High Energy Physics Libraries Webzine

Issue 5 / November 2001

http://library.cern.ch/HEPLW/5/papers/2/

From Fulltext Documents to Structured Citations: CERN's Automated Solution

CERN-ETT-2001-003

Jean-Blaise Claivaz (*), Jean-Yves Le Meur (*), Nicholas Robinson (*)

15 Oct 2001

Abstract

For many years, CERN has been collecting all High Energy Physics documents to make them easy to access to physics researchers. A repository of up to 170,000 electronic documents is available via the CERN Document Server [1], the main gateway to CERN's digital library. On top of the creation of this digital archive, the computing support has been looking towards possible improvements in order to ease the task of searching through and reading the articles kept. In addition to the acquisition, cataloguing and indexing of standard metadata, specific treatments have been applied to the data itself.

In this paper, after a brief description of process applied to fulltext data, we shall focus on the specific work done within a collaboration between CERN, Geneva University and the University of Sunderland in order to successfully achieve the automated acquisition of structured citations from fulltext documents.

Introduction

When fulltext documents are received on the CERN Document Server (CDS), either via a direct submission [2] or via a batch upload procedure [3], the standard process is triggered which runs a set of procedures according to the rules shown below. This is completely separate from treatments done directly on metadata.

- First of all, all documents are converted to the Portable Document Format (PDF), whatever their original format may be. Automated conversions have been developed to handle all TeX, MS Word, MS Powerpoint and PostScript documents [4].
- Second, PDF files are archived on the CDS in specific locations. This archiving step enables our link manager [5] to give access to any file, according to its type. Access to a given file can be controlled so that it may be made public, password- or CERN-protected. This archiving step also allows the provision of services, for example GIF views of each page of the document.

- Third, the fulltext is indexed by the CERN ultraseek robot enabling the retrieval of this document by searching any of its text strings [6], which is unique for all HEP eprint servers worldwide.
- Finally, the procedure for acquiring the citations made within the article starts, and this is the step with which we are concerned in this article, and upon which we shall now focus.

Objectives

Automatically enriching the metadata of a document with the set of all citations made within it is interesting for many reasons. Various similar enrichments have been made in the same way in the past. For example, a program developed at the Deutsches Elektronen-Synchroton DESY (called Giva) helps library cataloguers to obtain all authors of a paper by extracting the complete list from the fulltext (which may even be up to several thousand people for large collaborations). The fulltext is also used in a more recent project, which is concerned with acquiring keywords from documents in an automated way [7].

The CERN Citation project is therefore not our first experience of working with fulltexts to enrich the metadata kept about a document. The specific issues to be addressed by the project are that:

1- It is not possible to ask of those submitting documents, the heavy task of submitting references separately (as is done for the abstract). This would be far too time consuming for them.

2- The position and format of references within a document is almost unpredictable, due to the huge variations in the ways in which different authors cite documents.

3- A complete and well structured database allows the development of a large number of useful applications and studies.

If the first point above is obvious (we propose as an option, the inclusion of citations when submitting documents to CERN Document Server and ... we never get them that way!), the second point is being improved thanks to a SLAC initiative, pushing HEP authors to use a standard format for citations [8]. This is also difficult to achieve, in particular at CERN where authors from very large world-wide institutes may not be easily convinced to follow a "citation-writing" policy. The third point will be detailed later, with its three main consequences: improving the calculation of eprints' impact, enabling one to search terms appearing in references, and allowing the navigation to cited documents.

Background

Of course, CERN is not the only organisation to be interested in the exploitation of document citations.

Beyond the task of isolating "well"-defined references, the automated linking of related information within collections of documents is being investigated in many places. Let us quote here a few initiatives, such as the LIGHT project, started at CERN and taken over by industry [9], the SFX technology, proposing a solution for reference linking [10], the Crossref repository [11] using the DOI identifiers [12] and the S_Link_S XML based linking system [13]. These are various approaches to help semi-automated linking between documents, based on meta-data and link management technologies. They do not focus particularly on the references written by authors in their documents.

Still, in the HEP domain where there is a long tradition of eprints (research papers freely available), various initiatives exist with the same and precise goal to reach a comprehensive coverage of eprints citations. Mentioned below are some of the projects in this field, but some may have been omitted due to its extremely active nature.

1- Science Citation Index, by Institute for Scientific Information:
The Science Citation Index [14] keeps citations of papers in virtually all scientific journals (not just physics) since 1982. It is accessible only to subscribing institutions, either electronically or in paper form. Academic libraries often subscribe to this professional tool. It does not cover papers which are not published (conference articles, etc.) and it is not free of charge.

2- SLAC:
On top of the (Los Alamos, then Cornell) ArXiv.org eprint archives[15], SLAC has built a database of references and a search system enabling the counting and ordering of the most cited papers [16]. The following is the warning that prevents abusive interpretation of the results.
"The citation search should be used and interpreted with great care. At present, the source for the citation index in the HEP database is only the preprints/eprints received by the SLAC Library, and not the (unpreprinted) journal articles. Citations to a paper during the months it was circulated as a (non-eprint) preprint may also be lost, because only references to journal articles and e-print papers are indexed. Still, the citation index in HEP (SPIRES-SLAC) is formed from an impressive number of sources. For example, in 1998, the citation lists were collected from almost 14,000 preprints."

3- OpCit:
This project, "Reference Linking and Citation Analysis for Open Archives" [17], is a collaboration between Southampton University, Cornell University and the Los Alamos National Laboratory. Among other goals, one of them was to enrich fulltext documents from the ArXiv.org mirror site of Southampton with all references linked inside the PDF files and also to derive rules related to the impact of eprints [18].

4- Research Index, from NEC Research Institute:
ResearchIndex is a digital library that aims to improve the dissemination, retrieval, and accessibility of scientific literature. Specific areas of focus include the effective use of the Web, and the use of machine learning. Autonomous Citation Indexing (ACI) [19] automates the construction of citation indexes (similar to the Science Citation Index (R)). It has a more general scope than High Energy Physics and it is not based on a special collection, as documents can be directly retrieved from the Web.

5- CERN
It may look as though the work carried out at CERN is redundant with the projects mentioned above! Actually, the first analysis and development started at CERN was in 1994, with the first construction of a "CIT" database, containing only raw references of electronic documents for which the automatic parsing was successful. The interesting feature of this database [20] is that it is possible to look for any term (author names, report number or title) and obtain details of the papers that use this term within their references. It is complementary to the SLAC citation system where not all of the text is kept and indexed, but only the pointer to the corresponding article. Another difference between CERN and SLAC citation treatment is that while SLAC is pointing the references to its own database (where the preprint and its publication information are available), CERN decided to link journal article references directly to the e-journal site, whenever available to CERN members.

The scope of the project covers all of the CERN Document Server, which not only contains documents from ArXiv but also many CERN preprints, internal and scientific notes.

At CERN, no human resource for manual editing was allocated to the project for the long term, making mandatory the building of a more and more complete and complex acquisition algorithm. A new step was reached this summer 2001 thanks to a successful collaboration with a librarian from the University of Geneva (CH) and a student in computing from the University of Sunderland (UK). The reference analysis, the technological choices and the software have been deeply studied and completely renewed.

This is described in detail in the next part of this article.

Methodology and Techniques for the Creation of Links to the Citations Made in a Document

The creation of links from the references of a preprint is a quite complex process that may be divided into three phases: first, the extraction of the reference section from the article text; second, the recognition of citations; and finally, the linking to the cited source.

The Reference Extraction Process

This is the first phase in the process of acquiring structured citations from a fulltext document. It consists in itself of three main stages:

- Conversion from PDF document to plain text format.
- Extraction of the references section.
- Rebuilding individual reference lines that may have been broken due to line wrapping.

1- Conversion from PDF document to plain text format

As already explained, upon receipt of a new document on the CERN Document Server, a conversion is made from its original format to the PDF format. This is mainly done because these files are of a platform-independent nature and because they are relatively small in file size allowing space on servers to be used more efficiently. They are therefore ideal for viewing by users of the service, but not when it comes to extracting references. At that point, it becomes necessary to have a more simple form of file to search through: the plain text format. Even if PDF is of a complicated nature, not organised in a linear fashion (as read by humans), but rather as a series of reference tables that point to different byte locations at which the various objects that make up the file are stored, it was deliberately chosen as a format from which to create the plain text document due to its stability over other file formats such as PostScript. Many conversion tools were tested to find the most suitable one, and the « pdftotext » tool from Foolabs [21] was eventually chosen. A paper about the comparison study of these different tools with their advantages and drawbacks has also been released [22]. Having converted the document into a plain text format, the extraction process can be started.

2- Extracting the references section

The way the extraction script works is easy to understand: starting from the very last line of the document, the program scans it upwards, searching for the beginning of the references section, usually indicated by words such as 'References', 'Bibliography', etc. Having found this references section title, the script reads down the text and extracts all the lines until it encounters the next section title (e.g. 'Figures') or until it reaches the end of the document. If no references section title

can be found, a second scan is done, this time looking for the first two reference lines, but only when those lines are numbered with 'square brace' style:

```
[1] .....
[2] .....
```

The reason is that it is a fairly safe assumption that when a line beginning with '[1]' is followed (within a few lines) by a line beginning with '[2]' the references section has been found. The same cannot be said however for other styles of line numerator such as '1.', as they are far more commonly found within the document body.

Often, the references section in a document is quite large and it can be split across several pages. If the document contains headers and/or footers, this can result in their accidental inclusion in the extracted references when lines are rebuilt, sometimes breaking up the information of a given citation instance, thus causing recognition problems. To overcome this problematic situation, the script attempts to match the patterns created by headers and footers around the page break (Form Feed) characters in order to remove these unwanted lines that are inserted into the document during the conversion to text at the point of each new page. Having recorded the line number of each page break character, the program tests for lines above and below each page break character line that are effectively the same. Similar lines above a page break character can signify the footer of the previous page, and lines below a page break character can signify the header of the current page. Perhaps this is best explained with an example. The following text shows the end of one page, followed by a page break character, followed by the next page's contents and its footer. It has used some of the real citations taken from the references section of this document.

```
[9]  Preparing the LaTeX List of Publications from the SPIRES BibTeX output.
http://www.slac.stanford.edu/spires/hep/bibliotex.html


Page 8
<FORM FEED>
Le Meur, JY et al. From Fulltext Documents to Structured Citations: the CERN Treatment

[10] LIGHT project, http://light.cern.ch/
.
.
Page 9
<FORM FEED>
Le Meur, JY et al. From Fulltext Documents to Structured Citations: the CERN Treatment
.
```

It can clearly be seen from the above that for each page break line there should be a page number (Page X) on the line above it, and a document title (Le Meur, JY et al. From Fulltext Documents to Structured Citations: the CERN Treatment) on the line below it. The program would be able to detect this pattern, and thus be able to remove this unwanted information. The technique is not perfect however. Often in documents, authors place titles for each section of a document in the page headers. This would mean that the page headers would differ throughout the document, and the program would not therefore be able to identify the headers - they must effectively remain fixed throughout the document. The point here, however, is that nothing is lost from this process when it fails, but when it succeeds, much is gained from it.

The process is also often able to prevent other recurring information traps, such as the repeated presence of the word "References" in the document (for example in the current title of the chapter). However, limitations are clear. A well formatted preprint with clearly indicated chapters and numbered references leads to a good result. On the other hand, if the document has bad pagination with figures inserted in the middle of the references section, which often occurs, then the result becomes unpredictable.

3- Rebuilding individual reference lines that may have been broken due to line wrapping

Having located and extracted the references section from the document body, there still remains one necessary task before the process of recognition of the cited items can begin. As is apparent to anyone who is familiar with academic papers, there are often many reference lines present in a document. Often, a given reference line can be rather long, as it can cite several documents, or simply details a large title or many authors of a paper. When a large line such as this is viewed on screen, it is broken across one or more lines of text, depending upon how long it is and how large the 'canvas' size of the document is set to be. When a human reads such a line, it is apparent to them that this reference line that stretches across several actual lines of text, is actually only one true line - it has simply been wrapped for convenience sake, as it would stretch off the printable boundaries of the page were it all shown on one line. Take for example, the following reference line:

[11] H. Van de Sompel, P. Hochstenbach. "Reference Linking in a Hybrid Library Environment, Part 2: SFX, a Generic Linking Solution" D-Lib Magazine (April 1999), http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt2.html

It is clear that it is really only one line, but has been broken for convenience sake. However, a computer program is not aware that this is really supposed to be one line, because during the conversion-to-text process, carriage return characters are inserted at the point at which lines are wrapped. Carriage return characters are also inserted at legitimate line break points as well, however, which means that it is not possible to determine between what is really a line wrap, and what is a genuine line break. This means that long reference lines are broken during the conversion-to-text process. At first sight, this breaking of long lines may not seem to be a problem. However, the process for the recognition of citations within reference lines operates on a line by line basis. This means that an accidentally broken reference line will be considered as more than one reference line by the citation recognition process. Consider the broken reference line shown above. It can be seen that the date of the article has been separated from the title. This would mean that the citation recognition would fail due to the absence of the date in the citation. In short, it is necessary to rebuild broken reference lines so that citation information is not destroyed and thus lost.

In order to rebuild reference lines, two main cases are considered. The first case is when the reference lines have no form of markers to identify the start of a new reference line. This is the most difficult case in which to rebuild the reference line, and basically involves attempting to use blank lines between reference lines (if present) to rebuild reference lines, of co-ordinate information taken from a PostScript version of the document (if available) to rebuild the individual reference lines. In this situation, if the reference lines cannot be rebuilt correctly, they are all simply joined to form one very large reference line. This solution is perfect from the view of citation instance recognition, as the recognition process simply receives one large line, and searches through it for citation information. However, from the point of view of human reading of this line, the solution is messy, as it is difficult to search through a huge block of text for one single citation item. It is not a terrible situation, however, as large lines can be split up into smaller, more manageable lines at display time.

The second situation is when the reference lines start with markers of some description (such as '*', '[1]', '(1)', etc.). In this case, the program can simply join lines together until encountering a line beginning with the identified marker type, at which point it can make a split, having identified the start of a new reference line. At this stage, having rebuilt all reference lines, the process of identifying and tagging cited items within the lines can begin.

The reference extraction script was run during summer 2001, extracting the backlog files of fulltext documents starting in 1994. From 102,530 PDF files stored in the CERN database, 94,221 references sections were extracted, reaching a total of 91.90% success. Non-English articles or documents without a references section were also counted as failures.

The Recognition of Citations Within Reference Lines

There are essentially three types of citation item recognised by the process at this stage. These citation types are the internet address of the item (the URL), the report number of an item, and the title of the journal/periodical in which the item appears. The details of each of these items are recognised and recorded so that the information can be used.

1- Recognising Internet Addresses

Since the World Wide Web has become a very common research medium for academics, due to the wide range of information available from it both easily and immediately, it has become more and more common for authors to include the internet addresses of items that they have cited. This is often done in one of two ways:

- The internet address of the item is simply given in its plain, unmarked-up form, e.g. "**http://www.cern.ch/public/**". This is undoubtedly the most common way of writing the internet address of a cited item.
- The internet address of the item is written in its HTML-tagged form, e.g. "**The CERN Web site**". This is a fairly rare form of writing the internet address of a cited item, but it has been encountered on occasion. It is believed that the author has given thought to the fact that his paper will one day be represented in HTML format, and wanted the links to the cited items to be 'active' in their HTML form, as opposed to simply ordinary text.

Internet addresses are easily recognisable, using a form of pattern matching. It is possible to search for the HTML anchor tags surrounding something which resembles an internet address, or simply to search for the plain internet address format (e.g. http://....., or ftp://....., etc.). When a citation to an internet address is found with the HTML tagging information, both the address and the description ("**The CERN Web site**" in the above example) can be recorded. When it is simply a plain internet address however, obviously only the internet address is recorded, as there is no description present.

2- Recognising Report Numbers

Report numbers are composed of two parts, the root and the numbering. The root is usually the name of the broadcasting institution or one of its subdivisions. The numbering is generally formed with the year and a current number. The script focuses on the roots that should match with a list of occurrences defined for our needs. For each matching root, it looks then for numbering. If

successful, it standardizes the output by inserting a hyphen between each element as decided upon by the CERN librarians. The following three report numbers give an example of this:

- CERN-ETT-1999-123
- SLAC-PUB-6100
- hep-th-0110005

When a report number is recognised by the program, its details are recorded for later use.

3- Recognising Journal/Periodical Titles

The biggest step during the citation recognition process is that of recognising the Journal/Periodical title details. The reason for this is that more elements must be taken into account. In order to make a link to an item cited in a Journal/Periodical, it is necessary to know the Journal/Periodical title, the series name, the volume, the year of print, and the pagination information. The title of the Journal/Periodical is identified using a knowledge base of Journal/Periodical titles that CERN either knows about, or subscribes to. This knowledge base is of a rather large nature, as for each Journal/Periodical title, there is a standardised way of writing the title, along with many non-standard ways of writing the title. Unfortunately, authors seem quite unaware or unconcerned with the standard ways of writing the titles, so it is necessary to record the non-standard forms of titles that have been encountered, along with the standard. This way, when a non-standard title form is encountered by the program, it can be recognised and replaced by a standard form, which is then recorded for later use. The standard form of writing a title is set out by the ISO 4 standard. Currently, about 1800 entries are kept in the titles knowledge base for about 800 actual titles. In addition to the recognition of the title information, it is of course necessary to recognise the series name, volume, year of print and pagination information (collectively known as the numeration information). This numeration information is identified and recorded by the use of several standard models. Unfortunately, with this numeration information, there are many non-standard ways of writing it (writing the components of it in different orders), and it is necessary to use the different models to recognise these non-standard forms, and change them into a standard format - the sequence "**volume (year) page**". Notice that there is no punctuation in this standard form.

The matching against this list allows the script to recognise and standardise the titles with a high reliability. Take for example, the following:

- JHEP 12, 139 (2000)
  ...which is transformed into:
- J. High Energy Phys. : 12 (2000) 139

Another stage of recognition identifies the word "ibid" used in place of a periodical title and transforms it by the referred full title if possible. This is because often a single reference line can contain citations to many documents, and when an author has cited an article in a given Journal/Periodical title, and then goes on to cite another article in the same Journal/Periodical, they simply use the word "ibid" instead of re-writing the title. Replacing this word "ibid" with the actual title intended, allows a great many implicit references to be turned into explicit ones, which enables their linking (10,000 of 15,000 in the CERN database). An example of the transformation of these "ibid" occurrences is shown below (the "ibid" transformation has been emphasised):

- [2] W.H. Zureck, Phys. Rev. D 24, 1516 (1981); W.G. Unruh and W.H. Zureck, **ibid.** 40, 1071 (1989)
  is transformed into

- [2] W.H. Zureck, Phys. Rev., D : 24 (1981) 1516; W.G. Unruh and W.H. Zureck, **Phys. Rev., D** : 40 (1989) 1071.

4- Recognition Results

Among the 94 211 eprints for which the reference section extraction was successful during summer 2001, almost 3 million (2 896 541) individual citations were collected and the recognition process led to the following:

Total No. Citations	**Total No. Recognitions =**	**Internet Addresses +**	**Report Numbers +**	**Journal Titles**
2 896 541	2 329 286	12 684	379 440	1 937 162
-	80.42 %	0.44%	13.10%	66.88%

These results give a good picture of the trends followed in terms of citing habits in HEP over the past seven years or so. It can clearly be seen that journal articles are massively quoted (~70%), URLs are very rarely provided (0.44%), and report numbers are often used (~13%). A closer look at report number recognition shows an obvious increase in the use of report numbers in citations throughout the years: during the years 1994 to 1999, 11.3% of citations contain report numbers, whereas during the years 2000-2001, this number has reached 18.6%. It is probable that this results from the increasing availability of free literature (eprints to which authors refer with report numbers).

Probably, the most important factors of this result are that we have now the opportunity to:
- create **379 440 links** from the references to the eprints
- create **12 684 links** with the provided URLS
- create up to **1 937 162 links** to the journal articles, if available on-line, or to their corresponding eprint, if available on the CERN Document Server.

The Linking of Citation Items to Their Respective Fulltexts

Data extracted during the recognition phase are registered in MARC sub-fields [23] of the bibliographical record of the document. The repetitive field 909 is retained as appropriate for CERN needs, as the MARC standard has not foreseen this case. So, for each reference discovered a field 909 with all necessary sub-fields is added to the record. The $x sub-field will contain, for example, the internet address of the cited electronic resource. It does not mean this URL necessarily exists. Some articles are not yet numbered and others can not be seen free of charge. In any case, all elements of the cited items are identified, and the linking is possible without any new processing of the text.

- Nucl. Phys., B : 528 (1998) 185
  has for sub-field $x
- http://www.elsevier.nl/IVP/05503213/528/185/

The actual creation of the links to the fulltext documents is done by another script that generates the URL [24], taking into account the periods of coverage of periodicals (existence of electronic articles), the state of CERN subscriptions to the various journals, and the parameters of connection

defined by the editors. Update of this data makes it necessary to relaunch the program that calculates all the URLs and inserts them in the HTML of each reference.

Examples of Citation Recognition At CERN

The following examples show that there has been a great improvement in the success rate of cited items recognised within documents. The first box shows the results of the older, (now obsolete) citation recognition process that was used at CERN until the creation of the current system. The second box shows the successes of the newer citation recognition system.

Older CERN Citation Recognition System:

[43] I. Lauer et al., Phys. Rev. Letters 81, (1998), p. 2052 [13] M.B.H.Breese, Nucl. Instrum. Methods Phys. Res., B : 132 (1997) 540 [2]W.H. Zureck, Phys. Rev. D 24, 1516 (1981); W.G. Unruh and W.H. Zureck, ibid. 40, 1071(1989) [6] J.L. Feng, Int. J. Mod. Phys. A13, 2319 (1998) [hep-ph/9803319] [47] D. E. Kaplan and A. E. Nelson, J. High Energy Phys. : 0001 (2000) 033 [arXiv:hep-ph/9901254]

Citations Recognised by the Recently Developed Citation Recognition System:

[43] I. Lauer et al., Phys. Rev. Lett. : 81 (1998) 2052 [13] M.B.H.Breese Nucl. Instrum. Methods Phys. Res., B: 132 (1997) 540 [2] W.H. Zureck, Phys. Rev., D : 24 (1981) 1516; W.G. Unruh and W.H. Zureck, Phys. Rev., D : 40 (1989) 1071 [6] J.L. Feng, Int. J. Mod. Phys., A :13 (1998) 2319 - hep-ph/9803319 [47] D. E. Kaplan and A. E. Nelson J. High Energy Phys.: 0001 (2000) 033 [arXiv [hep-ph/9901254]

The end result of this project can be seen on the Web as it has now become part of the CERN Document Server. From the CDS Search interface (http://weblib.cern.ch/), when retrieving eprints, the "Show References" links point to these automatically linked references.

Conclusion: Exploitation/Future

Considerable time and effort has been invested in the area of reference extraction and recognition at CERN. The result has been the development of a system that is considerably more robust, and able to extract references from many more documents than was possible under the older system. Under this new system, the number of cited items within given reference lines has also been greatly increased, resulting in a better situation for both CERN and its users. The citation information that has been extracted is organised in a very structured and efficient manner. Not only does this make for better system performance and less wasted space in the CERN bibliographic information databases, but it also allows the exploitation of the data in order to provide many more services to the user.

These services are now being investigated. Possible examples of them are:

- Allowing users to easily jump from one document to another.
- Allowing authors to see statistics relating to how many people have been making citations to their papers (this comes as an add-on to the option of knowing how many people have read a given document [25]).
- Allowing users in general to view statistics about the most cited papers and authors, etc.
- Allowing users to search any text string that can appear within reference sections.
- Providing CERN library staff with information about which journals are cited most by physicists, and thus possibly helping with administrative decisions, such as whether or not subscriptions to certain journals should be pursued.

With the future will certainly come greater volumes of papers submitted to CERN. It seems that, today, authors are writing their documents in a wider range of document formats such as TeX, MS Word, etc. It is beneficial that this system can simply rely upon the fact that all files will be converted to the PDF format before their references are extracted, thus allowing references to be extracted automatically from a greater number of documents than in the past. Hopefully, this new system will serve well towards extracting and storing the references of documents at CERN, as well as bringing an increase in the quality of service offered by CERN to its users.

References

[1] CERN Document Server, http://cds.cern.ch/
[2] T. Baron, J-Y Le Meur. "From Desk to Web: the Electronic Document Submission", presentation at CERN, Aug 1999
[3] N. Pignard, I. Geretschläger, J. Jerdelet. "Automation of electronic ressources in the Scientific Information Service at CERN", High Energy Phys. Libr. Webzine : 3 (2001) 003
[4] D. McGlashan, J-Y. Le Meur. "The CERN Conversion Service", presentation at CERN, Aug 1999
[5] D. McGlashan, J-Y Le Meur. "SetLink: the CERN Document Server Link Manager", CERN-ETT-2000-001, High Energy Phys. Libr. Webzine : 1 (2000) 1
[6] Fulltext searching of HEP eprints, http://weblib.cern.ch/Fulltext
[7] D. Dallman, J-Y Le Meur. "Automatic keywording of High Energy Physics", CERN-AS-99-005, 4th International Conference on Grey Literature : New Frontiers in Grey Literature, Washington, DC, USA, 4 - 5 Oct1999 / Ed. by Farace, D J and Frantzen, J - GreyNet, Amsterdam, 2000. - pp.230-237
[8] Preparing the LaTeX List of Publications from the SPIRES BibTeX output. http://www.slac.stanford.edu/spires/hep/bibliotex.html

See also: H. B. O'Connell. "Physicists Thriving Paperless Publishing", physics/0007040
[9] LightObject, sharing knowledge, http://www.lightobjects.com/
[10] H. Van de Sompel, P. Hochstenbach. "Reference Linking in a Hybrid Library Environment, Part 2: SFX, a Generic Linking Solution"
D-Lib Magazine (April 1999)
[11] Crossref, The central source for reference linking, http://www.crossref.org/
[12] The Digital Object Identifier System, DOI, http://www.doi.org/
[13] E. Hellman. ScholarlyLink Specification Framework (S-Link-S), http://www.openly.com/SLinkS/
[14] Science Citation Index, http://www.isinet.com/isi/
[15] arXiv.org e-Print archive, http://arxiv.org/
[16] SLAC Spires Collection of References, http://www.slac.stanford.edu/spires/hep/references.html
[17] Reference Linking and Citation Analysis for Open Archives, http://opcit.eprints.org/
[18] P. Ginsparg, J. Halpern, C. Lagoze, S. Harnad, W. Hall, L. Carr. "Integrating and navigating eprint archives through citation-linking: the open citation (OpCit) linking project"
[19] S. Lawrence, C. Lee Giles, K. Bollacker. "Digital Libraries and Autonomous Citation Indexing", IEEE Computer, Volume 32, Number 6, pp. 67-71, 1999
[20] The CERN CIT database, References in E-prints, http://weblib.cern.ch/Home/References_in_E-prints/
[21] A PDF Viewer for X. http://www.foolabs.com/xpdf/
[22] N. Robinson. "A Comparison of Utilities for Converting from PostScript or Portable Document Format to Text", CERN-OPEN-2001-065 ; Geneva : CERN , 31 Aug 2001
[23] MARC 21 - MARC formats for the 21st century
[24] E. Lodi, M. Vesely, J. Vigen. "Link managers for grey literature", 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
[25] Direct Impact: Request on the frequency of download for documents on CERN Document Server, http://doc.cern.ch/impact/

Author Details

Jean-Blaise Claivaz
Geneva University, Switzerland
***Email**: Jean-Blaise.Claivaz@adm.unige.ch*
Jean-Blaise Claivaz works at present at the University of Geneva for the libraries coordination Service, where he is mainly responsible for the on-line electronic resources (journals and databases). He also participates in the project "Cybertheses", which is making doctoral theses defended in Geneva available on the Web.

Jean-Yves Le Meur
CERN, European Organization for Nuclear Research, Switzerland
Email: Jean-Yves.Le.Meur@cern.ch
Jean-Yves Le Meur works at CERN in the Education and Technology Transfer division to develop and promote the CERN Document Server: a set of databases and Internet & Intranet software for storing and delivering scientific documentation to the HEP community. He is the Deputy Group Leader of the Document Handling Service leading the Photo, Video, Desktop Publishing, Printshop and CDS sections.

Nicholas Robinson
University of Sunderland, UK

***Email**: nicholas.robinson@sunderland.ac.uk*
Nicholas Robinson is currently studying in the final year of a BSc Computing course at the University of Sunderland in the UK. As part of
the third year of this course, he spent fourteen months working as a technical student with the CDS team at CERN. His final year of studies at
university includes work in areas such as Artificial Intelligence, Object-Oriented Development and Software Engineering Methods.

For citation purposes:
Jean-Blaise Claivaz, Jean-Yves Le Meur, Nicholas Robinson, "From Fulltext Documents to Structured Citations: CERN's automated Solution", High Energy Physics Library Webzine, issue 5, November 2001
URL: <http://library.cern.ch/HEPLW/5/papers/2/>