
Bibliography

- L.Ciocoiu, I. Petre, D. Smada, D. Barbu, D. Nicolau – Multimedia Portal dedicated for people with disabilities for National Authority for People with Disabilities – 2006; <http://www.anph.ro>
- L.Ciocoiu, L.Constantinescu, I. Petre, D. Smada, D. Barbu, D. Nicolau - eBiMuz - Integrated multimedia system for access to the multicultural thesaurus of the areas inhabited by Romanians, as integrated part of the European culture – CEEEX 142/2005; <http://ebimuz.ici.ro>
- L.Ciocoiu, I. Petre, D. Smada, D. Barbu, D. Nicolau – eMeditur – A tool for delivering information for online services in the medical assistance and tourism - R2130/2005 – <http://emeditur.ici.ro>

Authors' information

Laura Ciocoiu – senior researcher; National Institute for Research and Development in Informatics; 8-10 Averescu Avenue, 011455 Bucharest 1 Romania; e-mail: ciocoiu@ici.ro; <http://intelligent-agents.ici.ro>

Ionuț Petre – researcher; National Institute for Research and Development in Informatics; 8-10 Averescu Avenue, 011455 Bucharest 1 Romania; e-mail: ipetre@ici.ro; <http://intelligent-agents.ici.ro>

Dragoș Smada – researcher; National Institute for Research and Development in Informatics; 8-10 Averescu Avenue, 011455 Bucharest 1 Romania; e-mail: dsmada@ici.ro; <http://intelligent-agents.ici.ro>

Dragoș Nicolau – senior researcher; National Institute for Research and Development in Informatics; 8-10 Averescu Avenue, 011455 Bucharest 1 Romania; e-mail: dragos@ici.ro; <http://intelligent-agents.ici.ro>

ELECTRONIC PRESENTATION OF BULGARIAN EDUCATIONAL ARCHIVES: AN ONTOLOGY-BASED APPROACH

Anna Devreni–Koutsouki

Abstract: *The paper presents an ongoing effort aimed at building an electronic archive of documents issued by the Bulgarian Ministry of Education in the 40ies and 50ies of the 20th century. These funds are stored in the Archive of the Ministry of the People's Education within the State Archival Fund of the General Department of Archives at the Council of Ministers of Bulgaria. Our basic concern is not the digitization process per se, but the subsequent organization of the archive in a clear and easily-searchable way which would allow various types of users to get access to the documents of interest to them. Here we present the variety of the documents which are stored in the archival collection, and suggestions on their electronic organization. We suggest using ontologies-based presentation of the archive. The basic benefit of this approach is the possibility to search the collection according to the stored content categories.*

Keywords: *digitization, archives, history of education, ontologies, SWP.*

ACM Classification Keywords: *H.5 Information Interfaces and Presentation, H.3.3 Information Search and Retrieval, H.3.7 Digital Libraries*

Introduction

Digitization of cultural and scientific content in European countries is important field of work which results should contribute to the development of The European Library portal (TEL)¹. Currently, there are numerous ongoing digitization projects and initiatives in libraries, archives and museums.

Within this general picture, extensive work on digital capture and exposure of educational archives has not been undertaken so far, according to our research. In the educational field most attention is concentrated on the development of e-learning applications while historical documents of the educational institutions and the governmental bodies shaping the policy in education and research field are still not digitized on mass scale.

¹ <http://www.theeuropeanlibrary.org/portal/index.html>, date of last visit March 21, 2006.

However, such documents could be of interest not only to the researchers who study the development of the educational system (in one country or on comparative basis). Educational archives contain documents which could be of interest to the local historians, and to the general citizen.

Therefore, we decided to undertake an effort which would present in the electronic space the documents from the archive of the Ministry of Education of Bulgaria. We decided to start this effort with practical work on the documentation from the 40ies and 50ies of the 20th century, since this was one significant period of reform of the educational system in Bulgaria.

The presentation of educational archives also imposes some challenges.

1. Digitisation and metadata.

This type of archive contains quite diverse documents - official documentation, letters, notes, photographs, various documents, newspapers. The text documents can be printed, typewritten or handwritten. On the one hand, the digitization requires the application of different workflows. On the other hand, the metadata for these various documents, if detailed, should follow different structures.

2. Presentation and use

There has been a standing issue coming from the past – the problem related to the storage and access provision to already created materials, which were not designed for computer processing. We envisage the vast amount of documents, forms, protocols, letters/correspondence, photographs, maps, images and other objects which could be found in private or public museum collections or state, local or personal archives. The educational archive is a typical example of such a diverse collection. How should this collection be organized in the electronic space? If it just follows the traditional archival structure, the search of documents would be very difficult – one would have to browse everything, or search for the exact document. The general user does not necessarily have this information, neither should he (she) be knowledgeable about the metadata used. Thus our work is directed to looking for better and more user-friendly ways to provide access to the electronic collection.

The Archive and its Presentation

The idea for this effort was coined within a group of historians and education specialists from the University of Ioannina, Greece who work on comparative study of the Greek and the Bulgarian educational systems in the middle of the 20th century. Till now, the archives of the Bulgarian Ministry of Education (Ministry of the Enlightenment in the studied period) have been studied within 1940-1945. The sources are stored in funds 798k and 177k. of the Ministry of Peoples' Education 1879-1944.

Digital copies of several thousands of documents have been made. They are not sufficient for the purposes of the comparative study of the educational systems, but are sufficient for our purposes to suggest the organisation of the electronic collection and its use. The collected materials are interesting for the variety of types of sources they present. The next table summarizes the available document types which can be found as separate archival units. Here we do not discuss the issues of creating metadata on the whole inventory of documents, but rather describe the issues of describing the separate archival units.

DOCUMENT TYPE	EXAMPLE	METADATA AND CONTENT PRESENTATION PROBLEMS
Handwritten texts (general purpose documents, orders, notes, etc.)	This type of documents is typical for all archival collections.	Metadata for describing archival units can be applied. If we aim full text presentation, we have to face a massive amount of hand text entry. Typical elements appearing in these documents are names (personal and place names), dates, affiliations. Such documents are interesting for study of the problems which circulated in the educational administration.

<p>Typewritten documents (general purpose documents, orders, notes, etc.) in some cases with handwritten resolutions</p>	<p>This type of documents is also typical for all archival collections. We place it separately from the group above, because digitisation and processing of typewritten documents may involve OCR and the workflow would be different.</p>	<p>The same as above; OCR can be tested for text recognition.</p>
<p>Handwritten documents presenting records related to the educational sector, in some cases with signatures and stamps</p>	 <p>Sample from Fund 798 k, inventory list 2, archival unit 98. Book of orders of the Seres High School</p>	<p>Here we can use again the general metadata. If we aim to present the full text we should re-create the structure. Additional issue is how to present structured data on stamps and signatures.</p>
<p>Typewritten documents presenting records related to the educational sector, in some cases with signatures and stamps</p>	<p>This type of documents is also typical for all archival collections. As with typewritten generic texts, we place these documents separately from the group above, because digitisation and processing of typewritten documents may involve OCR and the workflow would be different.</p>	<p>The same as above; OCR can be tested for text recognition.</p>
<p>Individual documents with signatures, postal stamps, state fee stamps</p>	 <p>Sample from Fund 798 k, inventory list 2, archival unit 114. Certificate for a completed educational degree, Seres High School</p>	<p>Here we can use again the general metadata. Again, an issue is how to present structured data on stamps and signatures. State stamps might be of interest, for example, to philatelists, i.e. in a very structured approach we should encode data on these objects too in order to make the information on them searchable.</p>
<p>Newspapers <i>(The newspapers contain orders of the Ministry of education, reports, letters of local administrations, materials about a cultural week of the village, etc.)</i></p>	 <p>Sample from Fund 177 k, inventory list 2, archival unit 2251. Certificate for a completed educational degree, Seres High School</p>	<p>Here we can use again the general metadata for the archival unit, but then we should decide how to present the contents of the newspaper. A highly structured approach would require to present the content in detail, and/or provide full text search capabilities. The photographs in the newspapers also should be considered as a separate object.</p>


Photographs	 <p data-bbox="485 454 970 539">Sample from Fund 177 k, inventory list 2, archival unit 2251. Certificate for a completed educational degree, Seres High School</p>	The description of photographs differs from description of documents. Currently we study projects which deal with electronic presentation of historical photographs in order to suggest what metadata to use within the frameworks of our endeavour.
-------------	--	--

Table 1. Samples of documents from the archives of the Bulgarian Ministry of Education, 1940-1945

This brief presentation illustrates some of the problems which we face:

- How detailed should be the presentations of the various types of documents? On the one hand, we might be tempted to provide full text for all documents, but is this effort justified?
- How exactly to present multimodal objects (as we see in the examples, we have special layouts in some cases; stamps; signatures; marginal notes, etc.).

We believe that one approach which makes such collections searchable even without the application of very detailed and fragments presentations is the proper use of ontologies. Below we will present briefly the concept of ontologies and then will present one possible practical solution, SWP.

Ontologies

In philosophy, ontology (from the Greek ὄν, genitive ὄντος: of being (part. of εἶναι: to be) and -λογία: science, study, theory) is the study of being or existence. It seeks to describe or posit the basic categories and relationships of being or existence to define entities and types of entities within its framework. Ontology can be said to study conceptions of reality. It is often confused with epistemology, which is about knowledge and knowing.

According to recent artificial intelligence research "an ontology is a shared and common understanding of some domain that can be communicated across people and computers" [Gruber, 1993], [Guarino, 1995], [Borst, 1997] and [van Heijst et al., 1997]. Ontologies can therefore be shared and reused among different applications [Farquhar et al., 1997]. "An ontology can be defined as a formal, explicit specification of a shared conceptualization" [Gruber, 1993], [Borst, 1997]. "Conceptualization" refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. "Explicit" means that the type of concepts used, and the constraints on their use are explicitly defined. "Formal" refers to the fact that the ontology should be machine-readable. "Shared" reflects the notion that ontology captures consensual knowledge, i.e. it is not private to some individual, but accepted by a group.

The concept of ontology is defined even narrower within the famous project *Ontolingua* of Stanford University [Ontolingua project]. It suggests that the ontology is an *explicit specification of some topic*. This approach presupposes formal and declarative presentation of a given topic, which includes the vocabulary (or names) for referring to the terms in that subject area and the logical statements that describe what the terms are, how they are related to each other, and how they can or cannot be related to each other.

Ontology describes the subject matter using the notions of *concepts, instances, relations, functions, and axioms*. Table 2 presents the requirements to which ontology has to be compliant.

Necessary properties of an ontology	Typical but not mandatory properties	Desirable properties, but not mandatory nor typical
Finite controlled (extensible) vocabulary Unambiguous interpretation of classes and term relationships Strict hierarchical subclass relationships between classes	Property specification on a per-class basis Individual inclusion in the ontology Value restriction specification on a per-class basis	Specification of disjoint classes Specification of arbitrary logical relationships between terms Distinguished relationships, such as inverse and part-whole

Table 2. Properties of ontologies.

From the practical point of view, in the simplest case an ontology is „a formal explicit description of concepts in a domain of discourse (classes (sometimes called concepts)), properties of each concept describing various features and attributes of the concept (slots (sometimes called roles or properties)), and restrictions on slots (facets (sometimes called role restrictions))” [Noy, McGuinness]. If we take accept this as a rule of thumb, an ontology together with a set of individual instances of classes indeed can be seen a *knowledge base*. However, in reality, there is a fine line where ontology ends and the knowledge base begins – the latter can be more sophisticated presentation of a subject domain while ontology is always hierarchical and follows certain requirements as described above. From technological point of view, ontologies can be seen as knowledge bases of special kind, which can be “read” and understand, and could be shared between users and/or developers.

The basic reasons to create ontologies are summarized in [Noy, McGuinness] as follows:

- To share common understanding of the structure of information among people or software agents
- To enable reuse of domain knowledge
- To make domain assumptions explicit
- To separate domain knowledge from the operational knowledge
- To analyze domain knowledge

The development of ontologies is still a difficult and challenging task, because so far there are no common platforms and verified methods which would prescribe what procedures should be followed in the process of creating ontology. As [Jones et al.] explains it, „at present the construction of ontologies is very much an art rather than a science”. This situation needs to be changed, and will be changed only through an understanding of how to go about constructing ontologies. In short what is needed is a good methodology for developing ontologies.

While there is no common methodology for building ontologies, there are principles for design and implementation suggested in [Gruber 1995]:

- **Clarity** – the ontology should present the terms included efficiently and without ambiguities. The definitions should be objective as much as possible, although the motivation for adding a term might be driven by the circumstances and the requirements for computability. A clear formalism should be used, and it is recommended to present the definitions in the form of logical axioms.
- **Coherence** – the definitions should be logically disambiguous, and all statements derived from the ontology should not be in disaccordance with the axioms.
- **Extendibility** – the ontology should be designed so that the dictionaries of terms could be enlarged without revision of concepts already defined.
- **Minimal encoding bias** – the conceptual abstraction implemented in the ontology should be developed on the concept level instead of the level of the symbolic representation.
- **Minimal ontological commitment** – the ontology should contain only the most essential assumptions on the modeled world, so that there is enough space for making it wider or narrower.

How do ontologies relate to our archival presentation task? We believe that the use of ontology could be a good solution which would allow users to make a variety of searches within the collection of electronic documents while these documents are still not available in searchable full text format. If we incorporate as an element of the data several relevant ontological references, based on the assumptions for typical requests for information, the results returned to a query would include all documents which metadata are matching the concept from the ontology.

Definitely, this requires extra human effort: first, to develop a subject domain ontology (covering *educational administrative documentation*) – to the best of our knowledge such ontology does not exist, and moreover it would be specific for the Bulgarian documentary system; and second, to add references to the concepts from the ontologies within the archival units metadata. Compared to the creation of full text and sophisticated search tools, we believe that this approach will lead to fast and reliable results and will implement it in the nearest future.

A Possible Practical Solution Involving Ontologies: SWP

Over the past few years, various approaches have been proposed to effectively organise digital content on the Web. Traditionally, these have included techniques such as building keyword indices based on image content, embedding keyword-based labels into images, analyzing text immediately surrounding images on Web pages, etc. Nevertheless, current Web technology presents serious limitations to make information accessible for users

in an efficient manner. The general problem to find information on the Web is summarized in [Ding, Fensel 2001]: „searches are imprecise, often yielding matches to many thousands of hits”. Moreover, users face the task of reading the documents retrieved in order to extract the information desired. These limitations naturally appear in existing Web portals based on this technology, making information searching, accessing, extracting, interpreting and processing a difficult and time-consuming task.

More recently, there has been a research focus on the Semantic Web technologies in different domains. The purpose of the Semantic Web is to create a universal medium for the exchange of sharable and processable data by automated tools, such as software agents, as well as by the users. “The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation” [Berners-Lee et al., 2001].

One solution to these problems is the use of Semantic Web Portals (SWPs), known also under the names *Knowledge Portals* or *Community Web Portals*.

There are different views in the research works on SWPs and their definition. An earlier study defines them as „portals for the information needs of particular communities on the web” [Staab et al., 2000]. According to [Gorcho, 2006] “a Semantic Web Portal is a Web application that offers information and services related to a specific domain, and that has been developed with Semantic Web technology”. The same author emphasizes that the primary difference with the traditional Web Portals “is based on technological aspects: traditional Web portals are based on standard Web technology (HTML, XML, servlets, JSPs, etc.); semantic portals are based on that technology plus the use of Semantic Web languages like RDF, RDF Schema and OWL”.

The SWPs which are well developed and functioning are not too many; also they are prone to various limitations. In [Karvounarakis et al, 2000] they are defined as Web applications that “provide the means to select, classify and access, in a semantically meaningful and ubiquitous way, various information resources (e.g., sites, documents, data) for diverse target audiences (corporate, inter-enterprise, e-marketplace, etc.)”.

[Lausen et al., 2005] and [Lara 2004] offer more strict definition, which states that SWP has the following characteristics:

- It is a web portal. A web portal is a web site that collects information for a group of users that have common interests
- It is a web portal for a community to share and exchange information
- It is a web portal developed based on semantic web technologies.

The briefest but clear explanation is to view SWPs as “portals that typically provide knowledge about a specific domain and rely on ontologies to structure and exchange this knowledge” [Hartmann, Sure 2004]. The accent here is on the most typical feature of SWPs – their application in specific subject domains and the use of one or more ontologies as a backbone of the application.

Currently „SWP are still at their very early stages” [Lausen et al., 2005]. The benefits of implementing these Semantic Web technologies can be easily identified or foreseen as Semantic Web technologies have the potential to increase the information consistency and the information processing quality of portals. On the other hand, Semantic Web technologies themselves are still under development and most of the theoretical issues are no easy to be employed into real world applications.

Conclusion

The national strategy of many countries, including private institutions, which possess such collections and archives, is making them widely-spread and accessible. The common practice is the creation of repositories of images or digital copies which can already be accessed through the Web [Hyvönen et al., 2004]. The management of such resources aims to reach maximum effectiveness of search in the sea of various forms of the stored information. Many of such collections currently exist and users are increasingly faced with problems of finding a suitable (set of) image(s) for a particular purpose. Each collection usually has its own (semi-) structured indexing scheme that typically supports a keyword-type search. However, finding the right image is often still problematic [Hollink et al., 2003].

In this paper we present a brief analysis of the types of documents in one particular Bulgarian archive (educational documentation from the 40ies and 50ies of the 20th century). We also made a brief overview of ontologies and SWPs which could help in structuring the electronic surrogates of archival documents. In our

future work we will suggest ontology designed especially to present the documents from this archival collection and the implementation via SWP of search tools for use of the archive.

This collection of documents in the archive presented is highly fragile – already now the documents are deteriorating as it could be seen from the illustrations in Table 1. We hope that our effort will help to preserve for the future these documents which could be of interest to various groups of users.

Bibliography

- [Berners-Lee et al., 2001] Tim Berners-Lee, James Hendler and Ora Lassila, The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, Scientific American, 2001 (Visited 02-03-07) <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>
- [Borst, 1997] W. N. Borst. Construction of Engineering Ontologies. PhD thesis, University of Twente, Enschede, 1997.
- [Ding, Fensel, 2001] Ding, Y.; Fensel, D.: Ontology Library Systems. The key to successful Ontology Re- Use. In: Proceedings of the First Semantic Web Working Symposium. California, USA: Stanford University 2001; S. 93-112.
- [Farquhar et al., 1997] A. Farquhar, R. Fikes, and J. Rice. The ontolingua server: a tool for collaborative ontology construction. International Journal of Human-Computer Studies, 46(6):707–728, June 1997.
- [Gorcho, 2006] Corcho, O.: A Platform for the Development of Semantic Web Portals In Proceedings of the 6th international conference on Web engineering, Palo Alto, California: ACM International Conference Proceeding Series Pages 2006; P 145 - 152
- [Gruber 1995] Gruber, T., "Toward Principles for the Design of Ontologies Used for Knowledge Sharing", International Journal of Human-Computer Studies, Vol. 43 (1995), pp. 907-928
- [Gruber, 1993] T. R. Gruber. A translation approach to portable ontology specifications. Knowledge Acquisition, 5:199–220, 1993.
- [Guarino, 1995] N. Guarino. Formal ontology, conceptual analysis and knowledge representation. International Journal of Human-Computer Studies, 43(5/6):625–640, 1995. Special issue on The Role of Formal Ontology in the Information Technology.
- [Hartmann, Sure 2004] Hartmann J., Y. Sure, "An Infrastructure for Scalable, Reliable Semantic Portals" IEEE Intelligent Systems 19 (3): 58-65. May 2004
- [Hollink et al., 2003] Hollink, L., Schreiber, G., Wielemaker J., and Wielinga. B. Semantic Annotation of Image Collections in Knowledge Capture - Knowledge Markup & Semantic Annotation Workshop (2003)
- [Hyvönen et al., 2004] Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., Kettula, S., MuseumFinland—Finnish museums on the semantic web in 3rd International Semantic Web Conference (ISWC2004, Hiroshima, Japan 07-11 November 2004)
- [Jones et al.] Jones, D., Bench-Capon, T., Visser, P., "Methodologies for ontology development", (Visited 06-03-2007) <http://www.iet.com/Projects/RKF/SME/methodologies-for-ontology-development.pdf>
- [Karvounarakis et al, 2000] Karvounarakis G, Christophides V, Plexousakis D, Alexaki S (2000) Querying community web portals. Technical report, Institute of Computer Science, FORTH, Heraklion, Greece.
- [Lara 2004] Lara R., Sung-Kook Han, Holger Lausen, Michael Stollberg, Ying Ding, Dieter Fensel, " An Evaluation of Semantic Web Portals", IADIS Applied Computing International Conference 2004, Lisbon, Portugal, March 23-26, 2004
- [Lausen et al., 2005] Lausen H., Ying Ding, Michael Stollberg, Dieter Fensel, Rubén Lara, and Sung-Kook Han, "Semantic web portals: state-of-the-art survey", Journal of Knowledge Management, 2005, Volume: 9 Issue: 5 Page: 40 – 49
- [Noy, McGuinness] Noy, N, McGuinness, D, Ontology Development 101: A Guide to Creating Your First Ontology, http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html
- [Ontolingua project] (Visited 06-03-2007) <http://www.ksl-svc.stanford.edu:5915/doc/frame-editor/what-is-an-ontology.html>
- [Staab et al., 2000] Staab S., J. Angele, Stefan Decker, Michael Erdmann, Andreas Hotho, Alexander Maedhe, Hans-Peter Schnurr, Rudi Studer, York Sure , „Semantic Community Web Portals”, In: Computer Networks (Special Issue: WWW9 - Proceedings of the 9th International World Wide Web Conference, Amsterdam, The Netherlands, May, 15-19, 2000), Elsevier.
- [van Heijst et al., 1997] G. van Heijst, A. T. Schreiber, and B. J. Wielinga. Using explicit ontologies in KBS development. International Journal of Human-Computer Studies, 46(2/3):183–292, 1997.

Author's Information

Anna Devreni–Koutsouki – PhD student, Sofia University, Bulgaria; e-mail: annadevreni@hotmail.com