

---

## METASPEED: Metadata ExTraction for Automatic SPEcifications of E-Documents

*Science-research Project: Automated Metadata Generating for e-Documents Specifications and Standards*

*Supported by: Bulgarian National Science Fund under contract D002-308*

*Thematic priority: 4. Information and Communication Technologies*

*Duration: 2009-2011*

---

### 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, Technical University of Sofia, New Bulgarian University and Institute of Mathematics and Informatics, Bulgarian Academy of Science.

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 by common schemas and rules for the purpose of 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. Thus building an object repository appears to be a very labor consuming activity carried out by highly qualified people. Consider the electronic resources e.g. tests, learning content, etc. in the National Educational Portal (<http://start.e-edu.bg>). 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	e-learning ,computer linguistic
	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 valorization 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 cultural heritage multimedia documents;
- c) prescriptions for Bulgarian standards in different subject areas;
- d) methods 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 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:

- critical analyses of existing methods and tools for retrieval of metadata from electronic texts;
- investigation of specialized technologies and methods for metadata retrieval for documents in different areas;
- Design and implementation of proper software prototypes;
- Experiments with the realized methods and tools for metadata retrieval on the documents in examined areas.

---

### *WP3. Automated 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 (<http://www.enterprisestrategygroup.com/management>) 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 we need to extend our 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 (<http://media-vibrance.itn.liu.se/>), Tiltomo (<http://www.tiltomo.com/>) and Alipr (<http://www.alipr.com/>), declare to index hundreds of thousands of images.

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

- *Low level* – this level includes basic perceptual features of visual content (dominant colors, color distribution, texture pattern, etc.);
- *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. The visual art is area, where these features play 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;
- *High level* – this level includes queries according to rational criterions. 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 *semantic gap*. The Project aims at finding and analyzing new content-based image retrieval methods to analyze, index, and retrieve the image 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 finding the description of some concept, which humans use in their everyday life.

---

***WP4. Automated creating 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 creating and studying of methods and tools for *automated metadata generating from Internet-pages (in corresponded thematic domains)*. There will be examined and adapted known technologies for search in Internet-pages, using tools for automated metadata generating from text and multimedia. As a result Internet-space will be used as a source for creating of digital repositories and for testing corresponded methods and tools. Other digitalized sources for carrying out the experiments are created during the years collections of electronic resources of the Project partners in the field of e-learning and scientific publications.

Basic interest during leading the *investigations of p.A* will be separate on studying and elaborating of search methods in document collections (contained text and graphic objects) and their automated classifying in appropriate ontological systems.

In *investigations of p.B* special attention will be focused on the question for *automated metadata retrieval* from learning resources (for instance: concepts, relations between them, degree of complexity, etc.) and generating of new learning objects (for instance test tasks and e-courses), rendering an account the results of leading e-learning (for different types of learners). For this purpose, except leading a linguistic analysis (processing of the content of used learning materials and tests), it is necessary (in order to data extraction for so cold *portfolio* of each learner) to trace and to analyze dialogue between learners and lecturers, learners and system for content management, etc.

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.

In case of *investigations in p.D* it is posed a question for creating repository of scientific publications review (using automated metadata retrieval from full-text articles). Referring of scientific articles compels referent-experts to review a lot of articles in journals and books, connected with given scientific area, to classify and to summarize on thematic directions. Main problem of the referring is a quantity of work, which referents have to do, to find important for the domain articles and results. Automating of the preliminary selection and classification of the incoming papers can be assist referring and decrease the probability for gaps. In the same time, automated solving of the opposite task – how is adequate one review to the content of the corresponding paper – is also important.

Automated metadata retrieval from the collection of scientific publications in different languages (in the case of specialized texts in concrete scientific domain) and theirs classifying in corresponded ontological system is important in the case of multilingual referring journals (for instance – containing reviews in English, Russian and Bulgarian languages) increase the quality of the issue.