report (2004) [Available at http://www.jisc.ac.uk/uploaded_documents/OLIVE_Project_Report.pdf]

[32] RICHARDSON, J. *Building Bridges between Learning Management Systems and Library Content Systems*. Presented at the 11th Australian World Wide Web (AusWeb05) Conference, Gold Coast. 2005 [Available at http://ausweb.scu.edu.au/aw05/papers/refereed/richardson/index.html]

[33] WILSON, S.; BLINCO, K.; REHAK, D. *Service-Oriented Frameworks: Modelling the Infrastructure for the Next Generation of e-Learning Systems*. Presented at Alt-I-Lab Conference, 2004

[34] POWELL, A. *A 'service oriented' view of the JISC Information Environment*. UKOLN Report. Nov. 2005 [Available at http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/soa/jisc-ie-soa.pdf]

[35] PAYETTE, S. *Choosing Technology that can Evolve With User Needs. A service-oriented approach to e-research, e-scholarship, and advanced scholarly publication*. VALA 2006 Conference. Melbourne, Australia. February 2006 [Available at http://www.valaconf.org.au/vala2006/papers2006/97_Payette_Final.pdf]

[36] HUNTER, J. *Scientific Models – A user-oriented approach to integrating scientific data and digital libraries*. VALA 2006, Melbourne. February 2006 [Available at http://www.valaconf.org.au/vala2006/papers2006/55_Hunter_Final.pdf]

[37] JOHNSTON, L. *Development and Assessment of a Public Discovery and Delivery Interface for a Fedora Repository*. In D-Lib Magazine, Vol. 11, Number 10. 2005 [Available at http://www.dlib.org/dlib/october05/johnston/10johnston.html]

[38] LAGOZE, C; KRAFFTI, D.; PAYETTEI, S.; JESUROGAII, S. *What Is a Digital Library Anymore, Anyway? Beyond Search and Access in the NSDL*. In D-Lib Magazine Vol. 11, Number 11. 2005 [Available at http://www.dlib.org/dlib/november05/lagoze/11lagoze.html]

[39] The JISC e-Framework for Education and Research: http://www.jisc.ac.uk/whatwedo/programmes/programme_eframework.aspx

[40] MOFFAT, M. *'Marketing' with Metadata - How Metadata Can Increase Exposure and Visibility of Online Content*. PerX Project Report. 2006 [Available at http://www.icbl.hw.ac.uk/perx/advocacy/exposingmetadata.htm]

[41] BRANTLEY, P., Greenbaum D. and Yee R. *Emerging Best Practices for Integrating Library Content and Services with Educational Technology*. Presented at EDUCAUSE Annual Conferences. 2003

[42] COLEMAN, R. *Publication, Business and the Digital Library*. Sydney University Press. 2006 [Available at http://www.valaconf.org.au/vala2006/papers2006/58_Coleman_Final.pdf]

[43] CrossRef: http://www.crossref.org

[44] Google Scholar: http://scholar.google.com/intl/en/scholar/about.html

[45] Directory of open access journals: http://www.doaj.org

[46] Scirus: http://www.scirus.com

[47] XIE, H.I. *Supporting ease-of-use and user control: desired features and structure of web-based online IR systems.* Information Processing and Management, Vol. 39, pp.899–922. 2003

[48] VILAR, P.; ZUMER, M. *Comparison and evaluation of the user interfaces of e-journals*. Journal of Documentation. Vol. 61, No. 2, pp.203–227. 2005

[49] JACSO, P. *Google scholar: the pros and the cons*. Online Information Review, Vol. 29, No. 2, pp.208–214. 2005

[50] MARKLAND, M. Embedding online information resources in Virtual Learning En*vironments: some implications for lecturers and librarians of the move towards delivering teaching in the online environment* Information Research, 8(4), paper no. 158. 2003. [Available at: http://informationr.net/ir/8-4/paper158.html]

[51] COYLE, K. *Open Source, open standards*. Information Technology and Libraries. Vol. 21, No. 1 pp. 33-36. 2002

[52] CORBA: http://www.corba.org

[53] LIPPINCOTT, J. *Net Generation Students and Libraries*. Chapter 13 in *Educating the Net Generation*. Diana G. Oblinger and James Oblinger Eds. E-Book. 2005 [Available at http://www.educause.edu/NetGenerationStudentsandLibraries/6067]

[54] BERTONCINI, M.; MASCI, M.; RONCA, A. *Paving the Way for the Next Generation Cultural Digital Library Services: The Case Study of 'Fortuna visiva of Pompeii' within the BRICKS Project*. Proceedings of the 10th International Conference on Electronic Publishing ELPUB2006. Bulgaria, June 2006 ISBN 978-954-16-0040-5, 2006, pp. 5-16 [Available at http://elpub.scix.net/cgi-bin/works/Show?245_elpub2006]

Designing Metadata Surrogates for Search Result Interfaces of Learning Object Repositories: Linear versus Clustered Metadata Design

Panos Balatsoukas; Anne Morris; Ann O'Brien

Department of Information Science, Loughborough University
Loughborough, Leics, LE 11 3TU, UK
e-mail: {p.balatsoukas; a.morris; a.o-brien}@lboro.ac.uk

Abstract

This study reports the findings of a usability test conducted to examine users' interaction with two different learning object metadata-driven search result interfaces. The first was a clustered metadata surrogate interface (where metadata elements were divided into sections), and the second a linear or single metadata surrogate interface (where all metadata elements were listed in a single record). The objectives of this research were: firstly, to investigate the time needed by learners to identify a relevant learning object, using both interfaces; secondly, to examine learners' subjective satisfaction for both interfaces; and finally, to study the impact of task complexity on users' interaction with both interfaces. To facilitate the objectives of the study, twelve postgraduate students participated in a user study which employed a multi-method approach and involved observation of users' interactions, subjective satisfaction questionnaires and semi-structured interviews. Data collected included the time needed for users to identify relevant learning objects in both interfaces and the rating of users' subjective satisfaction. In addition, qualitative data were collected based on interviews and the think aloud protocol. Parametric analysis (ANOVA tests) was conducted to identify statistically significant differences between the two interfaces in terms of time, user satisfaction and the impact of task complexity. The data analysis revealed that users needed less time to perform the tasks using the clustered metadata surrogate interface. This difference, however, was not significant. In addition, there was no significant impact of task complexity on user's performance. In terms of subjective satisfaction, however, the participants perceived the clustered metadata surrogate interface to be significantly more satisfying, stimulating and easy to use (F=89.690, p.<0.01). The findings of this study provide useful recommendations for the design of search result interfaces in learning object repositories.

Keywords: user-centred metadata; metadata design; e-learning; interface design

1 Introduction

Research on user-centred metadata surrogate design in search result interfaces can be divided in two research strands: Firstly, the presentation of metadata in search result interfaces, and secondly, the content of metadata in search result interfaces. The former covers aspects such as, interface layout, presentation, display format, interactivity and 'tailorability' of metadata in search result interfaces, while the latter examines the users' level of understanding, the usefulness and the quality of metadata semantics and vocabularies for relevance judgment.

1.1 Presentation of Metadata in Search Result Interfaces

Research in this area has been focused on a comparison between list, categorical/clustered and dynamic displays of metadata [1]. These studies employed a controlled – laboratory based experimental design. Most researchers concluded that users were more satisfied, selected fewer non-relevant documents and performed the task of judging the relevance of documents faster using the category-based interfaces [2-6]. Other researchers, however, did not observe significant differences between category-based and list-based interface designs [7], while other researchers revealed that users preferred dynamic rather than category-based interfaces [8].

Research has also investigated the way metadata elements should be arranged within metadata surrogates. A logical pattern has been identified, such that, metadata elements providing access or arranging access to the resource should follow content related elements, such as, the title, abstract, subject headings or keywords [9].

1.2 Content of Metadata in Search Result Interfaces

In addition, to these studies, the research area of user-centred relevance judgment has significantly advanced knowledge about the content and types of metadata that should be included in metadata surrogates in search result interfaces.

Researchers in this field have advocated the importance of the presence of an abstract in the metadata surrogate when users judge the relevance of documents [10-12]. More recent studies by Drori [13] and Paek et al [14] have also revealed that abstracts should include information related to the users' search query (contextual information) rather than simply presenting the first sentences of the document. Researchers have also revealed that topical or subject relevance is not the only criterion users employ to judge relevance when examining metadata surrogates or documents. Other criteria, were: the purpose and scope of the document, objectives, recency, source quality, reputation of the author, accessibility information and cost [11, 12, 15].

1.3 Design of Learning Object Metadata Surrogates in Search Result Interfaces

Although previous research has covered a wide range of issues related to metadata design in search result interfaces, the study of the presentation and content of learning object metadata has been neglected in the e-learning literature with only a few studies attempting to explore the phenomenon in some depth.

For example, the 'MetaTest' project investigated users' interaction with GEM-based metadata records. The findings of the study revealed that users were significantly more satisfied and made better relevance judgments when an abstract with information about the contents of the learning object was included in the metadata surrogate. Further clues that could be included in metadata surrogates were: relevance rankings, reviews and comments from others who did similar searches [16-18].

In another study, researchers evaluated the usability of the SearchLT learning object repository [19]. The study confirmed that well established heuristics, such as, the need for visibility and user-centred terminology should be applied in the design of learning object metadata surrogates in the search result interfaces of learning object repositories. The study also suggested that the contents of metadata records should be divided into clusters/categories and not be displayed as a list within a single and information cluttered surrogate. This issue had not been raised in earlier studies. Researchers in the SearchLT study did not compare different metadata interface designs (for example linear versus clustered metadata surrogates), did not employ parametric techniques for testing the statistical significance of the finding and neglected the impact of task complexity on user performance when using both linear and clustered metadata surrogate interfaces.

2 Aim and Objectives

The aim of this research was to examine users' interaction with two different learning object metadata-driven search result interfaces. The first was a linear metadata surrogate interface (where metadata elements were included within a single surrogate in a linear form), and the second a clustered metadata surrogate interface (where metadata elements were divided into sections). In particular, the objectives of this study were:

- To investigate the time needed by learners to identify a relevant learning object, using both interfaces;
- To study the impact of task complexity on users' interaction with both interfaces; and
- To examine learners' subjective satisfaction for both interfaces.

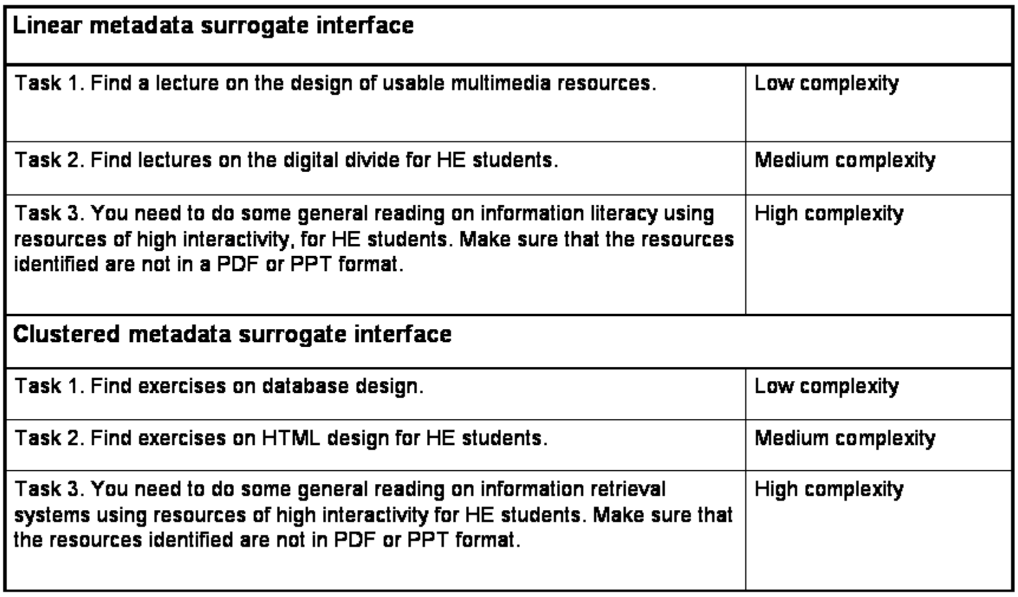
3 Methodology

3.1 Research Design

To address these objectives a usability test was conducted. A total of 12 postgraduate students in Information Science participated in the study. All participants were frequent users of Electronic Information Services (EIS) and the WWW. Students were recruited by means of e-mails and announcements on University notice-boards. A background questionnaire was completed by candidate participants before the usability tests. The background questionnaires facilitated the final selection of the participants in the study based on their familiarisation with EIS and the WWW.

The usability test employed a "within-subjects" design that required all participants to perform a similar set of tasks with both interfaces. The sequence with which the interfaces and the tasks were presented to the subjects was randomly altered for counterbalancing the effects of 'learning transfer'. Users had to perform three tasks in each interface. The tasks differed in terms of complexity. Table 1 presents the tasks assigned to the linear and clustered metadata surrogate interface and the level of complexity that each task represents.

During the testing, the 'think aloud' protocol was employed to elicit further qualitative and concurrent data about how users used both interfaces for identifying relevant learning objects. In addition, users' actions were captured through the use of screen recording software (Camtasia studio, v.4). After each 'task test' session users were asked to complete a subjective satisfaction questionnaire for each interface and took part in a semi-structured interview. In the interviews users were asked about their satisfaction with the learning object metadata elements, what additional metadata elements could be included in the records for facilitating relevance judgment, which style of presentation users liked the most, and how they would like metadata records to be displayed in search result interfaces. The research implements were piloted and then the testing took place during November 2006, at the Research School of Informatics of Loughborough University.



Linear metadata surrogate interface	
Task 1. Find a lecture on the design of usable multimedia resources.	Low complexity
Task 2. Find lectures on the digital divide for HE students.	Medium complexity
Task 3. You need to do some general reading on information literacy using resources of high interactivity, for HE students. Make sure that the resources identified are not in a PDF or PPT format.	High complexity
Clustered metadata surrogate interface	
Task 1. Find exercises on database design.	Low complexity
Task 2. Find exercises on HTML design for HE students.	Medium complexity
Task 3. You need to do some general reading on information retrieval systems using resources of high interactivity for HE students. Make sure that the resources identified are not in PDF or PPT format.	High complexity

Table 1: Task table

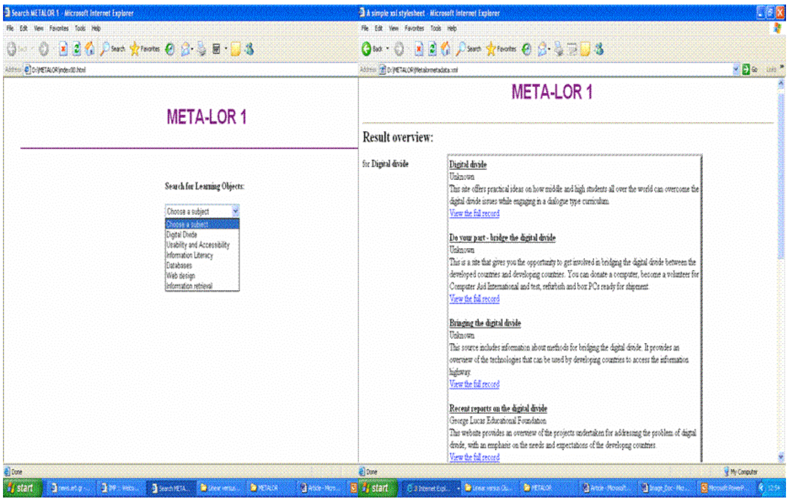
The data analysis included estimation of the means and statistical analysis (ANOVA tests) for the time needed for users to complete the tasks and users' ratings of subjective satisfaction. In addition, content analysis was performed on the qualitative data, such as the data collected from the interviews and the transcripts of the think aloud protocols.

3.2 The Meta-Lor 1 Interface

For the needs of this study a prototype learning object metadata repository system was set up using HTML, JavaScripts and XML technologies. The system stored and provided access to 60 learning object metadata records coded in XML. Metadata records included an identifier that provided access to the learning object itself. The data structure of the META-LOR 1 system is based on 19 elements derived from the LOM standard. The number and terminology of the elements was finalised after the pilot testing.

The prototype consisted of three main interfaces. The first was a simple search interface (see Figure 1). The second was the 'the search result overview' that provided a list of the retrieved results. The list included only the title of the learning object, the name of the author of the learning object, an abstract of the contents of the learning object and a link to the metadata record preview (see Figure 1). The third interface was 'the metadata surrogate preview' which included all 19 metadata elements. This interface comprised two different designs. The first presented metadata elements in a list (linear metadata surrogate interface) (see Figure 2) and the second divided metadata elements into three sections: General, Educational, and Technical metadata (clustered metadata surrogate interface) (see Figure 3).

In order to improve users' understanding of the metadata vocabularies, a definition of each metadata element was provided in a pop up box. The pop up boxes were contextually displayed every time the user selected a particular element with the mouse.


Search METALOR 1 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address D:\METALOR\index00.html
A simple xsl stylesheet - Microsoft Internet Explorer
Address D:\METALOR\Metalormetadata.xml
Done
My Computer
start

META-LOR 1

Search for Learning Objects:

Choose a subject
Choose a subject
Digital Divide
Usability and Accessibility
Information Literacy
Databases
Web design
Information retrieval

META-LOR 1

Result overview:

for Digital divide

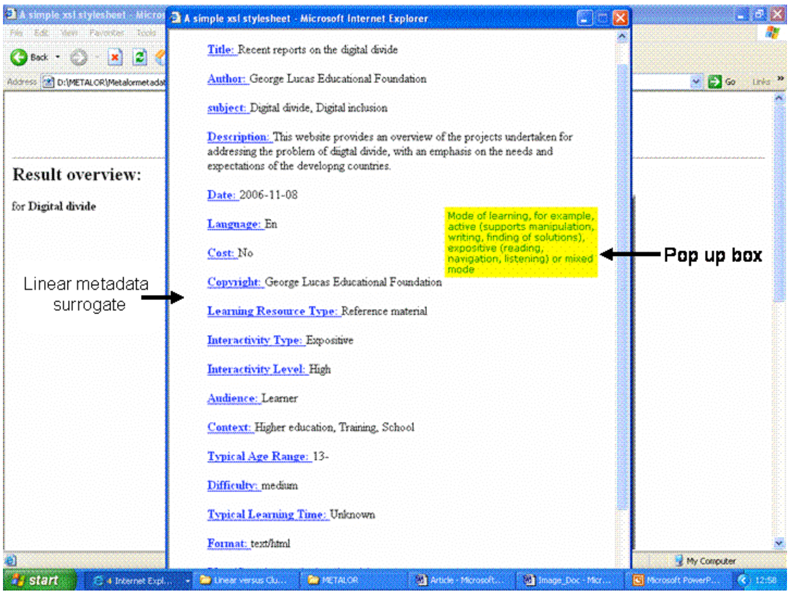
Digital divide
Unknown
This site offers practical ideas on how middle and high students all over the world can overcome the digital divide issues while engaging in a dialogue type curriculum.
View the full record

Do your part - bridge the digital divide
Unknown
This is a site that gives you the opportunity to get involved in bridging the digital divide between the developed countries and developing countries. You can donate a computer, become a volunteer for Computer Aid International and test, refurbish and box PCs ready for shipment.
View the full record

Bringing the digital divide
Unknown
This source includes information about methods for bridging the digital divide. It provides an overview of the technologies that can be used by developing countries to access the information highway.
View the full record

Recent reports on the digital divide
George Lucas Educational Foundation
This website provides an overview of the projects undertaken for addressing the problem of digital divide, with an emphasis on the needs and expectations of the developing countries.
View the full record

Figure 1: The search and search result overview interfaces


A simple xsl stylesheet - Microsoft Internet Explorer
Back
Address D:\METALOR\Metalormetadat
Go
Links
My Computer
start
Linear metadata surrogate
Pop up box

Result overview:

for Digital divide

Title: Recent reports on the digital divide

Author: George Lucas Educational Foundation

subject: Digital divide, Digital inclusion

Description: This website provides an overview of the projects undertaken for addressing the problem of digital divide, with an emphasis on the needs and expectations of the developng countries.

Date: 2006-11-08

Language: En

Cost: No

Copyright: George Lucas Educational Foundation

Learning Resource Type: Reference material

Interactivity Type: Expositive

Interactivity Level: High

Audience: Learner

Context: Higher education, Training, School

Typical Age Range: 13-

Difficulty: medium

Typical Learning Time: Unknown

Format: text/html

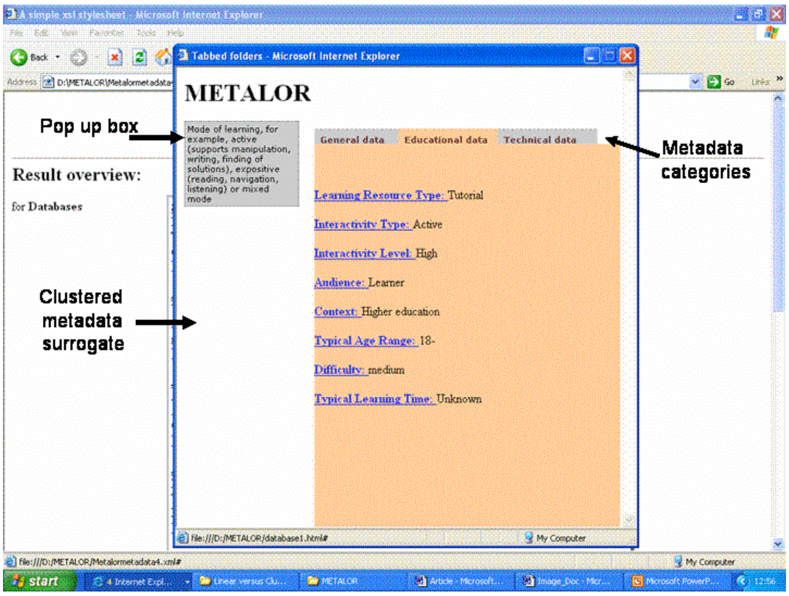
Mode of learning, for example, active (supports manipulation, writing, finding of solutions), expositive (reading, navigation, listening) or mixed mode

Figure 2: Linear metadata surrogate interface

4 Results and Discussion

4.1 Differences Between the Interfaces in Terms of Time

The analysis of time data revealed that participants performed the three tasks slightly faster using the clustered metadata surrogate interface. Participants needed an average time of 314 seconds to perform a task using the linear metadata surrogate interface and 301 seconds using the clustered metadata surrogate interface. This difference in time, however, was not statistically significant (p.>0.720).


A simple xsl stylesheet - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address D:\METALOR\Metalormetadata
Go
Links
Tabbed folders - Microsoft Internet Explorer
file:///D:/METALOR/database1.html#
My Computer
file:///D:/METALOR/Metalormetadata4.xml#
My Computer
start
4 Internet Expl...
Linear versus Clu...
METALOR
Article - Microsoft...
Image_Doc - Micr...
Microsoft PowerP...
12:56

Pop up box

Result overview:

for Databases

Clustered metadata surrogate

METALOR

Mode of learning, for example, active (supports manipulation, writing, finding of solutions), expositive (reading, navigation, listening) or mixed mode

General data | Educational data | Technical data

Metadata categories

Learning Resource Type: Tutorial

Interactivity Type: Active

Interactivity Level: High

Audience: Learner

Context: Higher education

Typical Age Range: 18-

Difficulty: medium

Typical Learning Time: Unknown

Figure 3: Clustered metadata surrogate interface

4.2 Impact of Task Complexity on the User Performance

Figure 4 summarizes the mean time needed for users to perform the three tasks in both interfaces. Based on these results it can be implicitly suggested that users completed the low and medium complexity tasks faster (Task 1 and Task 2) when they used the clustered metadata surrogate interface. On the other hand, users found it more efficient to identify relevant learning objects using the linear metadata surrogate interface (Task 3). These differences, however, were not statistically significant and further research is needed to investigate this further (p.>0.203).

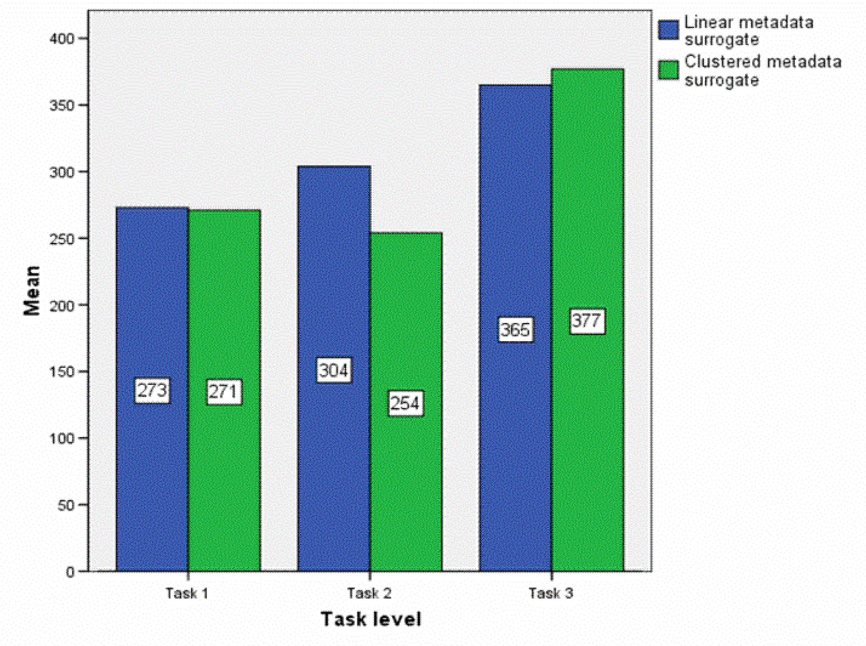


Figure 4: Difference in time across the three tasks

4.3 Subjective Satisfaction Questionnaire

Users' overall satisfaction with the two types of interfaces was measured in terms of four subjective measures: 1. Satisfaction, 2. Stimulation, 3. Easy of use and 4. Satisfaction with the presentation of metadata elements in the surrogate. The results revealed that participants' overall satisfaction (across the four subjective measures) was significantly higher in the case of the clustered metadata surrogate interface (mean overall satisfaction = 7.8) rather than the linear metadata surrogate interface (mean overall satisfaction = 6.3) (F=105.308, p.<0.01). Figure 5 presents the statistically and non-statistically significant differences between the two interfaces for each subjective measure ('Satisfaction': F= 6,796, p.>0.05; 'Stimulation': F=68,200, p.<0.01; 'Easy of use': F=35.200, p.<0.01; 'Satisfaction with metadata presentation in the surrogate': F=22,184, p.<0.01).

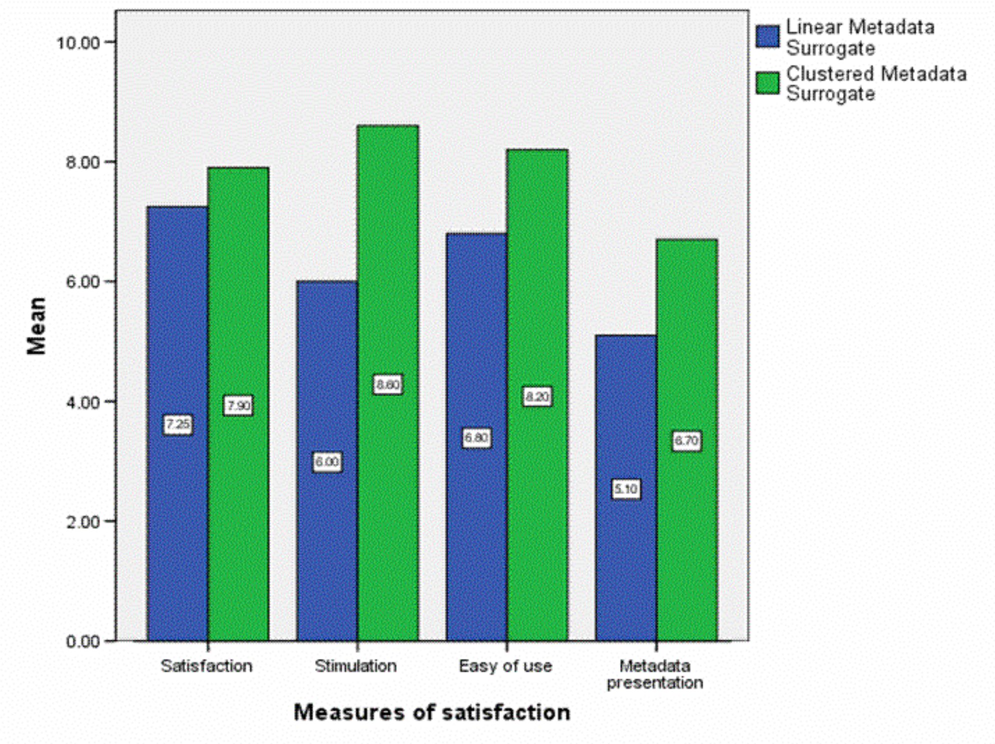


Figure 5: Differences in subjective satisfaction.

4.4 Qualitative Data (Interviews and Think Aloud Protocol)

4.4.1 Which Metadata Elements Users Liked The Most?

Participants in the study found the following metadata elements most useful for judging the relevance of learning objects: Title (12 users), Subject (12 users), Description (12 users), Difficulty level (12 users), Cost (10 users), Format (10 users), Identifier (10 users), Interactivity level (7 users), Audience (6 users). Based on these findings, it can be concluded that users favoured most of the 'General' and 'Technical' metadata categories. The former category was useful for users to judge the topical relevance of the learning object, while the latter provided users with information about how to access or download it. This observation agrees with previous studies that revealed users' preference towards content related metadata for evaluating the relevance of information and technical metadata for accessing it (Fraser and Gluck, 1999).

It is also worth mentioning that seven participants liked the use of pop up boxes that displayed information about learning object metadata elements and improved their understanding of some complex educational metadata elements, such as the 'interactivity type' and 'context' metadata. On the other hand, most of the educational elements were mentioned by only a few participants as useful. This, however, does not hold in the case of the 'difficulty', 'interactivity level' and 'audience' elements. These elements were regarded as useful by half or more participants in the study. This finding may be explained in two ways:

1. The laboratory and controlled nature of the study. This study employed a controlled task test with three predetermined tasks that were focused on specific metadata elements in order to be accomplished. More naturalistic studies that reflect real user needs in a variety of learning situations are needed to enhance the knowledge about how students use educational metadata to search for and judge the relevance of learning objects;

2. The semantic ambiguity of most of the educational metadata elements. During the task test a total of eight users found the meaning of some educational elements difficult to understand even though an explanatory pop up box accompanied these metadata elements. Future research should be focused on the investigation of the user-centeredness of the terminology and semantics of educational metadata. Such metadata should take into account learners' vocabularies and learning experiences.

4.4.2 What Other Metadata Elements Users Would Like to be Included?

The think aloud transcripts revealed that users did not like the inclusion of many metadata elements in the preview surrogates. This was more evident in the case of the linear metadata surrogate interface. Some of the comments users made regarding this issue were:

"There is too much information [...] Should I read all of it?"

[Male participant, linear metadata surrogate]

"I'll only select to read some of them [...]. Is this information really needed?"

[Male participant, linear metadata surrogate]

Although the length of the metadata records was criticized by some participants, few subjects mentioned that the metadata records should be enhanced with some additional information. Some of the information identified by participants is already included in the LOM standard. This includes: information about other similar resources available (Relation metadata category of LOM) and other people's comments about the resource (Annotation metadata category of LOM). It was interesting, however, that users identified the need for some information that is not explicitly covered by the existing LOM standard, such as, information about the time it takes for a learning object to be downloaded, accessibility needs (for example, how the learning object meets the needs of an heterogeneous community of learners) and information about the quality of a learning object.

4.4.3 Which Style of Metadata Presentation Users Liked the Most?

The majority of participants in the interviews (n=10) liked the way metadata was presented in the clustered metadata surrogate interface. The reasons that justify this preference are summarised in two categories:

1. *Plausibility and engagement*. The clustered metadata surrogate interface was characterised as more pleasant on the eye and more engaging. In addition, learners had the opportunity to focus on a specific category of metadata rather than the whole record;
2. *Structure and organisation of information*. The use of categories provides more structure to the metadata record. This minimises users' cognitive and memory load by making metadata more visible.

Two participants preferred the linear interface; they were more familiar with linear search result interfaces and metadata records and they did not like the way information was categorised in the clustered metadata surrogate interface. Further research needs to examine how metadata elements should be categorised in clustered metadata surrogates and whether the a priori categorisation of metadata elements as 'General', 'Educational' and 'Technical' should change in order to be better aligned with the mental model of learners.

4.4.4 Alternative Ways For Displaying Metadata Records

Although the majority of participants in the study preferred the clustered metadata interface, many users (n=9) mentioned that learners should be provided with the opportunity to control the display and the content of metadata records. The personalisation of metadata displays in search result interfaces could improve the relevance judgment process. For example, learners could select the number and type of metadata elements in search result interfaces. Such information could be either stored in learner profiles or generated dynamically during a search. In the latter case, the methodologies applied in relevance judgment research, sense-making and information foraging theory can provide a useful framework for predicting the dynamic nature of users' preferences.

5 Conclusions and Recommendations

This study suggests that learning object metadata elements in search result interfaces should be grouped into metadata categories. Although there is no significant impact of metadata interface design (linear or clustered metadata surrogate) on users' performance, users were significantly more satisfied with the clustered metadata surrogate interface which they perceived as stimulating and engaging to use. Future research should investigate the impact of different learning and cognitive styles on the design of learning object metadata-driven search result interfaces as well as employ more naturalistic research design methods. In addition, research is needed to challenge the a priori triple categorisation of learning object metadata as "General", "Educational" and "Technical". Although the findings of this study have implications in the design of metadata surrogates in other types of repositories and e-publishing systems, such as e-prints, institutional repositories, digital libraries, portals, library OPACs and WWW search engines, future research should focus on the needs of users of these systems as well. The paper concludes with some recommendations for the design of search result interfaces of learning object repositories. These are:

- The provision for alternative displays of metadata surrogates, for example, both in linear and clustered forms;
- The design of adaptive interfaces that present the content and format of metadata surrogates according to learners' needs;
- The use of pop up boxes for documenting and presenting the meaning of learning object metadata elements to users;
- The population of learning object repositories with less conventional educational metadata elements of LOM, such as, the 'difficulty level', 'interactivity level' and 'Audience'. In addition, there is a need for extending the LOM standard with information that users perceive as important to judge the relevance of learning objects, such as, 'the time it takes for a learning object to be downloaded', 'accessibility needs' information, as well as 'information about the 'quality' of a learning object.

References

[1] IHADJADENE, M.; CHAUDIRON, S. The effect of individual differences on searching the web. In: *Proceedings of the 66th Annual meeting of the American society for information science and technology*, vol. 40, 2003, pp.240 – 246.

[2] GRANKA, L.; HEMBROOKE, H.; GAY, G.; FEUNSER, M. Eye-Tracking Analysis of User Behavior in WWW Search. *In: Proceedings of the 27th annual international conference on Research and development in information retrieval*, 2004, pp. 478-479. Available at: <http://www.cs.cornell.edu/People/tj/publications/granka_etal_04a.pdf>, [accessed 24.06.2006].

[3] RESNICK, M. L.; MALDONADO, C. A.; SANTOS, J. M.; LERGIER, R. Modeling On-line Search Behavior Using Alternative Output Structures. In Proceedings of the Human Factors and Ergonomics Society 45th Annual Conference,

Minneapolis, MN, USA, 2001, pp. 1166-1171.

[4] DUMAIS, S.; CUTRELL, E; CHEN, H. Optimising search by showing results. In: *Proceedings of CHI 2001*, 2001. Available at: <http://research.microsoft.com/~sdumais/chi2001.pdf>, [accessed 25.10.2006].

[5] CHEN, H.; DUMAIS, S. Bringing order to the Web: automatically categorising search results. In: *Proceedings of CHI-00, ACM International conference on human factors in computing systems, 2000, p.145-152.* Available at http://research.microsoft.com/~sdumais/chi00.pdf, [accessed 23.10.2006].

[6] ZAMIR, O.; ETZIONI, O. Grouper: a dynamic clustering interface to web search results. In: *Proceedings of the 8th International WWW conference. Toronto, Canada*, 1999, pp. 1361-1374.

[7] RELE, R.S.; DUCHOWSKI, A. T. Using eye tracking to evaluate alternative search results interfaces. In: *proceedings of the Human Factors and Ergonomics society, September 26-30 2005, Orlando, FL, HFES*, 2005. Available at:

<http://andrewd.ces.clemson.edu/research/vislab/docs/Final_HFES_Search.pdf>, [accessed 09.07.2006].

[8] PRATT, W.; FAGAN, L. The usefulness of dynamically categorizing search results. *Journal of the American Medical Informatics Association*, 7 (6), 2000, pp. 605-617.

[9] FRASER, B.; GLUCK, M. Usability of geospatial metadata or space-time matters. *Bulletin of the American Society for Information Science*, vol.25, no.6, August/September 1999. Available at: <http://www.asis.org/Bulletin/Aug-99/fraser_gluck.html>, [accessed 12.07.2006].

[10] SARACEVIC, T. Relevance: a review of and framework for thinking on the notion in information science. *Journal of the American Society for Information Science*, no.6, November-December 1975, pp.321-344.

[11] BARRY, C. User-defined relevance criteria: an exploratory study. *Journal of the American Society for Information Science and Technology*, vol.45, no.3, 1994, pp.149-159.

[12] TANG, R.; SOLOMON, P. Use of relevance criteria across stages of document evaluation: on the complementarity of experimental and naturalistic studies. *Journal of the American Society for Information Science and Technology*, vol.52, no.8, 2001, pp.676-685.

[13] DRORI, O. How to display search results in digital libraries : user study. *In proceedings of the New Developments in Digital Library, 3rd International workshop*, 2003. Available at: <shum.huji.ac.il/~of...drori042003c.pdf>, [accessed 24.07.2006].

[14] PAEK, T.; DUMAIS, S.; LOGAN, R. WaveLens: a new view onto Internet search results. *In: proceedings of CHI2004, April 24-29, 2004, Vienna, Austria,2004, pp.727-734*. Available at: <http://research.microsoft.com/~timpaek/Papers/chi2004.pdf>, [accessed 25.10.2006].

[15] RIEH, S. Y. Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology*, vol.53, no.2, 2002, pp.145 – 161.

[16] LIDDY, et al. MetaTest: evaluation of metadata from generation to use. In: *Proceedings of the 3rd Joint Conference on Digital Libraries*. IEEE., 2003, pp.398. Available at: <http://csdl2.computer.org/comp/proceedings/jcdl/2003/1939/00/19390398.pdf>, [accessed 18.07.2006].

[17] LIDDY, L.; FINNERAN, T. *Meta Test: automatic generation of metadata and preliminary evaluation of its utility in information seeking*. [power point presentation], [n.d]. Available at: <http://nsdl.comm.nsdlib.org/meeting/poster_docs/2003/1093_MetaTest.pdf?nsdl_annual_meeting=d0e759bf75a8ccda313c2639cb72be81>, [accessed 03.07.2006].

[18] DIEKEMA, A. R. Evaluating metadata from different perspectives. [power point presentation]. *In: Metadata tools for digital resources repositories, JCDL Workshop,* June 15, 2006. Available at <http://www.ils.unc.edu/mrc/jcdl2006/slides/diekema.pdf>, [accessed 12.07.2006].

[19] *FAILTE's SearchLT evaluation.* FAILTE, 2002. Available at: <http://www.failte.ac.uk/documents/eval_report.rtf>, [accessed 12.07.2006].

Disclosing Freedom of Information Releases

Ann Apps

MIMAS, The University of Manchester, M13 9PL, UK
e-mail: ann.apps@manchester.ac.uk

Abstract

The Freedom of Information (FOI) Acts passed in 2000 in England and Wales and in 2002 in Scotland require organisations, including UK Higher Education Institutions (HEI), to provide requested information within certain conditions. The JISC Information Governance Gateway (JIGG) project aims to provide a single online gateway into information and resources related to HEIs' compliance with information governance legislation, including FOI. One of the project's objectives is to provide dissemination of the FOI disclosure logs by a web search within the gateway and also using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). It is hoped this will assist HEI records management practitioners in sharing their experiences of dealing with FOI requests, and lead to future collaborations within a wider community. This paper describes the development of a JIGG FOI Application Profile as a 'template' for FOI disclosure log entries, and its subsequent translation into a practical application.

Keywords: Freedom of Information; information governance; records management; OAI-PMH; Dublin Core Application Profile

1 Introduction

The Freedom of Information (FOI) Acts passed in 2000 in England and Wales [1] and in 2002 in Scotland [2] require organisations to provide requested information to enquirers within a given timescale, unless the requested information is exempted under the legislation. The organisations covered by this requirement include UK Higher Education Institutions (HEI). To show just a few examples, people have asked The University of Manchester for information about the University's coat of arms, the amount of money taken in library fines, prospectuses supplied on recycled paper, and the awarding of honorary degrees to the Bee Gees.

HEIs are encouraged to publish disclosure logs that summarise the FOI requests they have received and the information they have released. Currently only a minority of HEIs maintain such public disclosure logs and in most cases these consist of simple lists on web pages, for example one page per year. Many other UK organisations do publish FOI disclosure logs but not in any consistent format or single place [3].

The JISC Information Governance Gateway (JIGG) project [4] aims to provide a single online gateway into information and resources related to HEIs' compliance with information governance legislation, including FOI, as well as data protection, environmental information, practical issues such as records management and related legislation such as copyright. The gateway also provides a private discussion area for HEI records management practitioners.

In addition to being a portal for relevant and up-to-date information about governance resources, JIGG publishes HEI Publication Schemes as defined under the FOI Acts, and Disclosure Logs where available. A project objective is to provide dissemination of the FOI disclosure logs [5] using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [6] as well as providing a web search within the gateway. Provision of a JIGG FOI OAI-PMH service will allow other applications regularly to gather new JIGG FOI disclosure log entries into their own databases. It is hoped this will assist HEI records management practitioners in sharing their experiences of dealing with FOI requests, and lead to future collaborations within a wider community.

2 Methodology

2.1 The JIGG FOI Domain Model

In order to disseminate the FOI disclosure log entries over OAI-PMH it was necessary to define a 'template' of appropriate data fields. This was developed by investigating the content of existing FOI disclosure logs, with subsequent agreement within the community, the 'template' being expected to conform to the requirements of the Office of the Information Commissioner, the Scottish Information Commissioner and the Department of Constitutional Affairs. Thus the main entity within the JIGG FOI domain, or application, model is an FOI disclosure log entry, comprising this identified set of properties. Additionally each disclosure log entry has an associated set of administrative metadata that describes information 'about' the disclosure log entry within the application.

A disclosure log entry is a 'closed' record of a request for, and the release of, information. It is not anticipated that there would be any changes to FOI disclosure log entries after they've been entered into the JIGG system, except for minor textual corrections.

2.2 The JIGG FOI Application Profile

The disclosure log entry and the administrative metadata are documented using a Dublin Core Application Profile, based on the European standard (CEN Workshop Agreement) Dublin Core Application Profile (DCAP) Guidelines [7]. Standard Dublin Core properties [8] are used where applicable. The Application Profile indicates the source Dublin Core definitions of and comments about these properties as well as the application specific variations. Additional JIGG specific properties have been introduced where there were no suitable standard properties. Each of these is defined in a JIGG FOI namespace, currently within a human readable 'mini application profile' with its URI (Uniform Resource Identifier, a unique persistent identifier within the global internet) grounded on its position in that document, and with an intention of persistence.

The DCAP Guidelines specify an application profile that captures a single entity. This corresponds to a single resource description within the Dublin Core Abstract Model (DCAM) [9], which specifies a flat set of properties for a single resource, with no provision for any composite properties according to any hierarchical model and syntax. Thus some extension of the DCAP has been made to capture both a disclosure log entry and its associated administrative metadata, which together make up a "description set" within the DCAM. Thus the DCAP is composite with a section for each "description", preceded by a section that specifies these entities.

2.3 FOI Disclosure Log Entry Properties

Some of the properties within an FOI disclosure log entry are available for discovery purposes through the application. These properties capture the content of the request, when it was made, the HEI, and the relevant legislation and exemptions. All the properties provide documentation of the course of a request, such detail being potentially useful for processing future similar requests.

Table 1 lists the FOI disclosure log entry properties taken from namespaces Dublin Core ('dc' and 'dcterms') and JIGG ('jigg'), the URIs of which are defined within the 'dcxm:descriptionSet' of Figure 2. Occurrence requirements are displayed as 'Min' and 'Max', which also implicitly indicate whether properties are mandatory or optional and whether they are repeatable.

An FOI disclosure log entry contains several properties that summarise in free text the information that was requested, the information that was released and how the request was processed. Each disclosure log entry has a title indicating very briefly the topic of information requested. A more detailed summary of the request is the free text value of a description property. Optional summaries of the information released and the process of answering the request are further text fields. The topic of a request may be indicated by terms taken from the JISC Function Activity Model Vocabulary [10]. Each of the textual fields can be tagged with a language code, included for possible future enhancement of the application.

The HEI to which the FOI request was made is captured as the publisher. The JIGG FOI data submission system aims to ensure consistency of institution names. The country where the HEI is based is captured with values taken from a JIGG-defined vocabulary, which currently contains only the four UK countries (England, Northern Ireland, Scotland and Wales). An optional local identifier may be included where it is used by the HEI to denote the log entry. Ideally this identifier should be a URI and mandatory. But it was felt that at this stage of the JIGG project, requiring a global identifier may present too high a barrier to inclusion of records from information governance practitioners who may not currently use a consistent identification system, and may not be conversant with URIs. HEIs may currently publish some details of their FOI requests in some way, possibly a web page describing several requests within some time period. Thus a link to this composite disclosure log is included. Some HEIs publish the full text of the information released, so the disclosure log entry has an optional, repeatable link to possibly several documents.

Property	Definition (Summary)	Content / Vocabulary	Min	Max
dc:title	Title of log entry	text	1	1
dc:identifier	Local identifier		0	1
dc:publisher	Organisation publishing log	text	1	1
jigg:country	Country of publishing organisation	England; Northern Ireland; Sotland; Wales	1	1
dcterms:isPartOf	Organisation's disclosure log	URI	1	1
jigg:dateReceived	Date FOI request received	W3CDTF	1	1
dc:description	Summary of information requested	text	1	1
dc:subject	Topic of request	Function Activity Model vocabulary (in jigg namespace)	0	unbounded
jigg:infoReleased	How much information released	no; partial; yes	1	1

jigg:legislation	Applicable legislation	Freedom of Information Act 2000;
		Freedom of Information (Scotland) 2002;
		Environmental Information Regulations 2004;
		Environmental Information (Scotland) Regulations 2004
			1	1

jigg:exemptionsUsed	Exemptions used when processing request	Exemptions relevant to above applicable legislation taken from vocabulary in jigg namespace	0	unbounded
jigg:requestHistory	How request was processed	text	0	1
jigg:responseSummary	Summary of information released	text	0	1
dcterms:references	Full text response	URI	0	unbounded

Table 1: FOI Disclosure Log Entry Properties (jigg:foiLog)

The date on which an FOI request was received is included within the disclosure log entry. It is probable that this date will be used for discovery, as well as providing a record of when the request occurred. Within the plethora of Dublin Core 'date' element refinements there was not one that exactly fitted the semantics of receipt date. It seemed a better option to define a JIGG-specific property rather than trying to shoehorn a definition into an inappropriate Dublin Core date property, or adopting a term from an obscure namespace. Theoretically it would seem appropriate to indicate closure in dealing with an FOI request by also capturing the completion date, especially as the FOI Acts specify time limits. However, in practice, practitioners were reluctant for this detail to be included, partly because of a belief that it would expose them to unwelcome levels of scrutiny. Because there are concerns about the initial level of engagement with the project by HEI FOI practitioners, it was decided to omit completion date from the set of properties in the Application Profile.

The FOI Acts define various exemptions under which it is permissible to refuse to provide information or to supply it only partially. There are some differences in the lists of exemptions or in their wording between the Freedom of Information Act 2000, which applies to England and Wales [11], and the Freedom of Information Act (Scotland) 2004 [12]. Another pair of differing exemptions lists apply to the Environmental Information Regulations 2004 [13] and the Environmental Information Regulations (Scotland) 2004 [14]. Thus it is beneficial to record under which legislation, of these four, an FOI request has been processed, which exemptions were relevant (a repeatable property), and how much information was released (possible values being 'no', 'partial', or 'yes'). Each of these properties has a JIGG-defined vocabulary. Although the exemptions associated with the various legislations are listed in several publicly available documents, there do not appear to be any existing formal vocabularies. Thus vocabularies have been defined within the JIGG FOI namespace for reference by property values within the JIGG FOI application.

2.4 FOI Disclosure Log Entry Administrative Metadata

Associated with an FOI disclosure log entry is a set of administrative metadata, listed in Table 2. This includes the URI of the publishing organisation, which is JIGG itself for the central JIGG FOI application, and the originating organisation, which is the URI corresponding to the HEI named as publisher within the log entry itself. A JIGG identifier, a URI, is assigned to each FOI disclosure log entry within the application.

There are various rights captured about the disclosure log entry. Copyright belongs to the publishing HEI. Creative Commons [15] rights cover subsequent use of the disclosure log entry, indicating that the information is freely available for non-commercial use, provided attribution of its provenance is maintained, but no derivatives may be made. This seemed an appropriate rights statement for information released from publicly funded HEIs. A further rights statement requires that this administrative metadata must always be retained with the disclosure log entry.

The date when a disclosure log entry was entered into the JIGG FOI database, or when it was last updated, is recorded as ‘dcterms:modified’, which “dumbs down” to ‘dc:date’ when the administrative metadata is supplied according to simple Dublin Core. This is the significant date used for the OAI-PMH application, which provides harvesting based on ‘last modification date’.

Finally there is a relation that ties the administrative metadata to the disclosure log entry. This relation is used when both entities appear within an XML document description set, with value an internal identifier of the disclosure log entry within that document. It is not used for OAI-PMH dissemination where the relation between the ‘about’ part of a record and the ‘metadata’ is implicit.

Property	Definition (Summary)	Content
dc:identifier	Identifier of log entry within JIGG	URI
dc:creator	Originating organisation	URI
dc:publisher	Publisher of disclosure log entry	“http://www.jigg.ac.uk”
dcterms:modified	Date log entry added to JIGG repository	W3CDTF
dc:rights	Copyright over log entry	text
dc:rights	Creative Commons rights over reuse	“http://creativecommons.org/licenses/by-nc-nd/2.0/uk/”
dc:rights	Administrative metadata requirement	“The JIGG administrative metadata must always be retained with its associated disclosure log entry description.”
dc:relation	Link to related FOI disclosure log entry	Local identifier within an XML document

Table 2: FOI Disclosure Log Entry Administrative Metadata (jigg:foiAdmeta)

2.5 The JIGG FOI XML Serialisation

Dissemination of records via OAI-PMH is by an XML serialisation of the data that is defined in the JIGG Application Profile [16] and conformant to an XML schema. Because the Application Profile is conformant to the Dublin Core Abstract Model (DCAM), it seemed appropriate to follow Dublin Core guidelines for the XML serialisation. A proposed ‘Dublin Core in XML’ [17] format that is consistent with the DCAM is under development by the Dublin Core Metadata Initiative. However the capability within this proposed XML format to support the full DCAM results in a rather verbose XML record for general usage. Although the XML data is intended for use by machines, which have no concerns about complexity apart from efficiency, there are also human considerations. A complex XML format requires more effort to both create and process, and so is consequently more error prone. Because of these concerns a restricted functionality version of Dublin Core in XML has also been suggested. The JIGG FOI XML schema follows this ‘Dublin Core in XML Minimal’ [18] as it is was proposed at the time of schema development.

```
@prefix dc: <http://purl.org/dc/elements/1.1/>
@prefix dcterms: <http://purl.org/dc/terms/>
@prefix jigg: <http://www.jigg.ac.uk/foi/terms/#>

DescriptionSet (
#FOILog
   Description ( DescriptionId (log-1234567-890123)
      Statement ( PropertyURI ( dc:title )
          ValueString ( "An Example FOI Disclosure Log"  Language ( en-GB ) ) )
     Statement ( PropertyURI ( dc:identifier )
         ValueString ( "02/2006" ) )
     Statement ( PropertyURI ( dc:publisher )
         ValueString ( "Somewhere University" ) )
     Statement ( PropertyURI ( jigg:country )
         VocabularyEncodingSchemeURI ( jigg:FoiCountryVocab )
         ValueString ( "England" ) )
     Statement ( PropertyURI ( dcterms:isPartOf )
         ValueURI ( <http://www.somewhere.ac.uk/foilogs/> ) )
      Statement ( PropertyURI ( jigg:dateReceived )
         ValueString ( "2006-08-01"  SyntaxEncodingScheme ( dcterms:W3CDTF ) ) )
      Statement ( PropertyURI ( dc:description )
         ValueString ( "Details of the information requested"  Language ( en-GB ) ) )
     Statement ( PropertyURI ( dc:subject )
         VocabularyEncodingSchemeURI ( jigg:FAMVocab )
         ValueString ( "D External Relations" ) )
     Statement ( PropertyURI ( jigg:infoReleased )
         VocabularyEncodingSchemeURI ( jigg:InfoReleasedVocab )
         ValueString ( "partial" ) )
     Statement ( PropertyURI ( jigg:legislation )
         VocabularyEncodingSchemeURI ( jigg:FoiLegislationVocab )
         ValueString ( "Freedom of Information Act 2000" ) )
     Statement ( PropertyURI ( jigg:exemptionsUsed )
         VocabularyEncodingSchemeURI ( jigg:FoiAct2000Vocab )
         ValueString ( "26 Defence" )
         ValueString ( "29 The Economy" ) )
      Statement ( PropertyURI ( jigg:requestHistory )
         ValueString ( "Details of processing the request"  Language ( en-GB ) ) )
      Statement ( PropertyURI ( jigg:responseSummary )
         ValueString ( "The answer" Language ( en-GB ) ) )
     Statement ( PropertyURI ( dcterms:references )
         ValueURI ( <http://www.somewhere.ac.uk/foilogs/02-2006.pdf> ) ) )
#FOIAdmeta
   Description ( ResourceURI( http://www.jigg.ac.uk/foi/ids/1234567-890123)
     Statement ( PropertyURI ( dc:identifier )
         ValueURI ( http://www.jigg.ac.uk/foi/ids/1234567-890123 ) )
     Statement ( PropertyURI ( dc:creator )
         ValueURI ( <http://somewhere.ac.uk> ) )
     Statement ( PropertyURI ( dc:publisher )
         ValueURI ( <http://www.jigg.ac.uk> ) )
      Statement ( PropertyURI (dcterms:modified )
         ValueString ( "2006-09-05"  SyntaxEncodingScheme ( dcterms:W3CDTF ) ) )
     Statement ( PropertyURI ( dc:rights )
         ValueString ( "Copyright Somewhere University 2006" ) )
     Statement ( PropertyURI ( dc:rights )
         ValueURI ( <http://creativecommons.org/licenses/by-nc-nd/2.0/uk/>  ) )
     Statement ( PropertyURI ( dc:rights )
         ValueString ( "The JIGG administrative metadata must always be retained with its
                           associated disclosure log entry description." ) )
     Statement ( PropertyURI ( dc:relation )
         DescriptionRef ( log-1234567-890123 ) ) ) )
```

Figure 1: A DC-Text Example of an FOI Disclosure Log Entry

As an interim stage, a DC-Text [19] hypothetical example was produced to illustrate conformance to the DCAM. DC-Text provides a formal but relatively syntax-free means to document a metadata description set that is ideal for the development and discussion stage. This DC-Text example, which informed the development of the JIGG FOI XML schema [20], is shown in Figure 1.

3 Results

3.1 Dissemination of FOI Disclosure Log Entries

FOI disclosure log entries stored in the JIGG FOI central database are disseminated over OAI-PMH according to an ‘oai_jiggfoi’ metadata format. The XML ‘metadata’ part of a ‘GetRecord’ response conforms to the JIGG FOI XML schema. Examples are shown in Figures 2 and 3, illustrating different styles of FOI records management by two HEIs, and the use of different properties taken from the Application Profile.

```
<metadata>
 <dcxm:descriptionSet xmlns:dcxm=“http://dublincore.org/xml/dc-xml-min/2006/09/18/”
  xmlns:dc=“http://purl.org/dc/elements/1.1/” xmlns:dcterms=“http://purl.org/dc/terms/”
  xmlns:jigg=“http://www.jigg.ac.uk/foi/terms/#” xmlns=“http://www.jigg.ac.uk/foi/”
  xmlns:xsi=“http://www.w3.org/2001/XMLSchema-instance”
  xsi:schemaLocation=“http://www.jigg.ac.uk/foi/ http://www.jigg.ac.uk/foi/schemas/2006/10/30/jiggfoi.xsd”>
  <jigg:foiLog>
   <dc:title>LJMU Review Magazine cost</dc:title>
   <dc:identifier>FOI 5/12</dc:identifier>
   <dc:publisher>Liverpool John Moores University</dc:publisher>
   <jigg:country dcxm:vocabEncSchemeURI=“http://www.jigg.ac.uk/foi/terms/#FoiCountryVocab”>
    England</jigg:country>
   <dcterms:isPartOf dcxm:valueURI=“http://www.ljmu.ac.uk/secretariat/75554.htm”/>
   <jigg:dateReceived dcxm:syntaxEncSchemeURI=“http://purl.org/dc/terms/W3CDTF”>
    2005-06-07</jigg:dateReceived>
   <dc:description>
    Costs of producing the LJMU Review Magazine and detailed accounts for the Marketing department for
     the previous 5 years
   </dc:description>
   <jigg:infoReleased dcxm:vocabEncSchemeURI=“http://www.jigg.ac.uk/foi/terms/#InfoReleasedVocab”>
    partial</jigg:infoReleased>
   <jigg:legislation dcxm:vocabEncSchemeURI=“http://www.jigg.ac.uk/foi/terms/#FoiLegislationVocab”>
    Freedom of Information Act 2000</jigg:legislation>
   <jigg:exemptionsUsed dcxm:vocabEncSchemeURI=“http://www.jigg.ac.uk/foi/terms/#FoiAct2000Vocab”>
    43 Commercial interests</jigg:exemptionsUsed>
   <jigg:requestHistory>
    LJMU disclosed information on the costs of producing the magazine, although not all the information
    requested was held. LJMU, after applying the public interest test, refused disclosure of the details
    accounts of the Marketing Department, citing Section 43 of the FOIA.
   </jigg:requestHistory>
  </jigg:foiLog>
 </dcxm:descriptionSet>
</metadata>
```

Figure 2: An Example FOI Disclosure Log Entry from Liverpool John Moores University

As required by OAI-PMH, records are also disseminated in simple Dublin Core for interoperability, informed by a mapping from the FOI Application Profile. The administrative metadata is disseminated in simple Dublin Core within an ‘about’ section of the ‘GetRecord’ response, an example being shown in Figure 4. Further ‘about’ sections detail metadata rights according to the OAI-PMH Guidelines for Conveying Rights, and the provenance of any records that have been harvested from elsewhere, conforming to the appropriate OAI-PMH Provenance schema.

```
<metadata>
 <dcxm:descriptionSet xmlns:dcxm=“http://dublincore.org/xml/dc-xml-min/2006/09/18/”
  xmlns:dc=“http://purl.org/dc/elements/1.1/” xmlns:dcterms=“http://purl.org/dc/terms/”
  xmlns:jigg=“http://www.jigg.ac.uk/foi/terms/#” xmlns=“http://www.jigg.ac.uk/foi/”
  xmlns:xsi=“http://www.w3.org/2001/XMLSchema-instance”
  xsi:schemaLocation=“http://www.jigg.ac.uk/foi/ http://www.jigg.ac.uk/foi/schemas/2006/10/30/jiggfoi.xsd”>
  <jigg:foiLog>
   <dc:title>Student accommodation landlords</dc:title>
   <dc:publisher>The University of Manchester</dc:publisher>
```

```
    <jigg:country dcxm:vocabEncSchemeURI="http://www.jigg.ac.uk/foi/terms/#FoiCountryVocab">
     England</jigg:country>
    <dcterms:isPartOf dcxm:valueURI="http://www.manchester.ac.uk/aboutus/documents/foi/disclosurelog/"/>
    <jigg:dateReceived dcxm:syntaxEncSchemeURI="http://purl.org/dc/terms/W3CDTF">
     2005-01-05</jigg:dateReceived>
    <dc:description>List of landlords on the student accommodation list</dc:description>
    <dc:subject dcxm:vocabEncSchemeURI="http://www.jigg.ac.uk/foi/terms/#FAMVocab">
     B Student Administration and Support</dc:subject>
    <jigg:infoReleased dcxm:vocabEncSchemeURI="http://www.jigg.ac.uk/foi/terms/#InfoReleasedVocab">
     yes</jigg:infoReleased>
    <jigg:legislation dcxm:vocabEncSchemeURI="http://www.jigg.ac.uk/foi/terms/#FoiLegislationVocab">
     Freedom of Information Act 2000</jigg:legislation>
    <dcterms:references
     dcxm:valueURI="http://www.manchester.ac.uk/medialibrary/foi/disclosures/studentadmin/landlords.pdf"/>
   </jigg:foiLog>
  </dcxm:descriptionSet>
</metadata>
```

Figure 3: An Example FOI Disclosure Log Entry from The University of Manchester

```
<about>
 <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
   xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
    http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
  <dc:creator>http://www.manchester.ac.uk</dc:creator>
  <dc:publisher>http://www.jigg.ac.uk</dc:publisher>
  <dc:date>2007-03-28</dc:date>
  <dc:rights>Copyright The University of Manchester 2005</dc:rights>
  <dc:rights>http://creativecommons.org/licenses/by-nc-nd/2.0/uk/</dc:rights>
  <dc:rights>The JIGG administrative metadata must always be retained with its associated log entry
    description.</dc:rights>
 </oai_dc:dc>
</about>
```

Figure 4: Administrative Metadata for the FOI Disclosure Log Entry of Figure 3.

3.2 Data Creation

Submission of FOI disclosure log entries into the JIGG central database was a significant factor for the project to consider. A Web-form based data Editor allows input of values of the various fields defined by the Application Profile. Use of a dedicated Editor ensures consistency of records, in particular with respect to the various vocabularies, and the generation of valid XML. The design of this system is based on that of the JISC Information Environment Service Registry (IESR) [21], as is the rest of the JIGG FOI application, thus reusing an established application developed as part of another JISC project.

A consideration is obviously the effort required to supply FOI disclosure log entries to JIGG, and the additional steps that may have to be included in existing workflows. It is hoped that use of the JIGG FOI data Editor for log entry submission will not be too onerous. It is thought that the majority of administrative operations within HEIs are based on Excel spreadsheets, thus necessitating 'copy and paste' into the JIGG submission form.

Where possible data fields are populated automatically, which also ensures consistency. For example, the publisher's name will be taken from an HEI's initial registration as a data contributor. Further data congruity is achieved by setting values from vocabulary term lists, such as exemptions, by selection menus within the Editor.

3.3 Data Harvesting

A future vision is automatic population of the central JIGG FOI database of disclosure log entries via OAI-PMH. If an HEI provided a harvest service onto their FOI disclosure logs, using the OAI-PMH standard and the 'oai_jiggfoi' metadata format, JIGG could gather and ingest them on a regular basis. Possibly an HEI could incorporate population of this OAI-PMH enabled database into the process of responding to FOI requests. If they were to publish their FOI disclosure log entries in this way, then submission to JIGG could become automatic.

4 Discussion

4.1 Incorporation into the JIGG Portal

The FOI disclosure logs application is just a part of the JIGG portal. The human user's view of the gateway is controlled by a Content Management System (CMS) to provide a consistent interface to all aspects, informed by a considered information architecture, and implemented by a JIGG-specific template and web style sheet. Thus it is necessary for the web search of the FOI disclosure logs, and their display to end users, to appear within the CMS, rather than as a separate, potentially inconsistent, application provided by the IESR-based implementation. This implies that the web interface to the FOI disclosure log entries will be provided by a server within the CMS. This server will maintain its data records by regularly gathering new records from the separate FOI disclosure logs application. OAI-PMH is the obvious choice for this data interchange, because of its capability for supplying new or changed records on a regular basis after an initial bulk data load.

4.2 Publication of FOI Disclosure Logs

So far this paper has focused mainly on the technical aspects of the JIGG FOI disclosure logs application. But there are, of course, social aspects. As yet few HEIs publish their disclosure logs. This reluctance may be simply because of insufficient staff resource. But there may be a lack of motivation because of a perception that there is no value in sharing this information. Or, further, there may be an active objection because of concerns about accountability.

The project hopes to encourage more HEIs to publish their FOI disclosure logs and to promote their publication in JIGG. One approach will be to hold workshops for HEI records management practitioners who are potential data contributors, to advertise and explain the facility. The JIGG project has engaged with, and has support from, a range of stakeholders, and has several UK regional Advisory Panels consisting of practitioners and representatives from relevant bodies. An 'Information Legislation and Management Survey' of HEIs [22], which portrayed their current handling of FOI requests was recently undertaken by JISC infoNet.

Hopefully, as JIGG is populated with a sizeable corpus of FOI disclosure log entries, the value of such a resource will become apparent. Publishing summaries of information released following FOI requests, and in some cases the full text of responses, will potentially reduce the number of requests for the same information. It will enable HEIs to share their experiences of responding to such requests. It should avoid 'reinventing the wheel' by individual HEIs as they consider aspects of legal compliance that apply to the whole sector. Currently they could potentially give differing responses. Thus the JIGG FOI database should both help to ensure a consistency of response to similar requests, and potentially reduce the resource requirements on records management staff. Essentially JIGG is providing a platform for accumulating and sharing 'frequently asked questions' and their answers.

5 Conclusions

At present these advantages of a central repository of HEI FOI disclosure logs are largely hypothetical. The JIGG project intends to provide the practical infrastructure to realise the vision as the project matures over the next eighteen months. But this does depend on engaging the participation of HEI records management practitioners.

One consideration, mentioned above, is obviously the effort required to supply FOI disclosure log entries to JIGG. The vision is a scenario where HEIs publish their FOI disclosure log entries in an OAI-PMH enabled database, incorporated into their business processes for dealing with FOI requests. JIGG would harvest these disclosure log entries into its central database on a regular basis. The use of OAI-PMH would remove the need for manual effort once the system is in place. But this scenario does imply knowledge of OAI-PMH and technical development capability by the HEI's administration department.

The experience of using an Application Profile within the JIGG project has shown it to be invaluable for developing and formally documenting a metadata schema. It proved to be an ideal format to assemble, communicate and discuss suitable properties during the process of gaining agreement, and for dissemination of the details to other interested parties. This was within a sector where there was not general awareness of metadata schemas and no previous knowledge of OAI-PMH. The Application Profile provides a clear specification even to those who are not conversant with metadata schemas. It affords a relatively 'syntax free' format understandable by non-technical people. The JIGG FOI Application Profile is a web document, so it includes hyperlinks between various sections and definitions, which hopefully enhance usability by readers. At the same time it is regarded as a formal specification with a persistent URI.

The JIGG project is utilising, and thus disseminating awareness of, OAI-PMH within a new sector. But the marriage of Open Archives and Freedom of Information seems apt. It is envisaged that sharing of FOI disclosure log entries may be broadened to other organisations beyond HEIs, if they adopt the JIGG FOI Application Profile. This interoperability is assisted by using a standards-based approach, in particular by employing OAI-PMH.

Acknowledgements

The JIGG project is funded at The University of Manchester by the Joint Information Systems Committee (JISC) of the UK Higher and Further Education Councils, with technical development by MIMAS. The author wishes to acknowledge the contribution of project colleagues to the developments presented in this paper, in particular Steve Woods and Chris Watts previously at Liverpool John Moores University.

Notes and References

[1] *The Freedom of Information Act 2000*. London : The Stationery Office Ltd., 2000. Retrieved, March 29, 2007, from http://www.opsi.gov.uk/acts/acts2000/20000036.htm

[2] *The Freedom of Information (Scotland) Act 2002*. Scotland : The Stationery Office Ltd., 2002. Retrieved, March 29, 2007, from http://www.hmso.gov.uk/legislation/Scotland/acts2002/20020013.htm

[3] WOOD, S. Freedom of Information & Open Government Blog Disclosure Log Index. 2007. Retrieved, March 29, 2007, from http://www.foi-directory.org/

[4] JIGG. JISC Information Governance Gateway. 2007. Retrieved, March 29, 2007, from http://www.jigg.ac.uk/

[5] APPS, A. JISC Information Governance Gateway: FOI Disclosure Logs. 2007. Retrieved, March 29, 2007, from http://www.jigg.ac.uk/foi/

[6] LAGOZE, C.; VAN de SOMPEL, H.; NELSON, M.; WARNER, S. The Open Archives Protocol for Metadata Harvesting. 2004. Retrieved, March 29, 2007, from http://www.openarchives.org/OAI/openarchivesprotocol.html

[7] *CWA 14855: Dublin Core Application Profile Guidelines*. Brussels : CEN/ISSS, 2003. Retrieved, March 29, 2007, from ftp://ftp.cenorm.be/PUBLIC/CWAs/e-Europe/MMI-DC/cwa14855-00-2003-Nov.pdf

[8] *DCMI Metadata Terms*. Dublin Core Metadata Initiative, 2006. Retrieved, March 29, 2007, from http://www.dublincore.org/documents/dcmi-terms/

[9] POWELL, A.; NILSSON, M.; NAEVE, A.; JOHNSTON, P. DCMI Abstract Model. 2005. Retrieved, March 29, 2007, from http://www.dublincore.org/documents/abstract-model/

[10] JISC. Function Activity Model. 2006. Retrieved, March 29, 2007, from http://www.jisc.ac.uk/whatwedo/themes/eadministration/recordsman_home/srl_structure.aspx

[11] *FOI Full Exemptions Guidance*. Department for Constitutional Affairs, 2006. Retrieved, March 29, 2007, from http://www.dca.gov.uk/foi/guidance/exguide/intro/annex_a.htm

[12] JISC, Freedom of Information (Scotland) Act 2002: Implementation and Practice: Exemptions. 2003. Retrieved, March 29, 2007, from http://www.jisc.ac.uk/publications/publications/pub_ib_fois.aspx#08exemptions

[13] *The Environmental Information Regulations 2004: 12, Exceptions to the duty to disclose environmental information*. London : The Stationery Office Ltd., 2004. Retrieved, March 29, 2007, from http://www.opsi.gov.uk/si/si2004/20043391.htm#12

[14] *The Environmental Information (Scotland) Regulations 2004: 10, Exceptions from duty to make environmental information available*. Scotland : The Stationery Office Ltd., 2004. Retrieved, March 29, 2007, from http://www.hmso.gov.uk/legislation/scotland/ssi2004/20040520.htm#10

[15] Creative Commons. Retrieved, March 29, 2007, from http://creativecommons.org/

[16] APPS, A.; WATTS, C.; WOOD, S. JISC Information Governance Gateway Freedom of Information Disclosure Logs Application Profile. 2006. Retrieved, March 29, 2007, from http://www.jigg.ac.uk/foi/profile/

[17] JOHNSTON, P.; POWELL, A. Expressing Dublin Core metadata using XML. 2006. Retrieved, March 29, 2007, from http://www.dublincore.org/documents/dc-xml/

[18] JOHNSTON, P.; POWELL, A. Expressing Dublin Core metadata using XML (DC-XML-Min). 2006. Retrieved, March 29, 2007, from http://dublincore.org/architecturewiki/DCXMLRevision/DCXMLMGuidelines/2006-09-18

[19] JOHNSTON, P.; POWELL, A. DC-Text: a simple text-based format for DC metadata. *DC2006: Proceedings of the International Conference on Dublin Core and Metadata Applications, 3-6 October 2006, Manzanillo, Mexico*. Mexico : University of Colima, 2006, p. 24-30.

[20] JIGG FOI XML Schema. Retrieved, March 29, 2007, from http://www.jigg.ac.uk/foi/schemas/xsd/jiggfoi.xsd

[21] APPS, A. Disseminating Service Registry Records. *ELPUB2006: Proceedings of the Tenth International Conference on Electronic Publishing, 14-16 June 2006, Bansko, Bulgaria*. Sofia : FOI-COMMERCE, 2006, p. 37-47.

[22] JISC infoNet. Information Legislation & Management Survey 2006 – Results. 2007. Retrieved, March 29, 2007, from http://www.jiscinfonet.ac.uk/foi-survey/2006/results

EPrints 3.0: New Capabilities for Maturing Repositories

Leslie A. Carr

School of Electronics and Computer Science, University of Southampton, Southampton, United Kingdom
e-mail: lac@ecs.soton.ac.uk

Abstract

There are now a large number of repositories in the world, contributing a significant amount of content to the world's scholars and scientists. The landscape has changed since the emergence of the Open Archiving Initiative: as well as Open Access, we have seen Multimedia Scholarly Collections, Teaching, Scientific Data, Preservation, Research Management and Assessment emerging as key drivers for repository adoption and development across the world. A new version of the EPrints repository software has been developed to address the growing demands on repositories to accommodate a wider variety of digital objects and metadata, to integrate with a wider range of services and applications and to support higher deposit rates to serve the needs of the whole institution. Already described by one reviewer as "a significant milestone towards ideal repository software", EPrints 3.0 provides new features for well-managed, high quality, high value repositories.

Keywords: repositories; EPrints; digital libraries

1 Introduction

There are now a large number of repositories in the world, contributing a significant amount of content to the world's scholars and scientists. The EPrints platform was developed as an outcome of the Open Archiving Initiative in 1999. This movement is still gaining momentum, as there are many institutions across the world that have yet to commit to providing a service for the dissemination and curation of their scientific and scholarly materials.

The landscape has changed since 1999: as well as Open Access, we have seen Multimedia Scholarly Collections, Teaching, Scientific Data, Preservation, Research Management and Assessment emerging as key drivers for repository adoption and development across the world.

EPrints has responded to the diversification of the repository ecology thanks to the flexibility and adaptability of its software platform (designed for easy customisation at every level) and its metadata and data management facilities (not limited to one schema or standard profile). Examples of EPrints repositories that support each of the agendas listed above can be seen at the Exemplars pages[1] at the EPrints web site.

In developing the latest version of EPrints (version 3.0) we responded to the community's requests for increased interoperability with other services and applications, improved user experience of depositing & managing a wider range of digital objects and better open source development opportunities. EPrints v3 is a more effective and efficient platform to underpin any kind of repository, and has recently been described as "a significant milestone towards ideal repository software" (Ariadne 50, January 2007).

All repositories aim to provide high quality information services to enable (or improve) administrative, scholarly or research tasks based on high quality records of research/scholarship. These tasks impose high standards of metadata and data curation – acquisition, transformation, maintenance and dissemination. Consequently repository managers face two challenges with their faculty staff: firstly in encouraging them to regularly deposit their research outputs or artefacts, and secondly in remedying deficiencies in the metadata provided.

The new EPrints software addresses the apparently irreconcilable issues of low-impact-deposit and high-metadata-quality by making data entry easier to get right. Firstly, a range of importers allow existing objects to be imported from other services or data sources. Secondly, when editing newly imported (or freshly created) objects, intelligent metadata assistance is provided for key information items. Auto-completion on the author names field means that a complete creator (surname, forenames, title and email address or staff ID) may be

[1] http://www.eprints.org/software/examples

entered in as little as three keystrokes and a menu choice. Not only does this lead to less effort for the depositor, it also means that author names are complete and consistently referenced throughout the repository. Other assistance means that the official journal name will always be used together with its ISSN and publisher and that the project identifier assigned by the item's funding agency can be relied on and that duplicate deposits are avoided.

Beyond these specific user interface features, the aim of EPrints is to fulfill three key objectives:

1. The *High Quality Repository*: a repository where the object metadata and object data is complete, correct and consistent – a repository where data entry is easy, errors and omissions are minimised at source and an ongoing quality management process is facilitated;

2. The *High Value Repository*: a repository whose items can be used and reused in many contexts for many tasks – not only dissemination and information discovery but also administrative reporting, bibliography management, CVs, institutional portals, the Semantic Web and desktop and webtop applications (such as Microsoft Office and Google Maps);

3. The *Well Managed Repository*: a repository that supports efficient curation and reporting of the objects that it manages, that supports effective monitoring and feedback of the workflow processes and operators who enact the workflow, and whose configuration and management is possible by the librarian manager rather than the technical system administrator.

As Open Access mandates from funding agencies and Research Assessment activities (both institutional and national) impose external deadlines with real financial consequences for failure to deliver then significant challenges emerge for staffing the repository processes that are needed to satisfy them, EPrints provides the solution for an enterprise repository that collects scholarly and scientific materials from thousands of its faculty staff and researchers without imposing onerous demands on individual researchers, library staff or information services.

DCMI-Tools: Ontologies for Digital Application Description

Jane Greenberg[1]*; Thomas Severiens*[2]

[1] School of Information and Library Science, University of North Carolina in Chapel Hill, USA
e-mail: janeg@ils.unc.edu
[2] Institut für wissenschaftliche Information e.V., Universität Osnabrück, Germany
e-mail: severiens@mathematik.uni-osnabrueck.de

Abstract

The growth in electronic and digital publishing on the World Wide Web has led to the development of a wide range of tools for generating metadata. As a result, it can be difficult to select the appropriate type of application and the best metadata tool to support a project's metadata needs. The Dublin Core Tools (DCMI Tools) Community recognizes this need and is developing an application profile and a taxonomy of tool functionalities for describing metadata applications. The community will use the application profile and the taxonomy to standardize access to information on metadata via the DCMI Tools and Software program. This paper reports on the DCMI Tool Community's activities to develop an application profile for describing the wide range of applications (algorithms; metadata templates, editors, and generators; and other software) fitting this rubric. The paper begins with an introduction to metadata application challenges, and introduces the DCMI Tools Community in order to provide important historical context. Next, the paper reviews the concept of application profile and emphasizes the importance of this approach for describing metadata tools. The paper reviews procedures to develop the application profile and presents the DCMI Tools application profile. The paper also presents a metadata tool functionality taxonomy (to be used with the application profile), a glossary (to assist people in learning about metadata tools), and the DCMI Tool Community's implementation plans. The final part of the paper presents several conclusions and highlights next steps.

Keywords: metadata tools; application profile; DCMI Tools Community

1 Introduction

Today's metadata tool environment includes offerings ranging from algorithms that plug in to various multi-functional software applications to fully developed tools specifically labeled as metadata editors, templates, and generators. Included in the mix are many software applications, such as word processing and publication software (e.g., Microsoft's WORD and Acrobats Adobe) and MP3 software that increasingly include functionalities supporting metadata generation. Tools in this category often include templates for storing summary metadata, such as "keywords" or "author name" or a brief "description". This type of software generally automatically generates a range of metadata, such as "date created", "date last modified", "size" and "format" [1]. There is also an evolution of blog software and social software (e.g., Flickr or Del.icio.us) supporting similar metadata generation, including tags. Metadata generated with any of these applications (designated metadata tools, software applications, and social software) can be harvested by metadata tools to create coherent or more substantial metadata records, which can be ported into a metadata repository to support resource discovery and other desired metadata functionalities [2].

Although these developments are exciting, they have complicated our view of the metadata tool landscape. That is the wide range and diversity of applications can make it difficult to select the appropriate type of application and the best metadata application to support a project's metadata needs. Should a digital library project invest in a fully functional off-the-shelf metadata generation application? What open source algorithms might be accessible that could be integrated with an institutions existing software suite to satisfy metadata needs? Catalogers, metadata professionals, information architects, and project managers are constantly asking these and other questions to determine which applications will suit their needs. Their inquiry is made difficult because of the absence of a single place providing unified and consistent descriptions of metadata tools.

The Dublin Core Tools (DCMI Tools) Community, a part of the Dublin Core Metadata Initiative (DCMI), is addressing this challenge [3, 4]. For the last several years this community has provided a Web page with access information and brief descriptions of applications supporting the generation of Dublin Core metadata records. As the metadata tool community has grown to include both developers and users, so too has the need to provide unified and collective information about metadata applications. The need expands beyond applications supporting Dublin Core metadata, to tools supporting metadata creation following:

- Standard schemes beyond the Dublin Core (e.g., ONIX or the EAD).
- Content value standards (e.g., *Library of Congress Classification* system) and authority files.
- Encoding schemes to standardize the use of content value standards even further (e.g., W3C Date Time Format standard).

The DCMI Tools Community is addressing this need via the development of an application profile and a taxonomy of tool functionalities—both of which can be used for describing metadata applications generally accessible for digital library and related initiatives.

This paper reports on the DCMI Tools activities to develop an application profile for describing the wide range of applications (algorithms; metadata templates, editors, and generators; and other software) fitting this rubric. The following sections of this paper are ordered as follows: section 2 introduces the DCMI Tools Community and provides some historical context; section 3 reviews the concept "application profile" and emphasizes why this approach supports a unified description of metadata tools; section 4 presents procedures to develop the current DCMI Tools application profile; section 4 presents the DCMI Tools application profile, section 5 presents a taxonomy of tool functionalities for classifying metadata applications, a glossary containing terminology that is important for the metadata tool community, and application profile implementation steps; section 6 includes several conclusions and next steps.

2 The DCMI Tools Community

The DCMI Tools Community is a "forum for individuals and organizations involved in the development and usage of tools and applications based on Dublin Core Metadata or other metadata standards that interoperate with and enhance functionality of the Dublin Core" [5]. The DCMI Tools community was initially a working group and was initiated at the 1999 Dublin Core conference in Frankfurt, Germany. The founding chairs were Roland Schwänzl (Osnabrück University) and Harry Wagner (OCLC). The working group initially focused on RDF-Tools and XML-Schema, as well as on DAML+OIL (which since that time has developed as SOAP web-services). At the outset, the DCMI Tools WG recognized the metadata community's need to access information about metadata applications. The Tools WG, therefore, took up the initiative of documenting and making accessible basic and important information about metadata applications via the DCMI Website through the "Tools and Software" program [6].

Although no formal descriptive standard was created to describe the applications, a broad taxonomy was developed to classify the range of applications being represented. Metadata tools being currently represented via DCMI's Tools and Software program are classed accordingly: Utilities, Creating Metadata (Templates), Tools for the Creation/Change of Templates, Automatic Extraction/Gathering of Metadata, Automatic Production of Metadata, Conversion Between Metadata Formats, Integrated (Tool) Environments, Application Profiles (Examples and Tools), and Metadata Search Engines. Details given for the tools represented via this site range from brief abstracts to more descriptive accounts documenting the metadata elements and schemes a tool supports.

During the Dublin Core 2006 conference in Manzanillo, Colima, Mexico, the DCMI Tools working group was transformed to what the DCMI refers to as a community [5]. The goal of a DCMI community is to facilitate the "exchange of information, general discussion within a specific area of interest" [3]. This change was very timely for the DCMI Tools WG, which had a year earlier Madrid, Spain, revised their charge to develop as a forum for two classes of users: tool developers and individuals interested in using tools. The DCMI Tools working group sponsored a workshop at the 2006 Joint Conference on Digital Libraries (JCDL), bringing together these users into a single community [7]. These developments, and the growing interest in metadata tools well beyond the immediate DCMI community, have motivated the reevaluation of the current classification of tools represented via the DCMI Tools and Software program [6]. This work has been a major focus of the DCMI Tools community via the last year, through a task group comprised of the DCMI Tools community co-chairs, with input from other members of the DCMI tools community. Our process of revision has required the creation of an application profile. The next section of this paper defines what an application profiles is, and why we selected this approach.

3 Application Profiles: A Practical Approach for Describing Metadata Tools

An application profile is a declaration of the metadata terms an organization, information resource, application, or user community uses in its metadata. In a broader sense, an application profile includes the set of metadata elements, policies, and guidelines defined for a particular application or implementation. The elements may be from one or more element sets, thus allowing a given application to meet its functional requirements by using metadata elements from several element sets including locally defined sets. For example, a given application might choose a specific subset of the Dublin Core elements that meets its needs, or may include elements from the Dublin Core, another element set, and several locally defined elements, all combined in a single schema, as for example the "Dublin Core Collection Description Application Profile" [8] does. An application profile is not considered complete without documentation that defines the policies and best practices appropriate to the application.

Application profiles are created for practical reasons. *First*, it makes no sense to reinvent the wheel. Why should a project create a metadata scheme from scratch, when there is already a scheme, or a series of schemes, that have already defined needed metadata elements, including implementation and use requirements? *Second*, it is recognized that often, a single scheme may not fully satisfy the needs of an individual initiative. For example, the Dublin Core metadata scheme is very useful for supporting resource discovery of digital resources in a digital library, although the elements do not adequately document and help manage resource preservation. A digital library wanting to facilitate the functions of both "resource discovery" and "preservation" might create an application profile, integrating elements from both the Dublin Core metadata standard and the PREMIS metadata standard [9]. *Third*, an application profile, pulling together elements from other schemes, facilitates greater interoperability on the World Wide Web. The reasons stated here, the growth in the availability of metadata tools, and the expansion of the DCMI Tools community, have motivated the development of the DCMI Tools application profile for representing algorithms, code pieces, and software tools. The next section of this paper presents a more detailed description on the methodology used generating the application profile.

4 Procedures for Developing the DCMI Tools Application Profile

The objective of our application profile activity is to describe algorithms, crosswalks, software, software-tools, and utilities collected in www.dublincore.org/tools/ in a coherent way. In moving forward, we have aimed to achieve best practices, resulting from discussion with participants from various fields. Our procedure has included the following steps:

1. An assessment of all elements in the available from the Dublin Core (ISO 15836-2003), DCTERMS (www.dublincore.org/documents/dcmi-terms/), and DOAP (Description of A Project) (usefulinc.com/doap/) and their applicability to the DCMI Tools community's goals. We selected these three schemes to cover the obvious needs for describing those applications being collected in our initial repository;

2. An initial three level ranking of each element's usefulness to our goals, with level *one* being necessary, *two* being potentially valuable *three* being not germane;

3. The composition of a DCMI Tools application profile, which included all level one ranked elements, and slightly over half of the level two items;

4. The development of a taxonomy of metadata tool functionalities—to be used with the application profile and for classifying metadata tools;

5. The development of a glossary to aide with tool classification and to facilitate communication among the metadata tool user community.

5 DCMI Tools Application Profile

The DCMI Tools application profile contains 17 elements, drawing from the Dublin Core, the DCTerms, and DOAP schemes. Nine of these elements contain qualifiers. Qualifiers can refine the meaning of an element, indicate where the value associated with an element came from, or the content formatting of an element (e.g, the format of year-month-date: YYYY-MM-DD). Table 1 presents an overview of our application profile, including examples for two applications, DC-dot and Picard Tagger.

Name-space	Element	Qualifiers	Example DC-dot	Example Picard Tagger
dc	contributor	doap:maintainer doap:developer doap:documenter doap:translator doap:tester	Rachel Heery	developer: LukasLalinsky developer: RobertKaye
dc	creator		Andy Powell	
dc	date	dcterms:created dcterms:dateCopyrighted dcterms:modified dcterms:issued	Created: 7 July 1997	issued: 2006-06-25
dc	description		Extracts and validates metadata from HTML resources and MS Office files. The generated metadata can be edited using the form provided and converted to various other formats (USMARC, SOIF, IAFA/ROADS, TEI headers, GILS, IMS or RDF) if required.	PicardTagger allows you to automatically look up the releases/tracks in your music collection and then write clean metadata tags (ID3 tags, Vorbis comment fields, etc.) to your files. It also allow syou to specify how and where to write cleanly tagged files to your hard drive.
dc	identifer	doap:repository	http://www.ukoln.ac.uk/metadata/dcdot/	http://musicbrainz.org/doc/PicardTagger repository: http://svn.musicbrainz.org/picard
dc	language		en-us, en-GB	
dc	publisher			
dc	relation	dcterms:hasPart dcterms:hasVersion dcterms:isPartOf dcterms:isReplacedBy dcterms:isRequiredBy dcterms:isVersionOf dcterms:replaces dcterms:requires doap: release	requires: Libwww-perl, soif.pl, Jon Knight's MARC module	requires: PyQt4 Mutagen (1.7) python-musicbrainz2 isPartOf: https://musicbrainz.helixcommunity.org/ release: 0.7.1
dc	rights	dcterms:accessRights dcterms:license	accessRights: open source license: http://www.gnu.org/copyleft/gpl.html	accessRights: open source license: http://www.gnu.org/copyleft/gpl.html
dc	rightsHolder			
dc	source	dcterms:URI		Workman, http://musicbrainz.org/doc/Workman
dc	title	dcterms:alternative	DC-dot	Picard Tagger
dc	type	dcterms:dataset dcterms:InteractiveResource dcterms:service dcterms:software	dcterms:InteractiveResource	dcterms:software
dcterms	audience	dctools:developer dctools:users		dctools:users dctools:developer
doap	location		Bath, UK	
doap	programming-language		Perl	Python
doap	operating-system			

Table 1: DCMI Tools Application Profile

6 A Taxonomy of Metadata Tool Functionalities

The application profile can be implemented within a semantic web framework, with automatic processes and requires the use taxonomy terms wherever possible. This will improve the representation of objects described, allowing for fairly complete the metadata descriptions. The most important part of the application profile is the classification of objects by genre, represented in our taxonomy.

Every object described may be in one or more of the following classes, which allows for sorting of tools by functionalities:

- Conversion
- Crosswalk
- Metadata Creation
- Metadata Encoding
- Metadata Extraction
- Metadata Generation
- Metadata Harvesting
- Metadata Templates
- Search Engines
- Translation
- Transliteration
- Validation

We will extend these classes as new types of software are developed. Classes not filled with latest software will be deleted, and the list will be revised as needed to allow for appropriate growth. We see this lists as being organic—in order to meet the needs of the tools community over time.

Some still open questions remain as part of our work in developing the profile. For example, location information requires additional attention. The most useful and precise approach is to give geographical coordinates, so a service can link to map serves. An alternative approach is to use a controlled vocabulary for geographic names. In this case, it would be desirable to allow for access and linking via international names (e.g., “Wien” (German version) versus “Vienna” (English version) versus “Wenen” (Dutch version). For the agent roles in the application profile we tried to use the roles defined in DOAP namespace (usefulinc.com/doap/) mostly reused from the foam-project results:

- developer
- documenter
- maintainer
- tester
- translator

To re-use the collected information in multiple frameworks, it will be requested to clearly define all vocabulary used. For use in semantic web framework this will be offered as RDFS, for human readability we restrict to textual representation in this article.

To assist with our work and further bring the metadata tool user community together, we have also developed a Glossary. This is presented in Table 2. The glossary is a new development produced by the DCMI Tools Community, and will be enhanced and modified as we continue our work.

Algorithm
a finite set of well-defined instructions for accomplishing some task which, given an initial state, will terminate in a defined end-state. (Wikipedia)

Application Profile
an assemblage of metadata elements selected from one or more metadata schemas and combined in a compound schema. Application profiles provide the means to express principles of modularity and extensibility. The purpose of an application profile is to adapt or combine existing schemas into a package that is tailored to the functional requirements of a particular application, while retaining interoperability with the original base schemas. Part of such an adaptation may include the elaboration of local metadata elements that have importance in a given community or organization, but which are not expected to be important in a wider context. (Duval)

Conversion
can refer to either

- conversion between schemas
- conversion of encoding (x/html to xml)

Crosswalk
a semantic mapping of metadata elements across metadata schema specifications. Crosswalks permit searching across multiple databases that use different schemas (Greenberg)

Metadata
An item of metadata may describe an individual data item or a collection of data items. Metadata is used to facilitate the understanding, use and management of data. (Wikipedia)

Metadata Creation
creation of metadata can be either

- by professional metadata creators; these include catalogers, indexers, and database administrators
- by technical metadata creators; these include webmasters, data in-putters, paraprofessionals, encoders and other persons who create metadata and may have had basic training but not professional level training
- by content creators; people who create the intellectual content of an object and the metadata for that object
- by community / subject enthusiasts; people who have not had any formal metadata-creation training but have special subject knowledge and want to assist with documentation (Greenberg)

Metadata Encoding
the syntax or prescribed order for the elements contained in the metadata description (NISO)

Metadata Extraction
synonym to Metadata Harvesting

Metadata Generation
the act of creating or producing metadata. Metadata can be generated by people, tools and processes (Greenberg)

Metadata Harvesting
a technique for extracting metadata from individual repositories and collecting it in a central catalog (NISO)

Metadata Template
Metadata format designed for some specific use or subject. (Severiens)

Namespace
In XML, a namespace is a collection of names, identified by a URI reference, that are used in XML documents as element types and attribute names. In order for XML documents to be able to use elements and attributes that have the same name but come from different sources, there must be a way to differentiate between the markup elements that come from the different sources. (Webopedia.com)

Schema
In general terms, any organization, coding, outline or plan of concepts. In terms of metadata, a systematic, orderly combination of elements or terms. In terms of DCMI term declarations represented in XML or RDF schema language, schemas are machine-processable specifications which define the structure and syntax of metadata specifications in a formal schema language. In terms of an encoding scheme, is a set of rules for encoding information that supports a specific community of users. See also Encoding scheme. (DCMI)

Search Engine
A utility capable of returning references to relevant information resources in response to a query. (DCMI)

Software
consisting of programs, enables a computer to perform specific tasks (Wikipedia)

Software-Tool
small piece of software, designed for developmental and laboratorial use (Severiens)

Translation
the interpretation of the meaning of a text in one language and the production, in another language, of an equivalent text that communicates the same message. Translation between may also convert meaning between semantics or schemes. (Wikipedia, Severiens)

Transliteration
Conversion of names or text not written in the roman alphabet to roman-alphabet form. (AACR Glossary)

Utility
software program that functions for a particular purpose. (Wikipedia)

Validation

- validating that syntax of element contents is correct (e.g. YYYY-MM-DD)
- validating the encoding (e.g., XML)

Table 2: DCMI Tools Glossary

The database, from which www.dublincore.org/tools is being generated, contains the following structure:

- Title: corresponding to the dc.title field in the app. profile.
- URL: corresponding to the dc.identifier field in the app. profile.
- Description: corresponding to the dc.description field in the app. profile.
- Classification: used to sort the service into the different classes.
- Free/commertial: this field is corresponding to the dc.rights qualifier dcterms:accessRights
- Online/download/webservice: corresponding to the dc.type field in the app. Profile and its qualifiers dcterms.InteractiveResource / dcterms.software / dcterms.service, a tag for dcterms.dataset may be added, if an entry is being included into the database.
- Country: corresponding to the field doap.laocation.
- Comment: This field allows some free text comments.
- Provider: corresponding to the dc.publisher field in the app. profile

Based on application profile developments, our plan is to add the following fields to the database:

- Information on the contributors, which can be
  - developers
  - documenters
  - maintainers
  - testers
  - translators
- Information on the creator(s)
- Information on the dates associated with the object, like
  - the date of its creation,
  - date of its latest modification,
  - date it was issued,
  - or the date of its copyright notice
- Information on the language of the object
- Information on the relations of the object to other objects in the database
- Information on the license like
  - a link to the licence text,
  - information on the licence holder,
  - while the date of the licence was already given with the dates above.
- Information on the source, if they differ from the compiled resource
- Information on the used programming language, if a source is available
- Information on the operating systems requested for running the software, if its not an webservice or online service.

7 Conclusions and Next Steps

The experience of creating the DCMI Tools application profile has been fruitful and resulted in an application profile that is ready for implementation. The DCMI Tools Community will be meeting at the DCMI-2007 Conference in Singapore this August to update members on this work. Prior to this conference, we will be testing the application profile and revising the DC Tools and Software program [6]. Our implementation will allow us to evaluate the overall effectiveness of the DCMI Tools application profile and identify areas requiring attention and revision. We will use our time in Singapore to share our findings and discuss any other outstanding issues, such as integrating location vocabulary from doap:location field. We will then begin to work on a collection and maintenance policy plan for keeping the DC Tools program up-to-date.

Acknowledgements

We would like to acknowledge Tessa Sullivan, doctoral students, School of Information and Library Science, University of North Carolina at Chapel Hill, USA, for her work on creating a table for the DC Tools Application Profile.

Notes and References

[1] GREENBERG, J.; SPURGIN, K;.CRYSTAL, A. Functionalities for Automatic-Metadata Generation Applications: A Survey of Metadata Experts' Opinions. International Journal of Metadata, Semantics, and Ontologies, (2006) 1(1): 3-20.

[2] GREENBERG, J. Understanding Metadata and Metadata Schemes. Cataloging & Classification Quarterly, (2005) 40(3/4): 17-36.

[3] DCMI Tools Community: http://dublincore.org/groups/tools/.

[4] Dublin Core Metadata Initiative: http://dublincore.org/.

[5] DEKKERS, M. Operational aspects of DCMI Work structure: Communities and Task. Date Issued: 2006-12-18: Groups: http://dublincore.org/documents/workstructure/.

[6] Tools and Software: http://dublincore.org/tools/.

[7] GREENBERG, J.; SEVERIENS, T., Metadata Tools for Digital Resource Repositories, D-Lib Magazine, (2006), Volume 12, Number 7/8, DOI:10.1045/july2006-greenberg.

[8] Dublin Core Collection Description Application Profile: http://www.ukoln.ac.uk/metadata/dcmi/collection-application-profile/.

[9] PREMIS: Data Dictionary for Preservation Metadata: http://www.oclc.org/research/projects/pmwg/premis-final.pdf.

DRIVER – Supporting Institutional Repositories in Europe

Mary L. Robinson[1]*; Wolfram Horstmann*[2]

[1] SHERPA, Information Services, University of Nottingham
Greenfield Medical Centre, Medial School, QMC, Nottingham, UK.
e-mail: mary.robinson@nottingham.ac.uk;
[2] Research and Development Department, SUB Göttingen
Universität Göttingen, Göttingen, Germany
e-mail: whorstmann@sub.uni-goettingen.de

Abstract

This workshop will provide an analysis of the current state of development of institutional repositories across Europe, how this compares to initiatives in the rest of the world and will explain how the DRIVER [1] project will promote and support the development of an integrated European repository network. The workshop will provide information on DRIVER technological developments and services and on the DRIVER test bed of repositories being used to test DRIVER software and services. The success of the DRIVER project depends, not just on the technical integration and enhancement of a European repository network, but also on the involvement and participation of all those actively involved in European research or in its publication, dissemination or access. Hence DRIVER has an active advocacy and community building programme to address and support key stakeholder groups in Europe. DRIVER draws on existing services within the DRIVER partnership such as OpenDOAR [2] and SHERPA/RoMEO [3] as well as developing new services such as the Mentor service [4]. This workshop will be of value to all involved in European research and for those keen to play a role in its future development. It will be of particular interest to those involved in the development of individual repositories, those co-ordinating national repository networks and those interested in the implications of a European repository network for European research. The workshop will provide a unique opportunity to learn about the DRIVER project, to meet DRIVER representatives, share best practice and discuss the current trends in the development and future of institutional repository networks.

Keywords: open access; repositories; European research

1 Introduction

The current system of academic publication developed as a means to disseminate the findings of research. However, this system can impede the very process it was set up to serve, with access to articles being limited by publishers to only those who can afford to subscribe.

Open access digital repositories provide a means whereby the traditional publishing model can co-exist with the needs of authors and their readers, as well as with the demands of research funders for research impact and hence, value for money. Subject to copyright, authors can deposit copies of their finished articles in open access repositories, in addition to publishing them in research journals.

The recent study of scientific publication markets in Europe funded by the European Commission [5] strongly recommends the development of a European policy mandating open access to EC-funded research. In addition, it recommends an exploration of interoperability issues and how open access repositories can be implemented Europe-wide.

DRIVER (Digital Repository Infrastructure Vision for European Research) is an EU-funded project with 10 international partners and reflects the growing awareness in Europe surrounding Open Access. DRIVER sets out to build a testbed for a future knowledge infrastructure of the European Research Area. It aims to deliver any form of scientific output, including scientific/technical reports, working papers, pre-prints, articles and original research data to the various user groups. The testbed is based on existing nationally organized digital repository infrastructures. Other work includes the support of new European repositories.

2 Objectives

The five objectives of DRIVER are:

1. To organise and build a virtual, European scale network of existing institutional repositories;
2. To assess and implement state-of-the-art technology, which manages the physically distributed repositories as one large scale virtual content resource;
3. To assess and implement a number of fundamental user services;
4. To identify, implement and promote a relevant set of standards;
5. To prepare the future expansion and upgrade of the DR infrastructure across Europe and to ensure widest possible involvement and exploitation by users.

3 Discussion

Thus far DRIVER has conducted focused research studies including an inventory of the type and level of OAI compliant digital repository activities in the EU [6], to facilitate the iterative development of DRIVER and is developing the necessary infrastructure middleware and user guidelines to meet the DRIVER objectives. The project is now actively advocating repository development - creating an informed and active environment for repository infrastructure development in EU countries with focused activities, information and contextualized support.

Acknowledgements

The authors would like to thank Bill Hubbard and Sophia Jones of SHERPA for their assistance in the preparation of this paper.

References

[1] DRIVER (Digital Repository Infrastructure Vision for European Research), http://www.driver-support.eu

[2] http://www.opendoar.org

[3] http://www.sherpa.ac.uk/romeo.php

[4] http://www.driver-support.eu/en/community/mentor.html

[5] European Commission. Study on the economic and technical evolution of the scientific publication markets in Europe, Jan. 2006. http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf

[6] DRIVER. Inventory study into the present type and level of OAI compliant Digital Repository activities in the EU, Apr. 2007.

Cultural Content Management at a New Level: Publishing Theater and Opera Details by Means of Open Technologies from the Web 2.0

Markus W. Schranz[1,2]

[1]Distributed Systems Group, Institute of Information Systems, Vienna University of Technology
Argentinierstrasse 8/184-1, A 1040 Vienna, Austria
e-mail: schranz@infosys.tuwien.ac.at
[2]Research Development, Culturall Handelsges.m.b.H.
Graf-Starhemberg-Gasse 37/4, A-1040 Vienna, Austria
e-mail: markus.schranz@culturall.com

Abstract

Creating Internet services for a specific auditorium involves technical, organizational and sociological challenges to developers, processes and used technologies. Authoring and distributing recent cultural news, opera schedules, notes to theater visitors or even cultural service maintenance can be handled by modern electronic publishing systems, ideally finding user-friendly solutions to the above-mentioned challenges. Although the Web has greatly improved its level of interactivity within the first ten years of existence, a significant gain in usability and value has been reached by introducing concepts summarized in the term Web 2.0. In this paper we provide an example of on-the-edge technology in managing and publishing cultural contents for Internet services, focusing on content management and workflow management features rom within the entire software framework concept. We demonstrate, that modern Web 2.0 technologies are well suited to increase the quality of electronic publishing for both the consumers as well as the producers of the content and the service itself by providing rich user experience enhancements at different levels of the content and service management process.

Keywords: content management; cultural content; Web 2.0; technical framework

1 Introduction

Most up-to-date software technologies and new applications in the Internet have initiated a change in the perception of what was generally understood as the „Internet“. The first decade of the World Wide Web was dominated by a strict role distribution between few content providers and experts that used complex technical mechanisms and powerful tools to publish „centralized“ data to the mass of content consumers, far away from open content interchange. As stated in [1] the shift to a new class of Internet applications has brought innovation in both technical and practical use of software applications for the Internet. Without being able to fasten it down to a single event or technology we sense a significant change in how applications provide information management, distribution and communication control to end users.

Increasingly, local proprietary solutions have been exchanged by open network services, desktop software is extending to integrated network solutions, programs and applications are serviced and updated in an open and self-contained way, single services are becoming ready for exchange and interoperability and even technical laymen are ready to use modern services to share information and facilities in an easy way. Despite the criticism in terms of the technological step contributed [2], Web 2.0 has become a synonym for openness, innovative technologies and user-friendly applications to integrate the abilities of all web users. Consequently and as described in the following sections also for dedicated services and technologies for specific user communities, media and content, the innovative character of the Web 2.0 can be identified in the area of electronic publishing.

Following thorough research basics and experiences from Web application engineering methodologies [3], the logically consecutive step is to construct user-oriented modern applications for specific application domains. Public contents in the area of science, education, news [4] or culture have been utilizing modern technologies to be distributed in a wide-spread manner. Resulting services are exemplary for innovation in the creation and consumption of Internet technologies for the discussed areas, e.g. online current contents for the Vienna Opera House [5]. Modern concepts like agile programming [6] and the utilization of dynamic languages, which were smiled at by software engineers a few years ago, have been the recent choice of innovators to create open services and access to digital publishing and content distribution.

In the following sections we focus on the innovative way to electronically publish and distribute cultural contents by facilitating modern technologies, integrating strong support of the collective intelligence as a basis of the Web 2.0. Since data and information are treated as the most central good, we describe the management, contribution and provision (publishing and distribution) of cultural contents such as theater programs and schedules, the ticket management and innovative clearing and control services for several theaters and opera houses in Europe, as well as the workflow coordination of the software engineers, working on the electronic publishing services that are based on modern standards and open content exchange technology. The paper outlines Web 2.0 technologies for the discussed application domain in section 2, specific content and service management approaches in section 3 and gives a brief summary of our results in the conclusion.

2 Modern Web 2.0 Technologies for Cultural Content Presentation

The term Web 2.0 is describing rather vaguely an updated perception and utilization of the World Wide Web. Renowned experts in the area of software development and Internet technologies criticize the Web 2.0 to offer little innovation in terms of technical development. For the use in the specific application domain cultural content management and publication we focus on the organizational view of Web 2.0: users create and manage contents in an increasing amount on their own. User-oriented Web interfaces facilitate simple theater and opera detail publishing and modern technologies support the content management, asset handling and schedule interchange based on open methodologies and standards such as RSS.

Cultural content as discussed in this paper include opera and theater programs, event and performance details in text and multimedia presentation, event schedules, ticket management and presentation, event access management and reservation services for single theatres, multi-client cultural organizations or several independent theatres and opera houses in Europe. As demonstrators we explain case studies of the modern implementations for the Hamburgische Staatsoper, Vienna State Opera and Symphonic Orchestra of Bern. These and a dozen more cultural providers are managed by Culturall, a technology innovator strongly supporting the research on Web 2.0 application technology for cultural content management and publishing.

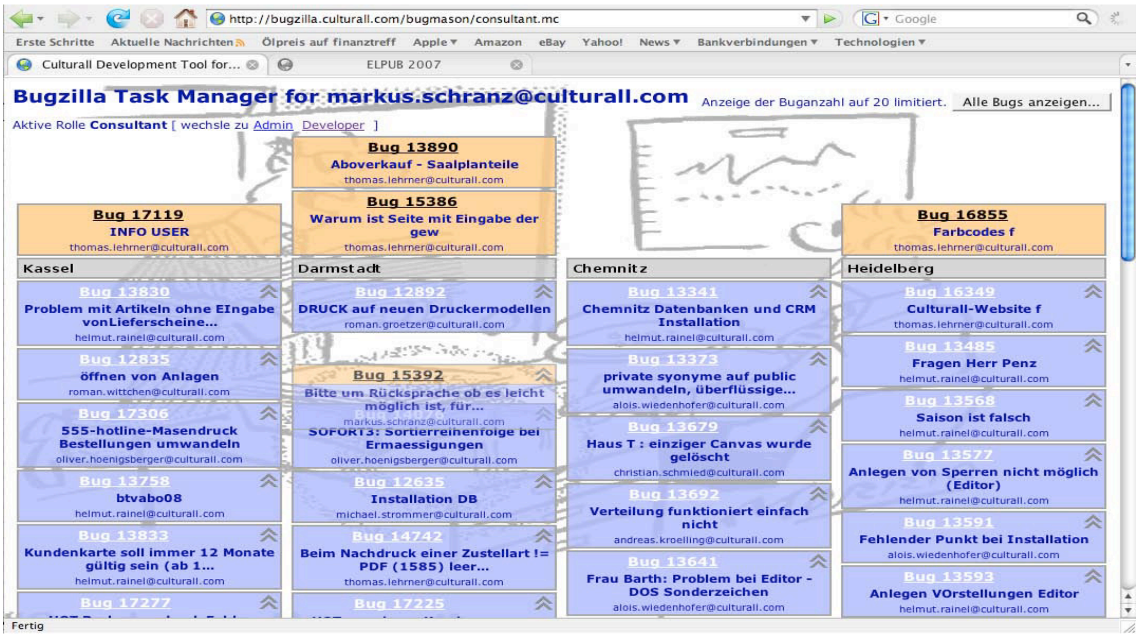
Most stimulative to the introduction of Web 2.0 technology in the cultural content management domain has been the important principle of supporting a Rich User Experience in the dedicated applications. The goal of software following the RUE principle is the creation of graphical interfaces that allow a handling that is comparable to that of local/desktop software implementations. Specific details of cultural content assets have to be exchanged frequently between user client and the providing server, thus interrupting the flow of the users visit by well-known brakes in between single Web pages. Based on the herein introduced approach we have developed a prototype that handles theater event descriptions, date scheduling, seat reservation and personalized ticket management has been integrated in a smooth and user-friendly way, following the basic principles of Web 2.0. Since all details of the concepts and a full description of the prototype is out of scope of this paper, we demonstrate the task management feature exemplarily to proof the usability and beneficial effects of Web 2.0 technologies in the application domain electronic pubishing.

Besides the technological advantages, the rich utilization of Web 2.0 technologies supports the researchers in extending the sense of community within the multiple users of the theater content management services. Since the information and the data/assets itself denote the highest value of the cultural services, content provision, management and publishing has to be simplified to a maximum extent. With the participation of all users, including the administrators, the theater experts, the content visitors, and the ticket bookers, the information is shared in a technologically well supported open way. Modern user-friendly content management interfaces as well as standardized open content exchange interfaces, such as RSS feeds for event schedules and ticket reservation assets for third party providers underline the open approach of the theater content management service in use.

3 Content Management and Digital Asset Management in Commercial Services

Based on the innovative application development methodologies outlined in the principles of Web 2.0 the content management and publishing services researched and developed for the cultural application domain includes mainly web-based services and database integration. Web software has been developed using dynamic development languages, which are well accepted in the Web 2.0 Lightweight Programming Models and the Agile Programming paradigms [7]. Based on open interfaces like AJAX [8] the web services have been developed and are currently under research investigations in terms of user-acceptance and scalability checks in current field studies. Particularly, the user interaction at the workflow management component of the cultural content publishing services is shown in Figure 1. Herein, a content developer can move a particular task (bug 15392) via Web 2.0 technology b simply dragging and dropping the item on the canvas. The service will not reload the page but instead send the alternated order via a ajax call to the server, where the new order is stored persistently in the database. With similar features, the visitors can pick theater and opera house seating, comment on published texts, etc.

The software implemented with ajax technology, object-oriented web application servers like Mason [9] provide access via user interfaces beyond device borders, so cross-platform and cross-device applications allow an open access to cultural contents via desktop computers, notebooks, handhelds and mobile phones. Furthermore the utilization of dynamic development languages like perl and java underline the principle of overcoming the software lifecycle in Web 2.0 applications. Instead of delivering version after version of a desktop application the cultural content management application is provided as a service that is under permanent development. Following this principles allow easier software maintenance and wide-spread service availability for a great mass of users. Providing services for a wide user group and offering open interfaces such AJAX/XML links and RSS feeds enables a distributed service enrichment similar to open source development. Specific features contributed by domain experts are demonstrated on the publicly available services [5].



Bugzilla Task Manager for markus.schranz@culturall.com

Anzeige der Buganzahl auf 20 limitiert.

Aktive Rolle **Consultant** [wechsle zu Admin Developer]

Bug 13890
Aboverkauf - Saalplanteile
thomas.lehrner@culturall.com

Bug 15386
Warum ist Seite mit Eingabe der gew
thomas.lehrner@culturall.com

Bug 17119
INFO USER
thomas.lehrner@culturall.com

Bug 16855
Farbcodes f
thomas.lehrner@culturall.com

Kassel

Bug 13830
Problem mit Artikeln ohne EIngabe vonLieferscheine...
helmut.rainel@culturall.com

Bug 12835
öffnen von Anlagen
roman.wittchen@culturall.com

Bug 17306
555-hotline-Masendruck Bestellungen umwandeln
oliver.hoenigsberger@culturall.com

Bug 13758
btvabo08
helmut.rainel@culturall.com

Bug 13833
Kundenkarte soll immer 12 Monate gültig sein (ab 1...
helmut.rainel@culturall.com

Bug 17277

Darmstadt

Bug 12892
DRUCK auf neuen Druckermodellen
roman.groetzer@culturall.com

Bug 15392
Bitte um Rücksprache ob es leicht möglich ist, für...
markus.schranz@culturall.com

SOFORT3: Sortierreihenfolge bei Ermaessigungen
oliver.hoenigsberger@culturall.com

Bug 12635
Installation DB
michael.strommer@culturall.com

Bug 14742
Beim Nachdruck einer Zustellart != PDF (1585) leer...
thomas.lehrner@culturall.com

Bug 17225

Chemnitz

Bug 13341
Chemnitz Datenbanken und CRM Installation
helmut.rainel@culturall.com

Bug 13373
private syonyme auf public umwandeln, überflüssige...
alois.wiedenhofer@culturall.com

Bug 13679
Haus T : einziger Canvas wurde gelöscht
christian.schmied@culturall.com

Bug 13692
Verteilung funktioniert einfach nicht
andreas.kroelling@culturall.com

Bug 13641
Frau Barth: Problem bei Editor - DOS Sonderzeichen
alois.wiedenhofer@culturall.com

Heidelberg

Bug 16349
Culturall-Website f
thomas.lehrner@culturall.com

Bug 13485
Fragen Herr Penz
helmut.rainel@culturall.com

Bug 13568
Saison ist falsch
helmut.rainel@culturall.com

Bug 13577
Anlegen von Sperren nicht möglich (Editor)
helmut.rainel@culturall.com

Bug 13591
Fehlender Punkt bei Installation
alois.wiedenhofer@culturall.com

Bug 13593
Anlegen VOrstellungen Editor
helmut.rainel@culturall.com

Figure 1: Task Management feature of the Web 2.0 cultural content management service prototype

4 Conclusion

In this paper we describe the research and development of a content management and publishing service for a specific application domain: theater and opera content management based on innovative Web 2.0 concepts and technologies. The shift to a new class of Internet applications has brought innovation in both technical and practical use of software applications for the Internet.

Beyond the thoroughly discussed area of technical innovation contributed by the Web 2.0 we especially extend the organizational principles and sociological aspects of this new direction in software development for the Internet. As a demonstrator application we have researched, conceptualized and implemented a framework for cultural content and asset management, including the publishing and distribution of theater event details, ticket assets and reservation management services, and a highly sophisticated workflow management service for the software development process. Modern principles such as the rich user experience for GUIs, end user integration for content management and open interfaces to exchange domain specific contents are well adaptable for the application domain of cultural content providers. Modern application frameworks can be well applied to other domains which we investigate in future work.

References

[1] O'REILLY, T., What Is Web 2.0?, O'Reilly Network, http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[2] BERNERS-LEE, T., developerWorks Interview podcast series, http://www-128.ibm.com/developerworks/podcast/dwi/cm-int082206.txt, July 28, 2006

[3] KAPPEL G, Pröll B, Reich S. Web Engineering. The Discipline of Systematic Development of Web Applications, Wiley & Sons, 12 May 2006

[4] SCHRANZ M., Pushing the quality level in networked news business: semantic-based content retrieval and composition in international news publishing, proceedings of the Elecronic Publishing conference, Bansko, Bulgaria, June 2006

[5] CULTURALL, Opera2007 - Vienna State Opera, https://www1.culturall.com/mirror_ctn/plsql/ctn_suche.spielplan, 2005

[6] THOMAS D., Heinemaier-Hansson D, Breedt L, Agile Web Development, Pragmatic Programmers, Dezember 2006

[7] FOWLER, Martin, The New Methodology, http://www.martinfowler.com/articles/newMethodology.html, 2005

[8] W3SCHOOLS, AJAX Tutorial, http://www.w3schools.com/ajax/default.asp, 2006

[9] ROLSKY, D. et al., Embedding Perl in HTML with Mason, O’Reilly, Oktober 2002

Ontologies at Work:
Publishing Multilingual Recreational Routes Using Ontologies

Bert Paepen

Centre for Usability Research, Katholieke Universiteit Leuven
Parkstraat 45 bus 3605, B-3000 Leuven, Belgium
e-mail: *bert.paepen@soc.kuleuven.be*

Abstract

Even though there is nothing new about the idea, ontologies are a hot topic. Built for many reasons and appliances, the use of ontologies in real-life applications remains limited. The WalkOnWeb project has developed ontologies in the area of recreational routing and applied them in a real application. This demonstration will show these applications and explain how they use ontologies. With the "Walk Planner" hikers can plan their trip by looking for trails, creating new routes and getting detailed information in print, web or mobile format. Authors can create and describe routes using the "Authoring Tool". By creating ontologies and using them in these applications the WalkOnWeb project has developed a system to publish electronic routes in a flexible and personalized way.

Keywords: ontology; navigation; XML; SVG

1 Introduction

Ontologies are well-structured representations of knowledge in a certain domain. They consist of concepts and relations between them, described in a computer readable form while still being descriptive for humans. Applied in many areas, their use often remains theoretical, so that the practical utility of ontologies in real applications remains unclear. The European research project WalkOnWeb [1] has tried to break with this tradition by applying ontologies in a real-life application area: outdoor navigation.

2 Publishing Recreational Routes Electronically

One of the problems the WalkOnWeb application is trying to overcome is the lack of flexibility offered to outdoor enthusiasts by traditional publications. Hiking guidebooks for example describe a route in only one direction, one language and from a fixed starting point. When switching to an electronic publishing paradigm a hiker can expect more flexibility: she should be able to get information about itineraries via the Internet, whichever country she is visiting. She expects to get information in her own language, to choose her preferred starting point and walking direction, and maybe even combine parts of existing routes to a new, personalized route.

For publishers this type of requirements poses huge challenges when publishing their material in an electronic form. In theory electronic publishing should be more cost effective, avoiding high fixed costs for printing books. In practice however the issue is not that simple. First, users expect to get up to date information, forcing publishers to continuously provide updates. Second, multilingual publishing involves high translation costs. Finally, the material used for paper publications does not support the type of flexibility expected in electronic publishing, both in terms of content and technology.

3 Information Model

Taking these considerations into account WalkOnWeb has defined a new information model for flexible electronic publishing of recreational routes. A walk ontology was developed for this purpose. Using an innovative software engineering process this ontology was then converted automatically to Java business objects and mapped to a relational database. This paved the way for practical application development.

The figure below depicts the information model developed during the project. Using topographic map data as a basis an author creates a networks of paths on which hiking is possible, nice, safe, legal, etc. The author then enriches the map data further by linking all kinds of information to a route: points of interest, pictures, texts, and others. The *walk ontology* includes concepts for many of these information items. For example: practical info could be “keep dogs on leash”, “hunting season” or “dangerous crossing”.

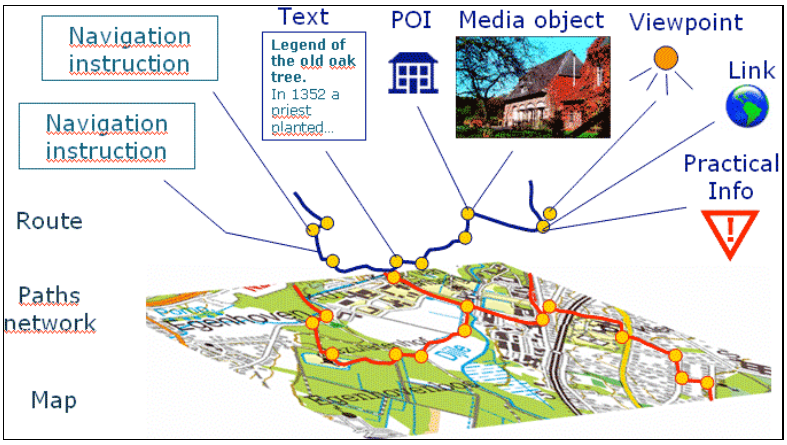
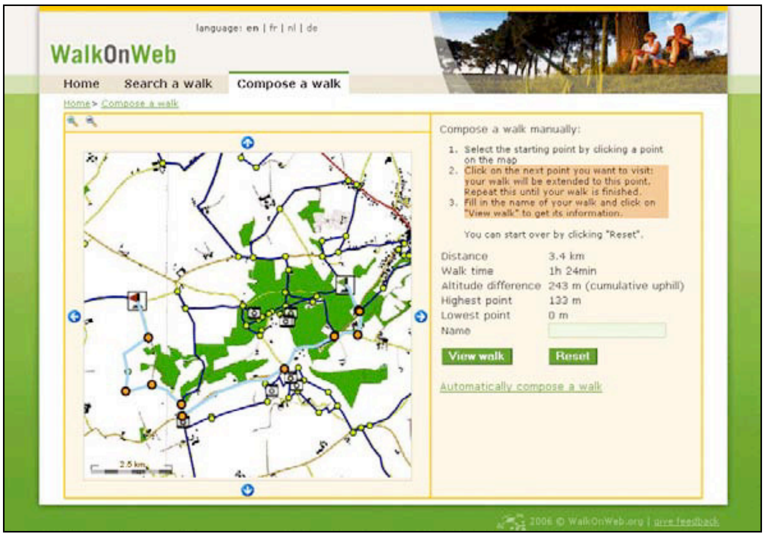


Figure 1: WalkOnWeb information model

In addition a navigation ontology was created to allow a new way of describing navigation instructions. Taking into account costs involved in traditional multilingual publishing this approach allows authors to describe a route in a language independent way, using a set of predefined ontology concepts (“building blocks”). We have first developed this ontology in theory (described in [2]) and then brought it to practice in two applications: the Walk Planner and the Authoring Tool.

4 Real-life Applications

On the "Walk Planner" website hikers can search for hiking trails using criteria like duration, difficulty, child friendliness and geographic location. All walks are shown on a geographic map using SVG. Hikers can also compose their personal walk based on parts of existing trails (see Figure 2, where the user has composed a walk from the green flag up to the red flag). Finally they can export detailed information for a walk to paper, electronic document or mobile device. This publication happens on the fly, based on user preferences: language, type of information, starting point and walking direction. For navigation instructions this means that the system generates a readable text from the navigation ontology concepts that the author has selected.


language: en | fr | nl | de
WalkOnWeb
Home
Search a walk
Compose a walk
Home > Compose a walk
2.5 km
View walk
Reset
Automatically compose a walk
2006 © WalkOnWeb.org | give feedback

Compose a walk manually:

1. Select the starting point by clicking a point on the map
2. Click on the next point you want to visit: your walk will be extended to this point. Repeat this until your walk is finished.
3. Fill in the name of your walk and click on "View walk" to get its information.

You can start over by clicking "Reset".

Distance	3.4 km
Walk time	1h 24min
Altitude difference	243 m (cumulative uphill)
Highest point	133 m
Lowest point	0 m
Name	

Figure 2: Walk Planner: compose a walk

With the "Authoring Tool" application authors are able to create and describe routes in a language independent way. This means that they apply the model depicted in Figure 1 and use ontology concepts for describing the details of the itinerary.

Notes and References

[1] PAEPEN, B et al. *WalkOnWeb project website*. www.walkonweb.org Leuven, 2006.

[2] PAEPEN, B; ENGELEN, J. *Using a Walk Ontology for Capturing Language Independent Navigation Instructions.* In 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 Dobreva. ISBN 978-954-16-0040-5, 2006, pp. 187-196

The PURE Institutional Repository: Ingestion, Storage, Preservation, Exhibition and Reporting

Bo Alroe

Atira A/S and 10 Danish Universities, Niels Jernes Vej 10, NOVI, DK-9220 Aalborg OE, Denmark
e-mail: ba@atira.dk

Abstract

Jointly developed over 5 years by Atira A/S and a number of university libraries, the commercial repository system PURE is a tool in the research administration and dissemination effort at 11 Danish and Swedish universities. Also the hospital sector, the pharmaceutical industry and other research institutions use PURE.

Keywords: meta-data models; long-term preservation; OAI-PMH; research portal; bibliometric analysis

1 Introduction

PURE is a commercial modular standard software platform for building institutional repositories at universities and other research institutes. The term CRIS [1] also applies. It supports the publication registration process in full - from ingestion and storage to publishing and reporting. PURE is a J2EE application running on most server operation systems and SQL environments and in connection with DSpace and FEDORA. This document offers a complete overview of PURE by addressing four basic repository issues: Data modeling, data ingestion, data storage and data exhibition.

2 Application Features and Architecture

2.1 Data Modeling

Apart from publications, the following content types can also be handled in PURE: Person, Organization, Project, Student Thesis, Activity, News Clip [2] and Clinical Trial. To some extent, these types are related to separate modules of the application, which are available as integral extensions to the basic module. The separate modules are: Reports, External Publications, Student Thesis, Bibliometry, News and Clinical Trials.

The terminology above is from the PURE meta-data model, which can be delivered with PURE - either as is or adapted to the individual research institution. But any meta-data model can be implemented in PURE; part of the application architecture is development frameworks that facilitate such implementations. During the last 5 years, 4 universities have specified their own meta-data model and have had it implemented.

Currently, the PURE meta-data model is the only default meta-data model available. However, a default implementation of the CERIF2006 [3] meta-data model is currently being planned. It is expected to be completed by the end of 2007. A key specification with both meta-data models mentioned is the use of relations objects and many-to-many relations between all primary content types.

2.2 Data Ingestion

PURE's user interface is browser-based. Both Firefox and Internet Explorer are supported on Windows and Mac, though only in relatively recent versions. Users are assigned roles, which define two things: First, what functionality a user have access to. The individual users interface is adapted accordingly. Next, what the user's rights are. About 10 pre-defined roles comes with PURE and custom roles can be defined and added.

Workflows in PURE allow several users to participate in different parts of the same work process - for example the process of registering a publication. Workflows allow different users to take part in a registration process by modifying, enriching, validating manually entered data.

Data from existing systems can be used in PURE by means of one or more dynamic integrations. Usually, data for Person-objects - e.g. a person's name, title, room number, employment number, direct telephone number, e-mail address, etc. - already exists in one or more systems, and using such systems as a data source for PURE will be desirable in many situations. Organizational data would be a typical example, too, originating from systems such as LDAPs or Active Directories. Integration between PURE and any number of local sources is possible. Also user authentication and Single Sign-On integration is possible.

Imports are different from dynamic integration in that they deal with historic data and are carried out only once per data set. Publication registrations from an old repository would be a good example. To facilitate import of such data, the PURE XML Archive format (PXA) was specified. PXA files are created outside of PURE and imported via an XML I/O unit. Relations between a) publications to be imported and b) organizations and persons already in PURE is possible, depending on available data.

2.3 Data Storage

Data access is encapsulated using an object/relational persistence and query service. This allows the use of most SQL-database environments with PURE. Microsoft SQL-Server, Oracle and PostgreSQL are the most used SQL-environments among current PURE-users. PURE also maintains an index to support searches.

A so-called file connector interfaces PURE with the OS file system. The connector stores full-text files to the OS file-system while still retaining relations to the relevant meta-data objects. Further, two additional connectors are available for storing to a DSpace environment and for storing to a FEDORA environment.

2.4 Data Exhibition

PURE's two web services are a Document/Literal web service, which will make XML easily available for re-styling, and an RPC/Encoded web service, which allows requisition of data from PURE as whole objects. In both cases rich libraries of methods are available. Parameters such as date-range or organization can be added to each methods. Further, PURE has a portal-framework called PUREportal, which is an internal framework for building customized websites for exhibiting data from PURE. The framework itself comes with each PURE license. Together, the two web services and the PUREportal framework are how PURE exhibits data to websites.

Two more services help exhibit data from PURE. One is an OAI-PMH data providing services, the other is a Z39.50 based service. Different formats ca be defined under OAI-PMH; Dublin Core and DDF-MXD is supplied by default. Z39.50 was added to PURE because many library systems interfaces nicely with it. SRU/SRW will be implemented as demand rises. In addition, Reference Manager exporting capabilities allows export of any data set from PURE in native RefMan format, saving double work in some cases.

Finally, the report generator in PURE is a reporting and statistics tool, that will respond to all data in the entire PURE repository. A number of standard reports are supplied and available in three categories: Lists, Analyses and Bibliometrics. To run these standard reports, only a time interval and an organization must be chosen. To that, each standard report can be customized. Finally, custom reports can be build from scratch.

Notes and References

[1] See http://www.eurocris.org

[2] Clips from media where researchers are mentioned. Such clips are usually supplied as an XML feed from a 3rd party supplier. Upon import to PURE, clips can be related to the appropriate researchers.

[3] See http://www.eurocris.org:8080/lenya/euroCRIS/live/

Access to Free e-journals via Library Portals: The Experience of the Shahid Chamran Ahwaz University in Iran as a Case Study

Amir Reza Asnafi

School of Education & Psychology, Department of Library & Information Science,
Shahid Chamran University
e-mail: aasnafi@gmail.com

Abstract

Journals as one of the most important information carriers are useful resources for libraries and information centers. As publishers more fully actualize the e-journal, it soon will be as insufficient to offer only print journals as it is to provide only print abstracts and indexes. Using journals especially scholarly journals, scientists can contact together in scientific communities. Impact of information technologies on journals has changed the format of these resources into electronic and has facilitated information storage and retrieval. Free e-journals are alternatives for non free e-journals and are useful for libraries that can not afford expensive costs to provide subscription-based e-journals. If libraries really support Free e-journals, then one would assume that we would be pretty aggressive about making free content readily available to our users through library systems and access points (e.g. ILS, knowledge base, web site). Finding free and non free e-journals in the web environment is more difficult than finding print journals, since they rarely are found in bibliographic resources, so their locating and retrieval is difficult. Librarians are responsible for information organization and retrieval and they must corporate in designing search engines and web pages to offer electronic articles that are needed for users. Portals are one of the tools that can be used for accessibility of free and non free e-journals to users. Portals as website are windows to World Wide Web and often have a search engine, links to useful pages, news or other services. In this article various literature and experiences about access to electronic journals via web pages has been reviewed. We decided to create a special portal of free e-journals for postgraduate students, masters & researchers of Shahid Chamran Ahwaz University to use of these resources. So, we provided a list of Shahid Chamran Ahwaz University courses and on the basis of this list, we selected free e-journals of each course via Directory of Open Access Journals (DOAJ)[1]. DOAJ has become "the" Open Access journal site for libraries because it is of a manageable size, Many librarians may think it is comprehensive, It is well organized and easy to harvest. In this survey, Researcher could extract 198 journals from this website that all of them were peer reviewed. Since we wanted to design a portal for free e-journals, we added selected journals to webpage that was linked to Shahid Chamran Ahwaz University Central Library website. This webpage was called free e-journals portal. All students and Masters could access to their needed articles freely. In this article we will discuss about importance of open access or free e-journals and their role in scholarly communication. Finally we will offer free e-journals portal of Shahid Chamran Ahwaz University Central Library as a tool for access to open access articles.

Keywords: open access; library portals; DOAJ; central libraries

1 Introduction

Using journals especially scholarly journals, scientists can contact together in scientific communities. Impact of information technologies on journals has changed the format of these resources into electronic and has facilitated information storage and retrieval.Free electronic jounrals are one of the types of electronic journals. These journals are accessible freely via Internet for users. Now, free electronic journals are the main part of scientific resources.

2 Research Aim

The major aim of this research, is designing a special portal for free electronic journals for Shahid Chamran University Ahwaz University on the basis of the attitudes graduate students of this university about these journals.

[1] http://www.doaj.org

3 Methodology

Data collecting tools were literature review, Checklist, questionnaire and Yahoo search engine. For data statistical analysis, descriptive statistics, Chi-Square, Anova One Way and Scheffe test were used. Results of checklist analysis, that was distributed among Iranian experts in Library & Information Science field, indicated 36 criteria can be used for selecting free electronic journals via Internet. By these criteria 198 journals, were selected from Directory of Open Access Journals.

4 Findings

This research indicated that graduate students of Shahid Chamran university of Ahvaz have little familiarity with free electronic journals of their special course and their use of these journals is in low level. On the basis of the attitudes graduate students of Shahid Chamran university of Ahvaz, evaluating criteria of free electronic journals quality and needed features for designing a special portal for free electronic journals were gained. In secondary findings part of this research, Chi-Square Test cleared that there is no significant difference among using full time and part time graduate students of Shahid Chamran university of Ahvaz of free electronic journals. In this research, by Webometrics method, highly cited free electronic journals were assigned. By this method, 63 highly cited free electronic journals were determined. Finally, by Microsoft Frontpage, that is a special software for designing web pages, primary version of special portal of free electronic journals for Shahid Chamran university of Ahvaz was designed and created. Free electronic journals of each university course will be accessible from this portal. Figure 1 shows portal of free e-journals in Shahid Chamran Ahwaz University. Its address is: http://www.cua.ac.ir/lib/central-library3/fire-home.htm

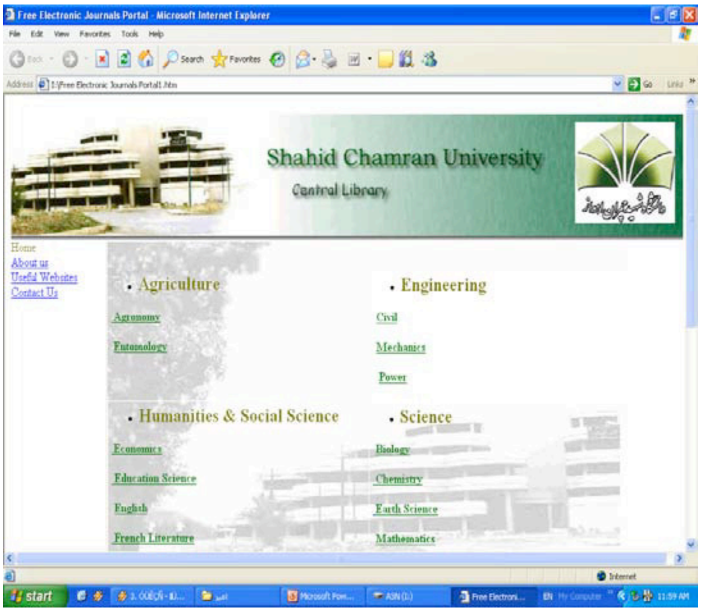


Figure 1: Portal of Free e-journals

Notes and References

[1] ANDERSON, R. 2004. Open access in the real world: confronting economic and legal reality. *College and Research Library News* 64(4). Available at: http://dlist.sir.arizona.edu/archive/00000351/

[2] SADEH, T.; WALKER, J. 2003."Library Portal : Toward the semantic Web". *New Library World*. 104(1185), pp. 19-11

[3] RICH, L. A.; RABINE, J. L.1999."How libraries are providing access to electronic serials: A survey of academic library websites". *Serials Review*. Vol.25, No.2, pp. 35-46.

Digital Archives at the University of Pisa

Cinzia Bucchioni; Zanetta Pistelli; Barbara Pistoia

Sistema Bibliotecario, Università di Pisa, Lungarno Pacinotti 44, Pisa, Italy
e-mail: bucchioni@angl.unipi.it; z.pistelli@bibant.unipi.it; b.pistoia@ing.unipi.it

Abstract

At the end of the '90s, the Library System Centre of the University of Pisa began to create a system of digital archives in order to enhance and promote the Institution activities regarding teaching, research and administration: 1. ETD (Electronic Thesis and Dissertations); 2. UnipiEprints.

Keywords: digital archives; open archives

1 ETD (http://etd.adm.unipi.it/)

Due to the lack of a national system for thesis management and access in Italy, in the case of both paper and digital theses, the need arose for the University of Pisa to create a local system. This led to the choice of the NDLTD platform, an open source system developed at the Virginia Polytechnic Institute and State University, as it is a widespread international system specific to theses. The writing and discussion of a thesis is a significant moment in a student's career, and in the Italian academic context involves a series of administrative issues and many laborious bureaucratic processes. Therefore, the ideal system would need to implement a single work-flow, integrating the entire process of presenting, revising, discussing, cataloguing, giving access and preserving an academic and doctoral thesis: with this aim, a number of important software developments were necessary:

- The introduction of managing different deadlines: date for official deposit of the final version; date for public discussion; date for "last content revision" (the author can enter minor corrections within up to 48 hours before the discussion: immediate notification is automatically sent to the academic supervisors);

- In conformity with Italian copyright law and academic practices, the author can decide whether his/her thesis is open access (full text, only parts of the text), or only metadata are available: the system enables the author to change this option when required

- the most important development (in progress) regards interoperability between ETD and the other software systems used at the University of Pisa:

    - ESSE3, the administrative system for student records management
    - Aleph, the bibliographic system

Thus, when a student accesses ETD for the first time, the system retrieves all relative personal and academic data from ESSE3; moreover when the thesis is deposited, a UNIMARC record is created and sent to the Aleph catalogue.

The project entered the production stage in 2006. It has been recently registered in the Open Archives Initiative (OAI) register, and currently contains almost 3300 theses.

The digital deposit of theses is not compulsory: it is a shared decision of the academic supervisors and the student. The project group has been working on widespread promotion of the system in the academic community, with very different feedback from the various disciplinary communities: the STM community makes wide use of the system (to date around 60% of theses are native digital), while the humanities community appears to be more sceptical (around 10% of theses are native digital).

2 UnipiEprints (http://eprints.adm.unipi.it/)

The University of Pisa officially signed the Berlin Declaration concerning open access to knowledge in May 2005: this prompted the project to create a system of institutional archives devoted to scientific and educational documents produced at our University. Eprints, the open source software of the OA Initiative was chosen, in order to be included in the OAI services. No serious customization was necessary, with the exception of translating the interface pages and adding the discipline categories specific to the Italian academic context (degree courses, departments, research teams, etc.).

The top-down start-up has found the librarian community much more aware of access and publishing issues than the academic community; nevertheless on personal or Faculty web pages freely available and interesting documents can still be found, often used for teaching; naturally these are in no way systematically organised, with no possibility of easy retrieval or permanence.

Since our institutional open archives aim not only to meet the global demand for new scholarly communication models, but also to meet the more specific needs of our academic community, an architecture with double archives has been chosen:

- UnipiEPrints, institutional archive devoted to research works of teachers and researchers and to institutional documentation (in the production stage since December 2006);

- UnipiEPrints Didattica, a repository for the different typologies of educational documents (production phase scheduled for April 2007; access will be restricted to the University network - some software development is required).

We are now entering the promotion phase: a first official presentation of the system will take place at the University Senate; a series of meeting and seminars in the faculties and departments will follow, aimed at explaining how UniEprints works and making academic authors aware of the economic models of scholarly communication, the policies of publishers, and the issue of copyright. We imagine a promotion model strongly based on libraries and librarian support. We have already observed that the main questions posed by Professors concern academic evaluation of works, with a number of distinctions:

- The STM community appears to be more interested in the assessment of quality, impact factor and referee etc.;

- The Humanities community also publish in monographs or in journals which are not included in bibliomethrics databases and sometimes have distribution problems; their main concern regards certainty of the publication status, above all in a situation where recent changes in legal deposit law are still awaiting case law and practices;

- We have found allies in a research group of the Political Science Faculty: their research field regards the social and philosophical aspects of knowledge production and communication, and the group has developed their own OAI disciplinary repository, also based on eprints software. This group is willing to interoperate and to support promotion.

The project group proposal to the academic government, which has shown interest, is to connect the EPrint system with "Anagrafe della ricerca", i.e. the official data which Italian Universities have to record as a base for fund distribution: this combination would significantly enhance use of the system.

A Survey on magiran.com: A Database for the Magazines of Iran

Mortaza Kokabi

Department of Library & Information Science, Shaheed Chamran University, Ahwaz, Iran
e-mail: Kokabi80@yahoo.com

Abstract

This paper present the design and function of magiran.com, a databse of periodicals published in Iran. It also attempts to answer the following questions: How many of the total periodicals published in Iran are covered by magiran? What is the subject coverage of the periodicals covered? Which subjects seem to have been given importance among the periodicals covered? How many of the periodicals are available full text? What is the subject coverage of the periodicals available full text? What are the languages of the periodicals covered? How many of the periodicals accredited by MSRT are found in magiran? What is the subject coverage of the accredited periodicals? Which subjects seem to have been given importance among the accredited periodicals covered? How is the general structure of the site in terms of colors, icons, pull-down windows, and so on?

Keywords: Iranian periodicals; subject coverage; user feedback

Although there have been some sporadic activities to index and abstract Iranian periodicals by some organizations responsible for the press in Iran, there has not been any complete source, employing the Internet and indicating the outcomes of the activities of people involved in country's press. The *ftāb Software Company* (ftābsoft.com) sponsored the design and development of magiran website [1], a database of periodicals of the country, simultaneous with the Press Festival held in May 2001 in Iran. The purpose of designing the site is producing an effective source for Iran's periodicals in the Internet.

The site has been able to cover and present services related to more than 1300 periodicals in publication, authorized by the Ministry of Culture and Islamic Guidance (MCIG). The site claims that it is used by more than 15000 users inside and outside of Iran. The free services offered by magiran are of two kinds: General and special. The general ones, besides the ones mentioned above include: the allocation of special address such as http://www.magiran.com/YOURMAGAZINE to provide quick access to periodicals information, the allocation of an email address by POP3 service with at least 10 Mb capacity such as: *YOURMAGAZINENAME@magiran.com*, the inclusion of subscription rates and forms for each periodical to subscribe from inside or outside of Iran, informing the users of the publication of new issue of each periodical, and the inclusion of the full text of new issues and the last ten issues of some periodicals. Free special services include: providing special pages for each periodical. There are some Extra services that are based on request and payment. Another useful service of the site is its acting as a dealer.

The subject directory of periodicals is a list that is continued via a link in another page. On the same page, the periodicals are presented according to a subject directory, the Ministry of Sciences, Researches, and Technology (MSRT)-accredited periodicals given importance by being at the top. On the same page, the periodicals can be searched through a search box, the search can be limited by some options, and the newspapers covered are also shown. There's a "sending message, viewpoint, and suggestions" page, on the same page, as "Introduce the site to your friends" and "Report the problems with the information" options. The site is totally independent and private and has no connection to any governmental or non-governmental institution or organization.

This paper tries to find answers for the following questions: How many of the total periodicals published in Iran are covered by magiran? According to the latest statistics belonging to 2005 [2], the total number of periodicals published daily, monthly, bi-monthly and quarterly in Iran is 1832. Thus the site covers approximately 70% of the periodicals authorized by MCIG; 2. What is the subject coverage of the periodicals covered? Literature, Art, Technical and Engineering, Society and culture, Areas and ethnicities, Industries, Agriculture, Information, computer and internet, General, Basic sciences, Islamic sciences, Humanities, Groups, Ecology, Commerce and economics, Health and treatment, Sport and entertainment, Education and research, Associations and NGOs; 3. Which subjects seem to have been given importance among the periodicals covered? The periodicals covered by Magiran are categorized in 20 subject groups, each subdivided in turn into some sub-categories with each of which the total number of periodicals in that sub-category is given. Of these subcategories, "Industries" with 31 sub-categories has the highest rank but "Health and treatment" with 192 titles has the highest rank in respect to the total number of periodicals in sub-categories; 4. How many of the periodicals are available full text? 92 titles; 5. What is the subject coverage of the periodicals available full text? No information on the subject coverage of these full-text periodicals is given; 6. What are the languages of the periodicals covered? No specific information on language coverage is given in the site, but periodicals are mostly in Farsi, the official language of Iran, and some are in English. For some periodicals, only the abstracts are given in English. Some periodicals are also bilingual, or in fact bi-dialectal, such Farsi-Kurdish; 7. How many of the periodicals accredited by MSRT are found in magiran? 130 titles (10%); 8. What is the subject coverage of the accredited periodicals? Agriculture, Medicine, Basic sciences, Art and Architecture, Humanities, Technical and Engineering; 9. Which subjects seem to have been given importance among the accredited periodicals covered? "Medicine" with 65 titles seems to be the most important subject; the "Agriculture" and "Humanities" both with 20 titles is the second; 10. How is the general structure of the site in terms of colors, icons, pull-down windows, and so on? The site was matched against some criteria [3], and the results are as follows: The site address and domain and the keywords in the site name are visible; the name of the site is short and informative; the text is readable but the fonts could be better; the illustrations, though not very frequent, are attractive; no site map; writing and grammar are acceptable; there's the possibility of navigation through the site; the name of the designer company is seen, but not the site administrator; the last date of updating is not seen; no FAQ provision; the introduction of new periodicals exists; no "help" provision but "about us". The site takes between one and two minutes to load, much longer than the standard 8 seconds.

Suggestions to improve the site are as follows: more beautiful Farsi fonts could be applied; and the inclusion of more illustrations; the site map; the total periodicals authorized by MCIG; more full-text periodicals; the last date of updating; FAQ provision; and "help" provision seem necessary.

Notes and References

[1] http://www.magiran.com

[2] http://www.sci.org.ir/portal/faces/public/sci/sci.gozide

[3] SABERI, M. *A comparison between the content and construct features of the central libraries' homepages of US, Canada, and Australia with those of Iran along with a survey on the viewpoints of users and experts to present an optimal model*. MLib. Dissertation, School of Education and Psychology, Shaheed Chamran University (in Farsi)

Developing National Open Access Policies: An Ukrainian Case Study

Iryna Kuchma

Social Capital and Academic Publications Program, International Renaissance Foundation
46 Artema str., Kyiv, 04053, Ukraine
e-mail: kuchma@irf.kiev.ua

Abstract

Since January 2007 Ukraine has a law mandating open access to publicly funded researches. It was widely supported by most of the Parliament members. And it is already the second parliamentary inquiry mandating the Cabinet of Ministers to take actions on creating favourable conditions for developing open access repositories in archives, libraries, museums, scientific and research institutions with open access condition to state funded researches. And for the second time the implementation of this law was interrupted by the political crises. Grass root initiatives of Ukrainian Universities and libraries as well as the political support from the principle legislative body in the country have still not resulted into a single well-functioning institutional/national repository. The poster highlights the developments that have taken place, actions for the years to come and recommendations for the countries that are in circumstances that can be compared to Ukraine.

Keywords: open access; mandating policy; publicly funded researches; institutional repository

1 Introduction and Developments

Mandating open access to publicly funded research in Ukraine was a movement launched by the scholars publishing their articles in open access journals, innovative librarians and University administrations. This movement was co-ordinated by International Renaissance Foundation (IRF, Soros Foundation in Ukraine), which since 2004 organised a number of awareness raising campaigns in mass media and regional seminars for the academic community. National Academy of Sciences (NAS) and International Researches and Exchanges Board (IREX) supported open access ideas and joined the movement.

The first public statement on open access policies in Ukraine was drafted during the international Open Access Scholarly Communication Workshop hosted by the National University Kyiv-Mohyla Academy (NAUKMA) and organised by IRF, Open Society Institute, NAS and International Association of Academies of Sciences on February 17-19, 2005. 140 researchers, administrators, librarians, information managers from higher educational institutions and scientific research laboratories involved in e-journal publishing and institutional repository development from 17 countries signed the Recommendations for Ukrainian authorities to ensure: the right of individuals and the public to access information and knowledge and to guarantee that intellectual property regimes are not the obstacles to the public access to knowledge, to encourage research and higher educational institutions to practice open access and to put an open access condition to state funded researches (except reasonable exceptions) and to provide state financing and technical assistance to research and higher educational institutions to set up and maintain open access repositories.

These Recommendations were endorsed by Ukrainian Vice Prime Minister. And on September 21, 2005, the Recommendations were presented at the first Parliamentary hearings on Developing information society in Ukraine. In December 2005 these hearings resulted into the Parliamentary Inquiry on Harmonisation of Governmental Educational Policies re open access movement [1]. Open access was one of the priorities in developing information society in Ukraine. The Cabinet of Ministers was responsible for creating favourable conditions for developing open access repositories in archives, libraries, museums and other cultural institutions and the Ministry of Education and Science of Ukraine – for encouraging development of open access resources in science, technology and education with open access condition to state funded research. Beginning of 2006 was also the time of parliamentary elections campaign, when the "old" Cabinet of Ministers didn’t feel any responsibility to start new activities like open access projects. And later on two "new" Cabinet of Ministers were busy trying to cope with political crises in summer and autumn 2006.

In September 2006 representatives of Parliamentary Committee on Science and Education, State Fund for Fundamental Researches, Scientific and Publishing Council of NAS, Ministry of Science and Education of

Ukraine, National Library of Ukraine after V.Vernadsky, State Department of Intellectual Property, Kyiv public administration, Association "Informatio-Consortium", Institute of Social Development and IRF created a working group on developing open access policies in Ukraine and pushing the Cabinet of Ministers to implement the resolution of Ukrainian Parliament on Open Access.

In November 2006 State Fund for Fundamental Researches commissioned IRF to develop an Open Access Policy for their grantees reporting publicly funded research. The goal was to require electronic copies of any research papers supported in whole or in part by Government funding to be deposited into an institutional digital repository immediately upon acceptance for publication.

Both initiatives turned previous parliamentary resolution into the law mandating open access to publicly funded research [2]. According to the law there should be six months of transition period (completed by July 2007). But the following political crises withdrew the attention of the Cabinet of Ministers from immediate implementation of this law.

Since October 2005 a grassroots initiative of the academic community undertook a project to create a network of open access repositories in Ukraine. Nine Ukrainian Universities reported this decision at the national conference for university and regional universal scientific libraries INFORMATIO 2005. The project has been implemented by Association "Informatio-Consortium", Scientific Library of National University Kyiv Mohyla Academy, Lviv Catholic University and Centre for the Humanities of Lviv National University after I.Franko. All these projects still lack financing and skilled staff. So far only pilot institutional repositories have been created.

Governmental institutions are still the unique donors of research and development in Ukraine. This is why a law mandating open access to publicly funded research plays a crucial role in open access initiatives. Delays with implementation of this law cause delays in the development of open access institutional repositories.

Nevertheless we will continue financial and expert support to Ukrainian network of open access institutional repositories encouraging Universities and research institutions to sign the Berlin Declaration and introduce self-archingl policies, develop model open access institutional repositories and providing training for the interested organisations. At the policy level we will keep pushing the implementing of the law of Ukraine mandating open access to publicly funded research. IRF implements open access projects in cooperation with the Information Program of the Open Society Institute and the Electronic Information for Libraries Consortia (eIFL).

2 Recommendations

Recommendations for countries that are in circumstances that can be compared to Ukraine: 1) alliances are crucial and local partners needed; 2) targeted web-sites and workshops proved to be useful tools for awareness raising and lobbying; 3) support from mass media is important to create public awareness.

Notes and References

[1] Decree of the Parliament of Ukraine "On Recommendations of parliamentary hearings on developing information society in Ukraine: http://zakon.rada.gov.ua/cgi-bin/laws/main.cgi?nreg=3175%2D15

[2] The Law of Ukraine on the principles of developing information society in Ukraine for 2007-20015 (Закон України "Про Основні засади розвитку інформаційного суспільства в Україні на 2007-2015 роки") at www.rada.gov.ua

Digitization of Scientific Journals in Serbia

Žarko Mijajlović[1], *Zoran Ognjanović*[2], *Aleksandar Pejović*[3]

[1] Faculty of Mathematics, University of Belgrade, Belgrade, Serbia
e-mail: zarkom@matf.bg.ac.yu
[2] Mathematical Institute SANU, Belgrade, Serbia
e-mail: zorano@mi.sanu.ac.yu
[3] School of Electrical Engineering, University of Belgrade, Belgrade, Serbia
e-mail: pejovica2@gmail.com

Abstract

A digitization project in progress carried out by the Mathematical Institute of the Serbian Academy of Sciences, Belgrade (http://www.mi.sanu.ac.yu) and the Faculty of Mathematics, Belgrade (http://www.matf.bg.ac.yu) is described. The projects aim is to build a database and an electronic presentation of digitized scientific books and journals printed in Serbia, particularly in mathematical sciences (mathematics, mechanics, astronomy, computer science and physics) and to make them searchable and available in the full-text mode on the Internet.

Keywords: scientific journals; retro digitization; digital archives

1 Introduction

There is a number of mathematical and mathematically-related books and journals printed in Serbia. Some of the published works have been digitally created (usually in TeX and its versions) and more-or-less they are accessible using ordinary web browsers, while the others, mostly older papers, have been born in print and are harder to obtain. Thus, we have also started the process of retro digitization of these printed works to produce their digital images that can be read or printed. The main achievement of the project in last two years is complete retro digitization of two leading Serbian mathematical journals: *Publications de l'Institut Mathematique* and *Publications of the Faculty of Electrical Engineering – Series Mathematics and Physics*. Besides these two completely digitized journals, five Serbian journals in mathematics, teaching and computer science are stored in the database. Some of them, for example *NCD Review*, a journal on digitization technologies, founded in 2002, are presented completely, while the others are partially digitized, mainly from the beginning of nineties of the last century. All together, there are more than 4000 digitized articles having about 30000 pages. Digitized items are displayed at the virtual libraries: http://alas.matf.bg.ac.yu/biblioteka/home.jsp, http://publication.mi.sanu.ac.yu, http://pefmath2.etf.bg.ac.yu/ and http://elib.mi.sanu.ac.yu/pages/browse_journals.php.

2 Serbian Mathematical Journals in Digital Archives

The journal *Publications Publications de l'Institut Mathematique* is the oldest Serbian scientific journal in the field of mathematics established in the year 1932 under the name *Publications Mathmatiques de l'Université de Belgrade*. It was founded with the help of two foundations of the Belgrade University foundations. Seven tomes were published until the World War II, the eighth tome was lost in the German bombing of Belgrade in April 6. 1941. Immediately after the founding of the Mathematical Institute in 1946, the publication of the journal was restarted in 1947 under the new name *Publications de l'Institut Mathematique*. More then 2000 articles were published in 102 volumes until these days. The scope of the journal in the beginning was broader, not only in mathematics, but articles referring to mechanics and astronomy were published in it as well. Most prominent Serbian and Yugoslav scientist in these fields published in the journal, including Đ. Kurepa, J. Karamata, M. Petrović, M. Milanković, A. Bilimović, J. Plemelj, S. Mardešić, and others. Some of the leading world mathematicians published in Publications as well: H. Lebesgues, P. Montel, P. Erdös, W. Sierpinski, S. Shelah, and others. Most papers in the journal are in English, but there are papers written in Russian, French, and German as well. The second journal, *Publications of the Faculty of Electrical Engineering – Series Mathematics and Physics*, was founded in the year 1956. In the beginning, each contribution appeared separately bound and numbered consecutively, several times a year. Since 1959, the issues have been appearing collected in one or more volumes per year. In the first years, the journal had contributions from different fields apart from Mathematics: Physics, Mechanics, and Electrical Engineering. Papers were written in the Serbian, French, Russian, German and English. In the course of time, the journal focused almost exclusively on Mathematics, especially convexity, functional equations and differential equations, and English language became dominant. The digitized version of the journal contains about 1000 papers. Both journals are reviewed and indexed in: *Mathematical Reviews* (MR), *Zentralblatt Math* (ZBL) and Russian *Mathematical Surveys*.

3 Digital Objects and Metadata

The digital object in the virtual library usually consists of several components: digitized image of the manuscript, some graphic components, and metadata. We developed a particular data base and Internet oriented software for handling digitized journals. It relies on three types of metadata: descriptive, structural and administrative. Special data and services important for papers published in scientific journals were also included: keywords, scientific classification of AMS (American Mathematical Society classification), numbers of reviewer reports in MR and ZBL, DOI numbers, and statistics of accessing and downloading papers. Descriptive metadata follow data contained in librarian printed catalogs, i.e., they obey librarian standards. One problem was that old issues do not have standard descriptive tags such as ISSN numbers so to classify them we needed particular solutions. Structural data explain how the components of the digitized object are interconnected. Administrative data describe exactly how an item is preserved: resolution, rate of compression, file type containing the digitized image, etc. The success of digital preservation efforts will rest to a significant degree on the scope and reliability of the metadata records. For example, metadata made possible the asset-management systems that back up and periodically duplicate digital records. Cataloging information enable one to locate what they are looking for in the library. Metadata help to make various internet presentations. Therefore full repository system required tens of metadata elements for each digitized item. Building such database systems and populating them is very labor-intensive and expensive. Creating the table of keywords and assigning them to articles was particularly complicated and time-consuming job since it could be done only by scientists. Some trade off needed to be found. For resolving these issues, cooperation between institutions working in the field of digitization was very important, in particular exchange and agreement of metadata formats. Particular attention was given to standards. Scanning was performed in 300dpi, in tiff format. Papers were converted into pdf format and in this form they are accessible on Internet. As curiosity, let us mention that all issues of *Publications Publications de l'Institut Mathematique* between 1980 and 1990 are retyped in TeX, and all issues of this journal since 1980 are accessible in dvi format, as well.

4 Implementation

We decided to develop our own software instead of using commercial, or open source software. We decided so, since we wanted to lower the development and maintain cost, then because of future upgrading and integration it in larger information systems, such as virtual libraries of wider scope. The software supports all usual functions, browsing, searching under various criteria, examining and downloading papers. Since papers were written in several languages, we decided to keep the multilingual feature. Therefore, we have chosen MySQL server for a database as it supports UTF-8 encoding. The multilingual support is embedded into the model of data, so information related to the corresponding languages are saved. JAVA programming language is used in developing a web application for administering and searching the database, especially advanced features like JAVA beans and strucs, which enable a high performance web application. Other technologies include PHP and Apache as a web server.

5 Conclusion

Digital archive of Serbian scientific journals will contribute significantly to the widespread accessibility of articles printed in these journals, particularly since they are obtainable on the web free of charge. One of the consequences will be the rise of scientific impacts of these journals and articles printed in them. A further plan assumes that our Virtual library will include once editions of all important Serbian scientific journals. It is difficult to estimate when this task will be finished, but a decade, we believe, is a good guess.

References

[1] MIJAJLOVIĆ, Ž., *On some undertakings in the field of digitization in the last decade*, NCD Review, 2002, http://elib.mi.sanu.ac.yu/pages/browse_article.php?cs=000001&rd=0000003.

[2] ARMS, W Y., *Digital Libraries*, MIT Press, 2001.

DRIVER - Digital Repository Infrastructure Vision for European Research

Mary L. Robinson

SHERPA, Information Services, University of Nottingham
Greenfield Medical Centre, Medial School, QMC, Nottingham, UK.
e-mail: mary.robinson@nottingham.ac.uk

Abstract

The current system of academic publication developed as a means to disseminate the findings of research. However, this system can impede the very process it was set up to serve, with access to articles being limited by publishers to only those who can afford to subscribe. This poster will explain the vision behind DRIVER and will describe how the various aspects of the project tie in together to form the knowledge infrastructure of the European Research Area. The poster will focus on the key aspects of the DRIVER project and the questions and needs that each addresses. The key aspects which will be addressed include: DRIVER technical developments and advice, the DRIVER Support website [1], community development, up-to-date news, and the benefits for various stakeholders. DRIVER is an ambitious and important project that will yield valuable results for individual researchers, the publishing community, funding agencies and the European Research Community as a whole.

Keywords: open access; repositories; European Research

1 Introduction

Open access digital repositories provide a means whereby the traditional publishing model can co-exist with the needs of authors and their readers, as well as with the demands of research funders for research impact and hence, value for money. Subject to copyright, authors can deposit copies of their finished articles in open access repositories, in addition to publishing them in research journals.

The recent study of scientific publication markets in Europe funded by the European Commission [2] strongly recommends the development of a European policy mandating open access to EC-funded research. In addition, it recommends an exploration of interoperability issues and how open access repositories can be implemented Europe-wide.

DRIVER- Digital Repository Infrastructure Vision for European Research- is an EU-funded project with 10 international partners and reflects the growing awareness in Europe surrounding Open Access. DRIVER sets out to build a testbed for a future knowledge infrastructure of the European Research Area. It aims to deliver any form of scientific output, including scientific/technical reports, working papers, pre-prints, articles and original research data to the various user groups. The testbed is based on existing nationally organized digital repository infrastructures. Other work includes the support of new European repositories and an active advocacy and community building programme to address and support key stakeholder groups in Europe.

2 Objectives

The five objectives of DRIVER are:

1. To organise and build a virtual, European scale network of existing institutional repositories;
2. To assess and implement state-of-the-art technology, which manages the physically distributed repositories as one large scale virtual content resource;
3. To assess and implement a number of fundamental user services;
4. To identify, implement and promote a relevant set of standards;

5. To prepare the future expansion and upgrade of the DR infrastructure across Europe and to ensure widest possible involvement and exploitation by users.

3 Discussion

This poster will provide key information on the DRIVER project including the ten project partners and the DRIVER Support website and logo. The poster will identify the various aspects of the DRIVER project and the questions and needs that each addresses. The poster will address the following: DRIVER technical developments and advice, the DRIVER Support website, community development, Up-to-date news, and the benefits for various stakeholders.

DRIVER has conducted focused research studies including an inventory of the type and level of OAI compliant digital repository activities in the EU [3], to facilitate the iterative development of DRIVER and is developing the necessary infrastructure middleware and user guidelines to meet the DRIVER objectives. The project is now actively advocating repository development - creating an informed and active environment for repository infrastructure development in EU countries with focused activities, information and contextualized support.

References

[1] DRIVER (Digital Repository Infrastructure Vision for European Research), http://www.driver-support.eu

[2] European Commission. Study on the economic and technical evolution of the scientific publication markets in Europe, Jan. 2006, http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf

[3] DRIVER. Inventory study into the present type and level of OAI compliant Digital Repository activities in the EU, Apr. 2007.

The Inclusion of Open Access Journals in Academic Libraries: A Case Study of Bioline International

Jen Sweezie[1]*; Nadia Caidi*[2]*; Leslie Chan*[1]

[1] Bioline International, University of Toronto at Scarborough, 1265 Military Trail, Scarborough, ON M1C 1A4
e-mail: {sweezie; chan}@utsc.utoronto.ca
[2] Faculty of Information Studies, University of Toronto, Toronto, ON, M5S 3G6
e-mail: caidi@fis.utoronto.ca

Abstract

Specialized open access digital collections contain a wealth of valuable resources. However, major academic and research libraries do not always provide access to them, and thus do not benefit from these unique resources. This case study of one such digital collection, Bioline International, surveys 76 academic libraries in Canada and the United States to determine how often libraries are linking to the collection. A follow-up questionnaire was sent to librarians at the surveyed institutions to determine their opinions about the use of open access journals. The findings suggest issues of poor adoption rates of open access journals, as well as some reasons why such journals may not be actively adopted.

Keywords: open access; journal collection; developing countries; bioline international; digital library

1 Introduction

Librarians and information professionals continue to struggle with a growing number of available publications and limited budgets, making the selection of library resources difficult. Open access (OA), and in particular, specialized OA collections provide libraries with access to a broad range of high quality academic research and offer the possibility to help alleviate what is referred to as the serials crisis [1]. This study examines why such valuable resources are often not incorporated into library collections. A case study of Bioline International (BI) is used to illustrate this problem.

Bioline International (http://www.bioline.org.br/) is a specialized OA collection that offers open access to over 50 bioscience journals published in developing countries. The BI website provides free access to regional journals in environmental and agricultural sciences, heath and medicine, that may be difficult to obtain elsewhere [2].

2 Methodology

Between October 2005 and April 2006, an exploratory study of Canadian and American academic libraries was carried out to determine how many BI journals were included in the libraries' collections. E-journal or e-resource sections of each library website were searched and a checklist was completed, indicating the presence of BI journal titles. 76 (46 Canadian and 30 American) libraries were surveyed for BI's journals. A list of all academic libraries in Canada was generated using Yahoo categories, resulting in the survey of 46 libraries. Due to the large number of academic universities in the United States, a sample of libraries was necessary. A ranking of health science libraries by number of total electronic materials from the Association of Research Libraries (ARL)[3] was obtained. 30 American libraries were selected from the table. Effort was made to select libraries from the top, middle, and bottom of the ranking list, in an effort to ensure that the libraries selected represent a variety of different institutions (both private and public) across the United States.

After the e-journal lists were surveyed, a questionnaire was e-mailed to librarians identified during the library investigation process. Efforts were made to send the questionnaire to librarians working in collection development or on electronic resources lists. Where such a contact was not available, the questionnaire was sent to the head librarian. 76 questionnaires were emailed and 17 responses were received. Librarians were asked about their opinions surrounding the use of open access journals in library collections, as well as about institutional policies and decision making surrounding open access journals.

3 Results

Preliminary findings indicate that in Canada only 33% of research libraries (15 /46 libraries examined) were linking to 50% or more of the titles available from BI (more than 29 out of the 59 available titles). In the United States, 56.7 % (17/30 libraries studied) of the research libraries surveyed were linking to 50% or more of BI's 59 journal titles. 89% (41/46 libraries studied) of Canadian libraries offered at least one BI journal title through their library collection. In the United States, 96.7% (29/30 of libraries studied) of libraries offered at least one BI journal title of the 59 available through their library collection.

Libraries were considered to be aware of the BI project if they linked to one or more Bioline journals, even if they linked to the publishers (or other website) directly, and not to BI. Journals participating in the BI project actively promote their inclusion in Bioline through their websites and on the covers of their printed journals, thus knowledge of the BI project is assumed. In both countries, most libraries were aware of BI titles, either through the BI website, through another project or the publisher's website. However, inclusion of all 59 BI journal titles was still relatively low (under 11%) for both countries.

The results of the questionnaire sent to collections development and electronic resource librarians indicated that a number of reasons may contribute to the low level of appearance of BI journals in the library collection. Lack of inclusion in major commercial or open access databases and indexes, the length of time a journal has been publishing, lack of librarian time to seek out and catalogue (or even keep up with) new titles, institutional policy and reputation or perceived credibility of OA journals were all cited as factors in why such OA journals are or are not included in library lists.

4 Conclusion

The results suggest that despite the open access nature of BI and the range of its offerings, librarians are not making effective use of the BI collection. Projects such as BI must address the concerns and needs of librarians in order to improve the rate at which journals are added into library collections. Though librarians are often aware of open access journals, time constraints may be considered one of the largest barriers to adopting these journals. The unique situation of open access journals is that they do not often have the large scale budgets of large, mainstream publications, making it difficult to develop compliance with a number of web standards being used by libraries today. Standards such as OpenURL and metadata linking protocols can greatly increase the ease of adoption of these journals in libraries, and improve the chances of having such journals eventually indexed in major databases. These protocols allow librarians to easily link new journal material into their catalogues, and in some cases, some commerical applications are already adding some OA journal titles as options in their databases – librarians merely have to toggle them on or off. Issues of sustainability and journal quality will likely improve gradually as more and more libraries link to and promote OA journals, and journals are included in more databases and indexes. Projects such as BI can provide a sustainable platform by working independently from their individual journals – ensuring web access that is reliable – while also working for the collection as a whole in developing protocols and linking systems that the journals may not be able to generate on their own.

Ongoing study into the trends of inclusion of OA journals in library collections lends itself to a number of conclusions about the barriers to open access publications, how they could be better promoted, and how librarians can be encouraged to make use of these valuable resources.

References

[1] OJALA, M. (2005) Open access: open sesame or opening Pandora's Box. EContent 28, 6, 31-32, 34-35.

[2] BIOLINE INTERNATIONAL (BI). "About Bioline" Retrieved Jan 24, 2006 from http://www.bioline.org.br/info?id=bioline&doc=about

[3] YOUNG, M.; KYRILLIDOU, M. (2005) ARL Academic Health Sciences Library Statistics 2003-2004. 29. Retrieved Jan 20, 2006 from http://www.arl.org/stats/pubpdf/med04.pdf

Developments in Publishing: The Potential of Digital Publishing

Xuemei Tian

School of Business Information Technology, RMIT University
GPO Box 2476V, Melbourne 3000, Victoria, Australia
e-mail: Xuemei.Tian@rmit.edu.au

Abstract

This research aims to identify issues associated with the impact of digital technology on the publishing industry with a specific focus on aspects of the sustainability of existing business models in Australia. Based on the case studies, interviews and Australian-wide online surveys, the research presents a review of the traditional business models in book publishing for investigating their effectiveness in a digital environment. It speculates on how and what should be considered for constructing new business models in digital publishing.

Keywords: digital publishing; business models; print-on-demand; publishers; Australia

1 Introduction

This poster session emerges from an Australian Government funded digital publishing research project. The research looks at the impact of technology and market change on traditional publishing practices, and the increasing imperative to digital publishing. For our purposes digital publishing is defined as publishing dependent upon the World Wide Web as its communication channel, producing digital content based on either domestic or global platforms, published and distributed online, with provision for the establishment of digital database facilities for future re-use. The process allows for links to e-commerce, for example, facilitating online payment, with all procedures in the process digitised. Based on customer requirements, the product (information) can be produced and provided in various formats, such as online, web, TV, CD Rom and if necessary, paper (Liu and Rao, 2005). Additionally, Print-on-Demand (PoD) and Video-on-Demand (VoD) are elements of digital publishing. There is a general consensus that the digital publishing production and supply chain incorporates authors, publishers, technology providers, databases, web distributors and end-users.

The Australian publishing industry has maintained a significant presence within the manufacturing and distribution sector over many years (IBISWorld, 2006). Over the past decade the publishing industry has undergone tremendous changes, including publishing markets, an increase in digital content formats, changes to distribution channels and supply chains. At the same time, revenues from online business have grown dramatically in recent years. Statistics reveal that major players in the Australian publishing industry have increased their online business revenues by about 30% from 2001 to 2005 (IBISWorld, 2006). In the process, publishers have been forced to re-evaluate their resources and capabilities, design new business strategies and re-engineer their business processes to take advantage of the potential of rapidly developing technology. This has led to the development and emergence of new supply chains in the publishing industry, and a need for new business models (IBISWorld, 2006)

This research investigates the practical implications of digitization for book publishing in Australia, focusing on aspects of re-engineering existing business models to maximize benefits to both the company and their customer base. Although not for profit publishers are already utilizing the Open Access system, this does not feature as a major element in this research.

2 Project Progress and Initial Findings

At the time of writing, the researchers have completed a national online survey of Australian publishers and are progressing through a series of eight case studies. The indications so far are that the Australian publishing industry has adopted a somewhat conservative approach to the challenges and opportunities presented by digital publishing. The majority of publishers believe however, that in the foreseeable future, there could be major changes in the industry. As publishing businesses have varying objectives, the pace of change from traditional to digital publishing will also vary, depending on their market and client base, as well as on take-up of technology. It seems clear however, that any reluctance on the part of publishers to embrace technological and other changes could be detrimental to their future.

3 Major Research Focus: Business Models

The major thrust of the research has ultimately been towards identifying the implications of digitization and related organizational and market changes for business models in the publishing industry. Our survey results indicate that subscription-based and content creation models maintain popularity, frequently in the context of niche markets. Faced with different markets and strategies challenging, businesses need models that fit with their particular circumstances. Rather than have recourse to, in this project we adhered to Weill and Vitale's (2001) approach to business models. They define business models as a description of the roles and relationships among a firm's customers, allies and suppliers that identifies the major flows of product, information and money and the major benefits to participants (Weill and Vitale 2001). Our case studies provide validity to this definition. To surface our perceptions of business models, we will display several business models at this poster session seeking feedback from conference participants. Figure 1 is a constructive example of a potential future business model for a book publisher, drawn directly from our research.

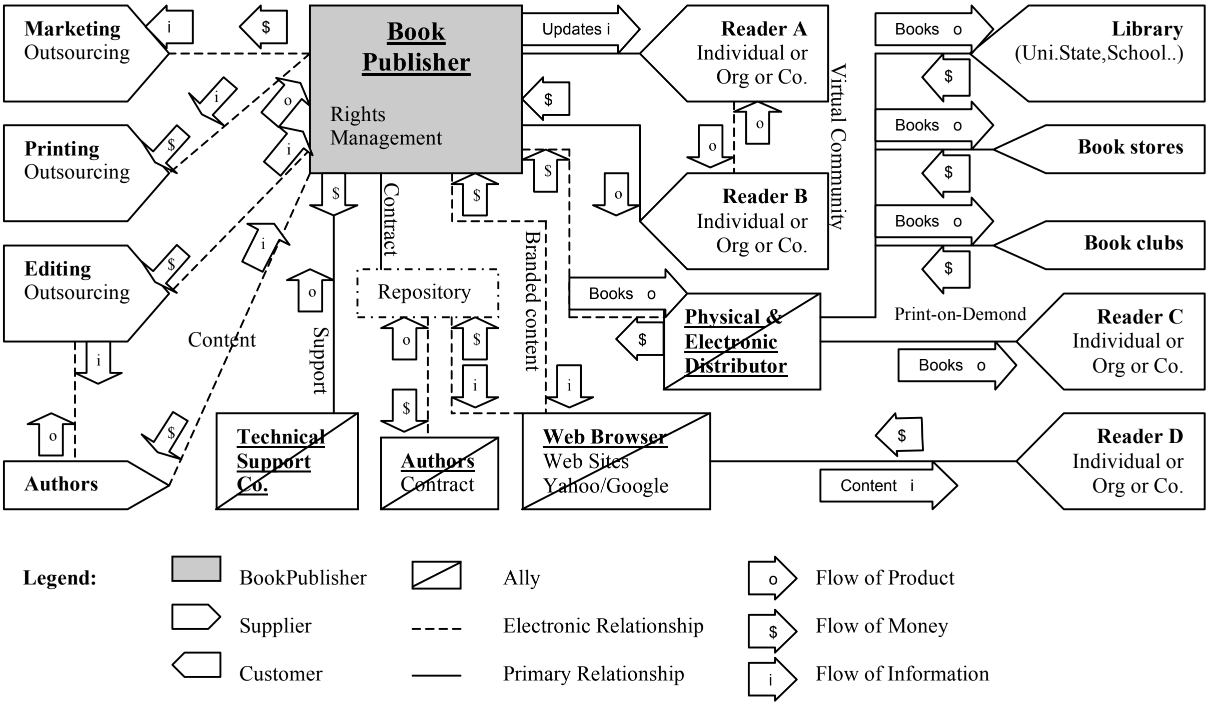


Figure 1: A constructive business model for a book publisher

4 Preliminary Conclusion

Although the research is being conducted in Australia, it has drawn widely in methodological terms from an international context. As such, it is likely that the findings, including those relating to alternative forms of business models, will be of wider relevance. Based on the models presented at this poster session we hope to add to our understanding and produce more refined models. At this stage preliminary conclusions are as follows:

- Currently book publishing models appear to be very familiar (i.e. largely traditional) with adding on digital elements.
- In 5 to 7 years time book publishing models could look vary different.

References

[1] Ibis World Industry Report. *Book and other publishing in Australia*. Sydney, Ibis World Pty, 2006.

[2] LIU, M. L.; RAO, B. H. *Operational concepts and changes of academic journals in digital age*, ChongQing University Publisher, 2005.

[3] WEILL, P.; VITALE, M. *Place to space*: *Migrating to E-Business Models*, Boston, Harvard Business School Press, 2001.

Index of Authors

Index of Keywords