DIGITARQ PROJECT

The Arquivo Distrital do Porto (www.adporto.pt) is a public and regional Archive dependent from the Direcção-Geral de Arquivos (Directorate-General of Archives www.dgarq.gov.pt). This web page is intended to present a synopsis of the DIGITARQ project considering the methodologies, tools and solutions adopted. We thank you for any feedback you would like to send us (info@adporto.pt).

Main goals of the project

1. To convert old finding aids, paper or digital, to archival standards.
2. To build an information repository to receive and manage the deliverables obtained on the previous process.
3. To develop a search engine on a web interface so as to allow queries in the built repository.
4. To develop an application, as well as new business processes, in order to manage digital objects and associate them to the information repository constructed.

General description of the project

The Archive owns electronic finding aids mainly in ISIS format but also in ACCESS, EXCEL and WORD. This represents an universe of c. 150.000 records descriptions considered all aggregation levels (fonds, class, series, compound document and document. The other finding aids were on paper: typed, published or handwritten, which represented c. 75.000 records descriptions.

The first step was to "clean" the DTD/EAD making some allowed adaptations, e.g., the way dates should appeared, or the description levels. The second step was to develop transformers capable to import all electronic data to EAD XML tagged structure. We developed a specific transformer to each electronic format (ACCESS, WORD, EXCEL, ISIS).

Meanwhile an archivists' team was annotating the existing finding aids with the appropriated EAD tags in order to give feedback to computer experts when the time would come to import all data. That same team proceeded with OCR, revision and annotation of the paper finding aids.

At the end, all data (electronic and paper) were converted to EAD resulting in c. 700 texts organized by fonds (each text corresponding to a fonds or records group). Stylesheets were developed in order to produce html from those texts, which allowed patrons to access the information and permitted a easier revision of those texts.

After that the development of a database where all the data might be imported and also that would permit the production of further descriptions took place. The database has an relational architecture combined with hierarchical XML structure, the interface with the user is articulated through two "lazy nodes" one of them interacting with XML and the other with SQL

The digital archive architecture is compliant with the following base documents:

1. The OAIS (ISO 14721:2003) Standard and
2. Project Interpares models and activity definition in Digital Preservation.

The basic management unit is the digital object which comprehends images, which are in fact the atomic units of the GOD (digital object management in portuguese:). Every digital object must be produced in the context of a specific project, being reflected in it's name. The Id string for each digital object is compliant to the following scheme: ***[nameOfCUstodialEntity][YearOfProduction][ProjectNumber][DigitalObjectNumber]***. The first, second and fourth block are automatically generated by the application, which requires from the operator manual insertion of the project's Id.

The DO (digital object) exists in the virtual world and it corresponds to a description unit that exists in the real world. This description unit can be digitally replicated at discrete levels:

- document,
- compound document and
- book (a particular kind of contentor)

The search engine (pesquisa.adporto.pt/) was developed in web environment and it interacts with the description and DO databases.

The project ended beginning May 2004. The next step was to develop an e-commerce interface to remotely provide products to patrons.
This goal was achieved in the project "Consulta Real em Ambiente Virtual", developed and implemented 2006 and 2007, in full operation since the beginning of 2008 at the Arquivo Distrital do Porto website (http://pesquisa.adporto.pt/cravfrontoffice).

This software was partially funded by POC (Programa Operacional para a Cultura/Operational Programme for Culture) promoted by the Portuguese Government.

Table of contents

Archival standards

The descriptions were stored in a data structure that ha as basis the ISAD descriptions areas and inside, all the suitable EAD elements were inserted:

- ISAD as the overall description standard
- EAD as a standardized (a standard *de facto* and not *de jure*) data structure.
- ISAAR (cpf) to produce authority files
- EAC as a exchange format for authority files

Several technical documents, guidelines were used in order to acquire proficiency on this area:

- The ISAD and ISAAR standards are available at the ICA/CDS website
- EAD cookbook
- EAC non official website
- EAD RLG guidelines

Metadata

A combination of several metadata schemes was compiled. We couldn't find a single scheme that entirely fited our needs. Some were too complete, other did not have elements that were suited for our goals. Therefore a compromise was found from the following metadata schemes:

- Library of Congress: Digital Repository Core Metadata Elements
- CEDARS (curl exemplars in digital archives) Project
- NISO Z39.87-2002

Other metadata schemes were consulted:

- Library of Congress - METS. Metadata Encoding and Transmission Standard
- National Library of Australia - Preservation Metadata for Digital Collections
- OCLC/RLG Working Group on Preservation Metadata - Preservation Metadata and the OAIS Information Model: A metadata framework to support the preservation of Digital Objects.
- Dublin Core elements set

The complete compiled scheme:

# element	elementName	# element	elementName
1	archiveDateTime	9.9	MIMEType
2	archiveId	10	captureEntityCorporate
3	archiveNextDateTime	11	captureEntityIndividual
4	archivingProfile	12	creationDateTime

5	captureDeviceID	13	depositDateTime
6	operatingSystem	14	externalDescriptiveInformation
7	operatingSystemVersion	15	handle
8	*deviceSource {abstract}*	16	reformattingGuidelines
8.1	scannerManufacturer	17	actionHistory
8.2	scannerModelName	18	reformattingMethod
8.3	scanningSoftware	19	structureInformation
8.4	scanningSoftwareVersion#	20	transformerObject
9	*captureDeviceSettings {abstract}*	21	Platform
9.1	compressionScheme	22	Parameters
9.2	compressionLevel	23	render_analyseEngines
9.3	colorSpace	24	inputFormat
9.4	sceneIluminance	25	revisionDateTime
9.5	spatialResolution	26	quantityOfTerminalObjects
9.6	imageWidth	27	Size
9.7	imageLength	28	preservationOriginalInformation
9.8	bitsPerSample		

Image capture configurations

From benchmarking tests made during previous digitisation projects we had defined "digitisation profiles" which consist on capture parameters obtains for a specific document typology. These profiles must be considered accordingly to hardware and software capture used. These typologies (or groups) were defined according to physical features of documents.

The profiles are intended to obtain matrix images with maximum archival quality and were obtained according to Cornell guidelines.(See also KENNEY, A, R.; CHAPMAN, S. - *Digital Imaging for Libraries and Archives*. Ithaca: Cornell University Library.1996

The derivatives configuration were obtained according to *NARA guidelines for Digitizing Archival Materials.1998)*

We often used as information resource the TASI (Technical Advisory Service for Images) website.

Conversion tools

Several ad-hoc annotations were developed to allow automatic import to EAD structure.

- Omniscan Pro 12.0 for optical character recognition
- Xmetal v 4.0 author and developer packages for marking digital texts and manage DTD

Scanning hardware

Two main devices were used in order to (1) digitise finding aids (which were after submitted to OCR process) and (2) digitise selected historical documents. The criteria to scan this last document group were based in conservation condition and access rate from patrons.

- Minolta PS7000 - Overhead scanner B/W, gray capture
- Zeutschell 10000 Hybrid - Overhead scanner+microfilm, B/W, Gray capture

- PC, OS Windows XP, HD60GB; RAM 512 MB
- Workstation, OS Windows XP; HD 80GB, RAM 512 MB

Scanning software

Interface Capture drivers:

- Omniscan 10.0
- Pixview 3.0

For image processing:

- Adobe Photoshop 7.0

Development software

- Visio 2000 to produce models for the databases architecture and processes reorganization
- Leadtools for image management inside the digital archive
- SQL server 2000 for database development
- Visual Studio.Net for Database development

Digital archive

A application was built according to OAIS (Open Archival Information System), now ISO 14721:2003 and Project Interpares deliverables guidelines on digital preservation.

Several normative and technical documents were used:

- BYERS, Fred - Care and Handling of CDs and DVDs. CLIR/NIST, (A Guide for Librarians and Archivists). 2003
- International Standards Organization - ISO 14721:2003, Space data and information transfer systems - Open archival information system - Reference model.
- International Standards Organization - ISO 12142:2001, Electronic imaging - Media error monitoring and reporting techniques for verification of stored data on optical digital data disks
- International Standards Organization - ISO 18927:2002 , Imaging materials - Recordable compact disc systems - Method for estimating the life expectancy based on the effects of temperature and relative humidity.
- PUGLIA, Steven; ROGINSKY, Barry - NARA Guidelines for Digitizing Archival Materials for Electronic Access. Washington: National Archives and Records Administration. 1998
- TWAIN Working Group Committee - TWAIN Specification, version 1.9. 2000
- W3C - PNG (Portable Network Graphics) Specification Version 1.0. 1996

The following functional requirements were implemented:

- Automatic production of derivative images according to the guidelines adopted
- Automatic identification and naming of digital objects
- Semi-automatic metadata association to images and digital objects
- Database image management
- Derivatives association to archival descriptions
- Visualization of Digital Objects from web search interface
- Media and file monitoring compliant with ISO 18927:2002 and ISO 12142:2001
- Migration alert
- Digital object integration (Ingest)
- Digital object hierarchical structure
- Digital object retrieval from off-line media
- Reporting