Electronic Publishing of Digitised Works

João Penas, João Gil, Gilberto Pedrosa, José Borbinha

INESC-ID – Instituto de Engenharia de Sistemas e Computadores
Rua Alves Redol 9, Apartado 13069, 1000-029 Lisboa, Portugal
e-mail: jpenas@ext.bn.pt; jgil@ext.bn.pt; gfsp@ext.bn.pt; jlb@ist.utl.pt

Abstract

This paper describes the automated process to create structured master and access copies for the digitised works at the BND – National Digital Library. The BND created during 2004 and 2005 nearly half million of digitised images, from more than 25.000 titles of printed works, manuscripts, drawings and maps. The resulting of the digitisation process is a group of TIFF image files representing the surfaces of the original works, which needs yet to be processed in order to be stored and published. Doing that manually would be a very complex and expensive task, with risks for the uniformity of the results, so it was need to develop an automated solution. To create the technical metadata, apply image processing actions and OCR, create derived copies for access in PNG, JPG, GIF, and PDF, we developed a tool named SECO. To create the master copies for each of those works, for preservation, and access copies in HTML, we developed a tool names CONTENTE, which exists as a standalone tool and as a library. Finally the copies are deposited and registered at the BND repository through the service PURL.PT, which assures also the WEB and intranet access control. This complex process is fully automated through several XML schemas for the control of the processes, description of the results (including the OCR outputs), descriptive metadata (in Dublin Core, MARC XML, etc.) and rights and structural metadata (in METS).

Keywords: digitisation: digital libraries; structural metadata; METS; image processing

1 Introduction

This paper describes the automated process to create structured master and access copies for the digitised works at the BND – National Digital Library [6]. The BND created during 2004 and 2005 nearly half million of digitised images, from more than 25.000 titles of printed works, manuscripts, drawings and maps. The result of the digitisation of each title is a group of TIFF image files representing the surfaces of the original work. From those images, it can create the technical metadata, apply image processing actions and OCR, and create derived copies for access in PNG, JPG, GIF, and PDF. To create the master copies for each work, for preservation, and also the access copies in HTML, we developed a tool names CONTENTE. Finally the copies are deposited at the BND repository through a specific service (RDD), and registered in the service PURL.PT, which assures also the WEB and intranet access control.

The overall architecture that processes, registers, stores and made available these results, follows the generic concept of SOA – Service Oriented Architecture [14]. The processes are fully automated through several XML schemas for description of the results (including the OCR outputs), descriptive metadata (in Dublin Core, MARC XML, etc.) and rights and structural metadata (in METS). For each object, besides the access copies, it is created one master copy that preserves the original images and all the structural and other metadata. This makes it possible, to manipulate automatically any object at any time, such as for example to create new access copies.

The next section describes the overall architecture for the system, stressing the main components. Following, we describe the main relevant sub-systems by their sequence in the workflow (SECO, CONTENTE, RDD and PURL.PT). Finally, we present some conclusions and ideas for future work.

2 Image Processing

When publishing digitised works at BND, a major concern is the online content availability. Both technical and temporal interoperability are important requirements [2] so only standard formats and platform-independent solutions are sought. For wide public access, technological limitations, such as low bandwidth, must also be addressed.

SECO is a service for the creation of the structural and technical metadata and applying of image processing actions. The master images, acquired through high quality digitization procedures, are processed by this system. Typically, these are uncompressed TIFF files at 24 bit colour depth and with a resolution of 300 to 600 dpi. To store all the data supporting its execution, the system relies on XML files allowing for dynamic and flexible data formats and the ability to keep work metadata linked to its main files.

The SECO server software uses the ImageMagick [3] library for many operations. Because TIFF files are not commonly supported by web browsers, SECO can generate JPEG, GIF, PNG and PDF image files. It also performs scaling, colour depth reduction and re-sampling, thus creating manageable image files for consumer-level hardware at 150 dpi and 72 dpi, at 8 or even 1 bit colour depth, many of which compress very effectively in PNG and the PDF formats. Other implemented advanced operations include a binary trim module to remove margins on good quality binary scans of text works. The image edges are found by examining an image put through a Despeckle filter based on the eight hull algorithm as described in [4], and as implemented in ImageMagick. All the operations in the SECO system are configured and executed through the web-based user interface WebSECO. The web-based interface displays bibliographic records, reports processing status and provides file and process management operations. It also includes a cropping tool to remove unwanted elements from images, such as colour charts and extra margins.

The process configuration page lists the available formats and sub-formats. The user can configure parameters related to compression quality, resolution and slicing, although the available presets seldom require extensive customization. To further enhance productivity, a single configuration can be shared by a group of works. The scheduling component can execute the process immediately or delay its execution to avoid overloading the server with simultaneous processor, memory and disk intensive image operations.

After the processing of the images, the process continues by generating XHTML bindings for the access copies. For this, SECO invokes the CONTENTE Service, described in detail below. The final stage is quality control. If the automated execution results are deemed satisfactory, which is true for the majority of the cases, the finished work is flagged as complete and the SECO system notifies the central storage service, which moves the data to the preservation and online access areas. On the other hand, if the results are not accepted, the user can reconfigure the process attempting to correct detected problems or export it to a workspace where it can be manually adjusted, retouched or corrected. This last option allows creating detailed indexes with the CONTENTE Local Application.

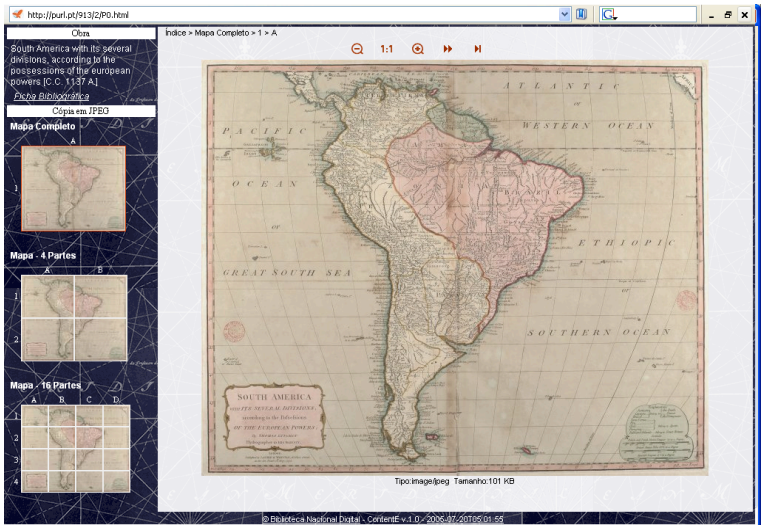

Figure 1: A large map, of only one original image, published in three levels of resolution details

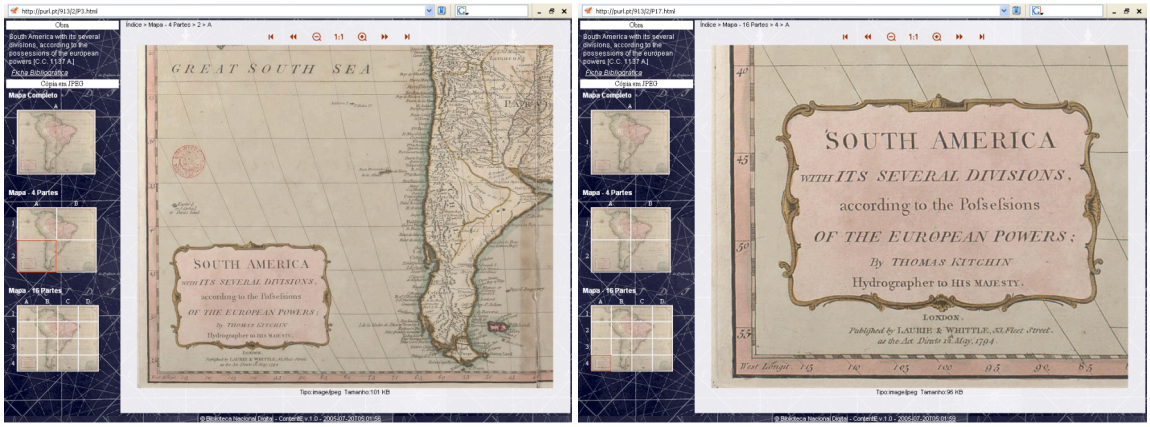

Figure 2: Details of parts of the digitised map easily downloaded at high resolution

```
<word recognized="ahi">
        <image bottom="236" left="457" right="484" top="219">1-64003-p_0033_64-65_t0.TIF</image>
        <image bottom="683" left="255" right="287" top="665">1-64003-p_0036_70-71_t0.TIF</image>
        <image bottom="508" left="181" right="206" top="491">1-64003-p_0037_72-73_t0.TIF</image>
</word>
<word recognized="ainda">
        <image bottom="628" left="666" right="711" top="611">1-64003-p_0034_66-67_p0.tif</image>
        <image bottom="731" left="681" right="728" top="714">1-64003-p_0035_rosto_t0.TIF</image>
        <image bottom="404" left="754" right="800" top="387">1-64003-p_0036_70-71_t0.TIF</image>
</word>
<word recognized="alarido">
        <image bottom="268" left="213" right="272" top="251">1-64003-p_0036_70-71_t0.TIF</image>
</word>
<word recognized="alcançado">
        <image bottom="706" left="396" right="480" top="685">1-64003-p_0037_72-73_t0.TIF</image>
</word>
<word recognized="alcançaram">
        <image bottom="242" left="271" right="368" top="222">1-64003-p_0037_72-73_t0.TIF</image>
</word>
<word recognized="alcácer">
        <image bottom="335" left="165" right="224" top="317">1-64003-p_0035_rosto_t0.TIF</image>
        <image bottom="585" left="710" right="770" top="568">1-64003-p_0035_rosto_t0.TIF</image>
        <image bottom="535" left="379" right="439" top="517">1-64003-p_0036_70-71_t0.TIF</image>
        <image bottom="831" left="146" right="206" top="813">1-64003-p_0036_70-71_t0.TIF</image>
</word>
```

Figure 3: Example of a word index resulting from the OCR process formatted in XML

2.1 Processing of Large Images

For very large and highly detailed images (especially maps and posters), more specific approaches must be found to handle standard network connections. A common solution for this is the DjVU technology [1], but this requires plug-ins to be installed, which should be avoided in order to provide broadly compatible and long-lasting content. SECO supports automated slicing of large images into trees of lightweight sections at progressively increasing detail levels. The result of this concept are standards-compliant web sites, as illustrated for the map in the Figure 1, originally only one image, but published so that detailed parts can be accessed as illustrated in Figure 2.

2.2 OCR and Text Indexing

Additionally, OCR can be applied to the master images with textual contents, producing standard text files as also textual PDF copies. These results are usually not good enough to be immediately published, and proofreading the extracted text can be very time-consuming. However, the data resulting from the OCR can be used to build valuable word indexes for the digitised pages, which can later be read by automated indexing tools such as search engines and complement standard search procedures. Word coordinates in the images are also stored and sophisticated presentation mechanisms can be built upon this information. These word indexes are kept in XML files, such as represented in Fig. 3.

3 CONTENTE

CONTENTE is a result of the initiative BND, promoted by the National Library of Portugal. This tool is made essentially of two components: a library, to be used by other systems, and a local application that uses the same library and provides a powerful user interface for advanced usage. Its initial purpose was the publishing of digitised collections, but its success made it a very practical tool to deal with a wide number of MIME types. It is a tool to create and edit structural descriptions of digital works using making it possible to save and reuse those results in formal structural metadata schemas, such as METS [10].

The structural descriptions edited in CONTENTE are defined as indexes, which are made of trees of nodes where each node can be an aggregator or a leaf. Aggregators represent structural concepts such as parts of books, chapters, volumes, sections, etc. A leaf node is a reference to a content file of any MIME format (such as image formats like PNG, GIF, JPEG, TIFF, etc., but also MS-Word, PDF, Postscript, ASCII, RTF, etc.). Aggregators can make reference to leaf elements, and can have also a list of descriptive and rights metadata files, such as MARCXML [9], Dublin Core [12], etc. CONTENTE can import those descriptions from local files, or it can retrieve them on-line from remote web-services.

CONTENTE manages also collections of style sheets, making it possible to create multiple publication copies of the objects, as XHTML sites, each one with its specific style. That makes it possible, from the same master object, to publish objects with different layouts (menus at the right or left, background colours, etc.), or with different contents (showing or not descriptive metadata, showing images at different resolutions for digitised books, etc.).

The METS schema, used in the CONTENTE, provides a flexible mechanism for encoding descriptive, administrative, and structural metadata for a digital library object, and for expressing the complex links between these various forms of metadata. It can therefore provide a useful standard for the exchange of digital library objects between repositories. In addition, METS provides the ability to associate a digital object with behaviours or services.

A METS document consists of seven major sections: METS header, descriptive metadata, administrative metadata, file section, structural map, structural links and behaviour. The METS Header contains metadata describing the METS document itself, including such information as creator and editor. Descriptive metadata section is used to point to descriptive metadata external to the METS document. The administrative metadata section provides information regarding how the files were created and stored, intellectual property rights, metadata regarding the original source object from which the digital library object derives, and information regarding the provenance of the files comprising the digital library object. All files containing content are listed in the file section. CONTENTE uses the structural map section to save the hierarchical structure of the object and links the elements of that structure to content files and metadata (ultimately, each image can have its own metadata descriptions).

3.1 CONTENTE Library

The CONTENTE library is used by SECO to process the results of the previous steps and produce master copies for preservation and copies for access. The master copies are just a folder organised inside in a set of other folders, one for each MIME type existing for the object. One typical MIME type that is always present is TIFF. Other types are usually JPEG, PNG, GIF, PDF and TXT. For all of this, CONTENTE creates also structural descriptions in METS, as also indexes, as shown in Fig. 4.

One index that is always created automatically is the original physical index, representing the images by their natural order. More complex indexes that can be also automatically created by SECO are tree indexes for images that are split in multiple areas, for better visualisation. Other complex indexes, such as authors, parts, chapters, etc., can be created only using the CONTENTE Local Application, or they can be also created trough SECO but in this case the respective XML descriptions (in METS) have to be provided with the submission of the images.

```
<?xml version="1.0" encoding="utf-8" standalone="no" ?>
<mets OBJID="cc-1134-a" LABEL="A new and general map of the southern dominions belonging to the United States of America, viz., London, 1794 [C.C. 1134 A.]" TYPE="ContentE v.1.0"
 xmlns="http://www.loc.gov/METS/" xmlns:xlink="http://www.w3.org/TR/xlink"
 xmlns:rights="http://www.bn.pt/rights/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://www.loc.gov/METS/ http://schemas.bn.pt/mets/v1.3/metsv1.3.xsd
 http://www.bn.pt/rights/ http://schemas.bn.pt/right/v1/rightsv1.xsd">
 <metsHdr CREATEDATE="2006-01-10T17:01:46" LASTMODDATE="2006-01-12T09:51:54"
  RECORDSTATUS="COMPLETED" />
- <dmdSec ID="dmd:PORBASE:UNIMARC:269622">
  <mdRef LOCTYPE="URL" xlink:type="simple" xlink:href="cc-1134-a_metadata/descriptive/cc-1134-a_unimarc.xml" MDTYPE="OTHER" OTHERMDTYPE="UNIMARC" MIMETYPE="application/xml" LABEL="C.C. 1134 A." />
 </dmdSec>
+ <amdSec>
+ <fileSec>
- <structMap TYPE="LOGICAL">
 - <div ID="w" LABEL="A new and general map of the southern dominions belonging to the United States of America, viz., London, 1794" DMDID="dmd:PORBASE:UNIMARC:269622" ADMID="r2" TYPE="BOOK">
  <fptr FILEID="jpg:cc-1134-a_JPG:cc-1134-a_JPG_24-C-W0140:cc-1134-a_0001_1_p24-C-W0140" />
  + <div ID="w_i0" ORDERLABEL="[Master]" LABEL="[Master]" ADMID="r1" TYPE="INDEX">
  - <div ID="w_i1" ORDERLABEL="Índice" LABEL="Índice1" ADMID="r2" TYPE="INDEX">
   + <div ID="w_i1_n0" LABEL="Obra Completa" ADMID="r2 t0" TYPE="MATRIX">
   + <div ID="w_i1_n3" LABEL="Obra - 9 Partes" ADMID="t1 r2" TYPE="MATRIX">
   - <div ID="w_i1_n16" LABEL="Obra - 25 Partes" ADMID="r2 t2" TYPE="MATRIX">
    <div ID="w_i1_n16_n17" LABEL="1" ADMID="r2">
     - <div ID="w_i1_n16_n17_n18" ORDER="10" LABEL="A" ADMID="r2">
       <fptr FILEID="jpg:cc-1134-a_JPG:cc-1134-a_JPG_24-C-R0150:cc-1134-a_0001A1_1_p24-C-R0150" />
      </div>
     - <div ID="w_i1_n16_n17_n19" ORDER="11" LABEL="B" ADMID="r2">
       <fptr FILEID="jpg:cc-1134-a_JPG:cc-1134-a_JPG_24-C-R0150:cc-1134-a_0001B1_1_p24-C-R0150" />
      </div>
```


Figure 4: Example of a XML METS created by CONTENTE

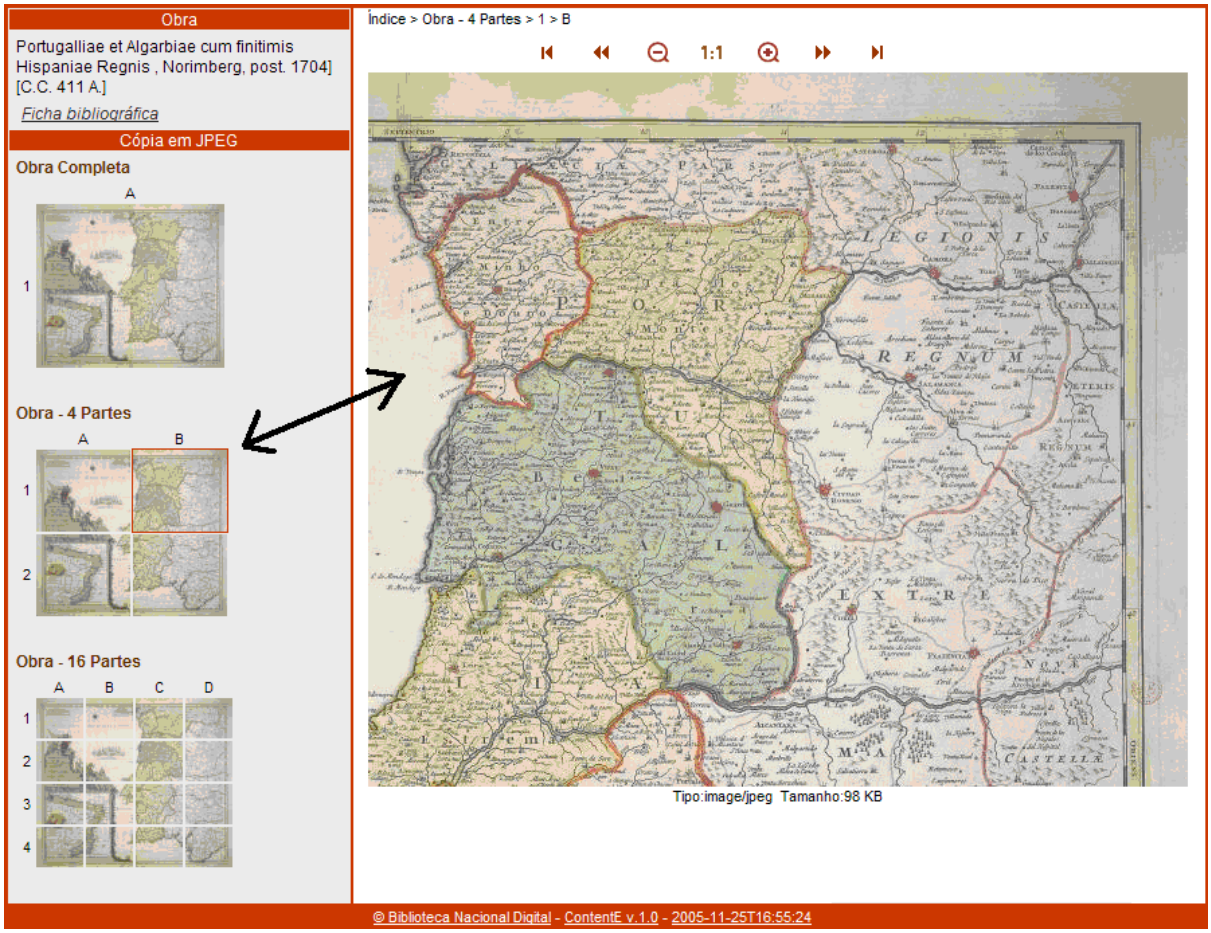

Figure 5: Example of an access copy of a digitised map structured by CONTENTE

The CONTENTE Library can retrieve descriptive metadata from external systems, such as PORBASE [11] and X-ARQ, or import it from local files (in Dublin Core or UNIMARC [13], coded in MARCXML). For rights metadata CONTENTE automatically assigned private right access to master copy, each access copy has also an access rights metadata structure, whose parameters can be chosen in the SECO's interface.

CONTENTE creates one access copy for each derived MIME type. Each of these copies contains not only a visualization XHTML site, but also a METS file describing the contents of each copy. This METS file also includes references to bibliographic, technical and access rights metadata.

To create these XHTML copies, CONTENTE can apply multiple XSL visual styles that are already integrated in the library. Figure 5 shows an example of an access copy for a work produced by SECO and CONTENTE using these procedures. It is a classical example of a map, for which it as asked to be created an index tree of three levels, where the first shows all the map, the second splits it in four parts (all with the same resolution of the image of the upper level), and the third splits it in 16 images. All was done automatically, once configured in SECO.

3.2 CONTENTE Local Application

In CONTENTE Local Application it's possible to do more complex works. The CONTENTE local application environment is split into two parts. The right pane presents the available content (the images), which are structured on the left pane (the indexes). Structuring is performed by means of a drag-and-drop mechanism. Fig. 6 shows an example of a work being processed using the CONTENTE Local Application (in this case one journal issue, on-line http://purl.pt/2065). Fig. 7 shows a result of a XHTML of another work, that was generated after structured in CONTENTE (this work is on-line at http://purl.pt/751).

In cases like these, it's possible to have detailed indexes or to associate different kind of metadata to different tree nodes, when specific node information is available. It is also possible to associate access rights to a specific node. All of this can only be produced manually, since it isn't automatically done through SECO.

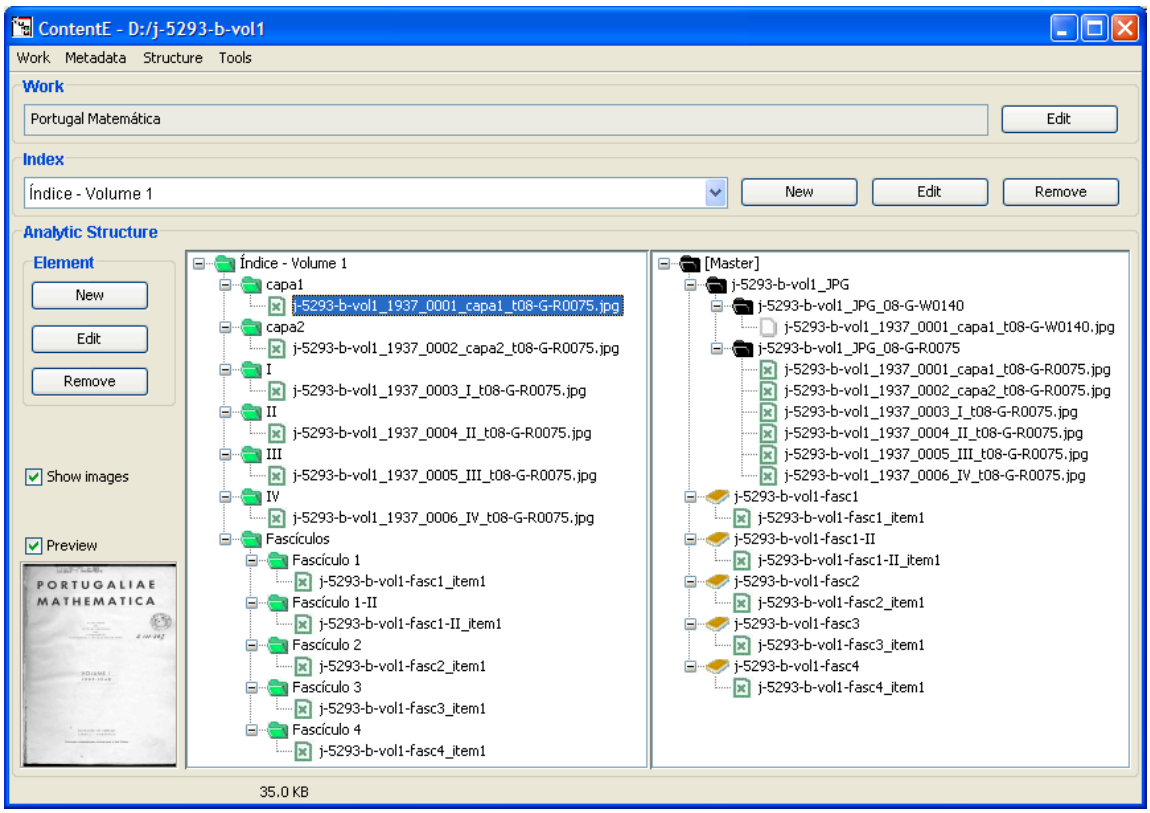

Figure 6: Example of a session using the CONTENTE Local Application to structure a digitised journal

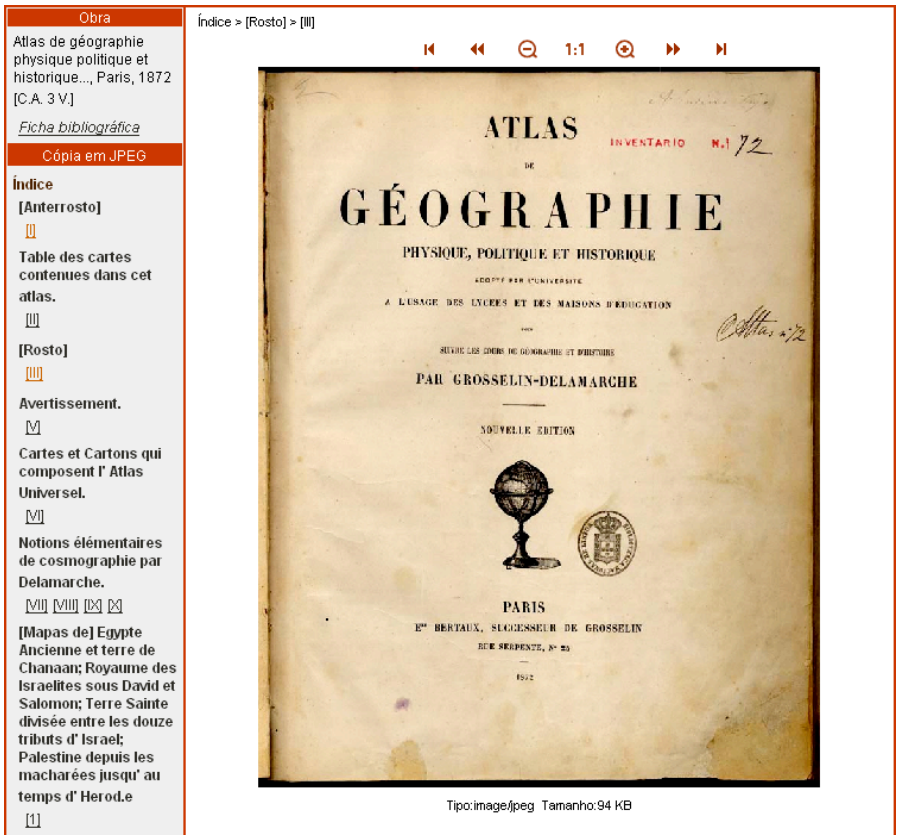

Figure 7: Example of an access copy of a digitised book created by CONTENTE

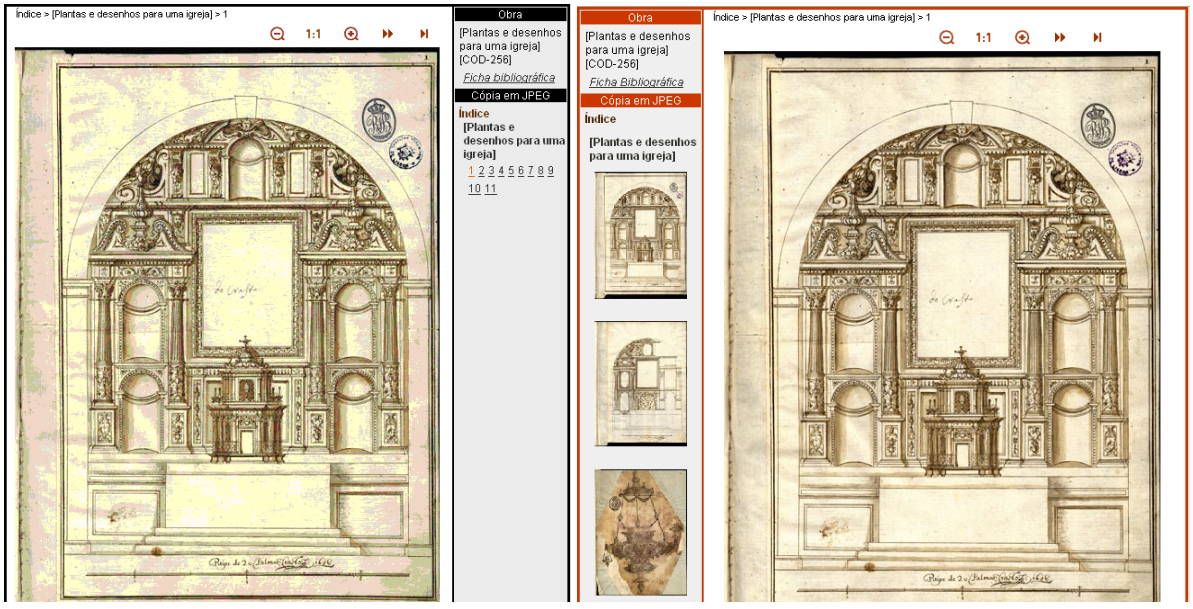

Figure 8: Example of multiple access copies (online at http://purl.pt/916)

Each style in CONTENTE is composed by an XSL file and an XML configuration file. This configuration file contains parameters for the XSL, enabling the user to perform modifications to the style configuration within the CONTENTE Local application itself. In Fig. 8 we show two examples of the same work where two different configurations were applied. For the CONTENTE library used in SECO, there are predefined styles for multiple genres of works, which can be applied automatically during work generation.

4 PURL.PT

If the automatic end result is acceptable, which experience has been shown to be true for most of the cases, the work can be tagged as completed and the SECO server notifies RDD service, an automated storage service. This service interprets a METS file designated METSItems, this file doesn't contain descriptive information about the copies; it only has references for them. Therefore, each copy has its own METS file with the description of its structure. The according to that METS it moves the data to the final storage spaces. These spaces are of the three kinds: a preservation space to which the master copy is moved, an internal access space for copies available only in the intranet, and a publishing space to which the public access copies are moved.

After this step, copies are also registered in the PURL.PT system. This system interprets again the METSItems file and registers each copy according to it. For bibliographic works already described in PORBASE, RDD service registers all other bibliographic information from copies on PORBASE; this is done only after confirmation of the registration of the works in PURL.PT service.

The PURL.PT service allows digital or digitalized works to be registered and accessed through unique and persistent URLs. These URL have the syntax "http://purl.pt/xpto", where "xpto" is a simple sequential number that started in 1 and grows up for each new work. These identifiers give access to a "home page" of the work, where a user can see descriptive and technical metadata (with references to the MIME formats of each copy), and have also access to the copies. The URLs for the copies have the syntax "http://purl.pt/xpto/copy", where "copy" is again a sequential number. The value zero is always reserved for the master copy, while the other copies are numbered sequentially without any particular order.

The system has an administration interface, a set of web services to interact with other applications. It works also as a proxy system to allow access control, with the architecture show in Fig. 9.

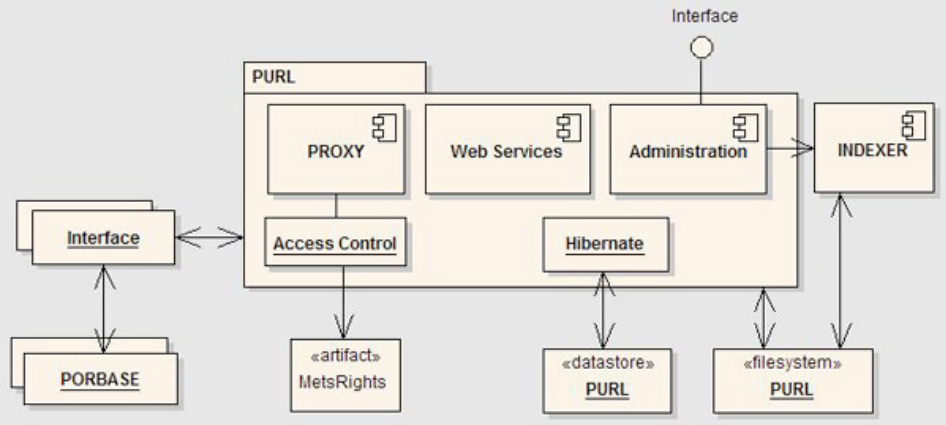

Figure 9: The architecture of the service PURL.PT

```
<?xml version="1.0" encoding="UTF-8" ?>
- <RightsDeclarationMD xmlns="rights" RIGHTSDECID="r2" RIGHTSCATEGORY="OTHER"
  OTHERCATEGORYTYPE="PUBLIC">
  <RightsDeclaration>Acesso público</RightsDeclaration>
  - <RightsHolder RIGHTSHOLDERID="BN">
    <RightsHolderName>Biblioteca Nacional</RightsHolderName>
    - <RightsHolderContact>
      <RightsHolderContactEmail>bndigital@bn.pt</RightsHolderContactEmail>
    </RightsHolderContact>
  </RightsHolder>
  - <Context CONTEXTCLASS="GENERAL PUBLIC" RIGHTSHOLDERIDS="BN">
    <Permissions DISCOVER="true" DISPLAY="true" COPY="true" DUPLICATE="true"
      MODIFY="false" DELETE="false" PRINT="true" />
  </Context>
  - <Context CONTEXTCLASS="INSTITUTIONAL AFFILIATE" RIGHTSHOLDERIDS="BN">
    <Permissions DISCOVER="true" DISPLAY="true" COPY="true" DUPLICATE="true"
      MODIFY="false" DELETE="false" PRINT="true" />
  </Context>
  - <Context CONTEXTCLASS="MANAGED GRP" RIGHTSHOLDERIDS="BN">
    <Permissions DISCOVER="true" DISPLAY="true" COPY="true" DUPLICATE="true"
      MODIFY="false" DELETE="false" PRINT="true" />
  </Context>
</RightsDeclarationMD>
```

Figure 10: An example of a rights metadata file

The PURL.PT service can process multiple descriptive metadata formats (UNIMARC, Dublin Core, etc.). Through the use of different style sheets (according to record format), a descriptive text is created for the presentation page of the work. This page also contains information regarding to the work's several digital items and its properties (total size of the copies, MIME type and access conditions), as well as for its physical items references; all that information is collected through each copies METS description. The METS' master copy also enables the creation of a HTML page with technical information about the different MIME formats and their characteristics that can be found in the copies (resolution, colour dept, etc.).

Each copy contains a XML rights file, referred through the METS file, from which the PURL.PT extracts the terms for access control. In BND we defined tree types of conditions: private, internal or public. Private items can't be accessed, which is the case of all the master items; internal copies can only be accessed at the intranet; and public copies are accessible to all users, including the Internet. The actual schema for this format is available at http://schemas.bn.pt/right/v1/rightsv1.xsd, and a sample is shown in Fig. 9.

In Fig. 11 we can see an example of the administrative interface of the PURL.PT service. In Fig. 12 we have an example of a descriptive page created automatically for a digitised work with four copies, one master and three copies for access (work available at http://purl.pt/724). It is interesting to see also that due to the interoperability with PORBASE, the national union bibliographic catalogue, we can see the references to the physical item that was originally digitised (note "Digitalizado em: PURL 724"), as also the references to other copies existing in other libraries members of catalogue.

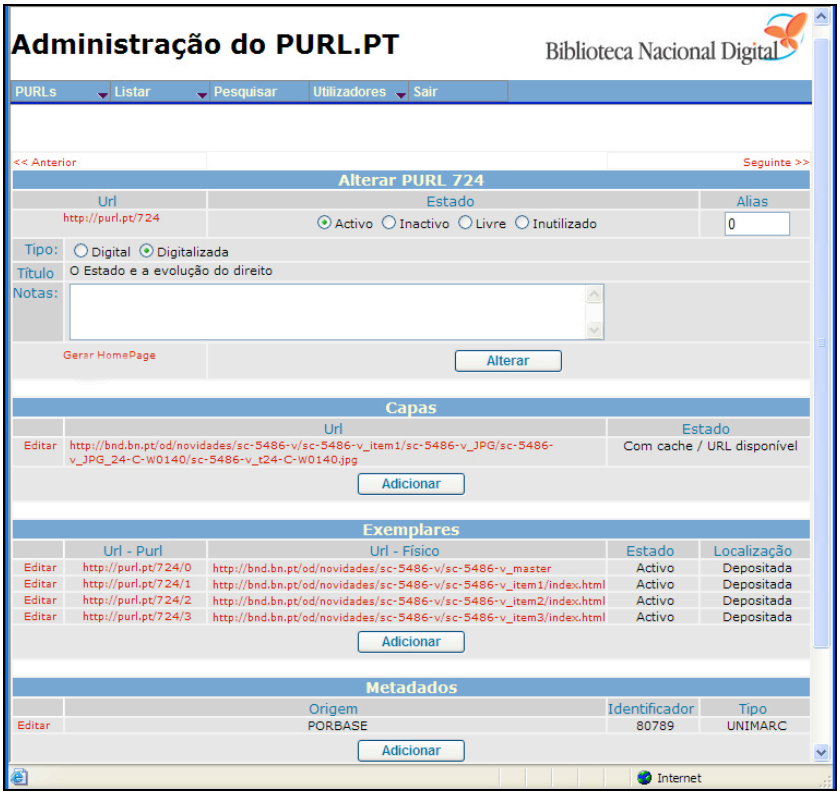

Figure 11: Administrative interface for the service PURL.PT

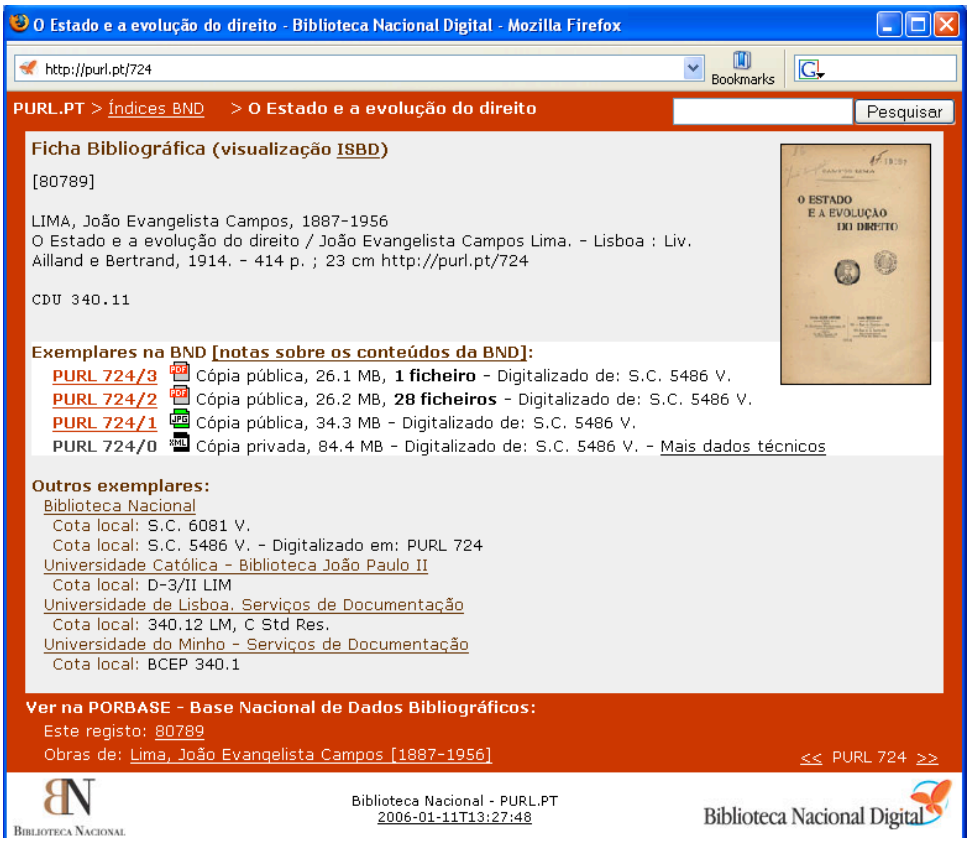

Figure 12: Descriptive page for a digitised work created automatically by the PURL.PT service

5 Conclusions

In this paper we explained how we harnessed a very complex problem with a strategy based on the development of a processing system integrating several functional blocks. The overall system is actually in production at the BND, and the experiences had so far make us to believe that it'll make it possible to publish, in a short time and with fair human intervention, the more than 20.000 digitised titles that are still waiting in the queue. The BND published already by this process more than one thousand of titles, and we expect to publish the remaining until middle of 2006. Meanwhile other digitisation programmes will be started, which will create new images and new titles for publishing. Another important aspect is related with the use of the METS in all the process, which demonstrated to be a simple, practical and flexible structural schema.

All the code was developed in JAVA, and all the results are available in open-source. We have also plans to continue the developments, especially to reinforce SECO with more image processing features (such as to cut margins and improve the legibility in images), add more schemas to CONTENTE (such as the DOCBOOK [8] and the Digital Talking Book [5] defined by the DAISY Consortium [7], which will make it possible to process also object with sound, for visual impaired persons). Finally, the PURL.PT service has many other features that were not mentioned in this paper, especially to support browsing in the BND and the publishing of collections and profiles, all supported easily thanks to the global usage of well defined XML schemas. All of this will make it possible to support new services, such as, for example, a user space under development, where users will be able to register and take advantage of new customised services.

References

[1] DjVU. <http://www.djvuzone.org/>

[2] HEDSTROM, Margaret. Exploring the Concept of Temporal Interoperability. Proceedings of the Third DELOS Network of Excellence Workshop on Interoperability and Mediation in Heterogeneous Digital Libraries. Darmstadt, Germany, 8-9 September 2001. <http://www.ercim.org/publication/ws-proceedings/DelNoe03/10.pdf>

[3] ImageMagick. <http://www.imagemagick.org/script/index.php>

[4] CRIMMINS, Thomas R. Geometric filter for Speckle Reduction. *Applied Optics*, 15 May 1985, vol. 24, no. 10.

[5] ANSI/NISO Z39.86-2005. Specifications for the Digital Talking Book. ISSN: 1041-5653 <http://www.daisy.org/z3986/2005/z3986-2005.html>

[6] BND. Biblioteca Nacional Digital. <http://bnd.bn.pt>

[7] DAISY Consortium <http://www.daisy.org/>

[8] DOCBOOK. <http://www.docbook.org/>

[9] MARCXML. MARC 21 XML Schema. <http://www.loc.gov/standards/marcxml/>

[10] METS. Metadada Encoding and Transmission Standard. http://www.loc.gov/standards/mets>

[11] PORBASE. Base Nacional de Dados Bibliográficos. <http://www.porbase.org>

[12] The Dublin Core Metadata Initiative <http://www.dublincore.org>

[13] UNIMARC <http://www.unimarc.info>

[14] Service-Oriented Architecture <http://en.wikipedia.org/wiki/Service-oriented_architecture>