Open Access and Semantic Web
SW Applications for Open publishing

Michele Barbera, Francesca Di Donato
barbera@netseven.it , francesca.didonato@sp.unipi.it

***Abstract*— *The Open Access movement, grown since the first Nineties and quickly developed in the last years, aims at enlarging the dissemination of scientific knowledge; based on the assumption that the Internet and the World Wide Web are able to offer the chance to constitute a global and interactive representation of human knowledge, including cultural heritage and the guarantee of worldwide access, the signatories of the Berlin Declaration on Open Access (2003), feeling obliged to address the challenges of the Internet as an emerging functional medium for distributing knowledge, pledged themselves to make the future Web sustainable, interactive, and transparent through the use of openly accessible compatible content and software tools. One of the encouraging applications of semantic web is devoted to the connection of the scientific knowledge in a unique global network where documents can be made machine-readable by annotating them with Dublin Core metadata expressed as RDF. Thanks to the metadata harvesting protocol of the Open Archives Initiative (OAI-PMH), the goal of obtaining a unique global network has yet become possible. Nevertheless, despite these developments are able to significantly modify the nature of scientific publishing as well as the existing system of quality assurance, nowadays the application of Web Semantic technology is limited to archiving and cataloguing; and the main issue of Semantic Web, selection by quality criteria, is lacking in application. This short paper describes a set of applications conceived in order to fill the gap; the aim of the HyperLearning project is to build a universal and free information space where researchers are able to use machines in order to apply their own selection criteria instead of using those enforced by publishers. These criteria often pursue different aims, which diverge from the real research. HyperJournal, in particular, is a web application that facilitates the administration of academic journals on the Web; it is based on four major features that will be further described, and on the idea of a shared linkbase based on a P2P technology.*

I. Introduction

From 1665 to the present with the help of the Print, scientific Journals multiplied and propagated; and it became easier and easier for scholars to establish a new Journal. A fundamental step in this story is the birth of the ”Science citation Index”. Especially after World War II, in a time of economic crisis, the problem for Libraries turns to be: How to keep track of thousands of citations? Meaning in practice: How to decide what to buy? At that time, publishers were a fragmented and isolated group. In the early Sixties of the last Century, three hundred years after the birth of Philosophical Transactions, Eugene Garfield of the American Institute of Scientific Information (ISI) recognized in the citation system the basics for the construction of an enormous net of knowledge. As a bibliographic instrument, the ”Science Citation Index” was born to provide a cartography of citations. So that, firstly, with the SCI the use of the ”Impact Factor”, a standardized measurement tool (introduced by the ISI) that allows to determine the impact of an article on later publications, took off. Impact factor is, obviously, a merely quantitative criterion, based more on the Journal than on the article itself. Nevertheless, promoted by the ISI and easily accessible, IF has quickly became a standard for Libraries. Secondly, and more importantly, Garfield reduced the entire set of little specialty ”cores” to one big ”scientific core” and used this set of journal titles as the basis of emerging Science Citation Index. The number of core journals was confined to a few thousand titles, and although has gradually grown, still remains a small fraction of all scientific journals published in the world. The restriction of the interest of Librarians from the wide park of Journals to a limited number of ”Core journals” produces important changes in the scientific publishing market, transferring Power to publishers. Journals, in the traditional publishing framework, have more than a function: they grant rights of intellectual property (right to be cited), working as a public register; they provide a brand and become an instrument for authority, more than a medium: ”Being published in a well-known journal, writes Guédon, is a bit like appearing on prime-time TV”[1]. So, they work as instruments for the evaluation and management of academic careers.

A revolutionary impulse came from the Internet and the Web. From its origins, the WWW idea is to set up a documentation system based on citations (links); to that idea, several years before, Vannevar Bush, Ted Nelson (and others) had devoted themselves. The first of them, in the well known essay As we may think? of 1945, supposed the action of a fotoelectronic engine, the Memex, able to make and to follow crossed links on microfilm, using binary code, photoelectric cells, and snapshots. The second one is the author of the famous Literary Machine[2] (written in the early eighties) and the inventor of the term Hypertext, an expanding grid potentially able to connect in a unique system all texts of the world literature, following the project of the utopic software Xanadu. Every citation would carry a link with its source, assuring a reward to cited authors. It’s interesting to notice that Vannevar Bush’s work inspired also Eugene Garfield, the inventor of the Science Citation Index. The essential difference between the information system of Garfield (based on ”Core Journals”), and the one inspired from the Internet and the Web, is that the second is based on decentralization. The advent of new technologies provides alternative and innovative solutions to disseminate low-cost scientific literature (and cultural Heritage in general), and provides complementary and not competitive strategies to assure open access to the results of Public-funded research.

Since Los Alamos archive in 1991[3], similar archives began to be contemplated and implemented in a variety of fields and according to various disciplinary and institutional schemes; the movement began to grow and expand until the need for some kind of federative action became obvious. The result is the Open Archives Initiative, financially supported by several -mainly U.S.- institutions, and the Open Access Movement[4].

As we read in Budapest Open Access Initiative[5]:

1) Open access is intended as a comprehensive source of human knowledge and cultural heritage that has been approved by the scientific community. In order to realize the vision of a global and accessible representation of knowledge, in the future Web content and software tools will have to be openly accessible and compatible.
2) This idea is related to Public-funded scientific results, that authors publish for free; many initiatives for OA are promoted by public and private associations as well as Library networks, Academic Institutions and Research Centres, and by the Soros Foundation.
3) ”open access” to this literature, means its free availability on the public internet in a wide sense. The only relevant constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited.

There are two primary vehicles for delivering open access: archives and journals; therefore, OA has two branches:

- In Self-Archiving, authors use public repositories to archive their work:
  This strategy protects information from censorship and monopolistic power: archives, acting as repositories and catalogs, include everything, and the inclusion criteria are transparent. From an ”impact factor” perspective, the Open Archives Initiative model is successful. Thanks to the OAI- PMH (Protocol Metadata Harvesting), data about documents are exchanged among a net of archives in a distributed system of peers, who share common information.
- Open Publishing means publishing in on-line free Journals. Open access journals, as traditional ones, do perform peer review.

Despite Open Archives ones, Open Journal software, especially for the Humanities, still need to be developed. And HyperJournal, as we’ll see later on, is a response to this need. This system can give Citizens access to peer-reviewed research (most of which is unavailable in public libraries), whose research theyve already paid for through taxes.

II. HyperJournal

HyperJournal[6] is a web application that facilitates the administration of academic journals on the Web. Conceived for researchers in the Humanities and designed according to an easy-to-use and elegant layout, it permits the installation, personalization, and administration of a dedicated Web site at extremely low cost and without the need for special IT-competence. HyperJournal can be used not only to establish an online version of an existing paper periodical, but also to create an entirely new, solely electronic journal. In comparison with existing software applications, HyperJournal introduces four major innovations:

A. Dynamic contextualization

Dynamic contextualization automatically transforms cross-references contained in journal articles into hypertextual, bidirectional links. When the reader views an article published in HyperJournal, a contextualization bar provides immediate access to a) all the articles the author has cited, and b) all the articles that cite the article currently being viewed.

B. The HyperJournal Network

Dynamic contextualization is not limited to one journal only: it connects all the journals that use the HyperJournal software in a distributed, semantically structured and scaleable peer-to-peer network[7]. Additionally, Compatibility with the Protocol for Metadata Harvesting of the Open Archives Initiative ensures maximal interoperability between the HyperJournal Network and other electronic publications. The HyperJournal Network thereby creates a space in which knowledge is freely shared and readily accessible. Rather than using mere keyword searching or importing artificial conceptual tables to organize this space, HyperJournal transposes the time-honoured system of scholarly citation into an electronic environment.

C. HyperJournals versus core journals

By clicking on an authors name, the HyperJournal system automatically searches the entire HyperJournal network and produces a citation list that includes all the articles written by the author, all the articles the author has cited, and all the articles that cite the author. Comprehensive bibliometric lists can thereby be composed without the need to rely on the manual consultation of a small set of core journals, often exclusively in English. In this system, by contrast, it will be the actual give-and-take of academic discourse, registered automatically on the network through citations, which will signal the prestige of a journal (even of small niche journals written in so-called minor languages) and establish the reputation of scholars. In addition, through the use of RDF describers, bibliometric lists can be constructed that distinguish, for example, between positive and negative citations.

D. Structured vs. Opaque Formats

The adoption of structured formats such as XML has enormous advantages over unstructured or opaque ones (such as MS Word or PDF)[8]. One of the major advantages is that structured formats are machine-understandable thus perfectly suited to be used in conjunction with Semantic Web technologies. The most widely adopted structured format is undoubtedly LATEX which is widespread within the scientific community. Unfortunately its usage within the Humanities is very limited. On one hand this is a drawback, on the other hand it leaves space for the diffusion of XML (who has even nicer

computability properties then LATEX) as the format of choice. Initiatives such as TEI has already gained wide acceptance within Humanities Scholars. TEI and other XML dialects such as DOCBOOK have the potential to be used to author articles, not only to encode existing texts[9]. Although HyperJournal allow the editorial board to choose which document formats are acceptable for submission, it also offers to the authors all the tools they need to use structured formats for writing their articles. In particular, the HyperJournal developers community is customizing and adapting some XML editors to facilitate the authors in their work. In any case, if the adoption of XML as a format for writing articles will be successful, we can expect searches to be easier and much more powerful than todays heuristic search techniques and even to remarkably reduce the cost of paper publication, as transforming XML to other formats suited for paper printing is a trivial task.

III. Semantic Web technologies in HyperJournal

Existing e-journal web applications organize information using hierarchical trees, although connections between published articles following citations could be better represented by directed labelled graphs. For this reason RDF is perfectly suited to represent citation information. Each node of the graph is an article, while an arc represents a citation. Additionally HyperJournal uses Dublin Core-like ontology to express metadata. The HyperJournal ontology is encoded following the OWL and RDFS recommendations. RDF is stored in a Sesame database and queries are performed via http. Each instance of HyperJournal maintains his own local storage, which is synchronized with the other instances via a P2P engine called HyperRDFGrowth.

IV. The subjectiveness of quality

The problem faced by researchers is not to find information but to find the information which adheres to each researcher's own quality criteria. In this case the notion of quality is heavily influenced by subjective, geographic and cultural factors, quickly changing and evolving over time. Let us now think for a moment about a researcher in a library during the pre-Electronic era. In a library researchers look up in catalogues, where they can find volumes; place these volumes on their desk and proceed their surfing in the sea of information following bibliographies, quotations, indexes. This sort of link selection has already been submitted to a filtering made by publishers and authors according to personal criteria. Nowadays the Web let researchers browse in a large library but it doesn't offer search and browsing tools that differ much from the old ones. Computers substitute paper and search engines substitute indexes. Searching tools are based on heuristics criteria but they neither can comprehend the demand of researchers, nor understand the nature of the links among texts. Obviously this is a restricted use of technologies, which repeats an already existing model, which has shown its limits during the past centuries. A model that led us to the so-called serial price crisis.

V. Conclusions and open problems

The HyperJournal Software has been initially founded by the Groupement de Recherche Europen(GDREplus) Hyper-Learning. Modles ouverts de recherche et denseignement sur internet which is a multidisciplinary research infrastructure promoted by the Centre National de la Recherche Scientifique (CNRS) regrouping 29 partners of 9 countries (universities and research centres, a large corporation (IBM), and three small enterprises)[11]. The software is currently being developed by both project members and volunteers. HyperJournal is distributed freely with an Open Source license. For these reason it is free to use and can be adapted to the exigencies of a large number of scholarly communities. A prototype of HyperJournal is expected in January 2005.

Nowadays, HyperJournal tackle the problem of selection at a federative level. Indeed each node of the federation acts as publisher and decides whether a contribution can be published or not, according to its own notion of quality. Nevertheless subjectiveness and variableness in the idea of quality and the possibility of collaborative annotations request a universal system. This system must be able to filter information according to trust criteria already existing within relationships among researchers. The selection of information happens through trust networks, by word of mouth, acquaintances of colleagues at congresses, trust in a publisher, reviews, or the institution the author belongs to; all those elements are evaluated when the researchers has to operate their selection. If the information is semantically structured on the Web, it is possible to repeat some of these processes or to introduce new ones.

The relations showed by the contextualization are the quotations contained in the articles approved by the peer review. As the number of journals adopting contextualization increase, the degree of quality control decreases. Everyone can start a journal and accept low quality contributions that quote contributions in other journals. The immediate consequence is that the quoted contributions are submerged by scarcely relevant contextual information, the so-called contextual noise. A possible solution could be to allow journal publishers to select trusted sources in order to draw the contextual information. So that every Journal could show the contextual information only drawing them from the trusted journal. Obviously this requires a notification mechanism through which publishers can receive announcement if a new journal has been started. Then they will be able to decide to include it into the trusted sources or not. Another mechanism to assign trust could be based on user decisions instead on publishers. So the users would select the sources they regard as trustworthy in order to draw contextual information. This second solution would surely raise a problem, that is the information surplus and then the need of a deeper process of selection. The authors should then constantly update their trustworthy sources list. This problem could be partially solved by exchanging the source list according to mutual trust among the researchers. If Alice trusts Bob, and bob updates his trusted-sources list with a new record, Alices list is automatically updated. In this case the peer review of each journal would play a filter role but a deeper filter is based on user mutual trust.

REFERENCES

[1] J.C. Guédon, *In Oldenburgs Long Shadow: Librarians, Research Scientists, Publishers, and the Control of Scientific Publishing*, Association of Research Libraries, Proceedings of the 138th Annual Meeting, 2001. http://www.arl.org/arl/proceedings/138/guedon.html

[2] T.H. Nelson, *Literary Machine 90.1* Sausalito, CA: Mindful Press, 1992

[3] *Los Alamos ArXiv*, http://www.arxiv.org ,2004

[4] P. Suber, *Open Access Overview* http://www.earlham.edu/ peters/fos/overview.htm ,2004

[5] *Budapest Open Access Initiative*, http://www.boai.org ,2004

[6] *The HyperJournal web site*, http://www.hjournal.org ,2004

[7] G. Tummarello, et al. , *RDFGrowth: a P2P annotation exchange algorithm for scalable Semantic Web applications*, in Proceedings of P2PKM 2004, Boston, 2004. http:/www.p2pkm.org/Camera_Ready/1568938872.pdf

[8] S. Hockey, *The Robert Busa Award Lecture 2004*, ALLC/ACH 2004, Goeteborg, 2004, video available at http://www.hum.gu.se/allcach2004/

[9] W. Piez, *Authoring Scholarly Articles: TEI or Not TEI?*, in proceedings of ALLC/ACH 2004, Goeteborg, 2004. http://www.hum.gu.se/allcach2004/AP/html/prop124.html

[10] F. Di Donato, *Verso uno European Citation Index for the Humanities. Che cosa possono fare i ricercatori per la comunicazione scientifica*, Bollettino Telematico di Filosofia Politica, 2004, http://bfp.sp.unipi.it/rete/ecih.html