CSECS 2011, pp. 000 - 000

The 7 Annual International Conference on
Computer Science and Education in Computer Science,

July 06-10 2011, Sofia, Bulgaria

THE METASPEED PROJECT – A PROGRESS REPORT

JULIANA PENEVA, STANISLAV IVANOV, PHILIP ANDONOV, NIKOLAY DOKEV

***Abstract**: The outcomes of the first stage of the Metaspeed project are represented. Our team participates in the national project "Automated Metadata Extraction for e-documents Specifications and Standards" funded by the Bulgarian National Science Fund, Ministry of Education and Sciences", contract N D002 (TK)-308/ 19.12.2008. The project aims 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 . In the context of general description of the project outcomes we represent here the contribution of our team.*

***Keywords**: digital repositories, automatic metadata generation, specification of e-documents*

***ACM Classification Keywords**: H3.5 On-line information services – Data sharing; H3.7 Digital libraries-Collection*

Introduction

In today competitive business environment the proper management of organizational digital resources is crucial for making timely decisions and responding to changing business conditions. Many companies are realizing a business advantage by managing successfully their business data. Resources include documents, images, video or audio clips, animations, presentations, online courses, web pages, etc. Organizations are of different types and sizes ranging from SME to international corporations. All of them exhibit an intensive use of digital resources because these e-documents 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 au par with finance and human resources.

In order to be easily retrieved, shared and used from different users and for different purposes the various types of e-documents have to be described following common schemas and rules e.g. specifications/standards and metadata. Depending on content and context standards for e-learning (SCORM, IMS, LOM, etc.), for multimedia data (MPEG-7), to name a few, have been proposed. As a rule, every standard requires too much metadata. Standardized metadata enables the easy choice of relevant e-documents. However poor quality or non-existent metadata means that resources remain invisible within a repository or archive thus becoming undiscovered and inaccessible. On the other hand quality metadata can be produced by experts in the subject domain only. So, building digital repositories with "standardized" e-documents appears to be a labor-consuming, highly qualified and expensive activity. With digital resources being produced in ever-increasing quantities, finding the time and resources necessary for ensuring quality metadata becomes a challenging task. Automation

seems promising to address this. We are convinced that automatic metadata extraction could be a right solution. Several approaches, including metatag harvesting, content extraction, automatic indexing or classification, text and data mining, etc. have been proposed. In [1] the quality of currently available metadata generation tools has been compared. The best scenario would be to auto-generate high-quality resource discovery metadata without any human intervention. Nevertheless most of the resource discovery metadata is still created and corrected manually either by authors, depositors and/or repository administrators.

At the same time we would like to mention that in Bulgaria there are no standards or even commonly accepted specifications to describe metadata for e-documents[1]. Again there exist an increasing number of newly created digital repositories in different subject areas.

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 our research efforts namely to automate the process of metadata generation for different in style e-documents. Taking into account the rapidly growing number of new digital repositories investigations in this area are promising.

This is the relevance of the Metaspeed project [2][2]. The goal of this project can be briefly summarized as follows: to investigate and create

[1] As a first step in this direction consider the recently adopted Bulgarian Law on e-documents.

[2] Bulgarian National Science Fund, Ministry of Education and Sciences, Agreement N D002(TK)-308/ 19.12.2008 “Automated Metadata Extraction for e-documents Specifications and Standards”

technologies, methods and tools for automatic generation of metadata thus facilitating the proper specification of documents with different e-format, content and location.

The rest of the paper is organized as follows. Section 2 deals with the project description. We present project partners, the project goals and the different working packages. In Section 3 we report the results of our participation in this project. We summarize project results in Section 4 and discuss further development beyond project end.

What Is the METASPEED Project?

Metadata **E**x**T**raction for **A**utomatic **SPE**cifications of **E-D**ocuments – **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. Project partners include Bulgarian researchers from state and private universities and Bulgarian Academy of Sciences. This project is carried out by a consortium composed of: University of Plovdiv, Institute of Mathematics and Informatics – Bulgarian Academy of Science, Technical University of Sofia and New Bulgarian University.

The METASPEED project is an interdisciplinary project. This justifies the participations of people interested in computer linguistics, e-learning, standards for e-documents, multimedia applications, archival sciences, database systems, etc. The partners and their competences are presented in Table 1:

Table 1 Partners of the METASPEED Project

PARTNER		COMPETENCES
	University of Plovdiv Department of Computer Informatics	computational linguistics, standards and systems for e-learning, theory of algorithms, programming languages
	Institute of Mathematics and Informatics - BAS Department of Information Systems	analysis, synthesis and retrieval of structured data from texts, images and video
	New Bulgarian University Department of Informatics	cognitive sciences, database systems, e-learning
	Technical University of Sofia Research laboratory "Technologies and standards for e-learning"	standards and systems for e-learning

The goal of this project is to achieve proper specification of different documents via an automatic generation of their metadata. The rationale behind this goal is that e-documents are to be described in a standardized manner to facilitate their retrieval, sharing and using. Usually documents in digital repositories are determined according to a particular specification and/or standard together with data about the document itself, i.e. metadata. As a rule the application of any standard requires too much metadata that are produced by experts in the subject area. Consider the electronic resources e.g. tests, learning content, etc. in the National Educational Portal [3]. These resources obey no unique

standard. That is why our research efforts towards a standardization and automatic generation of metadata for different format e-documents are an economically motivated activity. Project findings will facilitate the access to different digital collections in a straightforward manner. This is the first stage toward the development of a uniform information environment in Bulgaria.

The project is built up of four work packages –see Table 2.

Table 2 Work Packages of the METASPEED Project

WORK PACKAGES		AIMS
WP1	**Standards of e-documents and tools for their automatic generation**	• to finalize the research analysis in the area • to prepare state-of the-art reports; concerning standards for e-learning and multimedia documents in the field of cultural heritage; • to develop prescriptions for Bulgarian standards in different subject areas; • to investigate tools for automatic metadata generation.
WP2	**Automated metadata generation from text documents**	• to develop methods, algorithms and tools to retrieve structured data from electronic text documents, written in different languages taking into account existing standards and specifications

WP3	**Automatic metadata generation from multimedia documents**	• to discover content-based image retrieval methods
WP4	**Automated creation and testing of digital repositories in different areas**	▪ to develop methods and tools for automated metadata generation from Web pages

Certainly significant dissemination and supporting activities are foreseen.

For NBU the tasks as well as the corresponding working packages they belong to are listed below:

1. Internal project management, reporting and participation in project meetings.
2. Internal report on needs analysis (WP2).
3. State of the art report on practices concerning the building of digital repositories (WP2).
4. Evaluation of digital repository solutions in industrial contexts (pilot studies of software, case studies) (WP4)
5. Design of institutional digital repository (WP4).
6. Set up of assignment for Bulgarian standards (WP4)
7. Participation in conferences & fairs
8. Participation in thematic monitoring activities.
9. Valorisation and dissemination activities.

The results are presented in the next section.

Project Outcomes and Results of the NBU Team

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 are reporting for own repository investigations as well. 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.

Repositories increase successfully very quickly. In this perspective, universities and scientific institutions demonstrate remarkable activity. Open access academic repositories marked a boost of 300 during the mid of 2006. Since the beginning of year 2007 the growth of such repositories listed in the *Open*DOAR Database [4] shows a constant increase of 300 repositories per year up to its present number of about 1900. This justifies the goal of the state-of-the art report we have prepared. The first tasks were systematizing some findings and discover positive research directions [5].

In the context of the above, we proceeded with design activities of an institutional repository of the Department of Informatics at New Bulgarian University. We discussed what do we need and determined the type of the material to be stored in the repository. Creating a proper digital collection that captures and preserves the department's intellectual output would increase its visibility, prestige and public value. This repository will support learning and administrative processes of our department. To build an effective repository the technical set up process is to be planned properly. So, we considered the requirements with respect to our institutional context. At the planning phase of building the

department repository we focus on service design, policies, technical and system issues. Taking into account that flexibility among the different collections is a key feature the goal of the department repository is to offer a proper infrastructure with a well defined range of services. We adopt a well established model in this area – OAIS (Reference Model for an Open Archival Information System) [6]. With limited staff resources for long-term maintenance and support we have chosen to apply the most popular approach i.e. to use a standard package nevertheless that external hosting becomes recently more popular. Digital repository solutions consist of hardware, software and open standards. A wide variety of available software with different features and strengths exists. Our investigations show that there are over 308 repositories using the EPrints software, about 711 – DSpace, 82 – Digital Commons. The rest of the software exhibits a limited (up to ten repositories) application. In order to justify the choice of the software, pilot installations of some open source packages has been undertaken. These are used as test beds for the overall repository development. A pilot system will be used to tune the software parameters. We intend to perform users' acceptance testing as well.

During the carried out research activities a specific task has been arisen, namely the development of interactive methods for group decision making [7, 8].

Our team took also part in several dissemination [9,10] and valorisation activities within the frame of METASPEED project.

Conclusions and Future Work

In this paper we report our participation in the project "Automated Metadata Extraction for e-documents Specifications and Standards"- a national project funded by Bulgarian National Science Fund, Ministry of Education and Sciences.

Bibliography

[1] Polfreman M., Rajbhandari S. MetaTools - Investigating Metadata Generation Tools. JISC Final report, October 2008

[2] www. Metaspeed.org

[3] http://start.e-edu.bg/

[4] www.opendoar.org

[5] Peneva J. Ivanov St., Andonov F., Dokev N. Digital Objects – Storage, Delivery and Reuse. International Conference i.TECH-2, Madrid, Spain, published in Int. Journal "Information technologies and Knowledge, Vol. 3/ 2009, pp.61- 70, ISSN 13131-0455.

[6] CCSDS 650.0-P-1.1, Reference Model for an Open Archival Information System (OAIS) (Pink Book, Issue 1.1, August 2009)

[7] Andonov F. Solving discrete multicriteria optimization problems in group environment. In Proc. of the First Int. Conference on Software, Services and Semantic Technologies, October 2009, Sofia, Bulgaria, ISBN 978-954-9526-62-2.

[8] Andonov F. Interactive Methods for Group decision Making, International Journal "Information technologies and Knowledge", Volume 3/2009, pp. 25-30, ISSN 1313-0455.

[9] Peneva J., Totkov G., Stanchev P., Shoikova R. Automatic generation for Specification of e-Documents – the METASPEED Project. International Conference i.TECH-2, Madrid, Spain, published in Int. Journal "Information technologies and Knowledge, Vol. 3/ 2009, pp.118 - 127, ISSN 13131-0455.

[10] http://www.nbu.bg/index.php?l=484

Authors' Information

Your photo here:
Height: 2,58 cm
Width: 1,84 cm

Juliana Peneva, PhD, Assoc. Prof,New Bulgarian University, IMI – BAS, 21 Montevideo St. 1618 Sofia, Bulgaria, jpeneva@nbu.bg.

Major Fields of Scientific Research: *database systems, software engineering, information systems, e-learning*

Stanislav Ivanov, PhD, Assoc. Prof, New Bulgarian University, 21 Montevideo St. 1618 Sofia, Bulgaria, sivanov@nbu.bg

Major Fields of Scientific Research: *object-oriented modeling, programming, pattern recognition*

Philip Andonov, Ass. Prof, New Bulgarian University, 21 Montevideo St. 1618 Sofia, Bulgaria, fandonova@nbu.bg.

Major Fields of Scientific Research: *operational research, group decision making, database systems*

Nikolay Dokev, PhD, Assoc. Prof, New Bulgarian University, 21 Montevideo St. 1618 Sofia, Bulgaria, ndokev@nbu.bg

Major Fields of Scientific Research: networking, operating systems, decision making

This research 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 N D002(TK)-308/ 19.12.2008.