

AUTOMATIC METADATA GENERATION FOR SPECIFICATION OF E-DOCUMENTS – THE METASPEED PROJECT

Juliana Peneva, George Totkov, Peter Stanchev, Elena Shoikova

Abstract: A Bulgarian research project funded by the Bulgarian National Science Fund under the thematic priority Information and Communication Technologies, contract D002-308/19.12.2008 is presented in this article. The main goal and tasks are outlined. Some already achieved results are pointed.

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-documents¹. Again there exist an increasing number of newly created digital repositories in different subject areas e.g.:

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

-
- cultural heritage [2] – collections, museum exhibits, old-printed books, science publications, etc.;
 - education [3] – e-learning courses, multimedia learning content, tests items, etc.;
 - spatial information systems [4].

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.

What is the METASPEED Project?

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. 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 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. 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 [5]. 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. We expect that the main contributions 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.

Partners

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

Project Work Packages

The project is built up of four work packages. Certainly significant dissemination and supporting activities are foreseen.

WP1. Standards of e-documents and tools for their automatic generation

The goal of this package is to finalize the research analysis in the area and to prepare state-of the-art reports concerning:

- a. standards for e-learning;
- b. standards for multimedia documents in the field of cultural heritage;
- c. prescriptions for Bulgarian standards in different subject areas;
- d. tools for automatic metadata generation.

It is expected that project technical prescriptions of national standards for e-learning and specifications for cultural heritage e-objects will be proposed. For example the adoption or development of a standard, in the field of e-learning, could provide sharing (including export/import), multiplication and adaptation of learning resources (courses, materials and tests) hosted in the Internet. During the adaptation of well-known standards and specifications a reasonable question arises: to what extent the national specifics such as language, educational system and traditions are to be considered.

WP2. Automated metadata generation from text documents

The goal of WP2 is 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. To realize this goal a review-analysis of the existing methods and algorithms in the world and in particular – for Bulgarian language will be carried out.

The following tasks will be performed:

- a. critical analyses of existing methods and tools for retrieval of metadata from electronic texts;
- b. investigations on specialized technologies and methods for metadata retrieval for documents in different areas;
- c. design and implementation of proper software prototypes;
- d. experiments with the realized methods and tools for metadata retrieval on the documents in examined areas.

WP3. Automatic metadata generation from multimedia documents

In the next three years, the world will create more data than has been produced in all of human history. It is well known that the search power of current searching engines is typically limited to text and its similarity, since less than 1% of the Web data is in textual form, the rest being of multimedia/streaming nature, particularly since a large portion of pictures still remains as "unstructured data". The Enterprise Strategy Group [6] estimates that more than 80 milliard photographs are taken each year. Using of digital images promises to emerge as a major issue in many areas, for instance Google answers daily more than 200 million queries against over 30 milliard items. Because of this it is necessary to extend next-generation search to accommodate these heterogeneous media.

Some of the current engines search the data types according to textual information or other attributes associated with the files. An orthogonal approach is the Content-based Image Retrieval (CBIR). It is not a new area – in current surveys can be counted more than 300 systems, most of them exemplified by prototype implementations. The typical database size is in the order of thousands of images - very recent publicly-available systems, such as ImBrowse [7], Tiltomo [8] and Alipr [9] declare to index hundreds of thousands of images.

The user questions in image search are partitioned into three main levels:

- a. *Low level* – this level includes basic perceptual features of visual content (dominant colors, color distribution, texture pattern, etc.);
- b. *Intermediate level* – this level forms next step of extraction from visual content, connected with emotional perceiving of the images, which usually is difficult to express in rational and textual terms. Visual art is an example, where these features play a significant role. Typical features in this level are color contrasts, because one of the goals of the painting is to produce specific psychological effects in the observer, which are achieved with different arrangements of colors;
- c. *High level* – this level includes queries according to rational criteria. In many cases the image itself does not contain information which would be sufficient to extract some of the characteristics. For this reason current high-level semantic systems still use huge amount of manual annotation.

Usually the existing systems for image retrieval are limited by the fact that they can operate only at the primitive feature level, while users operate at a higher semantic level. This mismatch is often referred as a semantic gap. The Project aims at finding and analyzing new content-based image retrieval methods to analyze, index, and retrieve images and video. The goal is to increase the retrieval effectiveness by a proper choice of image features from the MPEG-7 standard and on that base to find the description of some concept, which humans use in their everyday life.

WP4. Automated creation and testing of digital repositories in different areas: A) cultural heritage; B) e-learning; C) spatial information systems; D) automated referring of scientific publications.

The main task is to investigate and create methods and tools for automated metadata generation from Web pages (in the listed above subject areas). Known technologies for search in Internet-pages, using tools for automated metadata generation from text and multimedia will be examined and adapted. As a result the Internet-space will be used as a source for building digital repositories in order to test the proposed methods and tools. In addition some collections of electronic resources being created from the Project partners for carrying out experiments in the field of e-learning and scientific publications will also be used to check the developed tools for automated metadata generation.

As it concerns cultural heritage the focus of investigations will be on search methods to be applied in collections of documents containing text and graphical objects. It is expected these documents to be automatically classified following proper ontology.

Investigations in e-learning will deal with automated metadata extraction from learning resources. The possibility for generation of new learning objects taking into account different learning styles will be examined closely. Besides a linguistic analysis of the created learning content, the interactions among trainees and "instructor-trainee" will be studied. The latter facilitates data extraction necessary to build the trainee's portfolio.

The INSPIRE standards for spatial meta-data are obligatory for the European Union member states. This means that it is very important the cultural space information to be described following these standards. Two main tasks are connected to this problem:

- to create Bulgarian standards and thesauri for spatial meta-data of the cultural heritage objects, which are corresponding to the INSPIRE standards;
- following these standards, to develop methods and tools for meta-data extraction from the cultural objects' descriptions.

As a result it is expected that a digital repository of scientific publications will be built via automated metadata retrieval from full-text articles. Referring is a time consuming task because experts have to review many papers and to classify them properly. Again automation seems to be appropriate. Automated metadata retrieval from the collection of scientific publications in different languages (in the case of specialized texts in concrete scientific domain) and their classification according to proper ontology is important in the case of multilingual referring journals (for instance – containing reviews in English, Russian and Bulgarian languages) as well.

Several dissemination and valorisation activities will also be carried out within the frame of the Project. The results from the PhD studies as well as the applied methods will be discussed during some Project specific workshops. It is supposed that most of the results will be presented and approved during the planned project meetings together with our scientific consultants. These results will also be presented as papers on different international and local conferences. When the Project is put into effect, it is presumed that not only extra young scientists will join these prospective investigations but these young people will receive moral and financial incentives to realize their work. At the same time the PhD students will gain the advantage to be in touch with leading researchers in order to acquire their experience. It is possible to provide students' mobility for a short period of time on the basis of a preliminary signed agreement. The organization of specializations for the Project staff and for university lecturers on the standardization in the area of e-learning and metadata generation is one of the aims of the Project management. At least five PhD theses will be promoted within the frame of this Project. In addition students enrolled in masters programs from Plovdiv University, Technical University of Sofia and New Bulgarian University will join the investigations. We expect to promote more than 20 master theses in the Project topics.

Project Deliverables and Overall Effect

The proposed in the Project scientific investigations are performed for first time in our country. Some of the proposed approaches can be also considered as an innovation as they have no analogue abroad. The qualification of the participants in the project team and their scientific achievements (including the successful participation in over 60 national and international projects) represent real preconditions for a successful realization of the Project. Last but not least it is expected that the research being carried out in different thematic areas will be very effective. So, the main goals of this Project can be definitely reached.

The project deliverables result from the application of intellectual technologies, in the making of proper models and methods followed by their testing and validation. The software prototypes that would be developed within the Project can be applied not just to test the theoretical results or to prove the models relevance. These prototypes allow for an objective analysis and further improvement of the proposed solutions. The Project deliverables can be summarized as follows: new methods and information technologies tools to be used both in theory and practice in four areas namely e-learning, cultural and historic heritage, spatial information systems, and scientific reviewing, will be worked out.

The proposed problems for investigation are innovative and their solution represents not only matter of a scientific interest. The findings will be used in practice to develop proper methods and tools for an automated metadata extraction and generation from different kinds of e-documents.

It is expected that one of the most important result from this Project concerns the making of technology and methods for automated maintenance of digital repositories (see WP2 and WP3) and all corresponding activities such as: generation and update of virtual catalogues; organization of content-based search; removal of discrepancies, etc. The developed methodology and tools will be tested in the project thematic areas mentioned above (WP4). The success of the Project investigations would overturn the traditional view on the ways of how to plan, organize and maintain a digital archive. Taking into account the Project results the efficiency in the development of new multimedia repositories as well as in updating existing collections of e-documents will be considerable increased.

It is worthwhile to notice the opened possibilities for a wide multiplication of the results. This is especially true as it concerns the application of the developed methods and proposed specifications to generate metadata of different types (content or/and context-of-use based). As an immediate consequence a significant improvement of quality for newly created or hared in different areas digital repositories, will be achieved. Two important deliverables from this Project could be mentioned: an easy-to-use model of a multimedia archive that enables metadata generation, and a research methodology to catalogue digital archives.

Researchers and lecturers from humanitarian and social spheres will be introduced to the problems and the obtained results. The dissemination will be done via specialized seminars and lectures and through Project members' participation in different kind of scientific events.

There are some objective risk factors for the successful finalizing of the Project and the fulfillment of the planned tasks:

- Differences in the organization and the specificity of the research, performed by the partner institutions;
- Insufficient national experience in creating systems and standards for e-documents and, as a consequence, a necessity for exploring and adapting good European and worldwide practices;
- Orientation of the research following in advance fixed goals instead of conforming to its internal logic and some result-driven investigations.

- Involvement of a comparatively large number of young scientists, without proper experience (and enough motivation) to perform long term scientific investigations, etc.

Overcoming the listed above risk factors appears impossible within the frame (and with the resources) of the Project only. However some Project activities included in different work packages e.g. wide publicity; motivation (moral and material) of young scientists and their affiliation to long-term scientific research; the involvement of affiliated scientific consultants with their experience and knowledge; regular (half-year) workshops; open seminars (delivered by experts and representatives of outer organizations), etc. could reduce the influence of the risk factors significantly.

In order to management the Project effectively and to evaluate the progress in every work package a fixed number of milestones are set. They favor a successful finalizing of the corresponding tasks and activities. Examples of milestones include surveys, specifications for a national standard, tools for an automated generation of metadata, papers, PhD theses, a Web portal, presentations, etc.

Dissemination of Knowledge and Expected Impacts

The planned and expected project results comprise innovation technologies and methods for generation of metadata. Among them prototypes of software tools that can facilitate the automatic building up of virtual repositories for digital documents, document sharing, repositories maintenance and management are of special importance. It is expected that similar tools possessing peculiar features will appear on the national and European markets for first time. Considering the Project findings a methodology how to build an integrated information repository for digital documents in Bulgaria will be set up. This methodology will be experimented on different thematic areas (see WP4). The main methodology components follow a systematic approach that allows for:

- development and introduction to proper national standards (see WP1);
- shared use and development of digital repositories for e-documents that conform to the proposed standardized specifications (see WP4);
- development of Digital Repository Management Systems to deliver functionalities instead of human experts (see WP3 and WP4), etc.

When the main goals of this Project will be achieved and the planned results become real, further work consist of building an integrated national informational environment to become part of the overall European one.

The Project deliverables (methods, tools and other findings) could serve a sound foundation and proper framework to develop national standards for e-learning and storage of digitalized objects belonging to the cultural and historical heritage. Moreover it becomes feasible to design and build a national informational network to share the repositories' digital content among different institutions e.g. universities, libraries, museums, public archives, community centers, etc.

In consequences of the project results, the efficiency in set up and maintenance of digital repositories will increase because of the possibility for an automated metadata generation and metadata replacement. The latter implies a new research problem: following the proposed methodology the existing standards being used for e-documents description are to be investigated. In addition the developed tools for metadata generation are to be tailored to the new thematic area (see WP2 and WP3).

After finalizing up the Project the team will make efforts to disseminate and valorize the project deliverables. Enterprises and institutions that attend to e-government, developers of digital repositories in various subject areas, civil organizations of the information society, to name a few, could be considered as prospective users of

the developed tools and technologies. It has to be mentioned that inherent conservatism in education, information services, etc. gives rise to possible problems in carrying out these valorization activities.

The benefit influence of our research activities on the quality of education in humanities becomes obvious. In humanities new subjects concerning the standardization, storage and digitalization of the object being studied, their representation and exploration via Internet can be introduced. A new possibility for classification and virtual representation of their artifacts via the developed within the frame of the Project tools becomes feasible.

Conclusion

In this paper we report about a Bulgarian research project: Metadata ExTraction for Automatic SPEcifications of E-Documents – briefly the METASPEED project, funded by the Bulgarian National Science Fund under the thematic priority: Information and Communication Technologies, contract D002-308/19.12.2008. The main project goal as well as the basic working packages, the overall effect and the expected impacts have been summarized.

Nevertheless that the project is in its very early stage some results have been achieved. Experimental tools and software prototypes to be applied for an automated metadata extraction from text and multimedia content have been developed. Some initial program tools, which will be used in the process of automated metadata extraction from text and multimedia content, are already made.

A system for extracting data from web documents based on simplified finite state automata is developed [10]. Using it custom data structures to be extracted from the documents are specified. The system 'learns' the automata from examples in the form of annotated texts and uses heuristics to expand and improve the automata-extractor. The system is expected to be suitable for the extraction of structured data and metadata in particular.

An environment for modeling and running multilevel processes, which can be used in further analysis of spatial and e-learning metadata extraction, based on expanded variants of Petri-net theory, is proposed. The system is found to be useful for upgrading existing learning management systems by introducing graphical modeling of the e-learning activities workflow [11]. The relevant standardization activities and initiatives, aiming to support the European-wide interoperability of e-learning systems are presented [12].

The main metadata standards of files containing digital photo images are discussed in [13]. Software system realizing extraction of metadata about image content and context is presented. The system gives the opportunity for searching photo images with different criteria using metadata standards EXIF, IPTC and XMP Possible applications of the system are intelligent Internet searching of digital photo images, automated filling in of SCORM metadata (if the corresponded learning material is a digital image), etc.

A classification machine learning system using multidimensional numbered information spaces is built [14]. Some practical implementation of MPEG-7 descriptors for definition and automatic extraction of color harmonies and contrasts, which cover intermediate level of image search are investigated [15]. The work over the area of classification system construction and experiments with trying for automated recognition of high level metadata, based on content based image retrieval, born some questions, which solving led to creating of specific theory that connects categorization/metadata and logic-combinatorial structuring/clustering of the descriptive part of the input table [16]. Some algorithms for group decision making has also been reported [17].

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] Polfreman M., Rajbhandari S. MetaTools - Investigating Metadata Generation Tools. JISC Final report, October 2008
- [2] Dobreva M., Ikonov N. The Role of Metadata in the Longevity of Cultural Heritage Resources. In Proc. of EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Greece, 2009.
- [3] Larran Aga. et al. Building Learning Objects from Electronic Documents. In Proc of the 16 International Conference on Computers in Education ICCE 2008, Taiwan, October 2008, pp.141-146.
- [4] INSPIRE'07, Directive 2007/2/EC of the European Parliament. Official Journal of the European Union, 25.4.2007, L 108/1. http://www.epsplus.net/content/download/3477/38314/ile/_10820070425en00010014.pdf.
- [5] <http://start.e-edu.bg/>
- [6] <http://www.enterprisestrategygroup.com/management>
- [7] <http://media-vibrance.itn.liu.se/>
- [8] <http://www.tiltomo.com/>
- [9] <http://www.alipr.com/>
- [10] Blagoev D., G. Totkov, Information Extraction by Learning Restricted Finite State Automata from Marked Web Documents, iTech'09 (this conference).
- [11] Indzhov Hr., D. Blagoev, G. Totkov, *Executable Petri Nets: Towards Modelling and Management of e-Learning Processes*, ACM International Conference Proceeding Series; Vol. 375, Proc. of the 10th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing 2009, Rousse, Bulgaria, June 18-19, 2009 (in print).
- [12] Doneva R., G. Totkov, N. Kasakliev, E. Somova, European Standardization: Mobility without Borders, ACM International Conference Proceeding Series; Vol. 375, Proc. of the 10th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing 2009, Rousse, Bulgaria, June 18-19, 2009. (in print).
- [13] Totkov G., E. Somova, Hr. Petrov, About Relationship between Metadata and Content of Digital Photo Images, 7th International Conference on Emerging e-Learning Technologies and Applications, Nov. 17-20, 2009, Stara Lesna, The High Tatras, Slovakia (accepted).
- [14] Mitov I., Ivanova Kr., Markov Kr., Velychko V., Vanhoof K., and Stanchev P.. PaGaNe – An Intelligent System Based on the Multidimensional Numbered Information Spaces. Fourth Int. Conf. on Intelligent Systems and Knowledge Engineering, 27-28.11.2009, Hasselt, Belgium (to appear).
- [15] Ivanova Kr. and Stanchev P. Color Harmonies and Contrasts Search in Art Image Collections, First Int. Conf. on Advances in Multimedia MMEDIA 2009, 20-25.07.2009, Colmar, France (to appear).
- [16] Ivanova Kr., Mitov I., Markov Kr., Stanchev P., Vanhoof K., Aslanyan L., and Sahakyan H. Metric Categorization Relations Based on Support System Analysis. Seventh Int. Conf. on Computer Science and Information Technologies CSIT 2009, 28.09-02.10. 2009, Yerevan, Armenia (to appear).
- [17] Andonov F. Interactive Methods for Group Decision Making. In Int. Book Series "Information Science & Computing" – Book No: 10. Intelligent Support of Decision Making. Sofia, 2009, pp. 25-30

Authors' Information

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

George Totkov – University of Plovdiv; chair of Computer Informatics Department; e-mail: totkov@uni-plovdiv.bg

Peter Stanchev – Kettering University, Flint, MI, 48504, USA / Institute of Mathematics and Informatics – BAS; chair of Information Systems Department; Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: pstanche@kettering.edu

Elena Shoikova – Technical University of Sofia; chair of Research laboratory "Technologies and standards for learning"; e-mail: shoikova@tu-sofia.bg