Weaving the Web of Science
HyperJournal and the Impact of the Semantic Web on Scientific publishing

Michele Barbera[1], Francesca Di Donato[2]

[1]Net7 – Internet Open Solutions
Via Marche 8/a, 56125 Pisa, Italy
email: barbera@netseven.it
[2]Dipartimento di Scienze della Politica, Facoltà di Scienze Politiche, Università di Pisa
Via Serafini 3, 56125 Pisa, Italy
email: didonato@sp.unipi.it

Abstract

In this paper we present HyperJournal, an Open Source web application for publishing on-line Open Access scholarly journals. In the first part (sections 1-3) we briefly describe the project and the software. In sections 4 and 5, we discuss the weaknesses of the current publishing model and the benefits deriving from the adoption of Semantic Web technologies, outlining how the Semantic Web vision can help to overcome the inefficiencies of the current model. In the last two sections, we present two experimental applications, developed on top of HyperJournal, with the purpose of demonstrating how the technologies can affect the daily work of scholars. The first application is a tool for graphically visualizing the network of citations existing between articles and their authors, and for performing bibliometric measurements alternative to the ISI Impact Factor. The second is a tool for automatically extracting references from non-structured textual documents, which is part of a tool-chain for the extraction of hidden semantics.

Keywords: HyperJournal; Open Access; impact factor; Semantic Web; electronic journals

1 Introduction

The HyperJournal [1] project was born in June 2004 as a spin-off of GDRE+/Hyperlearning [2], and has been supported by the University of Pisa (Department of Political Sciences). The project received from these two organizations a total founding amount of 25.000 euro. The residual amount has been borne by Net7, a small Italian enterprise [3], which sponsors the project by providing hardware and human resources. In fact, the project core team is also composed by volunteers, which in August 2005 founded the HyperJournal Association [4].

The aim of the association is to promote the growth and spreading of the Open Access movement. Apart from the development of HyperJournal, which is distributed with an Open Source license, the association activities include:

- Open Access advocacy;
- training courses and consulting for private and public bodies;
- participation to the Open Access policy development debate and awareness.

Individuals and organizations can become members by paying a low annual fee. Additionally, the HyperJournal Association has another category of partners, called "enterprise commercial partners", who are small enterprises who give a substantial contribution in terms of labour to the development of HyperJournal and possess a strong know-how about the application. The enterprise commercial partners donate 30% of the HyperJournal related revenues to the HyperJournal Association. We look forward for them to become a major founding source for the development of HyperJournal, since relying entirely on public founding cannot guarantee the necessary financial and organizational stability. One month after its first stable release, in November 2005 HyperJournal has been installed by 5 Italian institutions (University of Pisa, CNR Pisa, CASPUR, Aidainformazioni, University of Messina), and more installations are ongoing.

2 The Software

HyperJournal is a web application meant to make the creation and management of Open Access Journals easy for everyone. Despite it includes many advanced and complex technologies, its focus is on the simplicity for the

user. We believe that most of the barriers to the spreading of the Open Access model, as well as the main barrier to entry many big publishers rely on to retain their monopolistic position on the editorial market, are an artificial complication of what a scholarly community needs to establish, manage and distribute their own Open Access Journal. Unfortunately, even within the Open Access community, some actors (software applications, policies) fall in the trap, with the negative side-effect of scaring away the keystone of the process, scholars. It is only by involving who produce knowledge, i.e. scholars, that the movement can succeed. In order to do that, we must be able to show them how the Open Access model, together with the new amazing possibilities offered by Semantic Web technologies [5], can practically impact on their everyday research work. Instead of describing the general features of HyperJournal, which are similar to other existing e-journal publishing systems, we will here focus on its distinctive features, the underlying ideas and research directions.

3 Weaknesses of the Current Publishing Model

The current Web was originated as a universal documentation system, and was conceived in order to connect different pieces of information, linked in a unique global network. In practice, this particular feature allowed the Web to grow and become the most widespread information system ever conceived. The web inventor explicitly stated that one of the major innovations introduced by the Web, the possibility to link everything to all, is exactly what scientists have done for centuries: "Tables of contents, indexes, bibliographies, and reference sections are hypertext links" [6]. Tim Berners-Lee anticipated the potential effect of links would have, while foreseeing that "suddenly scientists [could] escape from the sequential organization of each paper and bibliography, to pick and choose a path of references that served their own interest". Unfortunately, as of today, the possibility to reach and access relevant information, has not become reality. As each scholar knows, it is seldom possible to click on a reference and be transported to the full text of the cited work. This mainly depends on three factors:

- documents are confined in "closed" spaces, which are not connected to each other;
- the semantics of references is hidden, therefore it is often impossible to resolve a reference to the target it points to;
- only a small part of the existing literature is freely accessible on the Web.

Additionally, a reference found in an article, which was originally conceived as a way to point to a relevant literature, that is, as an instrument for the reader, holds today a richer and somewhat distorting meaning. Citations are the main ingredient for the calculation of the so-called impact factor, the metric by which the productivity of scholars -and consequently their career advancement- is measured. In an enlightening article [7], Jean Claude Guédon explains the weaknesses of this model, which is based on the printing culture.

Among these weaknesses, we will focus on the problem that the set of journals on which the impact factor is measured, is predefined and decided by actors that are external to the scientific community. Failing a new method to measure the impact and to assess the productivity of a scholar, the Open Access movement lacks a powerful incentive to attract publications; authors and publishers will simply refuse to embrace a new model if this means harming their careers. Both the inclusion of some well-known Open Access journals in this set of core journal, as well as convincing people that Open Access literature has an higher impact (which remains measured on closed journals and a few Open Access ones), are only palliative solutions. They do not tackle the problem to the core: that the impact factor system entails an objective notion of quality.

At the origins of Scientific Publishing, back in the 18th century, being the variable costs proportional to the number of printed copies, the system was conceived to enforce a single -or at least a less variable- notion of quality. Accordingly, in the current scientific publishing system, still grounded on the practices of printing, selection comes before publishing; until it passes a review performed by peers, content is unavailable to the public. Even more controversial, is the system through which a journal is selected for the inclusion in the set of core journals on which the IF is calculated. This further selection is kept outside the boundaries of the scholarly community, and it is decided by a commercial monopolistic vendor.

Otherwise, on the Web, selection comes after publishing. Everyone is free to publish without asking permission to anyone. This allows the reader to apply its own notion of quality, by choosing what to read [8]. It is obvious that the two visions -the traditional scientific publishing model and the Web publishing model- do not match. However, the model offered by the web is gaining momentum even within the scholarly community; while doing their daily research work, scholars search the Web for relevant literature, read the articles they found even if they are not published in journals (e.g. preprints), apply their own criteria to judge their quality, and eventually use them to create, explain and expose their own thoughts (selection *a posteriori*). Conversely, when their research is done, and scholars need to publish their results, not only the references they put in their articles

tend to be different to the sources they selected in the previous step but also, more importantly, scholars strive to publish their articles in core-journals. Clearly, there is a tension between scholars as readers and scholars as authors. The result is that core journals become used for career management and not as a mean to participate in the scientific debate, while the Web becomes an ever growing collection of unfiltered literature. This suggests that using the web as a medium for electronic publishing at its full potential allows (and requests) an important change of paradigm, especially in the concept of quality. Nowadays, the scholarly community is so much larger than in the 18th century and the architecture of the web so different than the one of the printing system, that the tools and models used before are no longer fitting.

4 Selection and Quality

One of the fundamental premises of the Web is not to enforce a single notion of quality, and this is the reason why selection must be made *a posteriori*. Doing the inverse (*a priori*) will simply lead to an information loss: suppose that a journal judges an article not worth of publication and then, after some time, maybe decades, another discovery suddenly makes the article relevant for the field; what happens is that the now relevant article is forever lost. The example is meant to show that the notion of quality varies and changes; it is affected by time, space and cultural factors. If we accept this fact, then it becomes unacceptable to make the selection *a priori*.

However, the model of *a posteriori* selection proposed by the architecture of the Web, has a big side effect: it makes very difficult to find what you are looking for. Then, is it possible to find the pearls hidden in a multitude of waste? We are back to the problem of selection, which is not different from the one we face for every type of information -not only scientific information- today published on the Web. The answer lies in the vision of the Semantic Web. Discussing the capabilities and the architecture of the Semantic Web is beyond the scope of this paper; suffices it to say that the Semantic Web allows machines to "understand" the meaning of the information published on the Web. This means that if we are able to explicitly publish the reviewing criteria used for evaluating a journal submission (either accepted or refused), then scholars can instruct their software agents to retrieve information according to their own selection criteria or by grouping selection criteria of other scholars they trust. Being this possible, journals will remain useful as a collection of articles selected by their board of reviewers' quality criteria, but the scholars will also have the possibility to select and group articles by their own criteria. A first step in this direction is to publish both refused and accepted articles, along with machine-readable information about the scientific committee of the journal that made the review and about the review itself. The more machine-readable metadata about an article, its authors, their institutions, the journal and its scientific committee is published along with an article, the easier it will be, for a software agent, to retrieve meaningful information.

If a reasonable amount of scholarly literature is published along with its machine-readable metadata, then it becomes possible for an agent to answer queries such as: "find me all the articles accepted for publication in the last 5 years, in journals whose scientific committee has at least one third of the members affiliated to an institution belonging to my trusted institutions list" or "find me all the articles written by any co-author of an author who published a paper in one of the journals that belong to one of my colleagues' trusted journal lists (you can find the list of my colleagues on my address book and their list of trusted journals on their homepages), please exclude John and members of his research group lists". More generally, an interesting consequence of being compliant to the standards of Semantic Web is that the metadata made available in RDFmay be easily collected by external services and re-used in unanticipated ways. This is perfectly in line with the spirit of both our project and the Open Access movement as a whole.

5 Benefits of RDF

During the last years, the Open Access movement made a significant step ahead, from a technical point of view, with the development and subsequent wide adoption of the OAI Protocol for Metadata Harvesting [9]. The OAI-PMH protocol is a medium for transporting metadata from data to service providers. Metadata is expressed in XML; the only convention is that the protocol mandates a minimum set of Dublin Core encoded metadata for a data provider to be considered compliant to the protocol. Although within the community there is a quite clear understanding about the semantics of the metadata transported via the protocol, the format chosen to represent them (XML), is not suitable to be easily understood by machines unaware of the domain and provenance of such data [10]. To overcome the limitations of XML, which is made to represent tree structures, the W3C consortium developed and recommends RDF for exchanging data in the Semantic Web. RDF, which can be serialized in many different and isomorphic formats, is a mark-up language whose peculiar characteristic is that is made to represent graphs (which are super-sets of trees). Along with RDF, there are two languages, RDFS and OWL, which can be used to encode machine-readable ontologies, that make the semantics of data expressed in RDF

explicit. An important feature of RDF is that two or more different RDF graphs can be merged and the result is another graph, so that the process can be repeated *ad libitum*. It means that different data, coming from two different data providers and expressed according to different schemes, can be merged with no need of normalizing the data to map the different schemes used by the two providers. This is possible because RDF encoded data follow machine readable ontologies, making semantics explicit, so that a machine is able to understand the meaning of a particular term and eventually match it with another term used in a different ontology. Once again, describing the properties of RDF, RDFS and OWL goes beyond the scope of this paper; suffice it to say that RDF is more suitable than any other format to represent bibliographic metadata simply because a graph is more expressive than a tree, and because data coming from different providers can be merged without the need of a previous one-to-one agreement on the semantics. The former point entails several consequences: it implies that we can add "properties of an element" (such as "article x has been republished in journal y") at any time, even if the use of that property had not been planned once the system was designed. With traditional relational databases this is impossible, or rather complex. This property has a fundamental impact, as it implies that we can keep enriching the semantics and the amount of metadata about a resource. Given the collaborative nature of the Web, the potential is immense.

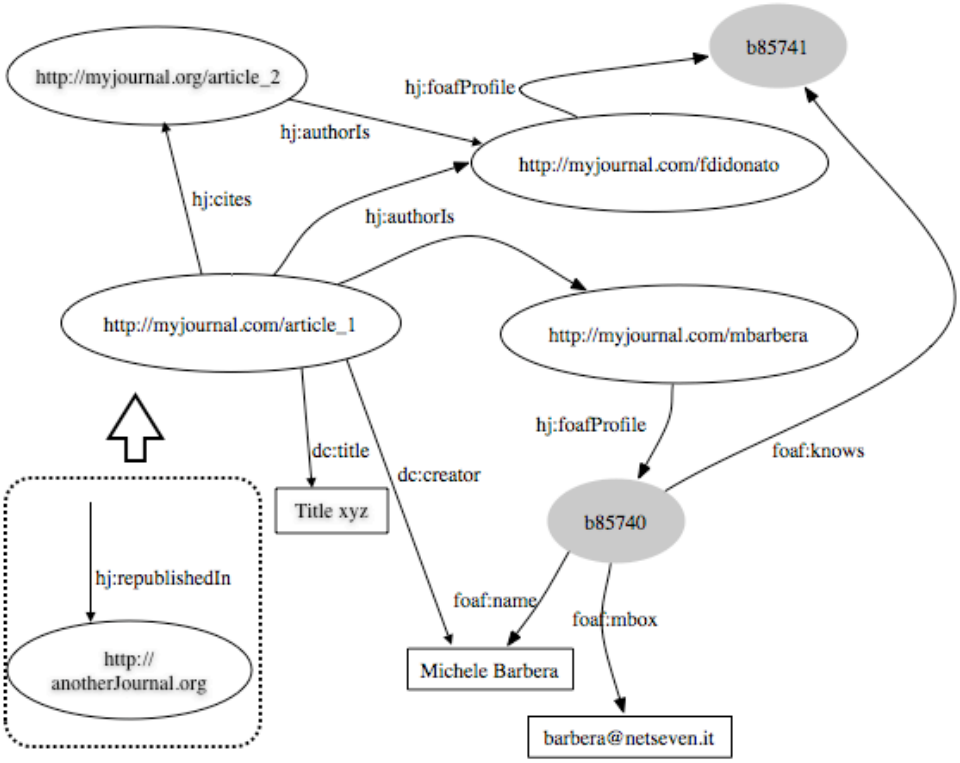

Figure 1: Merging two RDF graphs [the model has been simplified]

HyperJournal is currently the only open source journal publishing application that is based on RDF storage and uses RDF encoded metadata. Despite being compliant with the OAI-PMH specifications, HyperJournal also offers different and semantically richer means for expressing metadata. These extended semantic annotations are obtained by using RDF to encode metadata and ontologies (expressed in RDFS and OWL) such as FOAF, and DC. The employment of RDF instead of plain XML helps to overcome at least one order of problems: firstly, XML and its Schemes only enable to express loose constraints; this is the reason why service providers are often forced to treat the harvested data in a "data provider specific" fashion. Secondly, a conceptual encoding tree can correspond to many XML serializations; this means that, in spite of the degree of detail given in a Scheme, XML does not discourage ambiguous semantic descriptions. Although the efforts of the OAI community to increase interoperability between digital archives have been extremely fruitful, they presently face the risk of remaining bound into a "close" system. If they do not join the Semantic Web, the potential of applying trust metrics and inference rules on a virtually unlimited dataset (the whole Web, not only a specific community) will be lost. Hence, even if the usage of non-OAI compliant interfaces is not a good practice, testing different approaches could help to identify the limitations of the current OAI-PMH, especially those deriving from the usage of plain XML to encode metadata.

In HyperJournal the RDF metadata are not only collected for accepted and refused articles and for all the human actors that participate in the publication life cycle, but they are also made available for everyone, human or machine, who wants to use them. They are made available through a web service and can also be queried using SPARCL or SeRQL. At the time of writing, an interface for exposing RDF metadata via RSS is under development. Such an interface will follow the BuRST specifications, developed by Peter Mika [11]. The BuRST RSS 1.0 module is today one of the most interesting ways of disseminating semantic-rich bibliographic metadata. Another similar, though less complete, extremely interesting experiment has been done by Ed Crewe by extending the instance of OJS used by African Journals Online, to enable the delivering of RDF via OAI-PMH and PRISM RSS [9]. In the next section of this paper, we will examine two sample applications of the Semantic Web technologies embedded into HyperJournal. The aim of these two examples is to demonstrate the potential of using semantic-rich metadata collected from heterogeneous sources to build value-added services. It is worth noticing that, thanks to the use of RDF and Semantic Web toolkits, it has been possible to develop these samples (though unstable and incomplete) in a few days of work.

6 A Sample Application of SW Technologies: Going beyond ISI IF

The ISI IF is just a measure of impact but not necessarily the most representative. In fact, ISI IF suffers from two major weaknesses:

- It is calculated on a pre-defined set of journals and the decision on which journals to include is arbitrary;
- It is calculated at journal level, meaning that the impact factor of an article depends on the impact factor of the journal it is published in.

As suggested by [12], it is possible to compute different metrics of impact based on algorithms of Social Network Analysis. [12] demonstrated the significance of these metrics by applying them within the document repositories of LANL, however it seems to be extremely interesting to apply those metrics in a wider context whose boundaries are not defined *a priori*.

It is worth noticing that the applications discussed hereafter are not meant to be used in production systems, they are just experiments we developed quickly, in order to show how easy it is to retrieve, process and re-use RDF metadata.

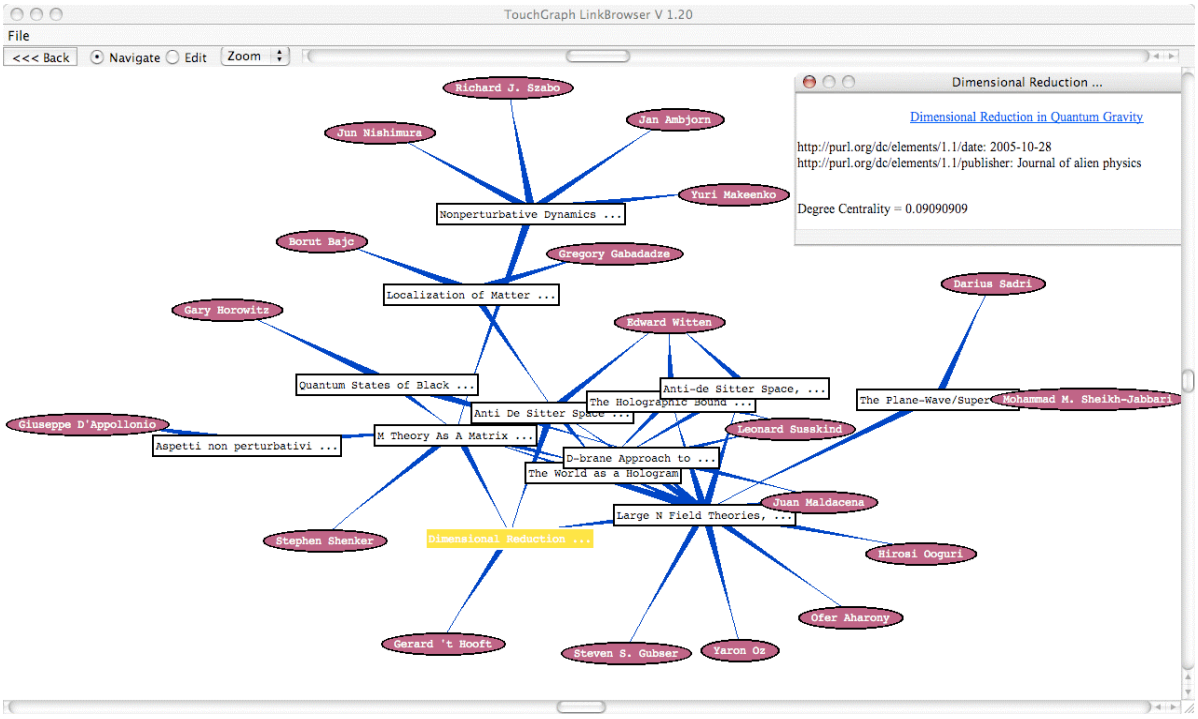

Figure 2: HJExplorer

The first application we will show is called HJExplorer. It a simple java stand-alone application that connects to

a set of HyperJournals, retrieve their RDF graphs, merge them in a unique local graph, and performs some calculations on it. In particular, it represents a network of articles interconnected through mutual citations; for each node on the network, HJExplorer calculates three measures of centrality -common in the research field of "Social Network Analysis"- based on the topological position of each node in the network. The centrality of a node in a network is a measure of the structural importance of the node. We consider, as suggested by [12], three measures of centrality:

Degree Centrality (normalized)

$$C_d(v_i) = \frac{\sum_{\substack{i=1 \\ i \neq j}}^{N} (x_{ij} + x_{ji})}{N-1}$$

Degree centrality indicates the number of links connecting adjacent nodes to a focal node. In other words, a focal node's centrality is calculated by summing the number of links directly connected to others. With asymmetric data, we can divide degree centrality into indegree centrality and outdgree centrality [13].
We normalize the result by dividing by the number of nodes in the network (N-1). Degree centrality is the measure, which is more similar to the ISI IF, although it is calculated directly on articles, not on journals (as in the ISI IF). Furthermore, the possibility to distinguish between indegree and outdegree leads to some interesting interpretations. Indegree can be interpreted as a measure of how much an article is directly cited, that can be a measure of its impact in respect to the network. An article with a high indegree is considered to be an authority. Conversely, outdegree is a measure of how much an article cites other articles; an article with a high outdegree can be considered a Hub.

Closeness Centrality (normalized)

$$C_c(v_i) = \frac{N-1}{\sum_{j=1}^{N} d(i,j)}$$

Where d(i,j) is the graph theoretic distance (length of shortest path) between nodes i and j.

Closeness centrality is the sum of geodesic distances between a node and all others [14]. Closeness is an inverse measure of centrality in that a larger value indicates a less central actor while a smaller value indicates a more central actor, thus we consider its inverse normalized by multiplying by N-1, where N is the number of nodes in the network. This measure indicates how far a node is from all others. A node with higher closeness score is less centralized than a node with lower closeness score. The interpretation of this measure in the context of bibliometrics is probably the most unclear of the three measures discussed here. However, it is possible that an article with a high closeness centrality can be seen as an article whose conclusions indirectly influenced many other writings. Thus, it can be considered as having a high impact, even if its indegree is not very high.

Betweenness Centrality (normalized)

$$C_b(v) = \left(\sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}\right) / (n-1)(n-2)$$

Where σst is the number of shortest geodesic paths from s to t, and $\sigma st(v)$ the number of shortest geodesic paths from s to t that pass through node v.

Betweenness centrality "measures the extent to which a particular point lies 'between' the various other points in a graph: a point of relatively low degree may play an important 'intermediary' role and so be very central to the network" [13]. An article with a high betweenness centrality score can be considered a bridge between two different disciplines or branches of a discipline, even if it receives few citations.

In addition to the use of the measurements shown above as alternatives to the ISI Impact Factor, it is important to remark that the HJExplorer and similar technologies can be used directly by scholars for their daily research work. The HJExplorer has a module to graphically visualize the network of citations between individual articles, as shown in the figure above.

Such tools can prove effective to visually explore the paths that led to the development of an idea or concept as well as to suggest the professional relations that incur within a group of authors. For example, it is easy to imagine the usefulness of such a tool while analyzing the history of science. With the help of such tools, "the inheritance from the master becomes not only his additions to the world's record, but for his disciples the entire scaffolding by which they were erected" [15].

7 Knowledge Representation and Extraction of Hidden Semantics

In a previous work [16], we discussed the potential benefits of adopting tools for extracting hidden semantics from unstructured textual sources and then enriching the published documents with these explicit semantics. Among these tools, the work presented in [17] is worth mentioning. Murray-Rust showed how tools for the extraction of hidden semantics can be used in the chemistry domain to identify bits of information whose semantics remains hidden in textual documents. Once identified, such a semantics can be made explicit and published along the text, so that it is possible to both rebuilding the raw data used to plot graphs and linking reference to molecules to online databases holding additional data about the molecule itself. The architecture of HyperJournal is designed in a way that permits a chain of filters to be applied sequentially over a textual document; filters can be plugged in at any time and each filter can be designed to extract information relevant to a specific domain.

In the experimental application described hereafter, called HJACI (HyperJournal Autonomous Citation Indexing) [18], we experiment with a particular type filter for the extraction of hidden semantics, which is not bound to a specific discipline: automatic identification of references from non-structured (e.g. PDF) text documents. This feature is highly desirable in an on-line journal application because it greatly simplifies the tedious task of creating metadata about the references contained in an article, which is a required effort in order to perform bibliometric analysis.

The task of automatically indexing citations is not new in the field of electronic publishing. There is a wide range of approaches and techniques used in a variety of different applications either stand-alone or embedded in publishing and archival software applications. However, the greatest part of these systems suffers from an important weakness: they are oriented to a particular discipline or field. Although some techniques are able to reach a high success rate [19-24], their efficacy is heavily reduced when applied to disciplines or languages that are different from the one they have been originally developed for. The distinguishing characteristic of the HJACI approach is to be natively not restricted to a particular discipline or language. Born as an autonomous citation indexing tool to be integrated into HyperJournal, whose intended use is not limited to a particular discipline or language, it cannot lack independence from language and discipline. HJACI relies on heuristics in order to extract potential citations from the texts and then match them with a knowledge base of bibliographic entries. At the time of writing, the implementation of the second part (matching) is currently under realization.
Once identified and matched, the citations found in an article will be stored in RDF using the model outlined in section 5 of this paper, so that it will be possible to include them in all the bibliometric measurements outlined in the previous section.

8 Conclusions

In this paper we discussed the possible benefits deriving from the use of Semantic Web technologies and concepts in the field of Open Access and electronic publishing. In particular, we discussed how the model of Scientific Publishing can evolve in order to benefit from the openness and universality of the Web. For this purpose, a deep revolution of both social and economic aspects of the current publishing model is required; we believe that a wide understanding and adoption of the Semantic Web vision can lead electronic publishing to a new era, where the Open Access model will play a major role. The HyperJournal project aims at representing a first step in this direction, for it can bring the new model to the broad public and can demonstrate the advantages of Open Access to all the actors involved in the life-cycle of a publication. As a conclusion, it is worth noticing that all the tools described in this article are released with an Open Source license, therefore can (and hopefully will) be adapted for other publishing systems, or used as a basis for derived and enhanced tools.

Acknowledgements

We would like to thank the core group of the HyperJournal project: Nicolò D'Ercole, Riccardo Giomi, Piergiorgio Pirro, Antonio Tolu. Without their contribution this work would not be possible. A special thank to Piergiorgio Pirro, who developed the HJACI tool in his graduation thesis.

References

[1] *HyperJournal website*. URL http://www.hjournal.org/

[2] *Hyperlearning*. URL http://www.hyperl.org/

[3] *Net7 – Internet Open Solutions website*. URL http://www.netseven.it/

[4] *HyperJournal Association website*. URL http://association.hjournal.org/

[5] *W3C Semantic Web Activity*. URL http://www.w3.org/2001/sw/

[6] BERNERS-LEE, T. *Weaving the Web. The original design and ultimate destiny of the World Wide Web by its inventor*. San Francisco : Harper, 1999. ISBN 0-06-251586-1.

[7] GUÉDON, J. C. In Oldenburg's Long Shadow: Librarians, Research Scientists, Publishers, and the Control of Scientific Publishing. *Arl Proceedings*, 2001, vol. 138. URL http://www.arl.org/arl/proceedings/138/guedon.html

[8] BERNERS-LEE, T. The World Wide Web - Past, Present and Future. Japan Prize 2002. *Commemorative lecture,* 2002. URL http://www.w3.org/2002/04/Japan/Lecture.html

[9] *The Open Archives Initiative Protocol for Metadata Harvesting*. URL http://www.openarchives.org/OAI/openarchivesprotocol.html

[10] CREWE, Edmund. Extending the Open Journals System OAI repository with RDF aggregation and querying [African Journals Online]. **In** *Proceedings 9th DELOS Network of Excellence thematic workshop*, *FORTH Crete*. 2005.

[11] MIKA, P. *Bibliography Management using RSS Technology (BuRST)*. URL http://www.cs.vu.nl/~pmika/research/burst/BuRST.html).

[12] BOLLEN, J.; Van de SOMPEL, H.; SMITH, J.; LUCE, R. *Towards alternative metrics of journal impact: a comparison of dowload and citation data. Information Processing & Management*. 2005. URL http://arxiv.org/abs/cs.DL/0503007

[13] SCOTT, J. P. *Social Network Analysis: A Handbook*. 2nd ed. : Sage Publications, 2000.

[14] WASSERMAN, S.; FAUST, K. *Social network analysis*. Cambridge : Cambridge University Press, 1994.

[15] BUSH, V. As we may think. *The Atlantic Monthly,* 1945, vol. 176, no. 1, p. 101-108. URL http://www.ps.uni-sb.de/~duchier/pub/vbush/vbush.shtml

[16] di DONATO, F. Designing a Semantic Web path to e-Science. **In** *Proceedings of SWAP 2005, the 2nd Italian Semantic Web Workshop, Trento, Italy, December 14-16, 2005*, *CEUR Workshop Proceedings*. Edit. by P. Bouquet, G. Tummarello. 2005. ISSN 1613-0073. URL http://ceur-ws.org/Vol-166/44.pdf

[17] MURRAY-RUST, P. Open Access and the Chemical Semantic Web. *ACS Spring Mtg*. 2005. URL http://wwmm.ch.cam.ac.uk/presentations/acs2005

[18] *HJACI*. URL http://www.hjournal.org/aci

[19] *Citeseer*. URL http://citeseer.ist.psu.edu/; OpCit, the open citation pro ject. URL http://opcit.eprints.org/.,

[20] *ParaTools*. URL http://paracite.eprints.org/developers/.,

[21] CONNAN, J.; OMLIN, C. Bibliography extraction with hidden markov models. 2000.

[22] BOLLACKER, K.; LAWRENCE, S.; GILES, C. Lee. CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications. In *Proceedings of the Second International Conference on Autonomous Agents*. Edit. by K. P. Sycara, M.Wooldridge. New York : ACM Press, 1998, p. 116–123.

[23] HUANG, I-Ane; HO, Jan-Ming; KAO, Hung-Yu; LIN, Shian-Hua. Extracting citation metadata from online publication lists using BLAST. **In** *PAKDD*. 2004, p. 539–548.

[24] BOLLACKER, K.; LAWRENCE, S.; GILES, C. Lee. Autonomous citation matching. In *Proceedings of the Third International Conference on Autonomous Agents*. Edit. by Oren Etzioni. New York : ACM Press, 1999.