International Spring School on the Digital Library and E-publishing for Science and Technology
CERN, Geneva, Switzerland, 3 – 8 March 2002

Electronic Journals and Electronic Publishing at CERN: A Case Study

David Dallman
Senior Scientific Information Officer, CERN (European Organization for Nuclear Research), Scientific Information Service, Switzerland

1 INTRODUCTION

In order to discuss the literature in the main research field of CERN (particle physics, sometimes still referred to as high-energy physics (HEP)), some very general remarks on this field may be useful to put the literature into its context.

Particle physics as a distinct field of physics is generally considered to have begun in the last decade of the 19th century with the discoveries of radioactivity by Henri Becquerel (1895) and the electron by J.J.Thomson (1897). Even after more than 100 years of further research in which much finer sub-structure has been found in many so-called "elementary" particle systems (such as nuclei and hadrons), it is ironic that the electron is still elementary even at today's level of understanding. The order of discoveries is not necessarily a steady progression from complex to simple.

Following the Second World War the particle physics field began to leave the areas of nuclear physics and cosmic rays where many advances in understanding had been made. The construction of particle accelerators allowed experiments to be carried out in much more controllable situations. As one moved to higher and higher energies, the sizes of the accelerators increased so that we now have the tunnel for the LEP and LHC machines at CERN which is 27 km in circumference. Even so, the energies are still far, far below some of the energies that are observed outside in the Universe, so it seems inevitable that there will be a return to cosmological sources at some time in the future. The new field of astroparticle physics has firmly established itself in the past ten years.

In particle physics today, there are about 20,000 active researchers worldwide, which is quite small compared to some other branches of physics. Although this is research requiring large and costly machines that single countries can no longer afford, this very fact enables physicists from many smaller countries to participate in the research. Only Africa is under-represented in the research community.

The research is carried out wholly in the academic field, because it has no current commercial applications (though there are many commercially-interesting spin-offs). As in many branches of science, increased specialization has meant that there are essentially no longer any people who are able to make major contributions on both the experimental and theoretical sides. Enrico Fermi (1930s and 1940s) is usually said to be the last person who managed this. The researchers are divided about equally into theorists and experimentalists. The theorists are scattered all over the world in rather small groups, mostly in universities. The database of

particle physics research institutes that we compile contains about 900 entries at present. The experimentalists also work in universities or national laboratories, but have to go to one of the major international laboratories like CERN in order to carry out their experiments. Their stays at these laboratories may range from a few weeks to several years, depending on the complexity of the project and their own particular role in it. Some commute from their home universities whenever their teaching schedule permits it. Some are permanently resident at the international laboratory and "anti-commute" back to their universities when they have to give their lectures.

CERN is the largest of the world's particle physics laboratories. Although the Member States financing CERN are all European, the laboratory does in fact welcome researchers from all over the world (about 80 countries). CERN constructs and operates the accelerators and provides the technical, computing and administrative infrastructure. But the costs of the experimental detectors have to be met by the research groups themselves. Of course those working on the accelerator and technical side are also at the forefront of their respective fields, so there is some pure research activity there as well.

2 PUBLICATION PATTERNS

Publishing the results obtained at CERN is of paramount importance. After all, this is the only legacy that will remain for future generations, once the accelerators have been closed down. The CERN Convention mentions in its second sentence that "the results of its experimental and theoretical work shall be published or otherwise made generally available". One aspect that is not an issue is that of translations. Most documents are written in English, but those that are in other languages (French, German, etc.) are not usually translated.

We do not however just deal with CERN documents, but with the complete world body of literature on particle physics. The most important class of documents contains the results of the particle physics research itself, both experimental and theoretical. The theoretical documents outnumber the experimental ones by perhaps 10:1, but they have typically only 1-3 authors, whereas the experimental publications tend to have much larger author lists of 50-500 people. After that there are secondary publications like conference contributions in which the primary results are discussed. One of the significant developments of the past two decades has been the enormous increase in the complexity of the experimental detectors, in order to cope with the precision and data taking rates necessary. With the long run-in times for future detectors (10-15 years for the LHC experiments), most experimental researchers are involved in detector development for a very long time, before there are any actual particle physics data to analyze. This activity also generates a substantial literature. Finally, the accelerator and technical areas also produce a large number of documents.

3 PREPRINT HISTORY AND THE START OF THE ELECTRONIC ERA

The CERN Scientific Information Service (which includes the Library) was created in 1955 soon after the creation of CERN. What we have done since the start of the electronic document era 10 years ago is very much a consequence of what we did in the 30 years before that.

Ever since the late 1950s, most primary particle physics documents were initially circulated as preprints before being published in the traditional journals. This preprint activity developed because researchers considered that the time lag before publication was too long, so that to wait for it would have drastically curtailed the activity in this fast-developing field.

During the period 1960-1990, many of the preprint series issued by universities and laboratories became very well known. They often had their own "look" with colorfully designed covers and many of the documents were refereed internally to ensure that the quality of the institute's work was maintained. They were distributed free to hundreds of other institutes in the field. As text-editing tools and better printers became available, they also developed from typewritten documents to have layouts that were comparable in quality with the traditional journals. In short, many of these preprint series gradually took on many of the characteristics of a journal.

CERN started issuing a weekly list of preprints in 1958. Already in the early 1960s, semi-automatic means (early computers) were being used to produce this list. Eventually the data were kept in a database from which the list could be produced, and in 1983 this database was made available to users for searching the library catalogue. Every bibliographic record was catalogued from the paper document and typed into the database. By 1991, the situation was that the number of paper preprints being catalogued had risen to about 10,000 per year and the system was absolutely at the limit of what could be achieved with the two staff members available. The publication references to the journal versions of the articles were also being added to the database by hand. There were also problems of space for storing all these documents. In fact, in 1990 we started a scanning procedure for preprints and several thousand were scanned. However this project had to be stopped when the person doing the scanning retired.

In August 1991, the Los Alamos electronic preprint server was launched. Actually it started at Princeton and is now at Cornell, though with the Internet it is perhaps hard to define what "where" really means. This was pre-Web, the full text documents were obtained by FTP. Even so, the server provided an enormous improvement in the speed and ease of preprint communication. By 1992, it was starting to become popular among particle theorists and it was not long before some institutes announced that they were stopping the costly distribution of paper preprints. We would have less preprints to handle! However it was clear to us that the electronic preprints would have to be treated in the same way as paper ones, except that here was a chance to cut down on the amount of manual input that was the bottleneck at that time.

4 HARVESTING THE METADATA

During 1993 we had a technical student write a program to read the daily e-mail alerts from the e-print archives (as they were then called), to decipher the bibliographical data (at that time there was just one block of free-text data) and to create database records with links to the full text documents. By this time, we had the Web but our database was not yet Web searchable. Instead we provided Web pages where one could navigate to the document using its archive number (like hep-th/yymmnnn).

The uploading of bibliographic data was started for the particle physics archives in December 1993, and from January 1994 onwards the records were all linked to full text versions. Although the terms had not yet been invented, this was probably the first example of a program to harvest metadata.

Manual input of paper preprints still continued and we scanned most of them so that we could offer an almost complete electronic library. Only theses and reports over 100 pages were not scanned. The number needing to be scanned decreased steadily as more and more were being submitted to Los Alamos or could be found on other Web sites.

In November 1996, the Los Alamos archives introduced some structure into its bibliographic data: the title and authors were separated into dedicated fields. A simplified version of the harvesting program was written to handle this and proved to be more efficient at extracting the various elements of bibliographic data. At the same time we expanded our coverage to the whole of the Los Alamos archives (all subject areas).

5 SUBMISSION OF ELECTRONIC DOCUMENTS

As at other institutes, CERN paper preprints were also being distributed in several hundred copies around the world. The printing shop had started to accept electronic documents but this was done on an ad hoc basis, the authors just sending the files with a message to the person responsible. The printing chain was completely independent of the bibliographic database. The CERN preprints were simply received by the Library after printing just as for other recipients, and catalogued by hand. The next step was to reorganize this procedure.

A submission procedure was set up using Web forms and launched in 1995. After some training, CERN divisional secretaries were able to submit their divisional preprints and reports electronically. The bibliographic information was entered on the Web form, together with the location of the complete set of files for the document. The submission program then copied the relevant files and put them onto our central server.

This was soon extended to other series of documents. One of the problems we had (and still have) is that too many CERN documents were being published without the Library being aware of them, even though this is technically against CERN's rules. Conference proceedings publications were the most common example of this. In the paper era, these were not always distributed to other institutes as preprints. Instead an author would simply write up his presentation and sent it to the editors for publication, without informing us. Theses were another example, being submitted to the university in question often after the student had left CERN, but sometimes being the only place where some critical pieces of information were recorded.

We often managed to pick up the existence of these documents by targeted searches in databases such as INSPEC, but that was usually too long afterwards and no electronic version was linkable this way. So in 1996, the two series CERN-Open and CERN-Thesis were introduced, to which authors could submit their own documents. There was also an "Ext" series to which documents from other institutes could be submitted, either by prior

arrangement with the institute concerned who did their own submitting, or by our own staff that submitted the electronic versions of documents received on paper, when they were able to locate them on the Web.

6 LOCATING THE DOCUMENTS IN THE DATABASE (WEBLIB)

Up until 1982 the CERN Library had used various in-house systems to manage its preprints. From 1983, we used ISIS (UNESCO) and the databases first became searchable by users all over CERN. In 1989 we changed to ALEPH (Ex Libris, Israel), which we are still using for most library functions. In 1995, the first Web interface developed by ALEPH was inadequate for our purposes, so from 1996 we developed our own interface (WebLib) that makes its own API calls to the central ALEPH query processor.

The philosophy of WebLib is to give the possibility of navigating in a tree structure of different collections of documents (at first the traditional library collections like books, preprints, journals, conference proceedings, but now many more). A search can be initiated at any point in the tree, to be performed in the collection defined at that point.

Once the WebLib interface was launched, we began to expand it to other textual documents and even to non-textual "documents". 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 had just been considered outside the scope of the Library when the database was simply the library catalogue. Now it was possible to cover many different electronic collections with a single search interface. So we included the internal notes of some of the larger experiments, the notes used in the 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. Among non-textual collections can be mentioned the CERN Photo collection, the HEP Institutes database, the CERN Historical Archive, videotapes, Webcasts and the objects contained in the Microcosm exhibition (no full text for these objects, some are a few metres high!).

As WebLib expanded into these other areas, the submission procedures were expanded in parallel, so that particular kinds of document could be submitted by the authorized people (password protected). 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 via 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. It can of course be replaced by a revised version, but one cannot compose one's document from scratch inside the system at present.

One of the advantages of our highly modular approach to the WebLib design is that we have great flexibility in developing it along the lines we want. Quite substantial modifications can be made extremely rapidly. Furthermore we now regularly receive enquiries from potential customers in totally different areas who express an interest in using our approach for their own data. The modular approach makes customization extremely easy and we are beginning

to enter into licensing agreements. The CERN Management has expressed the wish that the ETT (Education and Technology Transfer) division in which we find ourselves becomes financially self-sufficient as soon as possible.

7 EXTENSION OF HARVESTING (UPLOADER)

In 1998, we made an agreement with INSPEC to upload CERN records free of charge in order to catch some of the "by-passed" documents mentioned earlier. For this we had another student write a program. We also started uploading from UnCover the publication references of already-existing preprint records using yet another program to match and upload (this one written by an information scientist on a year's visit). It was clear that the number of independent external sources of data was going to increase and that large parts of all these programs were doing the same basic jobs: reading records, converting data into database fields, and uploading records into the database. Only the detailed format description was specific to each source. So, in 1999 a technical student from the Czech Republic (now again with us on his third visit, he has been summer student, technical student, and is now doctoral student) wrote a general UpLoader program with an input module having a simple language in which the precise format of the source data could be expressed. These configurations, as we call them, can now be created and modified by our library preprint staff, without involving the library computing support.

The Uploader program was introduced in February 2000 and we now harvest data on a regular basis from about 100 different sources such as institute home servers and databases. About 50,000 records per year are now uploaded from these sources. Not only that, but we invest a lot of effort into continually enhancing the quality of the existing metadata in the database, using a whole series of programs. The data do not just stay frozen: they are standardized, get enhanced structure and much new metadata (indexing) is also added.

8 ELECTRONIC JOURNALS

Although I have said that the collections of preprints do share many of the attributes of journals, by electronic journals we mean of course the electronic versions of the commercially published journals. The first publisher in our field to launch an electronic journal was the IoP in March 1997, followed soon after by the AIP, one of the major publishers in the particle physics field.

Since we already had a substantial database of preprints containing references to the corresponding journal articles for the core particle physics literature, it was clear that we wanted to access the individual articles directly from the database records and not to have to navigate each time via the journal site using the volume and page information. It was also clear that the manual addition of a URL to each database record would be too cumbersome. It would not scale with increasing quantity. Furthermore, if the URLs of a publisher changed globally (which they sometimes did), all the records would need to be modified.

Instead, we noticed that some publishers had regularity in the definition of their URLs in which the volume and page could be seen. We gave this task to a summer student in 1997

(yes, the same one!) and he wrote an algorithm to generate the required URL on the fly, whenever a user clicked for the full text. All that was then necessary was to maintain a small set of data for each journal defining the volumes that were currently available electronically. We contacted the publishers whose URLs were not of this form, and were able to persuade many of them of the usefulness of this approach. This program was incorporated in the WebLib interface in 1998.

Of course, there are also many other electronic journals of interest to CERN that our database does not cover, and this has increased over the years as more and more journals have gone electronic. To help our users get quickly to the articles they want, we used the same program to calculate the URL once the user has typed in the volume, year and page. This Go-Direct feature is available on the page that leads to our electronic journals. The e-journals can of course also be browsed in the standard way by locating them via title or subject category.

Currently we have access to nearly 1000 electronic journals of interest to people at CERN. Of these about 300 are those to which we have a paper subscription as well. Among the other 700 are many that we would have liked to subscribe to on paper but were prevented from doing by budget constraints (they now come as part of package agreements) and also some of those "old friends" which we had been obliged to cancel in the past due to the same constraints.

Concerning journal subscriptions, we find the situation for some particle physics core journals (such as Physics Letters B) rather bizarre. The community pays large sums of money to build the accelerators and carry out the experiments, and the editors and referees are also researchers in the field who carry out these tasks in their research time. But then we are asked to pay more and more to buy back the results we have produced, without which there would be no journal to publish in the first place.

9 CITATION NAVIGATION AND FULLTEXT SEARCHES

For all electronic documents on our server (now about 200,000) we have extracted the block of references at the end of each article, converted it from PostScript to text and have indexed the complete text of each reference. 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.

Firstly, this database permits the making of citation searches using the journal reference, first author, or in fact any text at all which is in the citation. But the user can also 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 and also any links to e-print archive documents. In fact, it has become commonplace for authors to quote both of these for each reference.

This project was started in 1998 and, after an analysis by a visiting librarian in 2000, we introduced an improved version in 2001. There are about 2 million linkable references inside our electronic documents. The handling of the citations in this way is carried out for all new documents as part of the daily procedures. There are further improvements underway to increase the proportion of references that can be linked to the full text.

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, that is, it cannot be combined with searches in the bibliographic data. What would be really good, for example, would be to combine a search for fragment of full text with a search for an author. There is more to be done in this area.

10 FREELY-ACCESSIBLE PARTICLE PHYSICS ELECTRONIC LITERATURE

If we look at the combined effects of the submission scheme, the UpLoader and the WebLib search engine we see that they have created a virtual library of electronic documents in the particle physics field. In the sense that a journal is also just a collection of documents, this can be considered to be a virtual e-journal. More than that, it covers the whole field of research, whereas a single e-journal only publishes a part of the complete literature.

The typesetting of the e-print articles is of course the responsibility of the authors, because they are not composing their documents inside our system. However since most authors use LaTeX, the layout is quite standardized in practice. The distribution aspect (getting the full text to the user) is handled by the WebLib linking facilities. No navigation inside the publishers' own pages (with their different interfaces) is necessary.

Journals have independent refereeing of the articles they publish. At present, this is clearly an advantage over the kind of refereeing that is done for preprints, since while the latter may well be stringent and fair it cannot be seen to be fair since it is after all internal refereeing. One thing that we want to introduce is a measure of the extent of the refereeing for each document, say on a scale from 0 to 5, which would be stored in the database record so that it could be displayed or even used in searching.

Access to WebLib is about 40% from CERN machines (even though the people themselves may be outside CERN) and about 60% from non-CERN machines. The links to the e-version of the preprint and the e-version of the journal article stand side-by-side. The user can choose whichever he wants. Of course those who do not have access to the e-journal will get the appropriate error message if they try to access the journal article. In a typical month, queries are received from about 11,000 different host computers around the world.

In spite of the open accessibility of e-prints, they are still being published in journals. The inertia of the current system is still pushing it forward. Universities still recognize only journal publications for job candidates and for evaluating the career paths of researchers. Funding agencies also use similar parameters when deciding where to spend their money.

11 FUTURE DEVELOPMENTS

Continuing along the lines described above, we still have a number of projects under development that will further enhance the possibilities.
We are using statistical and linguistic analysis of the complete full text in order to attribute

keywords and keyword phrases to the documents. 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 WebLib, it is now possible to generate a list of DESY Thesaurus terms for any document. This project is currently in its first phase and more development is necessary. We have a Spanish student writing his Ph.D. on this topic.

We will soon be moving the user search interface from the ALEPH-based one to a MySQL one we have developed ourselves. This will give us a better response time and also independence. One of the reasons contributing to this change is the presence of some serious bugs in the ALEPH search engine that causes many rather simple searches to fail completely. We first encountered this problem about three years ago, and it has got worse with time. In spite of the fact that we spent some considerable effort in reverse engineering to locate the origin of the problem rather precisely, ALEPH were never able or willing to correct it. Instead they directed us towards their new version. Unfortunately, this new version is still not stable in the centers that have it, so we still have not been able to make the changeover. Because of the power of Internet search engines like Google, we are planning to create permanent Web pages for each of our documents so that they would be picked up by the various Internet indexing programs. This would solve the problem of users who say "if it's not found by Google, it doesn't exist".

We are developing an encyclopedic 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 fulltext of any document with the description of the term and links to original documents that defined it.

All of these are just extensions to the way in which information in documents can be linked. 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. We have managed to replace most of the classical operations that could be performed manually with the paper literature, such as searching, looking up cited papers, and making citation searches (using a database for this last one, of course). However, 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. I think this is just a carry-over from the time when documents were on paper and had to be self-contained. In the long term, I think there is a much more efficient way of organizing the information. It would be as a sort of massive electronic textbook, where any new piece of work could be hooked on and linked to the relevant places.

12 CONCLUSIONS

We have managed to develop 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.

I think several factors have contributed in allowing us to get where we are today. The most

important is the strict modularity we have imposed on the design. It was not like this from the beginning. In the period 1994-1997 we branched out in many directions at once, developing the functions I have described. But then followed a period of consolidation, which has now led to a second phase of increased development which we expect to continue for a long time.

Then on the personal side we have worked together very well as a team.

We have been able to develop most of our staff to carry out work at a much higher level than their qualifications might have suggested. This has led to an enhancement of job interest and therefore excellent results.

We have had a regular supply of highly able students and other visitors from the Member States, without whom we would not have been able to launch these projects, while at the same time keeping our daily services running.

We have been fortunate to have a superb team of young people giving us the dedicated computing support that we need, in particular implementing a highly modular design where almost every piece of code is used in many different places.

