Implementing a funder repository with heterogeneous material and advanced presentation capabilities

Ioanna-Ourania Stathopoulou, Nikos Houssos, Panagiotis Stathopoulos, Despina Hardouveli, Alexandra Roubani, Alexandros Soumplis, Chrysostomos Nanakos

National Documentation Centre / National Hellenic Research Foundation
{iostath, nhoussos, pstath, dxardo, aroub, soumplis, cnanakos}@ekt.gr

1 Introduction

The present paper briefly describes the development of a funder repository aiming at the dissemination, reuse and preservation of digital material of diverse types, providing an enhanced user experience, which was produced under the auspices of large scale (multi-billion Euros) funding programmes of the Hellenic Ministry of Education (co-financed by the European Union). This involved the handling of a wide range of content, like (among others) educational material, books, peer-reviewed scientific articles, conference proceedings, theses, videos and studies/reports. The DSpace platform has been selected and used for the implementation of the repository and fulfilled the challenging requirements at hand, for example the heterogeneity of material; several DSpace extensions were also developed to assist both the back-end procedures for cataloguing and digital file processing as well as providing an enhanced end user experience. The project has been successfully completed and the system is publicly available since spring 2011 at http://repository.edulll.gr.
The rest of the article at first explains the overall project scope and the followed methodology and approach, including important issues encountered and relevant decisions that had to be made. Implementations aspects, particularly regarding metadata schema and workflows are presented in Section 3. DSpace enhancements are described in Section 4 and the paper concludes with summary, lessons learnt and ideas for future work.

2 Project scope, methodology and approach

Since 1994, the Hellenic Ministry of Education has been executing a series of large scale (at the range of billion of Euros) funding programmes, co-financed by the European Union through Structural Funds, under the names Operational Programme for Education and Initial Vocational Training (EPEAEK I & II, 1994-1999 and 2000-2006, respectively) and Operational Programme for Education and Lifelong Learning (2007-2013). The entity responsible for the management of the programmes is a Special Managing Authority within the Ministry [www.edulll.gr].
The programmes finance a variety of activities related to all level of education (i.e. K-12, Higher Education and Lifelong Learning) including, among many others, restructuring and enhancement of educational programmes, production of educational material, scholarships and fellowships for doctoral students and post-doctoral researchers, collaborative research projects with the participation of Greek Higher Education Institutions, as well as a wide range of studies and reports, conferences and events, seminars, etc.
In autumn 2010, the respective Special Managing Authority assigned to the Hellenic National Documentation Centre the task of organizing the digital material that has been produced in the course of the aforementioned funded activities into a publicly available repository that would provide a single point of access to the programmes results, ensure wide dissemination and reuse of content through enforcement of interoperability standards and also address certain long-term preservation aspects.
The input material available from the Special Authority had the form of digital files, arranged in folders mainly according to project, however without any metadata attached to them. A considerable portion of these primary sources concerned administrative reporting during project execution and related files (e.g. administrative reports, detailed schedules of courses and seminars, participants'

lists for seminars / events, etc.) It was decided to ignore this kind of material so that the repository includes only items that contain content of archival value.
Processing of digital files before publication via the repository was a challenging issue due to the inherent heterogeneity of the material and the non-uniformity in formats, naming and structure. A considerable part of the overall project concerned this aspect. An important decision was to store and preserve both initial and processed digital files within the repository so that potential problems in the sometimes non-trivial processing workflow could be corrected a posteriori. Notably, providing perpetual persistent access to the the digital material and handling certain preservation issues was an important goal of the project; among the main reasons for that was the fact that a significant part (probably the majority) of the content concerned grey literature that would not be handled through official publishing and preservation channels.
Regarding metadata schema, a custom application profile was created, based on existing common international standards. This has been the result of an iterative process. The first phase consisted of a detailed mapping of the input digital material during which the most common types of items were identified and candidate fields per type were defined. This activity was necessary due to the diversity of the material and the limited available information initially on types and formats of content. It has been a tedious – despite assisted to some extent by automated tools – process of going through the directories/files and capturing types of items, formats and doing an initial distinction of content candidate for inclusion to the repository and administrative documents. The first phase produced a first version of the metadata schema that was to a large extent stable for the most common document types and enabled the beginning of the cataloguing process early in the project, followed by subsequent phases of refinement. Further details on the metadata schema are provided in paragraph 3.1.
Certain important cataloguing decisions were made early on. First, there was a considerable allocation of effort to subject cataloguing and appropriate user interfaces for browsing by subject (see paragraphs 3.1 and 4.2), since the diversity of the material would make it inadequate to rely only on navigation/browsing based on fields like keywords, type, project or collection. Secondly the idea of self-archiving was abandoned, at least for this first phase project and metadata entry was assigned to experienced information scientists that could safely distinguish among archival and administrative material, identify and remove duplicates (probably scattered across different folders of primary material), perform subject cataloguing, produce abstracts of items (where appropriate) and provide clear guidelines to the digital files processing staff (e.g. for file merging/splitting).
Furthermore, the heterogeneity of the material created a requirement of various ways of browsing to be applied (e.g. not only listings but also tag clouds for presenting subjects and keywords; image-based browsing of material types) and also accessing the digital material itself (e.g. both the ability to download a PDF file and to stream it via an online reader; video material embedded into item pages).

3 Implementation

3.1 Metadata schema

Due to the heterogeneity of the material to be included in the repository, a custom application profile has been developed, utilizing elements from various metadata standards, in particular Dublin Core, Learning Object Metadata (LOM) [http://ltsc.ieee.org/wg12/] and PREservation Metadata Implementation Strategy (PREMIS) [http://www.loc.gov/standards/premis/]. Furthermore, a new EKT namespace/schema was created to capture custom elements that were required for our implementation but were not available in standard schemata.
The EuroVoc thesaurus [http://eurovoc.europa.eu/] has been utilized for the thematic indexing of the material, as well as for spatial coverage. This particular vocabulary was selected due to the following properties:

(a) Breadth and depth of coverage and inter-disciplinarity. EuroVoc is a detailed thesaurus across scientific disciplines and domains, so it is perfectly suitable for the diverse material of our case.

(b) Multi-linguality. EuroVoc is available in 24 European languages. Therefore, the thematic index can be made automatically available in all these languages.

Furthermore, regarding education levels, the controlled vocabulary of the Eurydice network (http://eacea.ec.europa.eu/education/eurydice/eurybase_en.php#greece) was adopted.

3.2 Workflow

As mentioned above, the content of the repository includes diverse material in terms of their format and type, therefore it was necessary to implement a workflow which would include the required processing of the digital data depending on their type and format. The workflow of the repository consists of the following steps:

1. Firstly, the item is submitted to the repository using an appropriate submit form. This step is performed by authorized users, specialized in the organization and management of information materials (in our case, information specialists who belong to the library staff of EKT/NHRF). In addition to providing the necessary metadata, the information specialists also upload the input digital files and provide, in separate fields and in a custom specification language, the necessary instructions for the processing of the digital file(s) by the digital material processing staff.
2. Once the submission by information specialists is completed, the item metadata are publically available online. The digital files are already within the repository, but not visible and available only to authorized personnel.
3. The processing of the digital files from the respective expert staff takes places. It follows the directives of the information specialists for the production of final digital items for publication of each record. In case of text material, the final digital item must comply with the following rules:
   i. The final document must be a full-text indexable and searchable pdf file. Optical character recognition (OCR) might be needed to achieve that, for example when the input file(s) are in pdf format, and they are either image pdfs or have improper encoding or font support (e.g. for Greek content).
   ii. If specified by the information specialist, the final document(s) may be the outcome of merging or splitting the input data files.
   iii. The name of the final document should follow a particular pattern, which includes the item's handle in the repository.

   In the case of video material, it has been selected to be hosted by an external FLV streaming video server and integrated to the repository using enhanced FLV players. Video has been received in various formats and the workflow is as follows:
   i. The video is encoded in H.264 format and the files are saved in the standard F4V file format.
   ii. If specified by the cataloguer and/or considered necessary by the digital material staff, some video editing or enhancement takes place.
   iii. The video is uploaded to a the dedicated video streaming server (WOOWZA) and the RTMP link is stored in a separate field in the repository.
   iv. The embedded video player is generated dynamically according to the stored RTMP link, which includes also time related information, i.e. start and stop of each video segment corresponding to a serarate metadata entry.
4. After the processing of digital files the respective digital files processing staff member records in the appropriate metadata field that the processing has been completed and the final files become publicly available.

4 DSpace extensions

DSpace platform was greatly customized and enhanced in order to accommodate the requirements of this specific digital repository implementation. The main issues that have been addressed include modifications in the submission form in order to simplify the process, supporting multi-lingual controlled vocabularies both in metdata entry and presentation to end user and context-sensitive presentation of material as well as a range of modifications in order to encourage efficient and

effective human-computer interactions and optimize the overall user experience in the repository.

4.1 Submission form modifications

The default submission process in DSpace platform comprises quite a few steps (collection selection, metadata entry, digital file upload, verify and grant copyright license). In order to somewhat simplify the process, we changed the submission form so as the user would be able to provide the item metadata and upload one or more digital files in the same step. Thus, the submission process is more simple and straightforward, as it consists of three steps: (1) collection selection, (2) describe (metadata entry including copyright, upload one or more files) (3) verify.

Moreover, the following new input-types have been developed: *'advancedName', 'refinementdrop_advancedName', 'refinementdrop_value', 'multiTextArea', 'onebox_lang', 'textarea_lang', 'startHeading' and 'endHeading'*. These input-types include capabilities like dynamic processing of persons' names which may be entered in various patterns and storing the value as expected by the repository (e.g. *{Surname, First and Middle Names}*), further refine the values by using controlled vocabularies, handle multiple inputs from the same textarea, or allow the user to choose the language of each metadata value.

Finally in order to improve the user experience and make the submission easier, the submission process was modified in order to be fully localised and support different languages. Specifically, all the labels and control vocabularies in submission form can be configured from the message properties and input-forms configuration file and be displayed in different languages. The item presentation page, also uses these configuration files in order to display the stored control vocabulary in the user's respective language. Also, help tips were implemented which are shown when the user clicks on one of the submission form fields and provide context-sensitive help.

4.2 Visual representation of browsing

DSpace browsing functionality was greatly enhanced by applying visual representation mechanisms. The visitor can browse through the subject classification metadata using a tag cloud which provides a quick perception of the most prominent EuroVoc terms. The browsing functionality with tag cloud can be fully customized through the DSpace configuration file. Specifically, we can define how the tag cloud will be displayed (e.g. maximum number of display terms, minimum acceptable occurences of a term in order to be shown in tag cloud, etc.) and select which metadata fields this browsing should be applied on.
Furthermore, a custom, more user-friendly browsing by type has been implemented using a characteristic image per type, in addition to string labels.

4.3 Video integration

As already stated video content is a considerable part of provided content. It included full length documentaries, short stories for educational use, online courses, etc. Various file formats and encoders were used in the content, from DVD MPEG2 VOD files to proprietary video files and encoding formats (H.263, MPEG-1, MPEG-4, DIVX). On the other hand a unified, meaningfull and intuitive experience should be provided to the end user when accessing this content. Relevant requirements were:

- Easy to use and intuitive user interface for streaming video.
- The option for downloading the video file should not be given, due to IPR issues.
- Integration of a web based video player to the repository and not an external video player program.
- Based on open file standards and codecs, if possible.

In order to cope with the various file formats and encoders the following decision were made:

- Content was batch transcoded to H.264 enconding using a F4V video file container, selected based on the encoder properties and its open nature.
- The video stream was provided by an appropriate video server external to the core repository system. The Woowza video server was selected based on means of versatility in the protocols used (RTMP and RTSP) and efficiency.

- The flowplayer embedded video player was selected based on features and configuration options. The flowplayer configuration is dynamically generated based on the video repository metadata. E.g. in case of a single video stream, with different repository entries, e.g. a DVD with multiple episodes, it was decided that instead of splitting the video file, rich metadata, including episodes start and end time would be included. This approach enables the capability to easily rearrange the video player configuration, be independent of particular video player and streaming servers and provide future enhancements, according to available metadata.

Relevant examples are available at repository.edulll.gr on the video content category, demonstrating the dynamically generated embedded video player according to the video's repository metadata.

4.4 Providing an ebook like experience as an alternative to PDF viewer

The initial format of most of the material imported to the repository was PDF files generated by a variety of tools. While PDF is a widespread format, a large percentage of the files included rich photographic material, multihundred pages books, or has been OCR processed, thus file sizes of tens, or even hundreds, of megabytes were common. Therefore an alternative was sought in order to not only make the time required for the initial viewing of the document shorter, but also to provide a more intuitive and user friendly user experience.

Initiatives such as the "Google Books" and "Google Art" project, the Internet Archive "Open Library" and advanced repository systems, e.g. Islandora, etc., have paved the way for novel online reading capabilities and experience, with features such as "page by page" reading of electronic resources and tile-based image viewing systems, exploiting advanced codecs such as JP2000.

In order to provide such page by page viewing experience the following backend components were integrated with DSpace:

- A multithreaded conversion management tool, in order to interface with DSpace and manage the batch conversion process from PDFs to page by page JP2000 files which wraps around existing standalone converters. This tool can batch convert selected PDF content from a DSpace repository to JP2000, with the conversion process transparently being initiated from the repository manager. It is available as open source software at code.google.com/p/jp2k-distiller/ and has been already been integrated also with the Open Journal Systems platform.
- The DJatoka image server for providing the page by page content in openurl format and in various sizes. This is set up on a shared clustered deployment, exploiting multiple virtual servers over a shared storage, utilising advanced caching, load balancing and failover. For scalability and performance, mapping of open URLs to local files is retrieved from a clustered PostgreSQL 9 installation.
- The online reader, based on the Internet Archive Open Library BookReader (http://openlibrary.org/dev/docs/bookreader) provides the end user interface. It features advanced end user capabilities such as interactive online reading with zooming, thumbnail view, full-text search and hit highlighting.

Overall the system provides a transparent manner to view and interact with PDF based content, using a more intuitive user experience, especially in the case of multi-MB PDF files.

5 Conclusions - future work

In this repository development use case, technical enhancements were used to provide the means to organise, process and import to the repository a wide range of heterogeneous material. Special facilities for the support of these workflows was included, along with enhanced user viewing capabilities for metadata, PDF files and video content, providing an end user experience suitable for the repositoty's users, which include the educational community, researchers and scientists and the general public. Future work, possibly in the frame of a sequel project, includes the inclusion of further material from past funded projects, and the integration of a workflow for the quick inclusion of content produced by currently running projects.