Using a Metadata Schema Registry in the National Digital Data Archive of Hungary

Csaba Fülöp, Gergő Kiss, László Kovács, and András Micsik

MTA SZTAKI,
Computer and Automation Research Institute,
of the Hungarian Academy of Sciences,
Department of Distributed Systems,
H-1111 Budapest XI. Lágymányosi u. 11. Hungary
{csaba.fulop, gergo.kiss, laszlo.kovacs, micsik}@sztaki.hu

Abstract. The National Digital Data Archive (NDDA) is an ongoing initiative of the Hungarian government that makes Hungary's national cultural assets available in digital form. The NDDA features a decentralized OAI-based network of archives and service providers facilitating discovery and access to digitized objects. Authors' participation in the project is described including the implementation of an NDDA service provider. This service provider is connected with an RDF-based metadata schema registry enabling the service to automatically adapt to the metadata schemas defined within the NDDA.

1 Introduction

The National Digital Data Archive (NDDA) is an ongoing initiative of the Hungarian government that makes Hungary's national cultural assets available in digital form [1]. Despite its name, NDDA is not a digital archive itself, rather a program that will enable a higher degree of co-operation at the archive levels, and a better integration at client levels. The motto is: „The NDDA is the nation's digital data store."

The NDDA places emphasis on self-organisation and co-operation among communities. The goal is to establish the technical infrastructure for the knowledge-based society where discovery and access to data and information in digital format becomes an everyday need. Actions are supported on different levels:

- Definition of key concepts and structures such as namespaces, metadata schemas, identifiers.
- Definition of the building blocks of the infrastructure and example implementations of these components.
- Digitization of cultural assets: museums, radios, music archives, libraries, etc. are supported in the process of converting their valuable material into digital format.
- Establishing digital archives and connecting them to the NDDA.
- Creation of new services for the easy and unified access to digitized assets.

MTA SZTAKI Department of Distributed Systems participates actively in the development of the NDDA at various levels: in the work of the advisory committee, in

A. Rauber et al. (Eds.): ECDL 2005, LNCS 3652, pp. 314 – 322, 2005.
© Springer-Verlag Berlin Heidelberg 2005

the design of metadata schemas, connecting archives to the network and implementing unified services. In this paper we present our view of the NDDA and our contributions to the initiative.

2 Architecture of the NDDA

The basic concept of the NDDA is a decentralized OAI-based network of archives and service providers [2]. In a broad sense, there are three types of information providers in the network:

- Data providers make metadata available for harvesting. They may also make their data downloadable either freely or with some restrictions.
- Service providers offer services for the public or for other NDDA components.
- Protocol providers serve core schemas, protocols needed for interoperability in the network.

Data providers, usually operated by the maintainers of the archives, implement the OAI Protocol for Metadata Harvesting (OAI-PMH) using NDDA metadata schemas instead of the plain old oai_dc schema. OAI-PMH does not contain any mechanism for the retrieval of the data. A usual solution was selected for this problem: the Identifier metadata element contains the URI for the downloadable resource. In this way archives provide metadata and data as well for the NDDA, thus facilitating both discovery and access to stored content.

Some service providers are planned to provide basic services usable by the whole network, not only by the clients. Descriptions for locations and persons are provided as separate services, as well as an ontology of Hungarian subject terms. These controlled databases can be harvested through OAI-PMH by service providers and archives. Archives may use this information for the improvement of their metadata definition process. Persons, locations or terms may be referred from these services in the subject, coverage, creator, contributor and other metadata elements. Service providers may use this information for the improvement of their query and browse facilities.

Protocol providers are primarily registries and repositories of internal standards for NDDA. These internal standards may include for example metadata schemas and protocol extensions.

Up to the date of paper submission, two regular service providers exist in the network, a generic service provider for authority records (service providers for geographical locations and Hungarian thesaurus are under development), and a protocol provider, which stores metadata schema definitions. 20 archives are available within the NDDA, with more than 500,000 metadata records. The number of resources available digitally is almost ten percent of all metadata records. Data types represented in our service provider are texts (cca. 60%), images (cca. 25%), video (cca. 2%), audio (cca. 1%) and statistical data (nearly 10% of the records contain no type information).

The initiative decided to create its own national metadata schemas. These schemas are based on DCMI Metadata Terms using element refinements and encoding schemas. This ensures more precise description of metadata than the standard, „unqualified” oai_dc schema. Furthermore, a specialized metadata schema is defined for each genre of stored items in NDDA. Currently, there are metadata schemas for textual documents, images, audio and video programmes.

3 Implementation of a Service Provider for the NDDA

Our department has implemented the first public OAI-based interconnection of libraries in Hungary [12]. Using the results of this pilot project, we decided to create a service provider for the NDDA. Metadata schemas did not stabilize until the development started which gave us the idea to automate metadata schema management in our prototype. Metadata schemas are thus dynamically loaded from a schema registry into the NDDA service provider (Fig. 2).

3.1 Schema Registry for Metadata

The authors previously participated in the CORES project, which provided a solution for automatic schema dissemination [3]. As part of the European Community funded IST Semantic Web Technologies programme, the CORES project has promoted the use of metadata schema registries to support the disclosure, discovery and navigation of information about metadata element sets stored as schemas distributed on the Web. Such a "schema navigation service" provides users (both human and software) with information about existing metadata element sets and the terms used within them. In particular, it assists implementers in locating and re-using existing schemas.

The CORES project implemented a registry infrastructure with the following components:

- A graphical schema creation tool: this tool facilitates the discovery and re-use of existing metadata schema elements and definitions by drag-and-drop and graphical editing. With this tool users are able to create proper schema definitions without learning schema definition languages such as RDF Schema.
- An API for the remote manipulation of schemas in the registry. The schema creation tool also uses this API to upload schemas to the registry.
- A schema registry, which stores schemas in an RDF database, generates various browsable and searchable representations of schemas and their relations.

Metadata schemas are constructed as RDF schemas [4] following the rules for creation of metadata application profiles [5]. Briefly, schemas are built using elements from existing element sets (e.g. Dublin Core), where each element can be refined in multiple ways:

- Permitted encoding schemes and values may be specified (e.g. W3CDTF for dates)
- Element definitions can be semantically narrowed
- The obligation and maximum occurrence of elements can be changed

The model used for storing metadata schemas is shown on Figure 1. Element sets and encoding schemes are the basis for schema construction. Application profiles select and refine these, thus creating new, specialised metadata schemas. Agencies play the role of publisher and maintainer of schemas, versioning and maintenance information of these schemas are stored as administrative data. It is possible to annotate all schema elements, which provides a way of sharing knowledge and practice among users and developers of schemas. The CORES toolkit is publicly available as a demonstration service on our department web server[1].

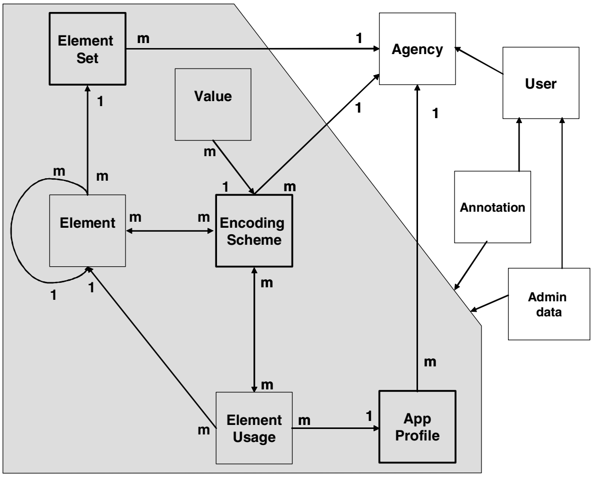


Fig. 1. The CORES model for metadata schema registry

3.2 The Harvester Module

The harvester periodically visits NDDA archives and collects metadata records. The metadata schema used for the available resources can be retrieved using OAI Protocol for Metadata Harvesting. The harvester prefers to fetch the record in an NDDA-defined schema, but plain Dublin Core records can be harvested as well. As the schema of the metadata record is known, the harvester can validate the record against the schema definition retrieved from the registry. During this validation we can identify the following problems with metadata records:

- Use of not allowed element or element refinement,
- Use of not allowed encoding scheme,
- Use of not allowed data types,
- Restrictions not fulfilled: for example a mandatory element is missing, a non-repeatable element is repeated.

When an improper element refinement is used, it can be replaced with its parent element. Inappropriate values can sometimes be automatically converted to the correct type or encoding scheme. Other types of the problems listed above cannot be corrected automatically in a safe way.

As our metadata storage and query engine are flexible enough to deal with such data inconsistencies, non-conforming elements are usually kept in their original format. Information loss is considered more disadvantageous than loose application of

[1] http://cores.dsd.sztaki.hu

schemas. In case of any problem with the metadata its publisher is notified about the non-conforming records.

Considering the small number of archives in NDDA so far and the close cooperation with these archives the use of metadata schemas is satisfactory. Generally, the problem is more about the incompleteness of metadata than about the conformance to metadata schemas. However, we make good use of automatic correction of values for Type, Language and Date elements. The most typical errors are:

- the use of Hungarian, Magyar, etc. as the value for Language element instead of ISO639-2,
- not using DCMIType vocabulary in the Type element,
- various non-standard date formats in Date elements.

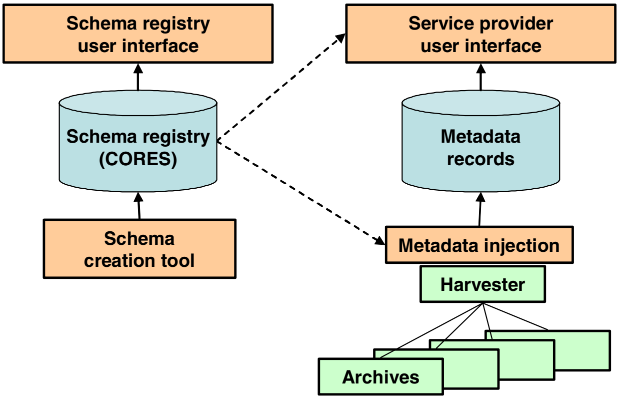


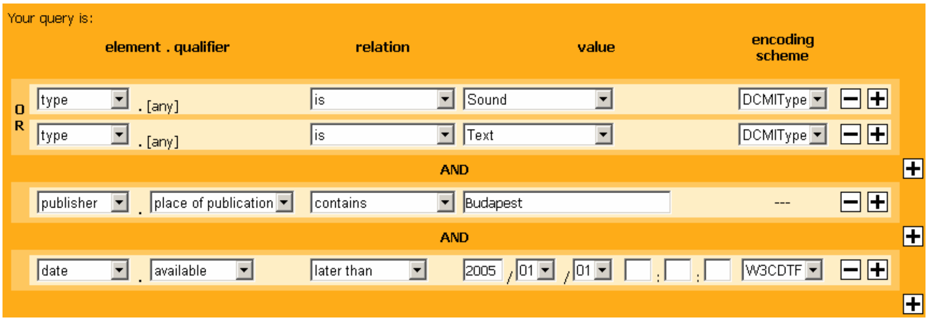
Fig. 2. Architecture of NDA@SZTAKI system

After validation metadata values are stored in a database with element refinements, encoding schemes and language identifiers all converted to an internal naming scheme. This schema-independent internal representation makes it possible to handle some of the schema modifications in the future.

Parallel to the introduction of new metadata schemas existing schemas are also bound to change. Such changes may break the schema-conformance of thousands of records. Besides automatic notification about schema changes to publishers, a system can also adapt to some of these changes. Renaming is considered as bad practice in this world, but with our extra mapping of names onto internal identifiers such changes can be handled transparently. A change in the allowed encoding scheme requires new data conversion algorithms, which normally needs human intervention. An interesting research question not elaborated here is the handling of semantic changes in a metadata schema, i.e. the shifting of the desired meaning of an element. In general, additions to a schema can be handled automatically in this system, while semantic changes may easily remain invisible. Reducing a schema presents the problems of information loss and backward incompatibility.

3.3 Query Interface

The user interface of the NDDA service provider[2] is accessible using Hungarian and English languages and offers various browsing and searching facilities for the nearly 500,000 metadata records. The service has been available since May 2004 and currently has an average of 5000 hits per day.



Your query is:

	element . qualifier	relation	value	encoding scheme
OR	type . [any]	is	Sound	DCMIType
	type . [any]	is	Text	DCMIType
AND				
	publisher . place of publication	contains	Budapest	---
AND				
	date . available	later than	2005 / 01 / 01 : :	W3CDTF

Fig. 3. An example query in the query construction window

Connecting the schema registry to the query interface enables us to provide the latest schemas and schema elements for query composition. This affects the list of elements in the simple search facility, which is a minimalist Google-like solution with one text input field and simple switches for query mode variants.

We also wanted to demonstrate the full potential of searching among metadata records of multiple schemas. The so-called advanced query interface enables the user to construct arbitrary Boolean query expressions in a simple and transparent way (Fig. 3). Queries are produced in conjunctive or disjunctive normal form, that is, queries are either in the form *(A or B) and (C or D)…*or in the form *(A and B) or (C and D)…*where *A,B,C,D* are simple (atomic) query expressions like *creator contains John.* Atomic expressions are composed of the selected schema element and the value sought. Elements may have a refinement selected and the value may be given using an encoding scheme. The type of match is a context sensitive selection box offering relations such as 'later than', 'earlier than' for dates, 'contains', 'begins with' or 'does not contain' for texts, etc. Negation is also possible as matching mode within atomic expressions.

A query construction process using all available features of the user interface contains the following steps:

- Restricting the query to run against selected archives
- Select metadata schemas for use during query construction
- Building the query expression
- Assign sorting order for the result
- Execute the query, browse results

[2] http://nda.sztaki.hu

- Switch back to query expression editor, refine query and execute it.
- Save the constructed query in personal profile for later reuse

The query expression editor provides a graphically emphasized view of the edited query where new atomic expressions can be easily added or removed using the plus and minus icons. Users may also include their saved queries into a newly built query: the selected query will be an atomic expression of the new query.

A functionality for users more familiar with Boolean logic is the ability to switch between conjunctive and disjunctive normal forms. This can be done in two ways: either the AND and OR operators are simply exchanged (syntactical) or the expression is transformed into an equivalent expression in the other format (semantic).

During 4 months of test operation, we found that most user activities are for browsing archives and sets (20%) and viewing metadata records (12%). Search results were retrieved in 5% for simple search and in 0.3% for advanced search of all accesses. 29% of accesses were initiated from Google hit lists as we allowed Google to index the contents of our service. 23% of user actions followed a link to the content or to another related webpage.

4 Related Work

The Stanford InfoBus [7] is a well-known architecture where archives with different metadata schemas (and query semantics) are unified as a single resource for the users. In this case user queries are automatically translated for each archive and their responses are merged into a query result. With the launch of the Open Archives Initiative the approach was shifted and the focus moved on standardized metadata export for unified query services. However, the use of a single metadata schema is not feasible in many scenarios. Instead of returning to the old world of multiple independent schemas, schema refinement and specialisation appeared in the evolution of Dublin Core [5]. A formal model for the description of schema reuse and schema refinement based on RDFS help to regulate this process and provide a mapping between schema elements [6]. A software framework supporting this model has also been developed [3]. Current paper describes the next step in this direction: connecting the metadata repository and the metadata registry in order to ensure metadata correctness and proper discovery with respect to registered metadata schemas.

Although best-match, ranked-output retrieval techniques are considered superior to exact-match systems based on Boolean queries in terms of recall and precision [11], Boolean queries are usually much preferred among professional searchers such as librarians [10]. Our system provides both solutions: ranked results using simple search and Boolean query construction within the advanced search interface. Precise query formulation and exact matches, for which the implemented interface is essential, are often requirements in libraries and archives, . We can mention Query By Templates (QBT) [9] or VQuery [8] to show how many different methods are available to help Boolean query construction. With QBT users can attach search terms to visually separated parts of a document (e.g. title, author and subheading) using a document template. VQuery also produces query expressions with the help of Venn diagrams: each search term is entered in a separate circle and the positions of the circles establish the Boolean operators between the terms. The main advantage of the query expression editor described in this paper compared to the above mentioned intuitive and experimental methods is its simple and minimalist design. It basically works with any web browser and does not require special graphical capabilities, yet it provides a view of the query, which is easy to understand and modify.

5 Conclusion

The NDDA has an essential role in the dissemination of Hungarian cultural objects and is a unique initiative both in its scope and in its architecture. It experiments with the principles of OAI and self-organisation at national level. Its distributed architecture stands as an example for new projects in tourism support, healthcare and administration.

In an initiative with such a broad focus the evolution of metadata schemas is a natural phenomenon. The presented service provides a working example for the connection of metadata registries and metadata repositories in real world settings. Furthermore, an easy-to-use interface for the construction of query expressions using multiple metadata schemas is also presented. This solution couples the results of the CORES project with our previous pilot experiments for library interconnections using OAI in Hungary. The connection provides the benefits of automatic checking of harvested metadata records and automatic adaptation of the query interface to the latest metadata schemas.

References

1. National Digital Data Archive (NDDA). http://www.nda.hu
2. A. Micsik: Open Archives Initiative: applications in Hungary and worldwide. Netties 2004 Conference, Budapest, 27-29 October 2004
3. R. Heery, P. Johnston, Cs. Fülöp, A. Micsik: Metadata schema registries in the partially Semantic Web: the CORES experience. 2003 Dublin Core Conference, DC-2003, 28 Sept - 2 Oct 2003, Seattle, Washington USA
4. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation 10 February 2004, http://www.w3.org/TR/rdf-schema/
5. R. Heery, M. Patel: Application Profiles: mixing and matching metadata schemas. Ariadne 25 (2000 September). http://www.ariadne.ac.uk/issue25/app-profiles
6. R. Heery, P. Johnston, D. Beckett, D. Steer: The MEG Registry and SCART: complementary tools for creation, discovery and re-use of metadata schemas. Proceedings of the International Conference on Dublin Core and Metadata for e-Communities, 2002.
7. M. Baldonado, C. K. Chang, L.Gravano, A. Paepcke: The Stanford Digital Library metadata architecture. International Journal on Digital Libraries, Volume 1, Issue 2, Sep 1997.
8. S. Jones, S. McInnes, M. S. Staveley: A graphical user interface for Boolean query specification. Int. Journal on Digital Libraries 2(2), Springer-Verlag, 1999, pp. 207-223
9. A. Sengupta, A. Dillon: Query by Templates: A Generalized Approach for Visual Query Formulation for Text Dominated Databases. 4th International Forum on Research and Technology Advances in Digital Libraries (ADL '97), May 7-9, 1997, Washington, DC

10. D. Byrd and R. Podorozhny: Adding Boolean-quality control to best-match searching via an improved user interface. Technical Report IR-210, Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, Massachusetts, 2000.
11. Nicholas J. Belkin and W. Bruce Croft: Retrieval techniques. In Martha E. Williams, editor, ARIST chapter 4, pages 109-145. Elsevier, 1987.
12. HEKTÁR project homepage. http://hektar.sztaki.hu