Why go via expensive solutions, when you can GoDirect?
A description of how CERN Library integrates digital content

Jens Vigen

CERN, European Organization for Nuclear Research, Geneva

Invited talk at the International Spring School on the Digital Library and E-publishing for Science and Technology
CERN, Geneva, Switzerland, 13-18 June 2004

Abstract
In the digital world the creative solutions have to be implemented far quicker than what was required in the era of printed publications. In such a context many vendors take advantage and offer the library community simple computer applications to outrageous prices as if the products were highly sophisticated information systems that would be impossible to develop and maintain in-house.

The CERN Scientific information Service has chosen not to join this game, partly due to that commercial solutions were not available at the time when first needed, partly due to that lots of the solutions actually can be implemented and run with already available resources within the library and the informatics support. Using only in-house developed solutions, digital content at CERN is today being integrated to the highest possible level in order to meet the high requirements of the particle physics community.

The paper gives an overview of the steps CERN has made towards the digital library from the day the laboratory conceived the World Wide Web to present.

Setting the scene
CERN, the European Organization for Nuclear Research, is the world's largest particle physics centre, used by half of the world's particle physicists. These scientists, altogether 6500 users, represent 500 universities and over 80 nationalities. The CERN staff, that comprises just below 3000 people, have as its global aim to support these scientists in their research. CERN staff encompasses a wide range of skills and trades - engineers, technicians, craftsmen, administrators, secretaries, workmen, ... and of course librarians who are there to meet all their information needs.

The CERN staff designs and builds CERN's intricate machinery and ensures its smooth operation. Then helps prepare, run, analyse and interpret the complex scientific experiments and carry out the variety of tasks required to make such a special organization successful. Constructing these highly advanced machines is an extremely costly operation [1]. The high costs imply that there is absolutely no financial room for research and development already carried out somewhere else in the world. This constrain lead the particle physics community into a culture based on preprints to accelerate the communication process more than 40 years ago.

Driven by the same requirements Tim Berners-Lee, a CERN computer scientist invented the World Wide Web, conceived and developed for the large high-energy physics collaborations which have a demand for instantaneous information sharing
between physicists working in different universities and institutes all over the world.

"In the beginning ..."
We are back in December 1990 and the Web was created. "... the earth was formless and empty, darkness was over the surface of the deep ..." and all the world's web servers were at CERN - maybe even in the same building. A prophet has however no honor in the prophet's own country, so instead of just moving one floor up, the later so famous web went off to Stanford Linear Accelerator Center where it shortly after was set up as a new gateway for accessing the bibliographic database SPIRES-HEP [2]. Looking back, information retrieval from SPIRES-HEP became the application that compelled the community to start using the Web. A few months later at Los Alamos National Laboratory, "in the middle of nowhere between two Indian pueblos", theoretical physicist Paul Ginsparg applied the new technology to set up arXiv, a system facilitating distribution of drafts to other theoretical physicists. These days Ginsparg is often credited by his peers for having revolutionised scientific communication by setting up this system [3]. At present the archive contains 280 000 papers, augmented with some 100 new papers every day. ArXiv is estimated to distribute about 25,000 daily e-mail alerts and there are probably at least 35,000 distinct daily users via the web.

Back at CERN
In spite of not having been the first library in the world with web access to it's catalogue, the various developments world wide were closely monitored by the CERN Scientific Information Service. Actually, an experimental initiative of scanning documents received on paper from laboratories and universities from all over the globe had started more or less at the same time, with the idea of diffusing the information to the whole of the community via the information networks. The bibliographic data was kept in the library catalogue while the fulltext was kept separately, so in order to retrieve a paper one first had to search for the key in the bibliographic database before one could navigate down to the paper itself on the preprint server [4]. January 1994 stands out as a paradigm change as from then onwards all new preprints were provided in an electronic format, a fully integrated service, or a "WWW GUI" as it so nicely was referred to at the time, would however not be launched until two years later.

Take-off for the CERN Document Server
The times with no linking capabilities between metadata and fulltext was perceived as having lasted for ages, at least by the library staff that had spent quite some energy in guiding lost readers through this initial period. There were however no major reason to complain, in 1996 most libraries still had not yet started thinking of having hyperlinks to external resources from their bibliographic catalogues - for most libraries there were basically no relevant materials to link out to, at least not materials considered to be relevant at the time. CERN librarians, having ALEPH as the in-house library system, realised quickly that the standard web interface of the system did not correspond to their ambitions of integrating digital content to the highest possible level. It was consequently decided to build a CERN specific interface, using extendable application program interfaces [APIs] enabling an expansion without expensive source-code modifications. The die
was cast, the CERN Document Server, was about to be implemented. To start with the "high ambitions" were made up of simply imagining a basic set of links between related records and links to the corresponding fulltext for preprints as the killer application ... The years went quickly by and the expectations of the library users augmented dramatically. After some time the killer applications were not any longer impressive and the response time of the system started to be unbearable when submitting queries with a certain length and with a high complexity. The programmers mumbled about abandoning ALEPH, but the librarians, many with experience from the early days of library automation, required a fully integrated library system as a basis for their operations. The solution that was identified was to keep ALEPH for managing the activities in the back-office and export the bibliographic data to an open source database made available to the user community. This solution was initially intended for just being applied at CERN. However, due to the arising Open Archive Initiative movement, it turned out to be a huge need for document management systems, outside of the traditional library systems, all over the academic world. The software was therefore packaged, labeled CDSware, and made freely available under GNU General Public License (GPL). It allows you to run your own electronic preprint server, your own online library catalogue or a document system on the web. It complies with the Open Archive Initiative metadata harvesting protocol (OAI-PMH) and uses MARC 21 as its underlying bibliographic standard. The database behind is MySQL, a database server that has quickly become the core of many high-volume, business-critical applications around the world. Earlie limitations implying stop words, maximal size of a retrieved set etc. became history. With the new setup the system could retrieve all records containing the word of, at present 458 000, in less than a second! Who knows, such results might become usefull one day … [5]

The publishers getting onboard
In parallel to the e-publishing activities in academia, all the major publishing houses were carrying out tests with the intention of launching electronic journals. The appearance of the web, which of course was a gift to anybody involved in electronic publishing, kind of wiped out all existing initiatives as being all of a sudden obsolete. "If there is no library system available or if it is decided to develop an entirely new system, the best option at this moment seems to be to use the very popular World Wide Web as user interface." was, not surprisingly, one of the conclusions from the TULIP project, initiated by Elsevier, in 1996 [6].

Institute of Physics (IOP) adapted quicker than the other publishers and was therefore the first publisher to offer it's entire journal portfolio across the web in the spring of 1996. The launch was well received by an innovative library community and a group of enthusiastic scientists, in spite of a response time one would wish not to think about - even back in 1996. The new resources were warmly welcomed, although not yet integrated into the libraries traditional retrieval tools. The level of integration of the electronic journals was simply restricted to setting up links to the available titles from the various libraries' website. Nobody seemed to think about cataloguing an e-journal. Why bother with cataloguing as the counterparts on paper were already in the catalogue?

One year later the American Physical Society (APS) entered the arena and now things finally started to happen in terms of integrating the electronic collections. For the architects behind the electronic version of Physical Review D it must have been clear
from the very beginning that they had to facilitate direct access to any article in the journal by using a URL scheme similar to the scheme used for article identification in the traditional library. This was an enormous progress as it permitted users to access articles without having to navigate through a whole set of pages before getting to the desired information. For information collectors it was thus possible to automatically generate links from their metadata repositories to the corresponding articles, but so far only for articles published in Physical Review D. The handling system developed at CERN to facilitate this feature was named "Go Direct" and caused quite some excitement, even though it could only, to start with, handle a few titles [7]. In addition to Physical Review D it could also handle some Springer titles using a set of cleverly thought out lookup tables that had to be manually updated whenever a new issue of the journal appeared. With "Go Direct" CERN had implemented SFX [8] - without knowing it – after all not so strange, SFX had still not yet been conceived ...

Triggered by the enthusiasm shown by the library's avant-garde users, it was now time for the CERN librarians to start a systematic lobbying of all physics publishing houses. At every occasion letters were sent, interventions were made during seminars etc. with the aim to make the publishers introduce URLs for their journal articles based on the triplet journal, volume, page. Surprisingly enough the idea did not generate the same amount of excitement among the publishers as it did at CERN, however, a few months later it was silently implemented for all APS journals [9] and the APS link manager has become a de facto standard for all publishers' handling systems. The philosophy of the link manager is simple, but powerful: the triplet which has served the world as a unique, or at least close to unique, identifier for articles since the appearance of the first journals centuries ago, should always be associated to a persistent and robust URL.

**Homework to be done**

The publishers had started doing their homework, so then it was just necessary to keep on going with more innovations on the library side in order to maintain the pressure. Entering into the new era it was rewarding to see the importance of having fully streamlined metadata. With an ISBN associated to each book-record, for example, it became straightforward to create links to the corresponding Amazon records, records which more and more often contain extracts of the books itself, the possibilities to search within the fulltext, the table of contents, reviews etc. Without clean data no links would have been created - finally the fruits of years of librarians' accuracy could be harvested. But how long was Adam in paradise? Library users were quickly getting acquainted and it was realised that in order to provide the most effective service it would be indispensable to collect a maximum of metadata for each document and these data would have to be added to the database with the shortest possible delay.

In the case of CERN this meant starting to add publication references to all preprint records as soon as the papers became published. Up to that moment such references had only been added to papers originating from CERN, so more automation was clearly required in order to absorb the additional workload [10] [11].

To identify the preprints' published counterparts is not straightforward as no publishers are willing to give the correspondence between the articles they publish and the "original" preprint numbers. Publishers even claim that the correspondence is unknown to them, this in spite of the fact that several of the major actors (Elsevier, IOP, APS etc.) even seem to prefer, or at least appreciate, the possibility of picking up
the fulltext of the submitted manuscripts from arXiv ... A comprehensive matching procedure, matching data from various sources of published material against the CERN collection of preprints, had therefore to be implemented. Matched records were updated with publication references, permitting "Go Direct" to automatically create links to the corresponding fulltext of the published articles.

At the beginning of 2000, the world's leading scholarly publishers made another step forward and joined to form the non-profit, independent organization, Publishers International Linking Association, Inc. (PILA), which operates CrossRef [12]. The general purpose of CrossRef is to promote the development and cooperative use of new and innovative technologies to speed and facilitate scholarly research. The specific CrossRef mission is to be the citation linking backbone for all scholarly information in electronic form. CrossRef is a collaborative reference linking service that functions as a sort of digital switchboard. It holds no full text content, but rather effects linkages through Digital Object Identifiers (DOI) [13], which are tagged to article metadata supplied by the participating publishers. The end result is a quite efficient, scalable linking system through which libraries and end users can gain access to cited articles via the simple triplet journal, volume and page. The handling system is used at CERN within GoDirect as one out of many linking systems. Libraries pay no charges to get a CrossRef account for the retrieval of DOIs or other metadata. It is thus rather attractive for libraries to link to digital material through CrossRef as it can be done without having to track the publishers' individual linking schemes. There are unfortunately still major publishers that are sloppy in reporting their data to CrossRef and there are also quite some mistakes in the CrossRef database, so in many cases it is still an advantage to use the 'native' link manager at the individual publisher's site in order to create robust links.

**Library system becomes CERN's main scientific information system**

Not only the library world profited from the developments of the digital media. Photographs, posters, newspapers etc. went all electronic, so they did at CERN. So far these materials had all been managed as isolated pieces of information, recorded in various local databases. The move into the digital world lead however naturally all database owners the same way and at a certain point all the data were brought into one single system, offering users the ability to search across all the various collections at once.

Within this optic the CERN Document Server has moved from being a pure library catalogue, towards becoming the CERN "global" scientific information system. The collection covers for the time being preprints, books, periodicals, reports, photographs, press cuttings, posters, exhibition objects and much more - in the foreseeable future it will cover, or point to, all scientific information resources needed by any particle physicist or CERN staff member to perform her or his work efficiently.

**What happened to the books?**

Apart from linking up the bibliographic records to Amazon, what happened to the books? Did they go electronic? Surely they did, but how will the electronic books make their way to the academic arena? Practically speaking, CERN Library has by today no collection of electronic books. Several signals indicate though that the electronic books are entering the arena. However the adoption of ebooks has been
sluggish. Existing offerings, often aimed primarily at the needs of retail end-users, is not sufficiently compatible with the expectations of the academic community. In the hope of building a product that meets the needs and expectations of the library users, CERN therefore participates in a project with eBooks Corporation to develop an eBook Library (EBL), a lending platform intended for release in 2004. The technology that enables the lending of ebooks is advancing. There is strong evidence of demand for a flexible lending model that can satisfy not only general lending requirements. There is also a strong business case for publishers to identify a model for making their content available for this medium. Observing the trends of electronic journals and feeling the effects of the electronic medium on the sale of monographs, publishers also recognize that ebooks in libraries, particularly in research libraries, are an inevitable evolution. Many publishers are steadily increasing the number of titles available as ebooks and more publishers are entering the ring. More progressive publishers have worked ebook formats into their production flow and are beginning to release titles simultaneously with their print releases. Developments in recent years have made many librarians feel a loss of control as journal publishers reconfigured the supply model to their extreme advantage. With the introduction of electronic journals, journal publishers enforced their concept of packaging bundled products. Acquisitions librarians are left with a minimum of choice and perceive more and more often their role simply as 'paying librarians'. With the introduction of electronic books in libraries, someone will still have to pay the bill. However, the philosophy of selling collections, or bundled subject packages, could be developed into something far more dynamic, actually resulting in a local user driven acquisition policy, carefully guided by librarians. So far, most library catalogues only include the description of documents that belong to a certain collection or a set of collections. This is probably due to the widespread concept of seeing a library catalogue more as a logistics tool than a portal to information in general. However, as long as information not held by a library can be obtained within a reasonable time frame, one could imagine expanding the catalogue to also include such items if the bibliographic data could be easily imported. The future role of the acquisitions librarian will be to identify sources, individual authors', publishers' or booksellers' catalogues, from which the description of the relevant documents could be extracted on a regular basis and later incorporated in the local library catalogue. Traditionally books not held by the library, have had to be borrowed from fellow libraries, normally including a small cost to cover the shipping and processing expenses. With electronic books the shipping costs and delays are eliminated, leaving room for a new concept of 'interlending'. Costs previously associated with traditional interlibrary loan could in this way be used to pay for temporary access to electronic books (renting them). A complete feedback system permitting the librarians to monitor the use of the collection, even interact with readers, could be put in place and be used as guidelines for general acquisition policy. ‘Rented items' could possibly later be upgraded to become part of the collection [14].

Additional features
Having reached the "maximum" of links based on the metadata, it was time to investigate what could be done with the actual content of the documents. As researchers spend quite some time just retrieving referenced papers, it was obvious to the CERN developers that this would be a field with great potential for savings. A program for automatic extraction of the references, using the fulltext documents, was
therefore implemented. The extracted data were later parsed through a "normalisation filter" to be made compliant to standard notations. The result was loaded into a devoted part of the database and links were automatically created to the corresponding texts [15]. This operation has in total generated more than 3 million links to fulltext, a number that is augmented with about 1500 per day, as the extraction is now a part of the regular routines. The results have become particularly striking with the recent developments of electronic journals and access to their so-called 'back-files' Just in the field of physics and the related subjects more than three million articles are directly available from our desktops at this moment in time! We can now all consult online a vast historical archive, including the famous Philosophical Transactions of the Royal Society of London that started in 1665. Literally speaking therefore nearly all references to journal articles have been converted into active links. 

The CERN-citation data have though not been exploited further to facilitate citation analysis or any other form of bibliometrics. With the growing coverage of citations in the database, such developments are though foreseen in the near future. In the meantime the citation data for each article can be obtained, for the users based at the CERN site, from Science Citation Index on Dialog DataStar. This functionality is enabled via links simply constructed with metadata already held in the record. Each record on the CERN Document Server, describing a published article, includes such a dynamic link based on the first author, year, volume and first page of the article. Similar services are also planed for SPIRES-HEP and NASA ADS [16], both databases including rich sets of citation data that are interesting to compare with the data in Science Citation Index.

Search in fulltext has often been pushed forward as a replacement of costly indexing. At CERN both approaches were considered to be important for providing an efficient information retrieval service. The CERN library however never indexed it's preprint collection in a systematic way due to the lack of resources, but entering the digital era opened new possibilities: Hosting a vast collection of electronic preprints made it natural to start experimenting with searching in fulltext. Ultraceck is used as the search engine and the fulltext search interface permits searching through a vast collection of fulltext documents stored on the CERN Document Server. A fulltext search can retrieve the one and only document describing the most rare concept, but it might also retrieve lots of noise. So in parallel to providing the fulltext search, it was decided also to look into the area of automatic indexing. Automatic indexing can be considered to be a branch of automatic summarization, which aims at the generation of abstracts from fulltext documents. The developers at CERN went ahead and made the HEPindexer that proposes a preliminary solution, which can open the way for further research into automatic indexing tools in the area of particle physics [17]. So far a first step has been achieved, namely the generation of main DESY keywords [18]. These keywords are generated following a statistical approach. Investing more resources in the development of the HEPindexer will certainly give results that will improve the precision and recall searching of the particle physics literature [19]. Among further developments, one could imagine introducing more enhanced reader service by for example linking the extracted concepts to the corresponding concepts or terms in relevant encyclopedias, handbooks, tables and dictionaries.

**From one central library to thousands of satellites**

The pendulum is swinging back, having centralised services for years the CERN
library is again establishing satellite libraries and even personalised libraries. These libraries are of course digital libraries, fully integrated on the readers’ desktop. So far the service is mostly restricted to information provision [alerts and searching] and private records management [personal e-shelves and loans]. Enhanced reading tools represents a path that is only partly explored so far, even if the links to the fulltext of the citations have a quite striking effect. What CERN Scientific Information Service has not yet started to explore is having authoring tools as a part of the library system. It is however clear, if librarians want to continue being advocates for integrating digital content, one has to start expanding in that direction, and then the ring will be closed.

Conclusion
Much has happened since the developments of the automated library systems in the 1970-ties. Those who experienced the initial development of these systems would however still not recommend any professional library to start in-house developments of such complex systems - and maybe, for the time being, not even consider any of the systems available as open source. However, this does not mean that libraries should not consider to add in-house developed solutions around or on top of their proprietary software. It is however not easy for library managers, being under continuous pressure to run their libraries on the edge of the technology at the same time as they have in mind the technological struggles they went through 30 years ago, to refuse offers from promising sales representatives offering their SFX, OpenURL, LinkSolver, CrossRef, DOI, OAI-PMH etc. compliant products - this in spite of their often highly excessive prices. There are however many free solutions "out there" that are worth considering and that actually does not take much time to maintain in terms of manpower. What has been implemented at CERN, we are ready to share.

After all, integrating digital content is only partly a technical challenge. The main challenge is to get all involved parties "to speak the same language". If you just believe in it strongly enough, nothing will be impossible. Just remember to ask yourself before making any important decision: “What is the advantage of using an expensive solution, or can I simply GoDirect?” Never forget that the interest of the commercial actors and the research libraries are after all quite different in spite of the fact that our end users are the very same ….
References
[1] Documents concerning the LHC Cost Review to Completion  
<http://info.web.cern.ch/info/LHCCost/>

<http://www.slac.stanford.edu/history/earlyweb/index.shtml>

[3] An Online Archive With Mountain Roots  
New York Times, August 28, 2001, Tuesday  
<http://www.physics.utoronto.ca/~peet/media/nytaspen/paul.htm>

[4] Electronic pre-publishing for world-wide access: the case of high energy physics  
Dallman, D P ; Draper, M ; Schwarz, S ;  

Vesely, M et al.  
http://cdsweb.cern.ch/search.py?recid=725772  
5th International Conference on Grey Literature  
Amsterdam, The Netherlands, 4 - 5 Dec 2003  
Ed. by Farace, D J and Frantzen, J - GreyNet,  
Amsterdam, 2003. – pp. 106 -111

Borghuis, M et al.  
Elsevier Science, 1996  
ISBN 0-444-82540-1  
<http://www.elsevier.nl/homepage/about/resproj/trmenu.htm#ToC>

[7] Link managers for grey literature  
Lodi, E ; Vesely, M ; Vigen, J ;  
CERN-AS-99-006  
4th International Conference on Grey Literature  
Washington, DC, USA, 4 - 5 Oct 1999  
Ed. by Farace, D J and Frantzen, J - GreyNet,  
Amsterdam, 2000. - pp. 116 -134

[8] SFX - context sensitive reference linking  
<http://www.sfxit.com/>

[9] Citing and Linking in Electronic Scholarly Publishing  
Doyle, M  
Smith, John W ed. ; Ardo, Anders ed. ; Linde, Peter ed. ;  
ISBN 1-891365-04-5  
<http://www5.hk-r.se/ElPub99.nsf/>
[10] Automation of electronic ressources in the Scientific Information Service at CERN
    Pignard, N ; Geretschläger, I ; Jerdelet, J ;
    <http://library.cern.ch/HEPLW/3/papers/3/>

    Vesely, M

[12] CrossRef
    <http://www.crossref.org/>

    <http://www.doi.org>

[14] E-books and interlibrary loan
    Vigen, J ; Paulson, K ;
    CERN-OPEN-2003-050
    <http://cdsweb.cern.ch/search.py?recid=682384>
    8th IFLA Interlending and Document Supply International Conference
    Canberra, Australia, 28 - 31 Oct 2003

[15] From fulltext documents to structured citations
    Claivaz, J B ; Le Meur, J Y ; Robinson, N ;
    <http://library.cern.ch/HEPLW/5/papers/2/>

[16] The NASA Astrophysics Data System
    <http://adsabs.harvard.edu/>

[17] Experiences in automatic keywording of particle physics literature
    Montejo Ráez, A; Dallman, D
    <http://library.cern.ch/HEPLW/5/papers/3/>

[18] DESY.
    The high energy physics index keywords
    <http://www-library.desy.de/schlagw2.html>

[19] Montejo Ráez, A
    Asignación automática de palabras clave en tiempo real.
    II Jornadas de Tratamiento y Recuperación de la Información.
    Madrid, Spain, 2003