A service-oriented national e-theses information system and repository

Nikos Houssos, Panagiotis Stathopoulos, Ioanna Sarantopoulou, Dimitris Zavaliadis, Evi Sachini
National Documentation Centre / National Hellenic Research Foundation

Introduction

In this article we present an overview of the information technology infrastructure that supports the operation of the Greek National Archive of Doctoral Theses (HEDI). The infrastructure, which has been recently re-developed replacing a legacy system, makes use of repository software, in particular the DSpace platform, as part of a service-oriented information system based on open source components.

The National Documentation Centre (EKT) of the National Hellenic Research Foundation has been granted by law 1566/1985 the responsibility of developing and maintaining the Greek National Archive of PhD Theses. The archive contains the doctoral dissertations produced in Higher Education Institutions as well as a number of PhD theses awarded to Greek scholars by universities outside Greece (USA, UK, Canada, and Germany, among others), in total about 24.500 theses as of February 2010, 2.75M pages of digitised and born-digital dissertations, with 1200 -1400 new dissertations arriving every year.

HEDI is supported by IT systems since 1986 when EKT developed the bibliographic database 'National Archive of PhD Theses" employing for cataloguing the home-grown library automation software, ABEKT [1]. Initially, EKT has been collecting from individual universities and cataloguing theses solely in print form. The database was made available to the public via the mainframe host computer 'Hermes' for more than a decade from 1986 until 1999. Thereafter, a new version of ABEKT has been used, including support for metadata standards like UNIMARC, UNIMARC Authorities and ISO 2709 and the Z39.50 protocol for search and retrieval of bibliographic records. This system, later integrated into the ARGO digital library portal [2][3] that is still in operation, provides free Internet access to metadata as well as advanced services to librarians, through a library catalogue-like user interface. Meanwhile, EKT proceeded with executing a major digitisation project for the majority of the dissertations in the – until then – print-only archive, which enabled open access to theses full text for Web users – realised through a specialised presentation application [4]. Furthermore, in later years universities have been submitting theses to the archive, in both print and electronic form. As a result, today more than 76% of the theses in HEDI are available online in full text.

In 2009, a decision was made to re-build and modernise the information infrastructure supporting HEDI – a project that was completed in early 2010. The following main choices were made:

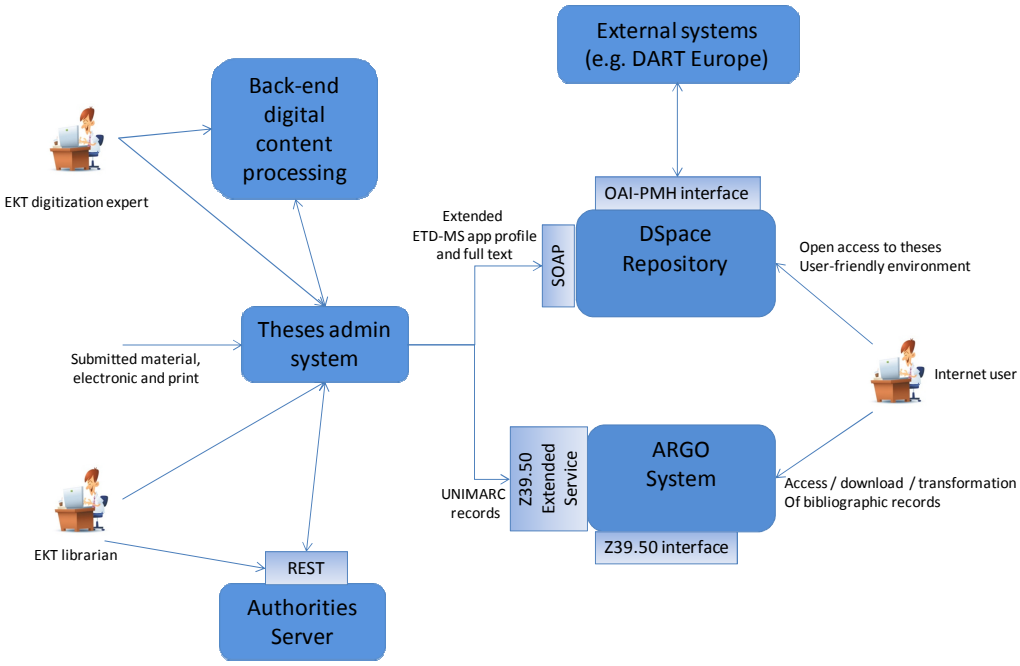
- Create an e-theses repository on the DSpace platform, mainly targeting end-users. The migration to a repository platform like DSpace, was considered extremely important due to user-friendly, highly configurable services (e.g., search/browse, alerting), easier interoperability with other systems both internal to EKT and external (e.g., document ordering, web 2.0 Internet sites for bookmarking, tagging, sharing content) and improved visibility due to high compatibility with major Internet search engines like Google and international repository and e-theses directories. Another factor for introducing DSpace, was to adopt an open repository platform for the storage and preservation of the digital content held in HEDI, which, until recently, have been entirely handled by proprietary, commercial software, namely the Filenet platform.

- Create a separate IT application for the administration and management of the EKT internal workflows that are necessary for the processing of the material submitted to HEDI. The output of these workflows for every thesis is a quality-controlled metadata record and a searchable full-text file (or set of files). An important aspect to consider is that these workflows should be able to handle dissertations in both print and electronic form – note that the print archive is still maintained although a thesis is normally submitted in both print and electronic form.
- Maintain and constantly update the ARGO version of HEDI as a system of choice for librarians. Notably, ARGO is a portal providing free access to a large number of bibliographic databases and union catalogs, both to users and librarians, offering to the latter additional important services like record export and transformation among various formats (e.g., MARC21 to UNIMARC). Ceasing to provide the HEDI database through this portal was considered to be a non sound option, given the popularity of ARGO among the Greek library community as well as the support of UNIMARC and related services.
- Select and reuse open source main software components, from the operating system to the repository and the middleware/database layer, while exploiting EKT's virtual infrastructure [6] in order to provide HEDI services with higher availability, scalability and total cost of ownership. The same software components and technology have been previously used by EKT while developing institutional and subject repositories.

System overview and architecture

The design of the overall infrastructure follows the Service Oriented Architecture (SOA) paradigm; no single monolithic system would be able to address all the aforementioned requirements. For example, DSpace is suitable for storing and providing end-user services but would not suffice to support workflows for handling of printed theses and monitoring / reporting of the internal procedures of processing material submitted to HEDI (e.g., throughput, time-to-repository, identification of processing bottlenecks).

An overview of the entire service-oriented IT environment for the support of the HEDI archive is depicted in the following diagram.



The theses administration/management system (PhDMS) is employed to manage the EKT internal workflows, mainly for the following tasks:

- Recording the submission of each thesis in print and electronic form and temporarily store the submitted electronic files for subsequent processing.
- Production of a quality-controlled metadata record for each thesis on the basis of the submitted material. Automatic export to both UNIMARC (for ARGO) and a custom ETD-MS based application profile (for DSpace) and update of the respective systems; note that even subsequent updates / corrections of the bibliographic record happen at a single place: the PhDMS; synchronisation with ARGO and DSpace is performed automatically with every modification in PhDMS. The ARGO update procedure is based on the open-source Ruby implementation of the Z39.50 Database Update Extended Service. DSpace update is based on a custom SOAP/WSDL based protocol – the Lightweight Network Interface was not employed since it was found to be possibly more cumbersome to implement the required functionality using it. Special mechanisms are used for attachments (full-text material) using socket-based communication developed using the Spring framework Web Services tools.
- Monitoring of the internal workflows for processing of the incoming archive material; on-demand generation of appropriate reports reflecting the archive status (e.g., material in the archive, processing throughput, etc).

The back-end digital content processing infrastructure contains systems such as batch OCR functionality as well as services for digital files quality control and transformations to achieve the following objective:

- Production of a searchable thesis full text file conforming to appropriate technical and quality specifications. This might involve manipulating and performing OCR on material that is even available in electronic form but not ready for inclusion to the repository (e.g., documents with non-Unicode encoding of Greek characters) and producing the page files for the required on-line reading application. Scanning / OCR and post processing could also be employed for old, not yet digitised dissertations.

The DSpace repository (available at http://phdtheses.ekt.gr) performs the following functions:

- Storage of the digital material associated with every thesis, essentially the thesis full text in searchable form.
- Keeping for each thesis descriptive metadata useful for the end user and technical metadata for preservation. The metadata schema employed is an application profile based on the ETD-MS interoperability metadata standard for theses.
- Offering end-user services like search, browse and alerting. Interoperation with other systems enables linking to other services like print-on demand and digitisation-on demand for theses and sharing and bookmarking repository pages in web 2.0 and social networking web sites.

Through the DSpace repository, HEDI has become a data source for DART Europe - the pan-European e-theses portal [7] and also the European Working Group of the Networked Digital Library of Theses and Dissertations (NDLTD), of which EKT is a partner.

The ARGO system provides search, retrieval and transformation facilities for the bibliographic records in HEDI.

A separate authorities server is employed to store established names for academic institutions, subject keywords and author / supervisor names. The system is compatible with the MADS standard authority files format. This server is accessible both from a graphical interface enabling data management (e.g., creation/update/deletion of records) by the EKT library personnel and a programmatic RESTFul web service interface to enable, for example, auto-complete in cataloguing forms within PhDMS.

In short, the developed system allows us to:

- Fully manage the administrative workflows for processing the incoming material.
- Achieve high degree of interoperability with a significant array of scholarly digital content. Since an e-thesis is for universities a distinct type of research output, it requires special treatment for publishing in repositories alongside other forms of research output, such as peer-reviewed journal articles.
- Make theses openly accessible in a user-friendly web environment, featuring both pdf and an online page-based reader, according to the user preferences.
- Make theses bibliographic records and related services freely available to interested Internet users, in particular librarians.

Design choices and technical challenges

The most significant design choice in our system was to follow a service-oriented architecture approach. ARGO, DSpace and the authorities server have been identified as the basic service building blocks that would support the PhDMS. A relevant decision was to build an entirely new administrative / management system – the PhDMS, separating the administrative functions from those of the repository, the bibliographic Z39.50 server and the authorities server. This separation of concerns is to our view important to achieve maximum future maintainability, flexibility and sustainability of the system – this approach is preferable than stretching existing systems to functionality that they are not designed for [5]. For example, trying to adjust DSpace for being the single platform in our system would be a non-optimal solution since our particular administrative needs significantly exceed the scope of any repository software platform. This choice naturally has the cost of providing interoperation among systems, for example among PhDMS and DSpace, however this has been achieved with reasonable cost – actually the only issue where we faced noteworthy difficulties was the transfer / update of the digital bitstreams to DSpace; nevertheless they have been overcome without much pain.

Another design decision with serious impact on the implementation was that the cataloguing would – as before – be UNIMARC-based and support of ARGO would continue. This required the creation of a detailed internal PhDMS data model that would fulfil the UNIMARC needs, the automated export of this model to UNIMARC and the subsequent update of ARGO with every newly completed thesis record. Despite the cost involved in achieving the above, we feel that it was worth it, since it kept HEDI, its detailed metadata schema and related services accessible to the Greek library community in a familiar form within a popular portal. Furthermore, EKT expertise both in the IT and the information science aspect of relevant standards (e.g., Z39.50, UNIMARC) made these tasks absolutely feasible with reasonable effort.

An important technical challenge concerned the development of the mechanisms for updating ARGO and DSpace as distributed transactions, so that cases of update failure are handled gracefully by the system (e.g., avoid the case that the synchronisation state among PhDMS, ARGO and DSpace is not fully clear). Given that ARGO and DSpace do not inherently provide transactions (e.g., no explicit rollback is supported), this functionality has been quite tedious to implement, however it has been successfully incorporated in the overall infrastructure.

The whole software stack for the delivery and management of HEDI was chosen to be based on either open source or home grown software. For the delivery of the HEDI a three tier model is employed, with each required server being a virtual machine running over at EKT's virtual infrastructure. Computing and storage resources allocated to the HEDI archive are flexible, and can be allocated dynamically based on the observed, by EKT's monitoring system demand. The monitoring system constantly checks crucial performance parameters of each tier and the proper execution and response time of each of the infrastructure's components and services. The Filenet proprietary software and Oracle database which were

used to deliver the HEDI archive and the online thesis reader were replaced with a more flexible open source approach. The migration of data to PhDMS and DSpace from the existing ARGO database and the legacy IT applications used by HEDI has also been a major challenge. A rigorous, semi-automated procedure has been followed before the migration for "cleaning legacy" data and thus, achieving higher metadata quality. This was an absolutely necessary step for making the migration feasible within a demanding time schedule.

Summary and future work

This article provides an overview of important issues and design decisions involved in the development of an e-theses IT infrastructure, in particular the Greek National Archive of Doctoral Theses (HEDI). The adoption of the service-oriented architectural paradigm, the identification of appropriate services and systems for each task and the use of a mature open repository platform (DSpace) allow the successful management of a large e-theses archive with both print and electronic material.

Future plans include, among others, the application of automated metadata extraction for the creation of bibliographic records, the automation of the procedure of digital file quality checking and the interoperation with CRIS systems [8] using the CERIF standard [9] connecting the theses with structured information about authors, organisations as well as research projects which have specifically funded PhD theses.

References

[1] ABEKT Library Automation Software, http://abekt.ekt.gr.

[2] ARGO digital library portal, http://argo.ekt.gr.

[3] Sfakakis, M., Kapidakis, S., 2003. An architecture for online information integration on concurrent resource access on a z39.50 environment. In: Research and AdvancedTechnology for Digital Libraries. pp. 288-299.

[4] Loverdos, C., Kapidakis, S. Flexible, service-based content presentation: The Hellenic Dissertations Presentation System. Demonstration on the 5th European Conference on Research and Advanced Technology for Digital Libraries, September 4-8 2001, Darmstadt, Germany.

[5] Atkinson, L. A., June 2006. The rejection of D-Space: Selecting theses database software at the university of Calgary archives. In: 9th International Symposium on Electronic Theses and Dissertations (ETD). Available at http://eprints.rclis.org/7938/.

[6] Stathopoulos, P., Soumplis, A., Houssos, N. 2009, June. The case study of an F/OSS virtualization platform deployment and quantitative results, 5th International Conference on Open Source Systems, Skövde, Sweden.

[7] DART-Europe E-theses Portal, http://www.dart-europe.eu.

[8] euroCRIS, organisation responsible for standards for Current Research Information Systems (CRIS), http://www.eurocris.org.

[9] Common European Research Information Format (CERIF), http://www.eurocris.org/cerif/introduction/.