DIGITAL OBJECTS – STORAGE, DELIVERY AND REUSE

Juliana Peneva, Stanislav Ivanov, Filip Andonov, Nikolay Dokev

Abstract: The development of a methodology and tools for an automatic extraction of metadata for digital objects deployed in various subject repositories is a potential research issue. This paper presents a general overview of topics concerning the building of repositories and the applied metadata schemas with respect to the objectives of METASPEED project. The main goal of this project is to investigate and create technologies, methods and tools for automatic generation of metadata thus facilitating the proper specification of documents with different e-format, content and location.

Key words: Digital object repository, metadata schema

Introduction

Nowadays the proper supervision of organizational digital resources is very important and many companies are realizing a business advantage by managing successfully their business data. Resources are built of different kind of documents ranging from images, video or audio clips, animations, presentations, online courses, web pages, to name a few. Organizations vary in types and sizes but all of them exhibit an intensive use of digital resources because these resources are stored, distributed, shared and reused without difficulty. Certainly some barriers like technical incompatibility or missing files are to be overcome to achieve an effective use. However digital resources are increasingly being recognized as a very important organizational asset on a par with finance and human resources. So, building repositories to manage the digital content is a very important activity that brings value in the inventive deliverables of the overall organization. Each time a digital resource remains undiscovered or simply not used the organization waste time or staff efforts, misses opportunities or looses possibilities to gain a competitive advantage.

The business managerial and technical benefits of digital resources are summarized in [1]. In order to examine their value [2] and to consider the opportunities for reuse, digital resources are organized in repositories that support the organizations' policy on digital asset management. During the last five years different types of repositories ranging from digital libraries through various institutional collections and e-journals up to collaborative learning environments have been built. Each of these systems contains thousands of digital objects in the form of data and/or metadata. Content is added to a repository via different workflows and tools, and represented to the repository clients via different mechanisms. Companies as Google and Microsoft [3] are reporting for own repository investigations as well. In addition there are many workshops and the annual Open repositories [4] conference that stress on important issues concerning repository creation and management. Nevertheless the disappointments for many organizations because of the resulted greater than expected costs for set up a repository, research effort in this area appears promising.

This justifies the goal of this survey, namely to systematize some findings and discover positive research directions in this area. We are convinced that a serious step towards an increasing spread of digital object repositories comprises standardization of their content. It appears that populating a digital repository with standardized e-objects is a time and labor consuming activity. However facilitating the retrieval, sharing and using (from several users and in different context) of these objects requires their unified descriptions. Up to now

documents, disposed in electronic repositories are determined by specifications and quality metadata. Various standards depending on the subject area, e.g. SCORM, IMS, and LOM - for learning; MPEG-series for multimedia, to name a few, have been developed as well. It should be mentioned that as a rule each standard deals with a huge amount of metadata. So, the development of a methodology and tools for an automatic extraction of metadata for digital objects deployed in various subject repositories thus facilitating clients' access is a potential research issue. At the same time repositories increase successfully very quickly. Fig.1 shows the growth of the *Open*DOAR [5] Database up to its present size and Fig.2 represents the number of repositories by country. Up to now about 1400 repositories all over the world have been reported.

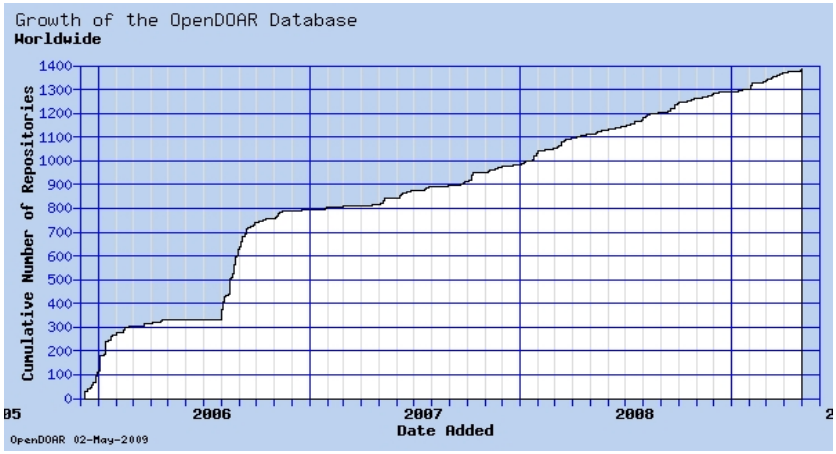

Fig. 1

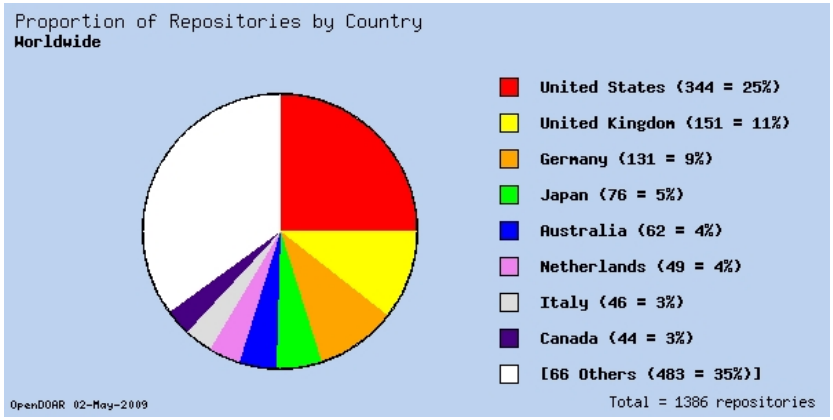

Fig. 2

In Bulgaria there are two open repositories only [6,7] registered in *Open*DOAR 2009 and one more is forthcoming to be available at the Institute of Mathematics and Informatics, Bulgarian Academy of Sciences [13].

Bearing these considerations in mind the paper is organized as follows: the next section introduces some of the key concepts of digital object repositories. Different classifications of repositories are briefly presented. Section 3 reviews some repository solutions. Section 4 discusses the importance of metadata and the variety of schemes/standards. Examples of widely applied schemas and their peculiarities are briefly reported. The conclusion presents the underlying project and determines some research tasks.

We have tried to do a general overview of topics concerning the building of repositories rather than to investigate particular issues in depth. Much more real work is required and it will be done within the project.

Basic Definitions

In [8] digital objects are defined as "a data structure whose principal components are digital material, or data, plus a unique identifier for this material, called a handle (and, perhaps, other material) ". This definition further evolved to capture access rules to use the object and metadata for description of the content [9]. Following these definitions digital objects can be referred as entities together with their metadata, and the services they offer to the clients.

There is no a clear definition of the concept of a repository as well. Usually any collection of digital objects is called a repository. Specialized repositories such as e-learning repositories, e-prints repositories, e-thesis repositories and subject-based repositories are being developed. However what's the difference from other datasets as directories, operational databases, catalogues, and portals? Defining institutional repositories as a "set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members" Lynch [10] focuses more on the services that a repository is supposed to offer. The type of the content and the different technological solutions e.g. free versus commercial software remain in the background or even lose to. The organizations differ in the underlying motivation to build digital repositories as well. Services that are expected from repositories range across several functional areas depending on the interest for different communities (digital libraries, research, learning, e-science, publishing, records management, preservation). Among these functional areas data sharing, preservation of digital resources, corporate information management and scientific collaboration are to be considered. Not surprisingly there is no a unique definition of the notion "digital object repository". During a special meeting of leading companies and associations [11] a repository has been defined as "a networked system that provides services pertaining to a collection of digital objects". Following this general definition repositories include: institutional collections, datasets, learning objects banks, cultural heritage artifacts, etc. Generally speaking a digital repository can be considered as means of handling digital content. Thus they may include a wide range of content for a variety of purposes and users. What goes into a repository depends on decisions made by each institution or administrator. The peculiarities of digital repositories that distinguish them from other digital collections are summarized in [12]. In addition an attempt to develop a classification of repositories is also proposed. According to Heery and Anderson [12] repositories can be typified by content (corporate records, e-theses, learning objects, research data), by coverage (personal, institutional, national, journal), by users (learners, researchers, teachers, etc.) and by function (access, preservation, dissemination, reuse). In [14] two more features of the repositories, namely policy (persistence, deposit, access) and infrastructure – centralized versus distributed have been taken into account. It is very important to determine the content and scope of any repository because this is the way to define the managerial policies. *Open*DOAR provides a list of major content types (publications, books, conference papers, theses, learning objects, multimedia items, etc.) as well. Some definitions of the various types are given in [15].

In our opinion the content type of the repositories (cultural heritage, e-learning objects and teaching materials, scientific papers, e-maps, etc.) makes a good distinction about the main groups of users and corresponding managerial policies.

Repository Solutions

Digital repository solutions consist of hardware, software and open standards. Recently the more commonly adopted software solutions fall into two broad groups: open source and commercial software.

Open source software is exemplified by DSpace, Fedora, and EPrints. DSpace is the software of choice for academic, non-profit, and commercial organizations building open digital repositories. DSpace preserves and enables easy and open access to all types of digital content including text, images, moving images, mpegs and data sets. It is applied for accessing, managing and preserving scholarly works [16] .Fedora (Flexible Extensible Digital Object Repository Architecture) was originally developed by researchers at Cornell University as an architecture for storing, managing, and accessing digital content in the form of digital objects [8]. Nowadays the Fedora Repository Project and the Fedora Commons community together with the DSpace project are under the supervision of the not-for-profit organization DuraSpace [17]. The Fedora Repository Project (simply Fedora) implements the Fedora abstractions [18] and provides basic repository services. This permits to express digital objects, to assert relationships among digital objects, and to link services to digital objects. Fedora ensures the durability of the digital content by providing tools for digital preservation. The Fedora Commons community deals with producing additional tools and applications that enlarge the functionally of the Fedora repository. The latter is extremely flexible and can be used to support any type of digital content. There are numerous examples of Fedora being used for digital collections, e-research, digital libraries, archives, digital preservation, institutional repositories, open access publishing, document management, digital asset management, and more [18]. Fedora Commons provides sustainable technologies to create, manage, publish, share and preserve digital content. EPrints [19] is an open source platform for building repositories of research literature, scientific data, and student theses. There are now over 210 repositories using the EPrints software, the repository at New Bulgarian University being one of them.

Commercial software could be based on an open source repository engine coupled with a proprietary application software layer – VITAL [20]. Other possibility includes openly accessible API's using XML interfaces - DigiTool [21] and DPS [22]. Because of the increased demand to manage digital assets, libraries need standard methods and tools to facilitate cataloging, sharing, searching, and retrieval of digital collections. Through highly customizable user interfaces DigiTool enables academic libraries and library consortia to manage and provide access to the growing volume of digital collections. Support for library standards and built-in integration with other Ex Libris products, e.g., Aleph®, Voyager®, MetaLib®, SFX®, and Primo, makes DigiTool an integral part of the library infrastructure and facilitates the incorporation of digital resources into library services. VITAL is an institutional repository solution built on Fedora, It is designed to simplify the development of digital object repositories and to provide online search and retrieval of information for administrative staff, contributing faculty and end-users. VITAL provides all functions such as storing, indexing, cataloging, searching and retrieving required for handling large text and rich content collections.

A functional comparison of repository software products is presented in [23]. Consulting services are available through Sun [24].

The Importance of Metadata

In order to be easily retrieved, shared and used from different users and for different purposes, various types of e-documents have to be described following common schemas and rules e.g. specifications/standards and metadata. The term metadata e.g. data about data is used differently ranging from machine understandable information through records that describe electronic resources. In a library, "metadata" applies for any kind of resource description. Metadata describe how and when and by whom a particular set of data was collected, and how the data is formatted. Metadata is essential for understanding information stored in data warehouses and has become increasingly important in XML-based Web applications [27]. In addition they ensure the accessibility, identification and retrieval of resources. Descriptive metadata facilitate the resources' organization, interoperability and integration, provide digital identification and support archiving. Poor quality or non-existent metadata mean that resources remain invisible within a repository or archive thus becoming undiscovered and inaccessible. In the case of digital assets, metadata usually are structured textual information that describes something about the creation, content, or context of an image [28].

There are several types of metadata:

1. Descriptive - title, author, extent, subject, keywords
2. Structural – unique identifiers, page numbers, special features (table of contents, indexes)
3. Technical - file formats, scanning dates, file compression format, image resolution
4. Preservation - archival information
5. Right s management - ownership, copyright, license Information

Metadata can be stored in different ways:

1. Separately as a HTML, XML or MARC21 (format for library catalogues) document linked to the resource
2. In a database linked to the resource
3. As an integral part of the record in a database or embedding the metadata in the Web pages

Nevertheless the importance of metadata has been recognized, means for efficient implementation still lack. Due to the rapid growth in digital object repositories and the development of many different metadata standards metadata implementation is complex. On the other hand quality metadata can be produced by experts in the subject domain only. So far, most of the resource discovery metadata are still created and corrected manually either by authors, depositors and/or repository administrators. It appears attractive to auto-generate metadata with no human intervention. Recent research findings are reported in [25, 26].

In order metadata to be processed via computer, proper encoding has to be applied. This is done by the addition of markup to a document to store and transmit information about its structure, content or appearance. Schemas comprise metadata elements designed to describe particular information. We can mention the following encoding schemas concerning how metadata is presented:

1. HTML (Hyper-Text Markup Language)
2. XML (eXtensible Markup Language)
3. RDF (Resource Description Framework)
4. MARC (Machine Readable Cataloguing)
5. SGML (Standard Generalized Markup Language

Metadata schemas can be viewed as standards describing the categories of information to be recorded. They ensure consistency in metadata application, support interoperability of applications and resource sharing. Schemas are built from individual components, i.e. metadata elements. Depending on the element definition each element contains a particular category of information. Certainly not all schemas contain the same elements as the needs of users differ.

There are widespread metadata standards (schemes) that are used in digital object repositories [29, 30]. Standards are being developed all the time. Bellow some well-known examples are listed.

1. Dublin Core [31]. The Dublin Core standard arose from a 1995 workshop held in Dublin, Ohio. The basic DCMES (Dublin Core Metadata Element Set) involves 15 elements. Each is optional and repeatable, and may appear in any order the creator of the metadata wishes. This simple generic element set is applicable to a variety of digital object types. It is used for the description of simple textual or image resources. For richer descriptions to enable more refined resource discovery, Qualified Dublin Core has been developed. This standard employs additional qualifiers to the basic 15 elements to further refine the meaning of an element. Qualifiers increase the precision of the metadata.
2. TEI (Text Encoding Initiative) [32]. The Text Encoding Initiative is an international project to develop guidelines for marking up electronic texts such as novels, plays, and poetry, primarily to support research in the humanities.
3. METS (Metadata Encoding & Transmission Standard) [33]. METS is maintained by the Network Development and MARC Standards Office of the Library of Congress. The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library expressed using the XML schema language.
4. MODS (Metadata Object Description Schema) [34]. This is an XML schema for descriptive metadata compatible with the MARC 21 bibliographic format. It includes a subset of MARC fields and uses language based tags rather than the numeric ones used in MARC 21 records. In some cases, it regroups elements from the MARC 21 bibliographic format. Like METS, MODS is expressed using the XML schema language.
5. EAD (Encoded Archival Description) [35]. (EAD) was developed as a way of marking up the data contained in finding aids so that they can be searched and displayed online. In archives and special collections, resources are described via a finding aid. Finding aids differ from catalog records by being much longer, more narrative and explanatory, and highly structured in a hierarchical fashion. They generally start with a description of the collection as a whole, indicating what types of materials it contains and why they are important. The finding aid describes the series into which the collection is organized and ends with an itemization of the contents of the physical boxes and folders comprising the collection.
6. LOM (Learning Object Metadata) standard (IEEE 1484.12.1-2002) [36]. LOM was developed by IEEE Learning Technology Standards Committee to enable the use and re-use of technology-supported learning resources such as computer-based training and distance learning. Learning Objects are defined here as any entity, digital or non-digital, which can be used, re-used or referenced during technology supported learning. The LOM defines the minimal set of attributes to manage, locate, and evaluate

learning objects. Where applicable, Learning Object Metadata may also include pedagogical attributes such as; teaching or interaction style, grade level, mastery level, and prerequisites.

7. MARC Standards [37]. The MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form. Today this is the most common standard format used by bibliographic and library catalogues to exchange information on their contents.
8. VRA Core Categories [38]. This is a scheme developed by the Visual Resources Association for the description of art, architecture, artifacts, and other visual resources. It consists of a metadata element set (units of information such as title, location, date, etc.), as well as an initial blueprint for how those elements can be hierarchically structured. The element set provides a categorical organization for the description of works of visual culture as well as the images that document them.
9. MPEG standards [39]. The ISO/IEC Moving Picture Experts Group (MPEG) has developed a suite of standards for coded representation of digital audio and video. Two of the standards address metadata: MPEG-7, Multimedia Content Description Interface (ISO/IEC 15938), and MPEG-21, Multimedia Framework (ISO/IEC 21000). MPEG-7 defines the metadata elements, structure, and relationships that are used to describe audiovisual objects including still pictures, graphics, 3D models, music, audio, speech, video, or multimedia collections. MPEG-21 was developed to address the need for an overarching framework to ensure interoperability of digital multimedia objects.
10. CSDGM (Content Standard for Digital Geospatial Metadata) [40]. This is a metadata schema for geospatial datasets comprising topographic and demographic data, GIS (geographic information systems), and computer-aided cartography base files. Geospatial datasets are used in many areas e.g. land use studies, biodiversity counts, climatology and global change tracking, remote sensing, and satellite imagery. An international standard, ISO 19115, Geographic Information— metadata was issued in 2003.

Conclusion

In this paper we examine some issues concerning digital object repositories and metadata. Besides the possibility of applying some well-known world standards such as SCORM, IMS, MPEG-7, it is not expected that the shared e-documents will be specified in a uniform way. This justifies any research effort to raise the productivity of the process of metadata generation for different e-documents. Taking into account the rapidly growing number of new digital repositories investigations in this area are promising.

Metadata ExTraction for Automatic SPEcifications of E-Documents – METASPEED is a Bulgarian research project funded by the Bulgarian National Science Fund under the thematic priority: Information and Communication Technologies. It aims to facilitate the development of Bulgarian standards and even commonly accepted specifications for the description of metadata for e-documents in different subject areas.

The goal of this project can be briefly summarized as follows: to investigate and create technologies, methods and tools for automatic generation of metadata thus facilitating the proper specification of documents with different e-format, content and location.

Project findings will facilitate the access to different digital collections in a straightforward manner. This is the first stage toward the development of an integrated information environment in Bulgaria. We expect that the main contributions will include:

- development of proper tools for an automatic metadata generation for collections containing digital documents of different shapes and types;
- building a framework to share European and Bulgarian e-resources;
- development of national standards for document sharing.

The work underlying the METASPEED project comprises:

- survey of standards and schemas for metadata;
- evaluation of current automatic metadata extraction and generation tools;
- compilation of recommended functionalities for automatic metadata generation applications;
- investigations on specialized technologies and methods for metadata retrieval for documents in different areas;
- design and implementation of proper software prototypes;
- prescriptions for Bulgarian standards in different subject areas;
- a methodology how to build an integrated information repository for digital documents in Bulgaria.

Acknowledgements

This work is partially granted by Bulgarian National Science Fund, Ministry of Education and Sciences in the frame of the project "Automated Metadata Extraction for e-documents Specifications and Standards", contract No: D002(TK)-308/ 19.12.2008.

Bibliography

[1] Duncan C. Digital Object Repositories Explained, an Intrallect White Paper, 2006

[2] Bluthe E., Chandra V. The Value Proposition in Institutional Repositories EDUCASE Review, September/ October 2005.

[3] http://dorsdl2.cvt.dk/

[4] http://openrepositories.org/

[5] http://www.opendoar.org/ - Directory of Open Access Repositories

[6] http://eprints.nbu.bg/ – New Bulgarian University Scholar Electronic Repository

[7] http://research.it.fmi.uni-sofia.bg:8880/dspace/ - Research at Sofia University

[8] Kahn R., Wilensky R. A Framework for Distributed Digital Object Services, 1995, http://www.cnri.reston.va.us/home/cstr/arch/k-w.html

[9] Lagoze Carl. 1995. A Secure Repository Design for Digital Libraries. D-Lib Magazine, http://www.dlib.org/dlib/december95/12lagoze.html

[10] Lynch, C. (2003). "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age." ARL, 226, February 2003, 1-7, http://www.arl.org/resources/pubs/br/br226/br226ir.shtml

[11] http://msc.mellon.org/Meetings/Interop/terminology.doc

[12] Heery R., Anderson, S.: Digital Repositories Review, AHDS, 2005.

[13] http://www.driver-support.eu/pmwiki/index.php?n=Main.Bulgaria

[14] http://www.ukoln.ac.uk/repositories/digirep/index/Typology

[15] http://www.rsp.ac.uk/content

[16] http://www.dspace.org/

[17] http://duraspace.org/

[18] http://www.fedora-commons.org/

[19] http://www.eprints.org/

[20] http://www.vtls.com/products/vital

[21] www.exlibrisgroup.com/digitool.htm

[22] www.exlibrisgroup.com/Preservation.htm

[23] http://www.rsp.ac.uk/repos/software/surveyresults

[24] Grant C. Delivering digital repositories with open solutions, a Sun white paper, Version 8.0, November 2007.

[25] Polfreman M. and Rajbhandaji S. Metatools – Investigating Metadata Generation Tools, JISC Final report, Oct.2008.

[26] Greenberg J. et al. Final Report of the AMEGA Project, UNC School of Information and Library Science, 2005.

[27] http://www.webopedia.com/TERM/M/metadata.html

[28] http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/metadata-overview/

[29] http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/putting-things-in-order-links-to-metadata-schemas-and-related-standards/, February 2009

[30] http://metadata-standards.org/

[31] http://dublincore.org/

[32] http://www.tei-c.org/

[33] http://www.loc.gov/standards/mets/

[34] http://www.loc.gov/standards/mods/

[35] http://www.loc.gov/ead

[36] http://ltsc.ieee.org/wg12/

[37] http://www.loc.gov/marc/marc.html

[38] http://www.vraweb.org/projects/vracore4/index.html

[39] http://www.mpeg.org/

[40] http://www.fgdc.gov/metadata/index_html

Authors' Information

Juliana Peneva – New Bulgarian University, Department of Informatics; e-mail: july_peneva@abv.bg

Stanislav Ivanov – New Bulgarian University, Department of Informatics; e-mail: sivanov@nbu.bg

Filip Andonov – New Bulgarian University, Department of Informatics; e-mail: fandonov@nbu.bg

Nikolay Dokev – New Bulgarian University, Department of Informatics; e-mail: n.dokev@nbu.bg