Aquitaine Patrimoines
&
Cyberdocs

4th Open Archives Forum Workshop
September 5, 2003

Rasik Pandey – AJLSM
rasik.pandey@ajlsm.com
http://www.ajlsm.com

Aquitaine patrimoines

PLATE-FORME Cyberdocs

Aquitaine patrimoines

Aquitaine Patrimoines

- **Cultural Heritage Portal (Service provider)**

  A portal encompassing diverse cultural information sourced from libraries, médiathèques, archives, museums, cultural heritage education centers, centers of documentation, etc.. The data, harvested by means of the OAI protocol, describes various heritage resources concerning the Aquitaine region of France. The contributors and actors from international to local levels are interested in validating methodologies and technologies for sharing resources in a distributed environment as well as investigating the services which can be derived from these sources.

- **Two development phases**

  ✓ **Prototype - August 2003**

  The portal entered its validation stage last week at which time it was presented to the concerned parties and a limited public

  http://ajlsm-sdx.hopto.org/sdx-22h/pa-portail/

  ❑ **Final Version – March 2004**

  Further development will focus on integration of more heritage resources while attempting to find commonalities in sources to make this service provider more usable, meaningful, and valuable for its users.

Aquitaine patrimoines

http://ajlsm-sdx.hopto.org/sdx-22h/pa-portail/

- **Simple search (full-text)**
- **Advanced search**
  - Field-level free text and limited value searching
- **Cartographic search**
  - Searching by geographic department followed by town (s)
- **Search Results**
  - Linking metadata records to their external resources
  - Ability to filter results based on type of resource or geographic department
  - Cartographic view
- **Thematic Guide**
- **Editorial Access using Keywords**
  - Text proposing themes on which keyword searches can be executed

Development of Portal

- **Heterogeneity of substance and formats**
  - Resources where provided in various formats: XML, HTML, and Delimited text
  - Resources describing different aspects of cultural heritage in Aquitaine; ranging from modern art to prehistory
- **Conversion version of raw data to XML**
- **Conversion raw XML into common metadata format PA (Dublin Core +7)**
  - Slightly richer format which allows for better building specific services
- **Integration of prepared data into OAI repositories**
  - Data which was not already provided via OAI-PMH
- **Harvesting of Data for use in Portal**

Development of Portal (cont.)

- **Uses Open Source Software**
  - SDX Documentary System in XML with OAI harvesting and repository implementations
    - Developed for the French Ministry of Culture
    - http://sdx.culture.fr/sdx/
  - Built upon Cocoon Publishing Framework
    - http://cocoon.apache.org
  - Integrates Lucene (Java based Search engine)
    - http://jakarta.apache.org/lucene/
  - Core technologies are Java, XML, XSL…

Aquitaine patrimoines

Conclusions

- **Technical aspects (programmatic, etc.) are simple as structured data and environments dictate**
- **The difficulty stems from contents**
  - Finding common threads in diverse content by which resources can be presented such that value is added in service providing is the greater challenge.
- **Why wasn't DC sufficient?**
  - Needed a more controlled structure by which specific data (geographic and resource type classification) could be described allowing certain commonalities to be exploited as services
- **Why convert to the PA metadata format?**
  - Not all resources were available in dynamic server environments
  - Conversion allowed the easy enrichment of resources as data was in a static environment
  - Two sources were provided via OAI-PMH so the PA format was not at all a limiting factor

Conclusions(cont.)

- **Why develop and use an OAI tool within SDX?**
  - SDX already used in cultural projects thus existing resource collections can be benefit from the OAI implementation.
  - SDX, being a searching and publishing utility, provides a interesting means of building services upon data harvested via OAI-PMH
- **Problems encountered using OAI**
  - Comprehension problem, some parties were hesitant as OAI was new for them and so it seemed complicated
  - Perceptions could be formed that PA is a poor metadata format based upon the services derived and therefore the limitations are not one of the metadata format, but rather OAI itself…
- **Advantages of using OAI for this Portal**
  - Simple integration of new resources (simplicity of OAI-PMH as well as use of « sets »)
  - Updating the portal is facilitated using OAI-PMH

Cyberdocs platform

Publishing structured electronic documents

http://sourcesup.cru.fr/cybertheses/

http://www.cybertheses.org/europe/

September 5th

PLATE-FORME Cyberdocs

Overview

- Background and objectives
- Results
- Description and demonstration: information processing platform
- Benefits
- Future development
- Conclusion: viable open source projects?

PLATE-FORME Cyberdocs

Historical backgrounds

- Cyber?
  - “Cyberthèses” is now a six year old project
  - “Cyberdocs” is a brand new open source publishing platform
- In 2002/2003, the Cyberthèses project has undergone a major upgrade to its processing tools
  - Completely open source
  - New dynamic publishing module
  - Not only theses, but any word processor document...
  - … Cyberdocs!

PLATE-FORME Cyberdocs

Historical backgrounds

- Origin: information processing platform for scholarly publishing, **Presses de l'Université de Montréal**, 1997
- Soon after:
  - **Université Lumière Lyon 2** joins the project, with other institutions
  - Financing from the *Fonds francophone des inforoutes* (Agence internationale de la francophonie)
- Processing and document model (first platform)
  - From word processor documents to logically structured documents in ISO-8879 (SGML), using TEILite DTD
  - Human styling of word processor documents
  - Processing mostly based on *Omnimark* scripts
  - Static HTML publishing on the Web, SGML publishing with Softquad Panorama

Cyberthèses project today

- Nine institutional partners
  - France, Canada, Egypt, Chile, Switzerland, …
- Around 40 institutional users
  - France, Morocco, Algeria, Lebanon, Burkano Faso, Vietnam, Mauritius, Madagascar, Senegal, Mali, …

But what about tomorrow…?

PLATE-FORME Cyberdocs

Objectives of the open source transition

- **Economical**
  - Increase the dissemination of the project and its platform at very low costs
- **Technological**
  - Review the platform and add functionalities where possible
    - XML?
    - Dynamic Web site?
    - Unicode?
    - MathML? SVG?
    - OAI?

PLATE-FORME Cyberdocs

Objectives of the open source transition

- **Long-term viability of the project**
  - Benefit from the four freedoms of *Free software*
    - Freedom to run the software, for any purpose
    - **Freedom to study the program, and adapt it to your needs**
    - Freedom to redistribute copies
    - **Freedom to improve the program**

(according to the Free Software Foundation)

Cyberthèses had already addressed the long-term viability of its documents, the **theses**, by using open and structured standards. It now addresses the long-term viability of its **tools** by using free software and by **building a community** around them.

PLATE-FORME Cyberdocs

Cyberdocs now

- A complete platform for publishing structured electronic documents
  - Conversion module: from word processor to TEILite XML
  - Management module: Web interface for driving conversions
  - Publication module: dynamic Web application for publishing documents
- First release available (http://sourcesup.cru.fr/cybertheses/)
- GNU Public Licence (GPL)

PLATE-FORME Cyberdocs

Description

- Conversion module: from word processor to structured documents
  - Word processor documents are styled to identify important contents and structures
  - Flat XML extraction using OpenOffice.org office suite
  - Conversion to TEILite based on XSLT transformations
  - Production of support files for the dynamic Web application
  - Production of static HTML, XHTML and PDF files from XML
    - Used in the dynamic Web application for *printing*
    - May be used to publish using other systems

PLATE-FORME Cyberdocs

Description

- Management module: Web interface for driving conversions
  - May serve multiple institutions
  - Administrators can create users
  - *Workers* can create workspaces for documents
    - Upload files (word processor documents, images, etc.)
    - Start the conversion process, at any step
    - See the results of the conversion, at any step, with adapted error and information messages
    - Make necessary corrections and restart if necessary
    - Forms to add metadata (Cyberthèses, Dublin Core, ETDMS, …)
    - Publish the document if the dynamic Web application is used

PLATE-FORME Cyberdocs

Description

- Publication module: dynamic Web application
  - Underlying infrastructure
    - SDX, an XML search engine and publishing framework
    - Cocoon, an XML-based infrastructure for building dynamic Web applications
  - Functionalities
    - Search documents using simple queries or complex forms
      - Search in metadata
      - Search in fulltext
      - Search in specific zones (section titles, figure legends, table captions, …)
    - List documents by institutions, dates, subjects, …
    - Browse documents
      - Interactive table of contents, list of tables, list of figures, …
      - Query terms highlighting
      - Search within a document

Description

- Publication module: dynamic Web application
  - Easily customizable
    - Translations
    - Skins
    - Metadata labels
    - Colours using per institution CSS
  - OAI-PMH repository
    - Built-in OAI-PMH support in the SDX platform
    - Sends metadata in Dublin Core (mandatory), ETDMS
  - One could easily add OAI-PMH harvesting capabilities

Benefits

- Technical
  - Some small indirect technical benefits
    - MathML
    - Unicode
    - Easier graphics capabilities
    - Dynamic Web site
    - OAI repository
  - Benefit from other dynamic open source efforts
    - In XML processing
    - In TEILite tools
  - Participate in the huge global effort within the **open** source community towards implementing **open** technologies based on standards

PLATE-FORME Cyberdocs

Benefits

- Organizational
  - Easier cooperation
    - The Cyberthèses program is already centred around cooperation
    - Now its platform favours such cooperation and exchange of expertise
  - Lower costs
    - More important in the long term
  - Attractiveness
    - New developers
    - New contributors
      - Documentation
      - Tutorials
      - Translations
    - New areas of application

PLATE-FORME Cyberdocs

Benefits

- Organizational
  - Easier cooperation
    - The Cyberthèses program is already centred around cooperation
    - Now its platform favours such cooperation and exchange of expertise
  - Lower costs
    - More important in the long term
  - Attractiveness
    - New developers
    - New contributors
      - Documentation
      - Tutorials
      - Translations
    - New areas of application

PLATE-FORME Cyberdocs

Future developments

- No specific plans yet, but should come before the end of the year
- Functionalities that could be added
  - Support for OAI-PMH harvesting
  - Better support for various document types
    - Reports
    - Electronic journal and newletters
    - Monographs
  - Support other DTDs
    - Conversion module
    - Publication module
  - Support a wider range of word processor features

PLATE-FORME Cyberdocs

Future developments

- Main objectives of future developments
  - Make Cyberdocs not only a great publishing platform for electronic theses and dissertations, but for **any kind of scientific literature**
  - Increase the number of supported languages in the various modules by translating messages, in order to broaden the distribution around the world
  - Attract new developers and financing institutions interested in various uses of the platform

PLATE-FORME Cyberdocs

Conclusion

- Three phases of open source projects
  - Phase 1: initial development
    - Few developers and sponsors
    - Few users
  - Phase 2: growing user base
    - Still few developers and sponsors
    - More users
  - Phase 3: sustainable development
    - Various developers and sponsors
    - More and more users

Cyberdocs will try to reach phase 3 by developing **user and developer communities**