This is a repository copy of *50 Years of Experience in Making Grey Literature Available: Matching the Expectations of the Particle Physics Community* .

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/75219/

Conference or Workshop Item:

O'Dell, Carmen, Dallman, David and Vigen, Jens (2008) 50 Years of Experience in Making Grey Literature Available: Matching the Expectations of the Particle Physics Community. In: Grey Lit 2008 (GL5), February 2008, Amsterdam.

50 Years of Experience in Making Grey Literature Available: Matching the Expectations of the Particle Physics Community

Carmen O'Dell, David Dallman, Martin Vesely, and Jens Vigen.
CERN, European Laboratory for Particle Physics, Geneva

Abstract:

The CERN Scientific Information Service has been active in the field of digital library research and in providing scientific information services to the high-energy physics community for almost five decades now. Most recently the research focus has been on interoperability issues in document storage and retrieval systems, metadata added-value services, digital library automation and networked information services. The achievements of this research and the implications for treating grey literature are presented, including practical implementation examples.

Introduction

CERN, the European Organization for Nuclear Research, was founded in 1954 and is the world's largest particle physics laboratory. The Member States financing CERN are all European but the laboratory has researchers from all over the world. In fact, there are some 6500 CERN users, representing 500 universities and over 80 nationalities. CERN itself employs around 2500 staff to provide the necessary

technical, computing and administrative infrastructure. The CERN Scientific Information Service was created in 1955 and one of its key tasks is to:

"acquire and manage information resources in all fields of relevance to the Organization, and make these accessible in the most convenient way to the worldwide particle physics community" [1].   To do this effectively the library has continually adapted and evolved its methods for collecting, treating and disseminating information resources.

1.   The Preprint Culture

In all fields of scientific research, there is a need to be the 'first' but research into the tiniest of particles ironically requires some of the biggest machines in the world, with price tags to match. When research is this expensive, there is simply no room to do the same science twice.   Over 40 years ago this drove the particle physics community into a culture based on preprints, to accelerate scientific communication beyond that of the lengthy journal publication process. Universities and laboratories produced their own series of preprints describing their research and experimental results, and distributed them free of charge to hundreds of other institutes in the field.

In 1958 the CERN library started issuing a weekly list of the preprints received and by the early 1960s, semi-automatic means were being used to produce it. Eventually each document was catalogued and the bibliographic data typed into a

database from which the list could be produced. In 1983 this database was made available to users for searching.

By 1991, around 10,000 paper preprints were being catalogued per year with publication references to the corresponding journal articles also being added to the database by hand. The system was absolutely at the limit of what could be achieved with the two staff members available. In addition, there were problems of space for storing all these paper documents.

Fortuitously, in August 1991, Paul Ginsparg set up the Los Alamos electronic preprint server, now known as arXiv. The server provided an enormous improvement in the speed and ease of preprint communication. By 1992, it was starting to become so popular among particle theorists that it was not long before some institutes announced they were stopping the costly distribution of paper preprints.

CERN librarians realized that while they would benefit greatly from having less preprints to handle, the e-prints still needed to be integrated into the library collection in a meaningful way. A program was written to read the daily email alerts from arXiv, decipher the bibliographic data and create database records with links to the full text documents on the preprint server. Although the terms had not yet been invented, this was probably the first example of a program to harvest metadata [2] .

Manual input of paper preprints continued and most were scanned so that the library could offer an almost complete electronic library. The number needing to be treated in this way though decreased steadily as arXiv grew in popularity.

2. Digital Collection Building

The library has used two different approaches to expand its digital collections beyond the data available at arXiv. One is library-driven and known today as metadata harvesting - other sources of relevant electronic documents are located and automatically uploaded to the bibliographic database. Once established, this takes place as long as the source continues to exist. The other is author driven with documents individually submitted to us via a Web form.

In both cases, if the fulltext of the document is available at the given source (not always the case, as we also harvest from databases/aggregators), our system automatically goes to retrieve it once the bibliographic record has been created. The digital document is then permanently stored on our server. When we consider that the source is at least as stable as ourselves, we may leave the full document untouched and provide an external link in our database. Clearly, this harvesting of digital documents and metadata only applies to material which is freely available.

2.1 Metadata Harvesting

We wrote our first harvesting program in 1993 and since January 1994 it has been running on a daily basis. Originally we only harvested from arXiv, which still represents about 50% of all our documents. From 1998 onwards though, we steadily added new sources until the current number is over 100. Initially a new script was written for each source but when we had several such scripts, we saw that we were heading towards a maintainability crisis.

Since large parts of these programs were doing very similar tasks, we developed a program called the Uploader which can handle a wide variety of external data formats. The part specific to the individual source format has a simple low-level language in which the external format can be described. Writing a so-called configuration file for a particular external source can then be done by a librarian, without programming knowledge being needed. The types of external data that can be handled range from highly-structured bibliographic database records to essentially unstructured HTML pages [3].

Before records are added to the database they pass through a simple matching procedure using terms from the title and authors. If a match is found, any new bibliographic fields are added to the already-existing record. Otherwise a new record is created. This reduces doublet generation quite considerably.

The frequency of harvesting is very varied, and is largely determined by the service offered at the other end. If a profile can be established, one can harvest data according to a daily or monthly schedule. If there are just Web pages listing documents, one has to find a way to signal the occurrence of new documents.

Most of the sources from which we harvest are freely available, for others we have made agreements about what we are allowed to upload. For example, INSPEC (one of the major databases in the physics field), allows us to upload all records connected with work at CERN.

Use of standards would be an enormous help. We are heavily involved with the OAI (Open Archive Initiative), having been host to the OAI European workshops in both 2001 and 2002 and have been one of the European representatives on the OAI technical committee. Our own system is OAI compatible [4], but even if this was true of the external sources, the protocol has not yet been exploited to such an extent that we could download data in the complexity we need.

2.2 Submission Procedure

Originally designed for capturing in-house CERN documents, this procedure was started in 1996 and has grown in complexity ever since to include more and more of the documents produced by the laboratory. Unlike arXiv which has a deliberately simple submission interface, there are more than 30 web submission forms, each one tailored to a specific document series or type of information resource [5]. The submitter, who may be one of the authors or a secretary responsible for a document series, fills out the necessary bibliographic information in the relevant form and gives the path to the full text files.

We have also introduced the possibility for external sources to submit documents, following requests from some outside universities and research centres. These

submissions are moderated to ensure that the submitted documents meet the scope and nature of our database.

If individual documents are located by us, the librarian may make a self-submission of the document. If it turns out that the source is regularly found to contain documents of interest, this may be the forerunner of a new harvesting configuration. It will depend on the quantity of interesting documents and the way the external source is structured. Although not planned as such, we also find that this submission procedure is taking over from the cataloguing module of our library system. It is especially useful for newcomers, who sometimes have little prior experience of cataloguing.

The submission tools are continually being updated and refined. We are able to handle documents that are confidential, access being allowed to only a specified group of people (such as a committee). Documents can be migrated from confidential to public status once they are released, by simply flipping a flag in the submission interface. We are also able to handle documents that have to pass through a refereeing procedure before they are made public. So far, we do not offer authoring tools: we expect a document to be complete by the time it enters the system.

2.3 Quality Control

On average around 800 notices are added to the database each week. This is far too many for each one to be manually checked but with author submission and

automatic uploading, errors will of course occur. So a number of in-house checking programs have been developed to identify common errors in the data input. Using associated knowledge-databases some of these errors can even be corrected automatically. For the rest, the librarians work mainly on files containing only the fields concerned, an individual record will only be accessed in the very small number of cases where the corrections are more complicated. This system allows us to handle a large number of documents on a daily basis with a small staff and still maintain the quality of the data.

2.4   Beyond Preprints

The CERN Library has used various library management systems in the past. Since 1989 it has been ALEPH (Ex Libris, Israel), which we are still using for most internal library functions. In 1996 though we chose to develop our own user search interface (CDSweb) in order to offer the functionality we needed. CDSweb made it possible to cover many different digital collections with a single search interface [6]. Once it was launched, we could integrate other types of textual documents and even non-textual "documents" into the database. So we included the internal notes of some of the larger experiments, the notes used in the Large Hadron Collider (LHC) construction project, the documents of the committees responsible for approving and scheduling the experiments (proposals, minutes, status reports), press cuttings about CERN and many other textual documents. Non-textual collections included the CERN Photo collection, the database of particle physics research institutes, the CERN Historical Archive, videotapes, webcasts and even the objects contained in the CERN permanent exhibition

(Microcosm). These items had not been handled by the Library before, either because of the lack of resources for doing so in the paper era, or because they were considered outside the scope of the Library when the database was simply the library catalogue

As CDSWeb expanded into these other areas, the submission procedures were expanded in parallel, so that each type of document could be submitted directly to the database by the authorized people.

3.   Linking to the Digital Documents and Other Information

3.1 Metadata Links

The most basic link is from the metadata describing a document to the document itself. For the electronic preprints, the URL has been stored as part of the metadata at harvest time. The corresponding article in an electronic journal is treated differently though since storing the URLs of every article in our database is not a scaleable solution. Instead we developed our own link manager, GoDirect, which takes advantage of the fact that in many cases the URLs for articles can be derived from the journal title, volume and page numbers. From the journal reference present in our metadata, GoDirect can create a link on the fly to the actual journal article [7].

We were able to convince other publishers to follow this example of a "friendly URL" but some publishers were only willing to offer us files of correspondences between volume/page and the DOI. This solution is unsatisfactory though as it means data has to be stored for each electronic journal containing the URL format and the volumes that are electronically accessible. This has since been solved by CrossRef, an initiative from a consortium of publishers which provides a guest interface that enables DOIs to be resolved from journal/ volume/page information.

Another useful link is to pass from our metadata to the metadata of the same document in another database. Whenever we harvest metadata from another database we store the system number of that record (if it has one) in our own database. Linking to a record in another database may give access to additional digital full text versions. For example, the particle physics laboratory KEK in Japan has made an enormous effort to scan paper preprints from the pre-electronic era. Linking to another database can also be useful if our own files become corrupted in some way.

Another possibility is to pass from one set of documents to a group of related items, by automatically generating a database search from the metadata. For example, we can go from a conference document to all other documents relating to the same conference, ask for all the publications of one of the authors or search for more documents having the same subject descriptors. We also have links from our metadata to non-bibliographic information resources in our field, enabling users to access explanatory or numerical information.

3.2 Citations Linking

For all electronic documents on our server (now about 250,000) we have extracted the block of references at the end of each article and have indexed the complete text of each reference [8]. For practical reasons, these have been stored in a separate citations database linked to the main bibliographic database on a record-by-record basis via the e-print number. In this sense, the citations have also become additional metadata.

This database permits the making of citation searches using any text which is in the citation. Additionally, the user can choose to display the list of references for a document. As this page is assembled, links to all e-journal references to which we have access are established using our link manager, as well as any links to e-print archive documents.

There are about 3 million linkable references inside our electronic documents. The handling of the citations in this way is carried out automatically for all new documents. A knowledge database with some 3000 entries is used in the standardization of the presentation of the citations. The project was started in 1998 and further improvements are underway to increase the proportion of references that can be linked to the full text.

This feature means you do not need to pass back through our metadata in order to arrive at the digital document. Apart from the obvious fact that it takes one click less to get to the full text, another huge advantage is that a document can be linked

to even when it is not in our database at all. Thus when more than 100 years of Physical Review were put online (back to 1893), all the links to these articles immediately became available in the citations of our documents, even though we do not have the metadata for the older articles in our database.

However, one big disadvantage arises from the fact that (for obvious reasons!) the journal article never contains any reference to an e-print. Thus, unless the author has also referred to the e-print as well as the journal reference, the e-print version will not be accessible from the citation list. If the user does not have access to the e-journal in question, they will then not have access to the digital document. There is also the reverse case when the original citation is only to an e-print because it had not been published in a journal at the time it was cited. The citation stays like this forever with no direct link to the later article.

4. Future Developments

4.1   Full Text Indexing

The full text of all our electronic documents has been indexed using the Internet search engine UltraSeek. This allows any text in any of these documents to be retrieved. This full text indexing is also part of our daily processing of new documents. At present it is a standalone feature so cannot be combined with searches in the bibliographic data. However this should be possible in the next version of CDSWeb which will be released shortly. In this case one could say that the whole text of the document has become metadata!

We are also planning to create permanent Web pages for all of our documents, so they will be picked up by the various Internet indexing programs, and hence by users completely outside of the library website. Conversely, we want to make it possible to launch an Internet search from our metadata. Very often this is a successful way of locating a digital document when all else has failed, for example when the document exists on the author's own Web site.

4.2 Automatic Keywording

We are using statistical and linguistic analysis of the complete full text in order to automatically attribute keywords and keyword phrases to the documents [9]. We have made a correlation between the full texts of a training sample of about 2000 documents and the thesaurus terms attributed to those documents by the documentation group at our sister laboratory DESY in Hamburg. From CDSWeb, it is now possible to automatically generate a list of DESY Thesaurus terms for any document. This project is currently in its first phase and more development is necessary.

4.3 Lexi

We are developing an encyclopaedic database of all terms used in the particle physics field (20,000 at present). When finished, we plan to link the mention of any of these terms (except perhaps the very common ones) in the full text of any document to a description of the term, with further links to the original documents that defined it.

4.4 GRACE

CERN is a partner in GRACE (Grid Retrieval And Categorization Engine). This project is developing a search engine that will allow users to search through heterogeneous resources stored in geographically distributed digital collections. GRACE will be run on the European Data Grid and will not have a centralized index as current search engines do [10].

Conclusions

Currently about 95% of the particle physics literature is available to us electronically. By treating paper and digital documents in essentially the same way from the very beginning, the transition from a paper-based to largely digital library was able to take place seamlessly at its natural pace, as more and more documents became available in digital form. Since 1994 we have moved from a collection of paper documents described by a searchable computer-based library catalogue, to a largely digital library. The scope has expanded enormously to take in types of document that were not present in the traditional library.

The links to the e-version of the preprint and the corresponding electronic journal article stand side-by-side. This is especially useful for the non-CERN users who won't have access to the e-journals via our site. In the year 2001, queries were received from about 150,000 different host computers around the world.

We have developed a lot of functionality in the way we can access our documents since the start of the electronic era, despite the fact that the Library staff has been reduced in this period. As is often the case, automatic techniques do not really save time, they just enable one to do more in the same time. The database contains around 500,000 documents and is now expanding at about 50,000 documents per year, five times more than we could manage in the paper era.

The electronic document era has changed *how* we do things quite a lot, but so far it is hardly changed *what* we do at all. Despite the success of preprint servers, authors still need to publish in journals which while they may be electronically accessible, are still essentially clones of the paper versions. Scientific communication still follows the path of producing completely contained documents containing an introduction and a description of the context in which the work is to be seen, only a part of the document deals with what is really new. But the possibilities for accessing and linking documents via widely different types of information continue to grow. When scientists decide to move beyond the limitations of the paper-publishing paradigm and exploit the possibilities offered by the digital age to the full, then the real scientific communication revolution will begin.

References

1. CERN Scientific Information Service Mission Statement
   http://library.cern.ch/library_general/group_mandate.html

2. Dallman, D. 'Electronic Journals and Electronic Publishing at CERN: A Case Study'. *Intern. Spring School on the Digital Library and E-publishing for Science and Technology, CERN, Geneva, 2002.*

3. Pignard, N et al. 'Automated Treatment of Electronic Resources in the Scientific Information Service at CERN'. *HEP Libraries Webzine, Issue 3/Mar 2001*
   http://library.cern.ch/HEPLW/3/papers/3/

4. Vesely, M. et al. 'Creating Open Digital Library using XML: Implementation of OAi-PMH'. *International Conference on Electronic Publishing, ELPub2002, Karlovy Vary, Czech Republic, 2002.*

5. CERN submission interface visible at:
   http://documents.cern.ch/EDS/current/access/index.php

6. CDSweb accessible at: http://cdsweb.cern.ch/

7. Vigen, J et al. 'Link Managers for Grey Literature'. *4th International Conference on Grey Literature, Washington, DC, USA , 4 - 5 Oct 1999.*

8. Claivaz, JB et al. 'From Fulltext Documents to Structured Citations: CERN's Automated Solution'. *HEP Libraries Webzine, Issue 5/Nov 2001.* http://doc.cern.ch/heplw/5/papers/2/

9. Le Meur, JY et al. 'Automatic Keywording of High Energy Physics'. *4th International Conference on Grey Literature, Washington, DC, USA , 4 - 5 Oct 1999, pp. 230-237.*

10. Haya, G et al. 'Developing a Grid-Based Search and Categorization Tool'. *HEP Libraries Webzine, Issue8/Oct 2003.* http://library.cern.ch/HEPLW/8/papers/1/.