Enhancing the Sustainability of Electronic Access to ELPUB Proceedings: Means for Long-term Dissemination

Bob Martens¹; Peter Linde²; Robert Klinc³; Per Holmberg⁴

¹Vienna University of Technology, Karlsplatz 13, A-1040 Vienna, Austria
e-Mail: b.martens@tuwien.ac.at
², ⁴Blekinge Institute of Technology, SE-37179 Karlskrona, Sweden
e-mail: peter.linde@bth.se, per.holmberg@bth.se
³University of Ljubljana, FGG KGI, SI-1000 Ljubljana, Slovenia
e-mail: rklinc@itc.fgg.uni-lj.si

Abstract

ELPUB can look back on a track record of a steadily growing number of conference papers. From a long-term perspective, access to this body of knowledge is of great interest to the community. Beyond this, extended preoccupation with the collected scientific work in the area of digital publishing has to be mentioned. Naturally, the authors are particularly focussed on the individual paper itself and possible connections with related efforts. Typically, conferences amplify and enhance opportunities of “getting-together”. A well-stocked repository may, however, serve in this respect as a fruitful complementary addition. In this contribution, the implementation of persistent identifiers on the existing ELPUB.scix.net-base is elaborated in detail. Furthermore, the authors present the result of efforts related to the harvesting of ELPUB-metadata and to the creation of a citation index. The paper concludes with an outlook on future plans.

Keywords: Digital preservation; shared information; self-organization; information retrieval; persistent identifiers

1. Introduction

After more than a decade of conferencing, the “threshold” of 500 published entries will be passed on the occasion of ELPUB 2008. There has been a long and persistent discussion on whether to have paper-based proceedings or not. The limited number of hard-copies of proceedings might cause difficulties, whereas parallel electronic release secures wider (and easier) dissemination. However, it can be regarded as an irony that 2003 was the first year in which all published papers were made electronically available in a repository (ELPUB.scix.net; [1]), which has since been extended consistently. In 2006, the repository was extended, i.e. once the set of proceedings was on its way for printing, it was made available on the web as a pre-publishing alternative in good time before the conference itself, accompanied by direct links in the conference schedule.

It has to be noted that the handling of ELPUB.scix.net is intentionally done on a shoestring budget to ensure archiving and collectioning in the long term. Furthermore, members of the ELPUB-community became intrigued by the wealth of stored information and focused on content analysis (see the work from S. Costa et al. [2,3]) in their work. At the start of ELPUB.scix.net, the policy of “Limited Open Access” was chosen, i.e. users had to register at no cost in order to have access to all stored PDF-papers. Open Access received a growing amount of dedication and it was decided to convert the repository accordingly. Supporting measures caused a minimal amount of programming efforts and led to registrations at www.openarchives.org, www.opendoar.org and roar.eprints.org. The paper will start out by exploring
opportunities for harvesting metadata.

The core part of this contribution focuses on matters of sustainability and investigates/describes the procedure towards the implementation of persistent identifiers for the ELPUB-repository. There are several schemes that apply to the concept of persistent identifiers - DOI, PURLs etc. But which scheme would be appropriate for ELPUB? The authors provide an analysis concerning the practical and technical experience of introducing URN-NBN as a persistent identifier scheme for the ELPUB repository.

A citation index has been made available in the ELPUB-repository, and the authors explore the situation at the time of writing. In the future, this option will support researchers with a view to impact analysis. The contribution concludes with an outlook on planned developments.

Figure 1: Screenshot Open Archives – ELPUB record

2. Harvesting ELPUB.scix.net metadata

On the occasion of ELPUB 2007 the decision was taken to switch to an Open Access mode. Given the theme of the conference (“Openness in Digital Publishing”), this did hardly come as a surprise. Participants provided the managers of the ELPUB-repository with valuable information and hints. Certain programming efforts had to be made in order to adapt the existing OAI-interface. Shortly after the conference, the authors registered the ELPUB Digital Library at www.openarchives.org, www.opendoar.org and roar.eprints.org was successfully concluded.

The Open Archives Initiative - OAI - develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. OAI has its roots in the open-access and institutional-repository movements. Continued support of this work remains a cornerstone of the Open Archives program. Over time, however, the work of OAI has expanded to promote broad access to digital resources for eScholarship, eLearning, and eScience.

OpenDOAR delivers a quality-controlled list of repositories, as submitted applications are checked by the project staff who visit the repository sites. The focus lies on academic repository contents. The listing shows an impressive growth, with currently over 1100 entries, and can be regarded as a key resource for the Open Access community. The services offered - such as text-mining - are continuously developed.

ROAR, an acronym for ”Registry of Open Access Repositories” promotes open access to the pre- and post-peer-review of research literature through author self-archiving. The registry monitors the overall growth in the number of eprint archives and also maintains a list of GNU EPrints sites.

It has to be noted that the entire repository (482 recorded papers – without ELPUB 2008) was made available to the general audience, which certainly boosted dissemination. By creating a personal login within ELPUB.scix.net, users can access the advanced search features and also store personal favourites.
etc. There are currently more than 500 registered-user accounts. Building on the developments of the previous year, the 2008 conference papers will be put online prior to the conference as soon as they become available and receive a direct link in the conference schedule. The work effort is manageable for the programme chair, who will derive metadata from the proceedings along with the full texts. If the idea of digital publishing is to be taken seriously, there should be hardly any delay in dissemination and, ideally, conferences attendees can seek timely advance information. In the end it is just a matter of handling and processing already created digital materials. Further options for harvesting may be explored as well and any proposal is welcome.

Figure 2: Screenshot OpenDOAR – ELPUB listing

Figure 3: Screenshot ROAR.eprints – Overview of the harvested ELPUB Digital Library
While the DBLP server initially focused on Database systems and Logic Programming, the scope of interest was gradually widened and could be interpreted as Digital Bibliography & Library Project. The system provides bibliographic information on major computer science journals and proceedings. As of April 2008, more than a million articles and several thousand links to home pages of computer scientists were indexed. The co-author index (see insert in figure 4) is an interesting feature. In the meantime ELPUB-paper have also been recorded in the DBLP-database.

3. Persistent Identifiers

The tedious “404 not found” message is a common experience for Web users. The problem with Uniform Resource Locators (URLs) is that they contain no information about their lasting reliability. Everything depends on the intentions of URL creators, and their level of commitment is hardly ever clear to the user. Hence, users cannot trust URLs.

The concept of URLs was developed for the first Web software in 1991. URLs are one kind of Uniform Resource Identifiers (URIs) and only serve for locating resources. From a preservation viewpoint, the weakness of URLs is the fact that they depend upon Domain Name System (DNS) information. This information is obtained locally, and passes through a national DNS registrar up to DNS root servers. It has to be noted that domain names are not stable and may change at any time. Furthermore, URLs depend on the fact that the local internet server, which is identified through the DNS, actually stores the object according to the information in the URL string.

In 1994, the idea was born to create a system for naming resources instead of addressing them, and the syntax of Uniform Resource Names (URNs) was formally described in 1997. The purpose was to provide a unique, Persistent Identifier in order to access the resource itself. The URN syntax uses a Namespace Identifier (NID) from a defined list maintained by the Internet Assigned Numbers Authority (IANA). URNs are designed to fit the definition of URIs.
Figure 5: Screenshot DBLP – Recorded ELPUB conferences and author’s details (insert)

A URN starts with a “urn” character string followed by the NID and, separated by a single colon, the last part of the URN which consists of a Namespace Specific String (NSS):

urn:nid:nss

Well-known standard concepts like ISBN, ISSN and MPEG can be found among namespaces on the IANA list. NBN, which stands for National Bibliography Numbers, is another namespace which is exclusively assigned to National Libraries. The Library of Congress acts as global registry for URN:NBN namespaces. The general syntax of the NBN-URN is as follows [5]:

urn:nbn:<ISO 3166 country code>:<assigned NBN string>

urn:nbn:<ISO 3166 country code>:<sub-namespace code>:<assigned NBN string>

NBN is a generic name referring to a group of identifier systems utilised by the national libraries, and only by them, for identifying deposited publications which lack an identifier. Each National Library decides independently to whom they will issue sub-namespaces [4]. It is quite common that Swedish university
libraries manage institutional repositories that include digital versions of doctoral and licentiate theses. In some cases these theses do not have any paper equivalent at all. It is reasonable to assume that in the future a growing number of theses will solely be made available in a digital format. Hence the wish – both from the Swedish National Library and from the data providers at the universities – that links to object sources remain persistent. The Swedish National Library (e.g. The Royal Library in Stockholm) is responsible for preserving all material printed in Sweden. However, collecting digital resources (e-books etc) requires an infrastructure different from the one used for traditional print materials. Adequate technical solutions have been developed at the National Library to meet these challenges and needs. The aim is to further develop technical solutions and to find practical solutions to the problems involved in administering legal deposits for digital material. One important task is to find out how to acquire metadata together with digital deposits. Swedish legal deposit legislation does not yet cover online publications, although the government has announced that a new Act which includes them will come in 2010.

One step in the preparation for legal deposits of digital publications is the implementation of URN-NBN. Since 2004, Swedish publishers, libraries etc. have been able to use the urn-nbn scheme and the national library resolver service in order to secure persistent links to digital resources.

For some time now there have been considerations about implementing some sort of persistent identifiers, both for the ELPUB repository and for the institutional repository of Blekinge Institute of Technology (BTH). The main reason for finding a preservation scheme was the ambition to keep the unique repository of conference papers, in ELPUB’s case, and theses, in BTH’s case, available over a long period of time. The two repositories also had another feature in common: a shoestring budget and a lack of administration resources. Hence, the major prerequisites for a persistent identifier scheme were the following: Persistence at no or at least low cost, involving as little technical and administrative effort as possible. With these requirements in mind, the horizon was scanned for plausible persistent identification schemas.

The Handle System was developed by the Corporation for National Research Initiatives (CNRI) within the framework of the Computer Science Technical Reports (CSTR) project. A first implementation of the system was made available in 1994. Participation in the Handle System requires registration along with the establishment of a Naming Authority. CNRI makes free software available, but requires the signing of a licence when registering a new Naming Authority. A small registration fee is charged.

In 1997, the Digital Object Identifier (DOI) initiative, based on the Handle concept, was launched at the Frankfurt Book Fair. Today, many scientific journal publishers are using the DOI. The International DOI Foundation (IDF) was founded in 1998 [11]. Although DOI was created for the publishing industry, it is used just as much outside that sector, e.g. by electronic commerce applications. The usage of DOI is not restricted to digital objects. One of the aims is to provide identification mechanisms for commercial transactions concerning rights management [6].

At the early stages of the DOI system, a central DOI depository directory was used with IDF as the only administrator or registration agency (RA). Today there are several RAs that will assign prefixes to new registrants in accordance with IDF standards. The RAs are free to set their own rates for registration costs. The IDF has fixed rates for different levels of participation [7].

One of these RAs is CrossRef, the reference-linking network of the Publishers International Linking Association (PILA). They have a preference for what is called “original work” and are rather cautious about how to define that. They often consider Open Access materials as insufficiently defined, and CrossRef only accepts material which is not published from institutional repositories. Furthermore, DOIs are not granted the right to pre-print or post-print materials [8].

Another persistent identifier schema based on the Handle system is incorporated in the DSpace software package. For persistent identification and linking, DSpace uses HDL Identifier and Resolution Services
The source code for DSpace is distributed with an API for CNRI’s Handle Server. Users of DSpace intending to use the handle-server announce this to CNRI and are then supplied with a handle-prefix against a yearly cost of approx. 50 USD.

The Persistent URL (PURL) is based on the URN specification and is developed by the Online Computer Library Center (OCLC) [9]. As one would assume, a PURL is a URL with some degree of modification. A PURL turns into a URL for the user after being associated with the correct URL at an intermediate resolution service using HTTP redirect. The PURL server software is developed by OCLC and can be freely downloaded. It contains group management for maintainers, tools to administrate PURLs, etc.

After a brief period of reviewing the system solutions described above, the authors decided to proceed with URN-NBN. The decision was mainly based on the lack of funding and an aversity to running administrative software. DOI involves negotiations with commercial entrepreneurs suspicious of institutional repositories and costs that are either to high initially or to unpredictable for the future for small-scale services. While alternatives such as PURLs are more or less free of cost, they demand installing new software and administrating the whole process of assigning identifiers to objects, which is not desirable in an organisation slimmed to the bone or relying on non-paid work. The URN-NBN scheme could be used by ELPUB.scix.net and the BTH repository at no cost and, on top of this, with only a minimal amount of administration.

In order to implement URN-NBN a contact with the National Library of Sweden was settled, which delivered instructions on how to proceed. This is work in progress and not all printed material and instructions for how to implement URN-NBN for customers are available at the library website yet. However, the staff of the Digital Library Department was very supportive whenever questions concerning implementation arose.

As for DOI, PURL etc., the general concept of URN-NBN is to assign a permanent identity to a digital resource (an article or thesis for example). A “resolver” translates the permanent identifier to the actual URL where the resource can be found at the moment. In order for the “resolver” to function it needs to be fed with the proper mapping instructions; in other words “Which information/data matches which URL?” This was arranged by creating a mapping file content of which is harvested on a regular basis and fed to the “resolver”.

Since the institutional repository at BTH [10] is based on the Lotus Domino platform, a Lotus application was constructed that generates an XML-file which maps the URN-NBN identifiers to the URLs. When generating the XML-file, the interface looks up all published theses and journal articles in the BTH-repository. There are two addresses where the BTH URN-NBN mapping can be harvested. One “cached” version exists, which is updated once a week:

http://www.bth.se/fou/forskinfo.nsf/urn-nbn-mapping/records/$file/records.xml

There also exists an option to generate the XML-file on the fly. The advantage of this option is, of course, that up to date content is received without delay. The downside is that the process is time-consuming and constitutes a burden on the server capacity:

http://saxofon.bth.se/fou/urn-nbn.nsf/on-the-fly

The BTH repository application was adapted in order to show URN-NBN identifiers in clear on the web for future reference purposes. Similarly identifiers can be retrieved at ELPUB.scix.net in the field <urn:nbn>. As indicated previously, every record in an URN-NBN scheme must have a unique identifier. In the BTH repository every record is assigned a unique serial number that can never be changed. These serial numbers, automatically generated by Lotus, were used for the NBN string after the sub-namespace. An example of a BTH URN-NBN using the “Urn:nbn:ISO country code:sub-namespace- nbn string” version
of the syntax is given below:

\texttt{urn:nbn:se:bth-da5e6bd8f0ee409ac1257424004d33c4}

It goes without saying that if, for whatever reason, the repository records were to change database to a new platform it would be necessary to migrate the Lotus serial numbers to the proper records in the new database. It took roughly one day’s work implementing the above schema. So far the implementation is running automatically without causing any noteworthy problems.

For resolving BTH or ELPUB URN-NBNs it is possible to use the Swedish Royal Library resolution service. Upon entering the persistent identifier generated at the repository level, one will receive the correct URL for the object. The service is available at:

http://urn.kb.se/start

Example: \texttt{urn:nbn:se:bth-e2b0aad6ae4c30f8c12573d8003b5b62}

returns: http://www.bth.se/fou/forskinfo.nsf/0/e2b0aad6ae4c30f8c12573d8003b5b62

It is also possible to use the demonstration resolver of the Deutsche Nationalbibliothek. This resolver is configured to redirect to external resolution services by analyzing the hierarchical parts of a URN (e.g. “urn:nbn:de” or “urn:nbn:se”):

http://nbn-resolving.de/ResolverDemo-eng.php

To give an example: An entry with the ID \texttt{urn:nbn:se:elpub-136_elpub2007} will be translated to

http://elpub.scix.net/cgi-bin/works/Show?_id=136_elpub2007

### 6. Citation Index

An index with references stemming from the conference papers was recently added to the ELPUB Digital library. While the idea as such is not new, the addition can be regarded as a way of making the context of digital publishing more tangible.

The procedure itself is fairly straightforward. First, the metadata has to be extracted from the PDF-documents and subsequently it is split into four parts (authors, title, source, year). This still requires manual work, and a higher degree of automated extraction would be desirable. Finally, the metadata is added to the existing records and may be viewed in both directions. The reference at the end of the record are leading to the citation index; From the citation database a direct link is offered to the recorded paper, where the reference is in place. Over 1,600 citations stemming from nearly 300 recorded entries (period: 2003-2007 conference papers) are available at the time of writing. Entries before 2003 are also likely to be processed in the near future. Not all the proceedings are, however, available in a machine readable format and their handling will depend on the disposable amount of volunteering resources. The citation index can be found at http://ELPUB.scix.net/cgi-bin/refs

This sample would allow for an analysis of the networks between the references, including the highly interesting phenomenon of \textit{self-citation}.

### 7. Future Plans and Outlook

With approx. 500 papers a critical mass is available. Collecting and archiving is to be regarded as a pragmatical issue and would, for example, serve as a starting point for ontology-based research. The efforts required are not limited to a certain group of “core activists”. Any member of the ELPUB-community is invited to use the metadata for harvesting issues.
Figure 7: Increase of the number of registered users with a trend line – clear decline after the shift towards open access

The shoestring-budget-idea as such has been mentioned already and constitutes a core principle. It aims at securing the ELPUB digital library in the longer term without making it dependent on a heavy input of human resources. The principle of self-organization should also be highlighted. In this respect, the procedure
related to the conference organization itself and the overlap of programme and general chair (year \(n=\) programme chair; year \(n+1=\) general chair) has been worked out well.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Title of the paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Openness in higher education: Open source, open standards, open access</td>
</tr>
<tr>
<td>2</td>
<td>DCMI-Tools: Ontologies for Digital Application Description</td>
</tr>
<tr>
<td>3</td>
<td>The Deconstructed Journal revisited – a review of development</td>
</tr>
<tr>
<td>4</td>
<td>Music Publishing</td>
</tr>
<tr>
<td>5</td>
<td>The Newer, the Worse: the Status of Farsi Word Processing Softwares in Iran</td>
</tr>
<tr>
<td>6</td>
<td>Scientific publication life-cycle model (SPLC)</td>
</tr>
<tr>
<td>7</td>
<td>Scholarly Communication in Transition: Evidence for the Rise of a Two-Tier System</td>
</tr>
<tr>
<td>8</td>
<td>Convergence and divergence in media: different perspectives</td>
</tr>
<tr>
<td>9</td>
<td>Peer-to-Peer Networks as a Distribution and Publishing Model</td>
</tr>
<tr>
<td>10</td>
<td>From print to web to e-paper: the challenge of designing the e-newspaper</td>
</tr>
</tbody>
</table>

**Table 1: Top 10 most downloaded full text papers from the ELPUB digital library**

The decision to support Open Access had an impact on the number of users registering to ELPUB.scix.net. Figure 7 clearly shows, that since the availability of this type of access which requires no user account to access full papers, the number of registered users remained more or less constant. Even more, since this change, users no longer log into the repository and it is almost impossible to identify groups of core users. It is rather difficult to make a clear distinction between active and inactive users.

On the other hand, the number of visitors coming from the different search engines is increasing and the number of visits to the collection of full text files is increasing. Table 1 shows the “Top Ten” papers that were accessed or downloaded from the ELPUB Digital Library. The paper in first place was downloaded more than 3000 times.

There is no doubt that the step towards “Open Access”, which was fully realised after the ELPUB 2007 conference.
conference, leads to enhanced user access. It is interesting to see is that the number of visitors coming from general (such as Google, AOL, Yahoo, Windows Live etc.) and specific purpose search engines (such as Google Scholar, Wikipedia, Open Archives, etc.) is growing and proves that the shift toward open access was worth the effort.

Finally, it has to be mentioned that some of the previous ELPUB-proceedings are already indexed in the ISI Web of Knowledge and those in charge will consider further steps towards having all the conference articles indexed there and in other important indexes such as Scopus and Google Scholar (making records visible in, for example, Herzing’s Publish or Perish citation analysis software program).

8. Notes and References


