Digital content sewed together within a library catalogue WebLib - The CERN Document Server

Jens Vigen

CERN, Geneva, Switzerland

Abstract. Aggregation, harvesting, personalization techniques, portals, service provision, etc. have all become buzzwords. Most of them simply describing what librarians have been doing for hundreds of years. Prior to the Web few people outside the libraries were c oncerned about these issues, a situation which today it is completely turned upside down. Hopefully the new actors on the arena of knowledge management will take full advantage of all the available "savoir faire". At CERN, the European Organization for Nuclear Research, librarians and informaticians have set up a complete system, WebLib, actually based on the traditional library catalogue. Digital content is, within this framework, being integrated to the highest p ossible level in order to meet the strong 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.

1 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, who comprise just under 3000 people, have as their global aim to support these scientists in their research. CERN staff encompass 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 design and build CERN's intricate machinery and ensure its smooth operation. Then help prepare, run, analyse and interpret the complex scientific experiments and carry out the variety of tasks required to make such a special organization su ccessful. 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 t he 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 w orking in different universities and institutes all over the world.

2 "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 honour in the prophet's own country, so instead of just moving one floor up, the la ter 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 co mpelled 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. 10 years later Ginsparg is often credited by his peers for having revolutionised scientific communication by setting up this system [3]. At present the archive contains 190 000 papers, augmented with some 100 new papers every day. ArXiv is est imated to distribute about 25,000 daily e-mail alerts and there are probably at least 35,000 distinct daily users via the web.

3 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 don to the paper itself on the preprintserver. [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.

4 Take-off for WebLib - 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 who had spent quite

some energy in guiding lost readers through this initial period. There were however no major re ason 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 materia ls 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, WebLib, 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 ...

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 pu blishing, 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 [5]. 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 g roup 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 integratet into the libraries traditional retrieval tools. The level of integratio n 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 alread y 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 artic les, 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 to start with it only could handle a few titles [6]. In addition to Physical Review D it could also handle some Springer titles using a set of cleverly thoughtout lookup tables which had to be manually updated whenever a new issue of the journal appeared. With "Go Direct" CERN had implemented SFX [7] - withou t knowing it - afterall 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 [8] and the APS Linkmanager has become a defacto standard for all publishers handling systems. The philosophy of the linkmanager 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.

6 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 record it became straightforward to create links to the corresponding Amazon records, records which more and more often contain samples of the books itself in addition to 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 provi de an effective service it would be indispensable to collect a maximum of metadata for each record 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 req uired in order to absorb the additional workload [9] [10]. 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 unknow n 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 pr ocedure, 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.

7 Library system becomes CERN's main scientific information system

Not only the library world profitted from the developments of the digital media. Photographs, posters, newspapers etc. went all electronic, so they did at CERN. This technological step lead naturally so that what so far had been managed as isolated pieces of information, were all at a certain point 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.

8 Enhanced reader services

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. The policy change actually implied that the long way from mainly being a document depository to become a r eal subject portal had started - based on Avi Saha'a definition : "A portal is a single integrated point of comprehensive, ubiquitous, and useful access to information (data), applications, and people." [11]

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 "normalization 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 correspondin g texts [12]. This operation generated in total about 2,1 million links to fulltext, a number which has since been augmented with about 1500 per day, as the extraction is now a part of the regular routines. For the astrophysics papers the results are part icularly striking due to the fact that a significant part of the journal literature in the field is available in fulltext, also retrospectively [13]. Literally speaking nearly all references belonging to this collection have been converted into active lin ks.

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 co llection 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. Ultraseek is used as th e search engine and the fulltext search interface permits searching through more than 90,000 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 als o 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 abs tracts 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 [14]. So far a first s tep has been achieved, namely the generation of main DESY keywords [15]. These keywords are generated following a statistical approach. Investing more resources in the development of the HEPindexer will certainly give results which will improve the the pr ecision and recall searching of the particle physics literature.

Having extracted keywords and references for each document, a range of possibilities for automatically connnecting related documents are opening up. The system can then in accordance with predefined rules, propose other sets of documents to the user as a function of what the user him/herself and similar users, have consulted earlier. In the case of few or zero documents as a result of a research, the system should verify the possibilities for miss spellings and propose alternatives when that would be considered as appropriate. The system should further propose to carry out the same qu ery in other relevant databases by simply transmitting the search arguments, applying the different syntaxes, to the various search engines.

Given that the system has captured knowledge about the nature of a certain set of documents related to specific users, the system should further be used to add additional links to other relevant sources. I.e. new-comers to the subject field can be propose d links from all keyword to the corresponding concepts in Encyclopedia Britannica or in any other general source, while the experienced readers will be directed towards sources as the Review of Particle Physics. The system will in the near future also lin k out to non-bibliographic information as hompages of authors, publishers pages etc.

In principle links can be created to any related entity provided that one find a scalable solution for introducing and maintaining the links.

9 From one central library to thousands of satellites

The pendulum is swinging back, having centralized services for years the CERN library is again establishing satellite libraries and even personalized libraries. These libraries are of course digital libraries, fully integrated on the readers desktop. So f ar the service is mostly restricted to information provision [alerts and searching] and private records management [personal e-shelves and loans]. Enhanced reading tools represent a path which is only partly explored so far, even if the links to the fullt ext of the citations have a quite striking effect. What CERN Scientific Information Service have 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, then the ring will be closed.

10 Conclusion

Integrating digital content is only partly a technical challenge, the main challenge is to get all involved parties "to speak the same language". However, if you just believe in it strongly enough, nothing will be impossible, even if the interest of a com mercial publisher and a research library are quite different in spite of the fact that the end users are the same.

References

1. LHC Cost Review ¡http://info.web.cern.ch/info/LHCCost/2001-10-16/LHCCostReview.html¿
2. Documentation of the Early Web at SLAC (1991-1994) ¡http://www.slac.stanford.edu/history/earlyweb/index.shtml¿
3. An Online Archive With Mountain Roots New York Times, August 28, 2001, Tuesday ¡http://college3.nytimes.com/guests/articles/2001/08/28/864834.xml¿
4. Electronic pre-publishing for world-wide access : the case of high energy physics Dallman, D P ; Draper, M ; Schwarz, S ; Interlending and Document Supply : 22 (1994) , pp.3-7
5. The TULIP Final report Borghuis, M et al. Elsevier Science, 1996 ISBN 0-444-82540-1 ¡http://www.elsevier.nl/homepage/about/resproj/trmenu.htm#ToC¿
6. Link managers for grey literature Lodi, E ; Vesely, M ; Vigen, J ; CERN-AS-99-006 ¡http://weblib.cern.ch/abstract?CERN-AS-99-006¿ 4th International Conference on Grey Literature : New Frontiers in Grey Literature, Washington, DC, USA, 4 - 5 Oct 1999 Ed. by Farace, D J and Frantzen, J - GreyNet, Amsterdam, 2000. - pp.116-134
7. SFX - context sensetive reference linking ¡http://www.sfxit.com/¿
8. Citing and Linking in Electronic Scholarly Publishing: A Pragmatic Approach Doyle, M 3rd ICCC/IFIP Conference on Electronic Publishing. 1999. Ronneby, Sweden. Smith, John W ed. ; Ardo, Anders ed. ; Linde, Peter ed. ; Washington DC : ICCC Press, 1999. - pp.51-59 ISBN 1-891365-04-5 ¡http://www5.hk-r.se/ElPub99.nsf/¿

9. Automation of electronic ressources in the Scientific Information Service at CERN Pignard, N ; Geretschlger, I ; Jerdelet, J ; High Energy Phys. Libr. Webzine : 3 (2001) , pp. 3 ¡http://library.cern.ch/HEPLW/3/papers/3/¿
10. Using Internet/Intranet Technologies in Library Automation Vesely, M Thesis : Univ. Economics Prague : 2000 ¡http://weblib.cern.ch/abstract?CERN-THESIS-2000-040¿
11. Application Framework for e-business: Portals Avi Saha IBM developerWorks, 1999 ¡http://www-106.ibm.com/developerworks/library/portals/¿
12. From fulltext documents to structured citations : CERN's automated solution Claivaz, J B ; Le Meur, J Y ; Robinson, N ; High Energy Phys. Libr. Webzine : 5 (2001) , pp. 2 ¡http://library.cern.ch/HEPLW/5/papers/2/¿
13. The NASA Astrophysics Data System ¡http://adswww.harvard.edu/¿
14. Experiences in automatic keywording of particle physics literature Montejo Rez, A; Dallman, D High Energy Phys. Libr. Webzine : 5, (2001), pp. 3 ¡http://library.cern.ch/HEPLW/5/papers/3/¿
15. DESY. The high energy physics index keywords ¡http://www-library.desy.de/schlagw2.html¿