The transition of CyberThèses to Open Source

Martin Sévigny
AJLSM
sevigny@ajlsm.com
17 rue Vital Carles, 33000 Bordeaux, France
http://www.ajlsm.com

Viviane Boulétreau
ERADD - Edition Electronique & Numérisation - Université Lumière, Lyon 2
Viviane.Bouletreau@univ-lyon2.fr
86 rue Pasteur, 69007 Lyon, France
http://sophia.univ-lyon2.fr/

Keywords: SGML, XML, Open source, CyberThèses

Abstract

CyberThèses is a platform for the archiving and dissemination of electronic theses and dissertations built on the use of structured document (initially SGML). An important step has been crossed with the evolution of the program towards XML and the evolution of the whole platform to open access.

We now propose to the community a set of tools covering the production of XML document from traditional word processing formats, their indexation which the enrichment associated to structured document and their dissemination. The whole platform is available freely (GPL license) on a collaborative development web site (http://sourcesup.cru.fr/cybertheses).

The benefits we aim from this transition are of different nature:

- the political one: dissemination of the results of research shall be free and we do agree on that point with some larger initiatives' recommendations OAI, BOAI, NDLTD, etc.,

- the financial one: there won't be anymore economic barrier to the implementation of many ETD providers,

- the practical one: we hope to enlarge the users and/or developers contributions to the CyberThèses program.

Introduction

During the last year or so, the CyberThèses project has initiated a major development effort in order to change the technical platform from technologies based on commercial solutions to an open source approach. The main results of this change are now visible: CyberThèses partners may now publish electronic theses and dissertations (ETD) with a dynamic Web application including a search engine, after bringing them to XML using conversion tools based on XSLT transformations.

The goal of this article is to present this transition and to understand the current or future benefits of the new open source approach. We will provide a short historical perspective of the CyberThèses project and its main realizations and explain what were the objectives and process for the open source transition. An overview of the new tools will then be provided, followed by benefits and a discussion about the future of the platform.

Historical backgrounds

The technical aspects of the CyberThèses project[1] have their origins in the work done at *Presses de l'Université de Montréal* (Canada) back in 1997. At that time, this institution was conducting a project concerning electronic publishing of scholarly journals in social and human sciences[2], and they looked at applying the same principles and technologies to provide tools for electronic theses and dissertations. Quickly, the *Université Lumière - Lyon 2* (France) entered the project, along with financing coming from the *Fonds francophone des inforoutes*, a major financing program of the *Agence intergouvernementale de la Francophonie*[3].

The documentary model behind this project was to use open formats based on SGML (ISO 8879) to archive and disseminate electronic theses. On the other hand, the production of these electronics theses should be as easy as possible for the authors, even though they would benefit from information and knowledge about structured information management and production. The results were processing tools that would let anyone convert a word processor document in RTF format to SGML using scripts written in a commercial programming language[4], the TEI Lite DTD[5] being used as the final format

[1] Readers not familiar with the CyberThèses project should go to the European main project Web site and follow the link "A propos" in order to read a general presentation and a list of participating institutions: http://mirror-fr.cybertheses.org/.

[2] Projet Erudit: http://www.erudit.org. Please note that this project entered a second phase with a new technical approach.

[3] See http://www.francophonie.org/fonds/.

[4] The Omnimark programming language : http://www.stilo.com/products/omnimark/buildingxmlmiddleware.html

of this process. These scripts can produce structured SGML documents because the source document had previously been *styled* using a specific word processor style sheet identifying important contents and the structure of the document. The publishing part was done with an automatic conversion of the SGML documents to HTML for *static* publishing on the Web. Along with this core set of technologies, various other tools have been provided to build a complete platform with a high usability.

With two initial partners and then seven more institutional partners of the project, and close to 40 institutional users worldwide, we can say that the technological foundations of CyberThèses have not retained institutions from participating into the project.

Objectives and process

Successful projects cannot stay with the same technologies and approaches for a long term without at least investigate if the tools used are still useful, efficient and open. On the other hand, even if the number of participating institutions is quite large, this user base could possibly be even larger if some constraints could be removed. So in 2002, the CyberThèses project leaders at Université Lumière - Lyon 2 decided to evaluate the feasibility of a major change in the platform: removing all commercial solutions (mainly Omnimark) to replace them with a completely open source platform.

The main reason for such a change is economical; user institutions were mainly located in countries where the cost of software licenses can be a major difficulty in implementing an ETD solution. It was becoming more and more difficult for the project leaders to attract not only other user institutions, but also financing and development partners. We must add that between 1997 and 2002, the role and the visibility of open source software has increased almost as fast as the Internet itself, including in France where public institutions have been strongly encouraged to make use of open source software and also to participate in the development of open source projects.

It is important to remember what is *open source software*; and although we use this term here, it would be better suited to talk about *free software*, defined as:

Free software is a matter of the users' freedom to run, copy, distribute, study, change and improve the software. More precisely, it refers to four kinds of freedom, for the users of the software: the freedom to run the program, for any purpose (freedom 0); the freedom to study how the program works, and adapt it to your needs (freedom 1, access to the source code is a precondition for this); the freedom to redistribute copies so you can help your neighbour (freedom 2); the freedom to improve the program, and release your improvements to the public, so that the whole community benefits (freedom 3, access to the source code is a precondition for this).[6]

This definition emphasizes the fact that open source - or *free* - software is more than software without acquisition costs, it is a freedom given to users, currently or in the future. In the case of the CyberThèses project, adhering to these principles was as important as providing a free solution - in the economic sense.

Changing to an open source solution was not a small evolution, so it was also a great opportunity to review the overall solutions and see if there were some better suited technologies or approaches. In particular, it was important to see if:

- The use of XML could replace SGML.
- The use of a dynamic application could provide more functionality to users.

In order to accomplish this transition, they mandated a private company to study the feasibility and then to develop a new platform using open source technologies. The study and the development have been realized between March 2002 and May 2003.

Results

The *CyberThèses platform* ("Plate-forme CyberThèses") is now a set of tools, methods and technologies aimed at providing interested institutions means of converting their electronic theses to XML and publish them on the Web or a local intranet. Most of the tools or there execution environment are Java-based, which mean they can be run on various operating SYSTEMs without any changes.

From word processor to documents

The original platform made use of scripts to convert word processor documents correctly styled to SGML documents respecting TEI Lite DTD. This approach has been kept, with the exception that now the target documents are in XML, but still respecting the same DTD. This small technical change makes a big difference in the available processing tools and libraries. Before entering a thesis in the processing chain, it must be prepared using a style sheet specifically designed for the task. This style sheet is basically the same as before, only a few adjustments have been made; new training for *stylers* is not mandatory.

The new process is divided into these general steps:

[5] See http://www.tei-c.org/Lite/DTD/.

[6] This definition comes from the Free software foundation: http://www.fsf.org/philosophy/free-sw.html.

1. The word processor document is automatically opened in *OpenOffice.org Writer*, and then automatically saved in OpenOffice.org native format, XML.
2. The XML from OpenOffice.org is converted to TEILite using XSLT transformations, performed in several steps to ease the development, debugging and maintenance tasks.
3. The TEILite reference document is converted to HTML using XSLT transformations, and to PDF using XSLT transformations and an XSL-FO representation of the thesis, these HTML and PDF versions are called *static*.
4. The TEILite reference document is processed in order to provide some specific information to the Web application.

There are three key aspects in this process needing some explanations here. First, most of the conversion process is done using XSLT transformations, a standard technology proposed by the W3C[7]. People developing applications with XML are familiar with this language, so we could expect a large number of contributors in this open source project. XSLT has proven to be a valuable tool for the task, even if *up-conversion* is not part of the target applications the language has been designed for.

Second, since XSLT requires an XML source to operate on, we had the need to convert the word processor document into XML without any loss in contents and styling. The use of OpenOffice.org is very helpful here, because not only it handles correctly most recent word processor formats, but it produces a high quality XML, using standards when available. For instance, styling properties are CSS-like, mathematical equations are in MathML, etc.

Third, the production of static versions of the thesis in HTML and PDF is helpful because it gives institutions something to publish directly, without using the Web application provided with CyberThèses. This Web application still uses these static versions when users ask for a printable version of the thesis.

Driving the conversion process

The conversion process can be driven using either scripts or a Web interface. The scripts are provided for expert users and the development process; the Apache Ant *makefile* SYSTEM is heavily used in this part of the tools. But we expect normal user to become familiar with the Web interface for driving the conversion process and not to use the command line scripts.

The Web interface has not changed much since the first version of the CyberThèses platform. Some adaptations have been made for handling the new steps in the conversion process, but these adaptations are fairly minor, and once again users already familiar with the Web interface provided with the old CyberThèses platform should make the transition without problem or training.

For completeness, we list here the main functionalities of this Web interface:

- Users and workspace management, in order to provide a single interface to many institutions and operators.
- Document and image upload on the server.
- Conversion process driving with efficient management of error or information messages.
- Metadata creation with specialized forms.
- Direct publishing of the converted thesis with the Web application.

Publishing in a Web architecture

The third part of the platform is a dynamic Web application that provide searching and browsing functionality for readers of electronic theses and dissertations. This application is based on the SDX platform[8], a powerful - and open source - solution for searching and viewing XML documents. SDX itself being based on the Cocoon project from the Apache software foundation, it provides a strong technological infrastructure to add new services in the future.

Here is a summary of the new Web application for CyberThèses:

- Users can search documents of browse lists by institutions.
- Searching can be performed with a simple fulltext query or with complex searching forms using highly structured metadata (title, authors, evaluators, dates, etc.) or specific parts of theirs contents, such as section headers, table captions, etc.
- Browsing is done via a standard Web browser, with interactive tables of contents, lists of figures or tables and word highlighting when the user has previously made a query.
- Printing can be done with specially prepared HTML or PDF versions of the theses.

The Web application is highly configurable using XML files. It uses skins to adapt the user interface to other projects or institutions. The SDX platform beneath the CyberThèses tools will have complete support for the OAI-PMH protocol during summer of 2003, which mean that any institutions publishing ETD with the CyberThèses platform will have the opportunity of exchanging metadata with other partners using a standard protocol.

[7] See http://www.w3.org/Style/XSL/.

[8] See http://sdx.culture.fr.

Benefits

The transition to open source brings important benefits to the responsible institutions but also for users. Some of the benefits are technical, others are functional, but most of them are organizational in the sense that they will make the project itself more viable on a long-term basis.

Organizational

Since the beginning, distribution of the knowledge and tools around CyberThèses to foreign countries has always been a major objective. The implication of institutions from Chile, Madagascar, Sénégal, etc. have always been opportunities to disseminate not only technologies, but most of all general approaches based on important concepts such as standardization, long-term preservation, cooperation, etc. On the other side, the foreign institutions have brought back into the project their own experience, needs, wishes and when put together have made the project and its tools even better.

The transition to open source will make all these exchanges of technologies, experiences and knowledge even easier, and with more institutions, since the economic barrier of purchasing expensive processing tools has disappeared. The cost of acquisition of software licenses is only a small part of the overall cost of publishing ETD, but once this cost barrier is removed and the number of participating institutions grow, helpful resources also grow and it makes the tools easier to use and understand, thus decreasing the overall costs of the project. So, when an open source project succeeds, the fact that the tools are available for free becomes more and more important with time.

Another important organizational benefit for the project is the opportunity to attract new developers, without any economic investment. Once the project is known and well established, once several people realize that it can correctly serve their needs, some of them will be interested in putting efforts in documenting or translating the tools, animating mailing lists or help desks, creating FAQ entries, preparing CDROMs for ease of distribution, testing and identifying bugs or new needs, and of course contributing code to correct bugs or to add news functionalities. So the concept of *developers* is really too restrictive here, we should talk about contributors, and not developers, since various tasks are needed, requiring various knowledge or skills, so a large number of people may be able to contribute.

Attracting new contributors is not just a way of making the tools and their documentation better. It is also an opportunity to change an economic model where large institutions make important investments in a project in order to have results that fulfils the requirements. When the time goes on and the number of contributors increases, the number of interested institutions will also increase, and, along with contributors giving their time for free, there will be some that will be mandated by other institutions to participate in the project. Overall, this increases the global amount of economic resources available, or if it stays the same it means that the original financing institutions may decrease their contributions.

Technical

The technical or functional benefits of the new platform could be analyzed one by one, but what is important here is to highlight the benefits implied by the open source approach. There are commercial solutions for all the functionalities in the platform; these solutions may be more, less or equally efficient, but in general it doesn't appear that the open source requirement has been a major technological constraint.

The future will better tell us what are the real technical benefits of the open source approach, because these benefits are mainly indirect. We believe that an open source methodology will help disseminate the overall technical and documentary objectives of the project, such as standardization with XML and TEILite, because not only the tools chosen are as standard as possible, but open source technologies make use, in general, or available standards and their implementations, not only because they are standards, but also because it increases the potential of fast and economical software development by making use of available standard libraries. Thus *open* source development and *open* technologies based on standards are closely related and this pair may become a winning technological scenario for CyberThèses.

From CyberThèses to Cyberdocs

The previous parts have shown how the new platform performs well with electronic thesis specifically prepared using the recommendations of the CyberThèses project. It was the main goal of the transition to open source: provide at least what was already available in the first version.

By doing so, the developers also made a special attention to let the tools be as *generic* as possible, in order to make the platform usable in other contexts, in a pure open source spirit. We can now say that this goal is achieved and the platform can be used as a basis for electronic publishing of *any word processor document*. The conversion tools handle correctly and document, even if CyberThèses specific styles are not used, and a valid TEILite XML document will result from the conversion.

After that, users just need to type a few metadata in an XML document, in order to provide basic information such as a title and a publishing date, along with internal identifiers for the institution and the thesis itself. Once this is done, the thesis with its accompanying metadata can be published using the Web application, which can easily be adapted using *skins* to provide a more specific interface for this new application.

Thus, the transition to open source has given a great opportunity to reviews some of the processes that were too specific to CyberThèses without any technical reasons. The result is a true electronic publishing platform, thus the new name **Cyberdocs platform** that will be used in the near future.

Conclusion

An open source project must live and grow before we can say that it is a true success. During its life, it will go through various phases that can be summarized as follow:

1. An initial development phase, where a few individuals or organizations are not only the soul developers, but also almost the only users of the tools.
2. Once the initial development accomplished, the same individuals or organization will still drive the development process, but other developers could join and, most of all, the user base will grow.
3. After some time and the growth of user base, the original developers and organizations are accompanied by others in the management of the project, because they see interest in its results, tools, methods, approaches, etc.

Starting an open source project and being in phase 1 is not a hard task; usually a good idea, strong technical backgrounds, and time and/or financing are the only resources needed. There a hundreds of such projects in open source development sites such as SourceForge[9]. Stepping to phase 2 may be done once the initial versions have been used and are proven to be suitable for the expected needs; the original developers will continue to make important contributions and still play a major role in its development. Once again, there are a great number of projects in this phase in the open source world.

But reaching phase 3 is another story. Most of the time, open source projects won't be able to attract a sufficient number of developers or institutions interested in investing time and money in the project. Although some projects may live a long time without ever reaching phase 3, we think that it is the only way to really insure a long-term viability to the project, which is so critical in, among others, projects in electronic theses and dissertations management.

The CyberThèses project and its open source tools took a different route, because the project and the technologies were available before going into an open source approach. Thus we can say that the project enters the open source world directly at phase 2, since users are there and the project is backed by several important international institutions.

Staying in phase 2 is not a solution for CyberThèses. The energy is now put into communication plans and specific tools in order to attract other users, developers, institutions, financing into this dynamic project and community. Translations to other languages than French, along with the generalization of the platform from CyberThèses to Cyberdocs are important part of these efforts. The next months and year will tell us if the project succeeds to become not only an open source project, but a **successful** and **viable** one.

[9] http://sourceforge.net