Using MlBibTeX to Populate Open Archives*

Jean-Michel HUFFLEN
LIFC (EA CNRS 6942)
University of Franche-Comté
16, route de Gray
25030 BESANÇON CEDEX
FRANCE
jmhufflen@lifc.univ-fcomte.fr
http://lifc.univ-fcomte.fr/home/~jmhufflen

Abstract

After a brief recall of the notion of open archive, we recall that in France, an 'official' open archive's repository is used by the Government to evaluate public laboratories and study how to budget them, so this repository should be updated as soon as laboratory members' papers are accepted. We show how we use MlBibTeX to update both bibliography files and this repository.
Keywords BibTeX, MlBibTeX, XML, XSLT, open archive, HAL.

Streszczenie

Po krótkim przypomnieniu znaczenia pojęcia „Open Archives" zwrócimy uwagęna to, że we Francji rząd używa „oficjalnego" repozytorium Open Archives do oceniania publicznych instytucji naukowych i na tej podstawie przyznaje im budżety. Tak więc repozytorium powinno być aktualizowane bezpośrednio po zaakceptowaniu publikacji pracownika instytucji. Pokażemy, jak MlBibTeX jest używamy do aktualizowania zarówno plików bibliograficznych jak i repozytorium.
Słowa kluczowe BibTeX, MlBibTeX, XML, XSLT, archiwa otwarte, HAL.

0 Introduction

For a decade, Digital Edition and Open Archives have deeply changed the ways of publishing books and journals, especially about research works. Now, such a work can be immediately available as soon as it is validated. This point may be crucial, in particular within domains like medecine: a new treatment can be immediately known as soon as it has been proven efficient. Readers interested in the *modus operandi* of sites devoted to open access archives can consult the manual of the EPrints program [2], frequently used to manage such sites. Aspects related to law and organisation are described — in French — in [1].

As another 'official' point related to using open archives in France, the AERES[1], the governmental agency in charge of public research decided to grade French institutions for research by looking into the HAL[2] site of open archive. HAL is a multi-disciplinary open access archive for the deposit and dissemiation of scientific research papers — whether they are published or not — and for PhD dissertations.

Since 1996, our institution, the LIFC[3], has managed bibliography data bases[4] as .bib files usable by the BibTeX bibliography processor [8]. The challenge was to enter as many bibliographical entries as possible into HAL, from our .bib files, with particular attention to recent entries. In Section 1, we make precise the specification of this task, then Section 2 explains how we proceed. Reading this article only requires basic knowledge about BibTeX, MlBibTeX[5] and XML[6].

* Title in Polish: *Zastosowanie MlBibTeXa do zasilania danymi* Open Archives.

[1] ***A**gence d'**É**valuation de la **R**echerche et de l'**E**nseignement **S**upérieur*, that is, 'agency evaluating research and university courses'.

[2] ***H**yper-**A**rticle en **L**igne*, that is, 'hyper-article on-line'. This site is located at http://hal.archives-ouvertes.fr/.

[3] ***L**aboratoire d'**I**nformatique de l'université de **F**ranche-**C**omté*, that is, 'Laboratory of Computer Science of the University of Franche-Comté'.

[4] Located at our institution's Web site: http://lifc.univ-fcomte.fr/lifc/publications/.

[5] **M**ulti**L**ingual BibTeX.

[6] e**X**tensible **M**arkup **L**anguage. Readers interested in an introductory book to this formalism can refer to [9].

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<HAL xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation="hal.xsd">
  <CONNEXION LOGIN="..." PASSWORD="..."/>
  <ARTICLE_RECENT>
    <META_ART>
      <REFERENCE_BIBLIO>
        <COMM_ACT>
          <TITOUV>Intelligent Tutoring System conference</TITOUV>
          <AUDIENCE>2</AUDIENCE>     <!-- International scope.     -->
          <DATEPUB>2002</DATEPUB>
          <PAGE>31-40</PAGE>
          <EDCOM>Springer-Verlag</EDCOM>
          <EDSCI>S. A. Cerri and G. Guardêres and F. Paraguaçu</EDSCI>
          <SERIE>LNCS</SERIE>
          <TITCONF>Intelligent Tutoring System conference</TITCONF>
          <DATECONF>2002</DATECONF>
          <VILLE>Paris</VILLE>
          <PAYS>FR</PAYS>
        </COMM_ACT>
      </REFERENCE_BIBLIO>
      <DOMAIN>...</DOMAIN>
      <TITLE>
        Social Network Analysis Used for Modelling Collaboration in Distance Learning Groups
      </TITLE>
      <ABSTRACT>
        We describe a situation of distance learning based on collaborative production occurring
        within groups over a significant time span. For such a situation, we suggest giving
        priority to monitoring and not to guiding systems. [...]
      </ABSTRACT>
      <LANGUE>en</LANGUE>
      <REF_INTERNE>rc02:ip</REF_INTERNE>
    </META_ART>
    <AUTLAB>
      <AUTEURS>
        <AUTEUR><LABIDS>3199</LABIDS><NOM>Reffay</NOM><PRENOM>Christophe</PRENOM></AUTEUR>
        <AUTEUR><LABIDS>3199</LABIDS><NOM>Chanier</NOM><PRENOM>Thierry</PRENOM></AUTEUR>
      </AUTEURS>
      <LABORATOIRES>     <!-- The institutions known by HAL are given a LABID identifier.     -->
        <LABORATOIRE LABID="3199"/>
      </LABORATOIRES>
    </AUTLAB>
    <DEPOTS>
      <RIGHT>1</RIGHT>
      <FULLTEXT>
        <DEPOT NOM="http://lifc.univ-fcomte.fr/RECHERCHE/P7/pub/ITS02/reffayIts.pdf" FORMAT="PDF"/>
      </FULLTEXT>
    </DEPOTS>
  </ARTICLE_RECENT>
</HAL>
```

Figure 1: Metadata handled by HAL.

1 Tasks' specification

When an end user enters the information related to a new bibliographical item into HAL, the address of a file containing the corresponding work must be provided — in which case this file is downloaded by HAL if entering this item succeeds — unless there is some public access — a URL[7] — to this resource.

[7] **U**niform **R**esource **L**ocator.

This file may be a PDF[8] or PS[9] file, other formats such as LaTeX or RTF[10] are accepted[11]. Some metadata — e.g., the work's title and its co-authors — must be provided. This operation may be performed by means of a graphical interface, or by sending an XML file grouping all these metdata. Such a file must be conformant to a *schema* — expressed using XML Schema [12] — given in [3]. Fig. 1 is an example of such an XML file suitable for HAL.

At first glance, most of these metadata could be deduced from information stored in the entries of .bib files. However, what is missing can be difficult to express in BibTeX. For example, an author must be affiliated to an institution. More precisely, HAL distinguishes a *recent* article — *all* the authors must be affiliated to an institution — and a '*retro*' article — the rule only holds on for one of the co-authors. In addition, authors may have written an article when they were members of our institution, and may have written another article when they were not, in which case this article is written in collaboration with other members of our institution.

Other information should be supplied for articles included in journals or presented at conferences: does the journal have an editorial board? does the conference provide proceedings[12]? Another information related to *scope* must be supplied: is a journal's scope or a conference's *national*, that is, limited to one country, or *international*[13]?

Last be not at least, the ADDRESS field of a BibTeX entry does not have a precise format. For a conference, it is mandatory for HAL to make precise the country, the town being optional. To sum up, most of information needed by HAL is included into .bib files, but converting .bib files into files suitable for HAL is not immediate.

2 Using MlBibTeX and XSLT

The .bib files used for the laboratory's bibliography were already split into 'national' conferences, international conferences, 'national' journals, etc. First we asked authors to add either an URL field for works publicly available or a PDF field for other works, in which case this PDF field gives the local address of the document. An abstract should be given, too, but the entries of our .bib files already included an ABSTRACT field[14].

Let us recall that we developed MlBibTeX [4], a re-implementation of BibTeX with particular focus on multilingual features. When MlBibTeX parses a .bib file, it results in an XML tree. Besides the nbst[15] language, used by the bibliography styles taking as much advantage as possible of MlBibTeX's features is close to XSLT[16], so it should be suitable for deriving XML texts using HAL format. Unfortunately, there was something missing in nbst: reading additional information from another XML file, as performed by the doc function in XQuery [6]. In addition, nbst is close to XSLT 1.0 [10], whereas some features of XSLT 2.0 [11] — using sequences of values [5] — were interesting for our task.

That is why we put a 2-step process into action: first, a .bib file is converted into an XML tree by means of MlBibTeX; second, this XML tree is processed by the Saxon program [7], Michael Kay's processor of XSLT 2.0. By comparison with 'original' MlBibTeX, some additional checks are performed about the ADDRESS field. Allowed values are:

- ⟨*country name*⟩
- ⟨*town name*⟩,⟨*country name*⟩
- ⟨*town name*⟩,...,⟨*country name*⟩

where ⟨*country name*⟩ can be given in French or English. The second step uses three additional XML files:

- the first file groups all the members of the laboratory, making precise the corresponding period of time:

```
<person-db>
  <first>Jean-Michel</first>
  <last>Hufflen</last>
  <period>
    <from year="1994" month="sep"/>
    <!-- Membership's end is specified by a to
         element.
      -->
  </period>
</person-db>
```

- the second file is used for the affiliation of external people appearing within .bib files;
- the third file is used to specify the scope of journals and conferences:

[8] **P**ortable **D**ocument **F**ormat.
[9] **P**ost**S**cript.
[10] **R**ich **T**ext **F**ormat.
[11] If no such file can be provided, the article can be only known by HAL as a *note*.
[12] For example, BachoTeX conferences provide proceedings. As a counter-example originating from TeX's conferences, the meetings of the German-speaking group DANTE (***D**eutschsprachige **AN**wendervereinigung **TEx**, e. V.*) do not.
[13] In fact, this scope information may be left unspecified. But let us not forget that these informations will be used to evaluate researchers: it is preferable for them to make precise that the scope is international, whenever accurate.
[14] Let us recall that when BibTeX builds a 'References' section, the fields not involved in this process, that is, unknown by the bibliography style used, are ignored.
[15] **N**ew **B**ibliography **ST**yles.
[16] e**X**tensible **S**tylesheet **L**anguage **T**ransformations.

```
<journal>
  <name>Biuletyn GUST</name>
  <scope level="3"/>
  <!--  ('3' is for 'national'.)      -->
</journal>
```

This last part required huge work. In fact, .bib files were populated 'manually', and the conventions for journals' and conferences' names were not uniform. A first pass allowed us to collect all the names, then we decided about a 'canonical' name for each journal and conference in a second step. Finally, we had to ask people involved in these journals and conferences whenever we did not know the corresponding scope.

3 Conclusion

The final action is planned for the middle of May 2010. Presently, tests are promising. We decided to enter recent articles as many as possible, that is, if we can find all the affiliations. Otherwise, that is a 'retro' article, provided that it has been written prior to 2006. Whilst we are performing the last tests, we encourage people to put the PDF files of their articles on the laboratory's central machine.

4 Acknowledgements

Some parts of this work have been done in collaboration with Fabrice Bouquet, Pierre-Alain Masson and Jean-Michel Caricand. Thanks to them since they trusted and helped me since the first steps of this work. Many thanks to Jerzy B. Ludwichowski, who has translated the abstract and keywords in Polish.

References

[1] Thierry CHANIER : *Archives ouvertes et publication scientifique. Comment mettre en place l'accès libre aux résultats de la recherche ?* L'Harmattan. Décembre 2004.

[2] *EPrints — Digital Repository Software*. March 2010. http://www.eprints.org/software/.

[3] *Documentation Import* XML-HAL. Novembre 2007. http://hal.archives-ouvertes.fr/.

[4] Jean-Michel HUFFLEN: "MlBibTeX's Version 1.3". TUG*boat*, Vol. 24, no. 2, pp. 249–262. July 2003.

[5] Jean-Michel HUFFLEN: "XSLT 2.0 vs XSLT 1.0". In: *Proc. BachoTeX 2008 Conference*, pp. 67–77. April 2008.

[6] Jean-Michel HUFFLEN: "Introduction to XQuery". In: Tomasz PRZECHLEWSKI, Karl BERRY and Jerzy B. LUDWICHOWSKI, eds., *Proc. BachoTeX 2009 Conference*, pp. 17–25. April 2009.

[7] Michael H. KAY: *Saxon. The XSLT and XQuery Processor*. March 2009. http://saxon.sourceforge.net.

[8] Oren PATASHNIK: *BibTeXing*. February 1988. Part of the BibTeX distribution.

[9] Erik T. RAY: *Learning* XML. O'Reilly & Associates, Inc. January 2001.

[10] W3C: XSL *Transformations (*XSLT*). Version 1.0.* W3C Recommendation. Edited by James Clark. November 1999. http://www.w3.org/TR/1999/REC-xslt-19991116.

[11] W3C: XSL *Transformations (*XSLT*). Version 2.0.* W3C Recommendation. Edited by Michael H. Kay. January 2007. http://www.w3.org/TR/2007/WD-xslt20-20070123.

[12] W3C: XML *Schema*. December 2008. http://www.w3.org/XML/Schema.