Adopting Open Protocols to Increase the Impact on Digital Repositories

Ligia Eliana Setenareski, Marcos Sfair Sunye, and Walter Shima

Universidade Federal do Paraná, Curitiba, Brazil
{ligia,sunye,shima}@ufpr.br

Abstract. Recently the discussion of technological standards became important due to the profusion of new technologies arising from the development of microelectronics, computing, telecommunications, etc. Sometimes, one standard can be adopted or not, depending on the way the competition unfolds. As it is well known, it is not necessarily the best standard which becomes dominant, neither are the consumers who choose it, but the standard from the firm that used the most efficient market strategies. Sometimes a standard takes a long time to become dominant or will not be established because the competition process forces the manufacturers to permanent innovation. Considering the development of digital libraries, the development of diffusion and preservation systems has followed another course, not based on competition, but through the exploration of the symbiosis between Free and Open Source Software (FOSS), Open Archiving Initiative (OAI), and following the Google dominance. From this initiative, the cost of interoperability among digital libraries has greatly diminished. The NSDL (National Science, Technology, Engineering, Mathematics and Education Digital Library) defines three levels of cooperation needed to achieve interoperability. The technical level is related to the capacity of each digital library for sharing their metadata and enabling a unified search. The level of correlated content allows distinct repositories to describe their contents uniformly. The organizational level allows the sharing of management and governance of the repositories. This case study describes the open protocols adopted by UFPR in the construction of its digital repository. The digital library's files on logs are used to analyse the increase in the accessibility and visibility of scientific production in this institution.

Keywords: Digital Library, Accessibility, First Mover, FOSS, Open Archiving, OAI-PMH, network externality, technological patterns, lock-in.

1 Introduction

The objective of this paper is to discuss the emergence and development of technological standards concerning digital repositories and their impact on the visibility and accessibility of scientific content. In addition, concepts such as Network externalities reached by technological standards and open source software used in digital repositories will be discussed as well.

C. Stephanidis and M. Antona (Eds.): UAHCI/HCII 2013, Part III, LNCS 8011, pp. 408–416, 2013.

Recently the discussion of technological standards became important due to the profusion of new technologies arising from the development of microelectronics, computing, telecommunications, etc. Sometimes, one standard can be adopted or not, depending on the way the competition unfolds. In other words, the release of a new product by many manufacturers implies the commitment to the greatest number of suppliers and co-developers and a strategy to achieve the greatest numbers of consumers in order to create lock-in and path-dependence.

According to David [1], a path-dependent sequence of an economic change means that important though temporary and remote events may influence the final output. Such events may be defined by elements of luck instead of systematic forces, being full of uncertainty and immeasurability. Furthermore, historical accidents can not be ignored or isolated. In 1873, for example, the layout QWERTY was important for the sellers to impress the consumers when two words were typed fast: TYPE WRITE. It was also considered the best layout to avoid keys to be jammed - even not being so efficient as DVORAK or DHIATENSOR (David [1]). Those two accidental circumstances (not planned as a strategy to dominate a new industry) made the industry move forwards the adoption of QWERTY as universal. Simultaneously, typists were not trained to work on Dvorak because no employers used it..

The social benefit of switching standards may be greater than its social cost, but it is unattainable unless all users do it, It is too costly to coordinate all users' simultaneous switch. Thus, each random decision in favour of QWERTY would raise the probability that the next selector would favour QWERTY, which would result in decreasing costs of selection. Therefore, the type machine begun to belong to an interrelated complex system involving buyers, typists, training organizations for typists, manufacturers and suppliers of parts that used to grow spontaneously due to network externalities.

Network externalities are the benefit for the growing numbers of users who incorporate a technology that brings success to a network sponsoring firm. The users have a collective benefit from network because they exchange information and enjoy the future generation of improvements in such a technology. Other systems may be better, but they do not have the same diffusion and, consequently, do not have the same critical mass that enables incremental improvements to its development. In other words, newcomers do not have scale for development. The sponsoring firm's profits with growing number of buyers (users) increase and so does its network as long as its costs of development for improvements decrease. Specifically, in economic terms, the firm acquires economy of scale and its marginal cost tends to zero (Shapiro & Varian) [2].

Despite the historical circumstances that involved QWERTY standard, modern standards were established for new equipment and technologies due to efficient strategies of competition from a specific firm, which also depends on the users' technological expectations. The first movers have their own advantages, but the most important issue to consolidate the standard is the capability to find some element in or within the technology that persuades the market to presume that such a technology is going to be the standard. From that point on, a virtuous circle with an increasing support from the market begins. On the other hand, the opposite is also feasible, i.e. a vicious circle (Shapiro & Varian) [2].

Even with some key elements that determine the strategies, another important issue for the dominant standard firm is the architecture's opening or closing. At this stage, the important point is the technology valorisation and not the control per se, resulting in the increase in the number of users. This technology will probably be open when there is no capable firm to establish the standard and the complexity of the technology demands coordinated work between multiple developers and several new interfaces. At this moment, the earnings will be generated by network externalities, when more and more users and other developers connect to this open system and, consequently, create opportunities for more development with decreasing costs (Shapiro & Varian) [2]. The GSM open architecture for mobile communication became a dominant standard because more and more users and developers connected to it and the network value increased exactly because of the network externalities. Nowadays, the same situation happens to Android OS with disadvantages to proprietary systems as Symbian, RIM, Palm and also Apple. A proprietary technology in complex systems is very risky since the uncertainty does not encourage interface developers to work with specificity and restrictions.

The technologies for digital repositories did not arise from a private competition strategy, but from social movements by the academy against the enclosure of private editors of scientific journals. In 1991, Paul Ginsparg developed and implemented the arXiv [3] as a tool to enable the access to preprints[1] in physics. The objective was to facilitate and accelerate the access to high quality scientific content by the scientific community, who was also in charge of producing it. This initiative gave rise to a well known movement called Open Access whose main goal was to build open access networks for academic content, differently from the high cost of access to contents in private publishers' journals that tended more and more to cartelized behaviour. In 2001, this Movement established the OAI-PMH protocol (Open Archives Initiative – Protocol for metadata harvesting) [4] which created a low-barrier mechanism for repository interoperability, which rapidly enabled the adherence by the academy. In 2006, the main site OAI (OAIster) [5] had already more than 500 registered institutions that integrated its network. Nowadays, OAI-PMH and Dublin Core metadata are the main standard of the Open Access movement. Two other important initiatives that arose from the Open Access movement - the community FOSS (Free and Open Source Software) – were the software OAI-PMH protocol (Open Archives Initiative – Protocol for metadata harvesting) [4] which created a low-barrier mechanism for repository interoperability2 , which rapidly enabled the adherence by the academy. In 2006, the main site OAI (OAIster) [5] had already more than 500 registered institutions that integrated its network. Nowadays, OAI-PMH and Dublin Core metadata are the main standard of the Open Access movement. Two other important initiatives that arose from the Open Access movement - the community FOSS (Free and Open Source Software) – were the software Dspace [6] and the software OJS (Open Journal

[1] A scientific paper still to be published by a journal with peer review.

[2] Data Providers are repositories that expose structured metadata via OAI-PMH. Then Service Providers make OAI-PMH service requests to harvest that metadata. OAI-PMH is a set of six verbs or services that are invoked within HTTP (OAI-PMH) [4].

System) [7] that are currently the main systems for digital libraries and scientific journals' management respectively. They enable the sharing and management of scientific content and are responsible for the dissemination of open access digital repositories.

Therefore the scientific community, who developed interoperable standards, was never interested in competing with private editors or being a kind of potential entrant in order to expand its market share. Indeed, these standards worked as alternative solutions against the market monopolization by private editors. In other words, the editors' cartelized behaviour (supply) led the scientific community (demand) to create an alternative solution that was possible through the development of information and communication technologies. Such technologies caused the demand to adopt new behaviour against a restricted supply. The most important objective was not the establishment of path-dependence, positive network externalities, or a virtuous circle, but the possibility to divert from the restrictions imposed by private editors.

In 2004, an institutional digital repository was created at Universidade Federal do Paraná - UFPR. From that year on, the software Dspace has been used for theses, dissertations and videos, and the software OJS has been used for journals. UFPR was the first university in Brazil to deploy such software. This Chapter will show how the network externalities that came from the Open Access movement contributed towards the adoption of an Open Software by UFPR and how this choice was instrumental in increasing its visibility.

2 Open Protocols for the Digital Library Federation

The turn of the century saw the emergence of the organization of federations such as the Digital Library Federation that has worked with the standardization of metadata for digital libraries since 1996. This initiative was consolidated in 2001 with the emergence of the Open Archives Initiative (OAI) [4]. Two important standards were defined: the Dublin Core, which describes the minimum attributes that make up metadata in any digital library, and the PMH (Protocol for Metadata Harvesting) that standardizes the sharing of metadata.

The first goal of the standards promoted and developed by the Open Archives Initiative (OAI) was to provide accessibility to scholarship and scientific information produced by Academic Institutions. Therefore, this set of standards can be actually used by any institution that aims to manage its digital content. These standards are a formal foundation that allows the scientific community to build a network of digital libraries. An OAI-PMH federation is composed by two main agents: the Data Providers (DP) and the Service Providers (SP).

Data Providers are digital repositories that can be implemented by any system capable of inserting and recovering digital documents based on a single identifier. Besides, a Data Provider is expected to be capable of linking a document descriptor a.k.a. "metadata" with its correspondent document source. In the architecture of the OAI, the link between metadata and its original document is made by the insertion of a Permanent Uniform Resource Locator (PURL) in the metadata. This approach keeps the link between the original document (that is maintained in its original Digital Library) and the metadata, which can be distributed and copied by other systems.

Service Providers (SP) are systems that harvest metadata from Data Providers, grouping and indexing them in a database, which allows a unified search. The protocol PMH allows the communication between Service and Data Providers, as well as providing the synchronization between all Federations' metadata.

The OAI also proposes the standardization of the minimum description of attributes known as "Dublin Core" [9]. The Dublin Core standard has 32 attributes and Figure 1 shows their use at UFPR´s Digital Library.

As the only correspondence between metadata and its original document is given through a Permanent URL stored in the metadata, it is very common the adoption of a handle server. The function of a handle server is to translate dynamically the permanent address into its physical address. The Handle System [10] composes a major standard for Digital Object Identifier called DOI (www.doi.org) and it is used by the majority of Data Providers.

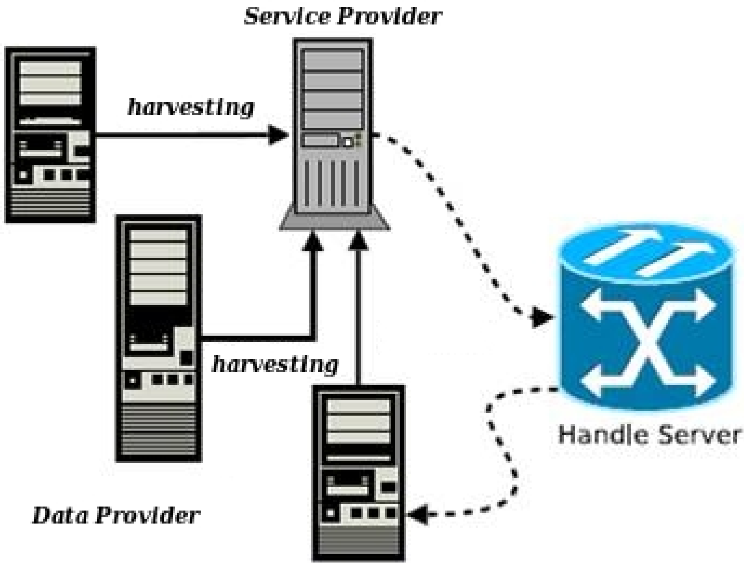

Fig. 1. OAI-PMH architecture and Handle System Server

2.1 Free Open Source Software (FOSS) for Digital Libraries

There is a great deal of Open Source Software that can perform the OAI protocols[8]: CDS-Invenio (Switzerland), DoKS (Belgium), DSpace (USA), EPrints (UK), FEDORA (USA), Greenstone (New Zealand), MyCoRe (Germany), OPUS (Germany), SciX (Slovenia). All of them use an Open Source Database Management System to store and search their metadata. The two most popular softwares for Digital

Libraries are Dspace that uses Postgresql and Eprints that uses MySql. The standard software for scientific journals is OJS (Open Journal System) that uses Postgresql.

As most Digital Libraries Systems run on Linux with an Apache server, the entire environment needed to maintain a Digital Library can be offered by the FOSS community.

3 The Adoption of Open Protocols by UFPR

The emergence of several open code systems for the management of digital libraries such as Dspace and Eprints as well as the implementation of these metadata searching systems (Google in 2008 started to index metadata through the standard OAI-PMH) have sustained the standard OAI-PMH, which has quickly become a reference. From this initiative, the cost of interoperability among digital libraries has greatly diminished. The NSDL (National Science, Technology, Engineering, Mathematics and Education Digital Library) defines three levels of cooperation needed to achieve interoperability. The technical level is related to the capacity of each digital library for sharing their metadata and enabling a unified search. The level of correlated content allows distinct repositories to describe their contents uniformly. The organizational level allows the sharing of management and governance of the repositories.

The Federal University of Paraná (UFPR) has adopted the standard OAI-PMH since 2004 when Dspace and OJS were implemented in its digital libraries, ensuring a technical interoperability of its digital collection. The insertion of a digital library for federations of theses and dissertations at UFPR, as it was the case of the server Oaister in 2005, has been ensuring the interoperability in sharing correlated content, allowing even the detection of duplicities of theses (when a researcher connected to the university defends his thesis in Europe and registers it in two collections). Finally, the construction of tools in order to integrate the physical and the Digital collection, started in 2008, has been allowing the backward scanning of works without increasing the cost of indexing at UFPR's digital library. .

The third level of interoperability, which includes the construction of agreements to enable the governance and the provision of coordinated services among various digital libraries, has been partially achieved by UFPR. Services such as the Handle System, which foresees a definitive and permanent identifier for each new digital object created, allows the repositories of UFPR to change its hardware and software without compromising any level of interoperability.

For instance, UFPR has changed its Library Management System (LMS) from Virtua [11] to Sophia [12] but the system remains interoperable with the Digital Library as the metadata of UFPR's Digital Library has been harvested from LMS through PMH.

3.1 UFPR's Choice of Open Protocols

The development of the digital library management software - from the symbiosis between the Open Access movement and the FOSS - rapidly expanded its positive network externalities with its dissemination in the academia, including university libraries. The university libraries, which are responsible for the preservation of printed scientific literature, were in search of suitable and affordable alternatives to store and manage content also in digital media. These positive network externalities were crucial for UFPR to build its digital repositories after analyzing available alternatives in the market. Thus, open and interoperable protocols started to be used to create a digital repository at UFPR taking the following aspects into account:

1. The increase in the visibility of scientific researches.
2. The publication of scientific production could be made by copyleft, with no costs of copyright.
3. The diffusion of knowledge produced at UFPR would be rapidly extended worldwide.
4. The absence or reduction of legal and institutional procedures in the adoption of open technologies in comparison to the vast number of procedures in the adoption of proprietary technologies.
5. A short estimated time to deployment.
6. A low estimated cost to deployment (whereas OJS and DSpace software are free and the only costs come from the hardware and the software customization).
7. The core competence inside the Department of Informatics at UFPR which was able to customize, deploy and train the first users of open technologies.

The purchase of the server, the software customization and the effective deployment of digital libraries for the first inclusion of digital objects took about a year to be finalized and cost approximately US $ 25,000.00. It is considered a short time and a relative low cost if compared to the time and cost of the computerization process of the UFPR´s library catalogs, which used proprietary software and took 7 years. This process demanded fundraising and specific bidding procedures - mandatory by the public sector in Brazil - for the acquisition of computer equipment and proprietary software, as well as for hiring the services of retrospective conversion of bibliographic records on paper (catalog cards) for the digital media. These steps along with the staff training had an approximate cost of US$ 750,000.00.

3.2 The UFPR's Digital Repository and Its Accessibility and Visibility

Due to the Open Access, the UFPR’s digital repositories were accessed by more than 150 countries between March 2012 and February 2013 - detailed access logs are available at [13]. CHART 1 describes the number of files downloaded per country last year. Accesses within the Brazilian territory were excluded, as well as the hits from the U.S., because they could be confused by Google’s Crawlers and, consequently would distort the number of hits to a higher level in these countries.

It is observed that the number of hits coming from Portuguese speaking countries is high. Besides being relatively high in numbers, hits from Portugal, Angola, Mozambique and Cape Verde have been continuous, revealing a constant habit of searching the UFPR’s library by these countries. It is also important to point out that language is not a barrier to access the UFPR’s digital repository. After Portugal, Germany, China, France, Netherlands, Spain and Great Britain, in spite of not being a Portuguese speaking country, present the highest numbers of accessed files. The Portuguese speaking country that has the largest number of accessed files after Portugal is Mozambique, which has the 7th position after all the countries with other languages.

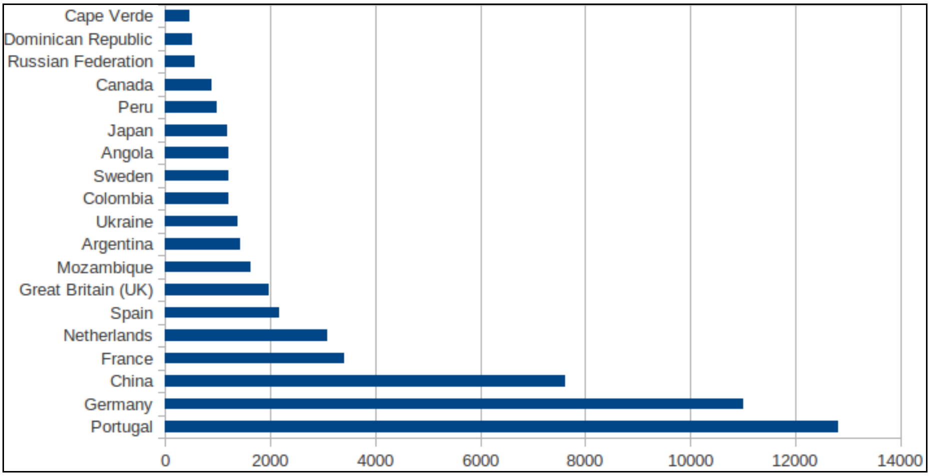

Chart 1. Number of accessed files per country, from March, 2012 to February, 2013

4 Conclusion

The main issue discussed in this paper is the level of accessibility to scientific content produced by the academia and not published by private editors, who more and more restrict the access by producers (researchers from universities who were expected to pay to publish their own papers) and to consumers (readers who were supposed to pay to access journals). Although OAI-PMH was a movement from the academia against the private editors' cartelized behaviour, it was not a strategy to establish a new technological standard as it can be seen in competitions when potential entrants establish new technological standards in emergent markets - for instance, the recent competition between Blu-ray DVD x HD-DVD). Therefore, the Open Access movement was not created in order to impose any kind of standard but it was a political and economic strategy that aimed a rapid diffusion of preprints and other scientific contents, which would never be accessed if they were in the hands of private editors. OAI-PMH rapidly expanded towards the universities, libraries and researchers and inevitably became one of the standard technologies for digital repositories. At the same time, FOSS also became an important interoperable standard in the Open Access movement.

The separation between digital documents and their metadata, as well as the free distribution of this metadata are the main features of the OAI-PMH approach. The simplicity of the protocol for metadata exchange quickly allowed large part of management software to implement digital libraries. Hence, most academic digital libraries became Data Providers, that is, an OAI federation.

Besides the free distribution of metadata that started with the separation of digital documents, the main feature of the OAI-PMH approach is the simplicity of the protocol for metadata exchange, which enabled a rapid deployment by most digital libraries.

Those technologies became the standard for digital repositories due to their open access and interoperability. They were marked by non-economic interests, and not meant to become proprietary technologies. Although the academic world had an interest in accessing and researching knowledge, it did not intend to appropriate technologies of diffusion as the private editors did. Thus, OAI-PMH became a standard with rapid positive network externalities as a virtuous circle, with an exponential growing number of digital repositories. This is the case of a positive path-dependence which brings about larger communities and high quality in the development of apps and solutions.

Due to the adoption of OAI-PMH standard in the digital repository at UFPR, there was a rapid and growing diffusion of its scientific content, not only among Portuguese speaking countries but also among countries of other languages.

UFPR has deployed more than 10,000 theses and dissertations as well as 38 scientific journals totalling nearly 20,000 titles since 2004. Undoubtedly, such figures would never be achieved without the adoption of open protocols, considering the characteristics, prices and enclosure from proprietary standards.

References

1. David, P.A.: Clio and The economics of Querty. American Economic Review 75, 332–337 (1985)
2. Shapiro, C., Varian, H.R.: A Economia da Informação: Como os Princípios Econômicos se Aplicam à Era da Internet. Elsevier, Rio de Janeiro (2003)
3. ArXiv, `http://arxiv.org/`
4. Open Archives Initiative Protocol for Metadata Harvesting, `http://www.openarchives.org/pmh`
5. The OAIster Database, `http://www.oclc.org/oaister`
6. Dspace, `http://www.dspace.org/`
7. Open Journal System, `http://pkp.sfu.ca/?q=ojs`
8. Madalli, D.P., Barve, S., Amin, S.: Digital Preservation in Open Source Digital Library Software. Journal of Academic Libraries 38, 161–164
9. Dublin Core Initiative, `http://dublincore.org`
10. Handles System Server, `http://www.handle.net`
11. Virtua VTLS, `http://www.vtls.com`
12. Sofia, Prima Informática, `http://www.primasoft.com.br`
13. Webalizer of UFPR digital library, `http://calvados.c3sl.ufpr.br/webalizer`