Share.TEC Repository System

Krassen Stefanov[1], Pavel Boytchev[2], Alexander Grigorov[3], Atanas Georgiev[4], Milen Petrov[5], George Gachev[6], and Mihail Peltekov[7]

[1,2,3,4,5,6,7] Faculty of Mathematics and Informatics, St. Kl. Ohridski University of Sofia, Bulgaria
[3] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
[1,2,3,4,5,6,7]{krassen, boytchev,alexander.grigorov, atanas, milenp, gachev, misho}@fmi.uni-sofia.bg

Abstract. The Share.TEC system has the main goal to establish a highly visible and functional portal with advanced brokerage services that will provide personalised access to a wide-range of Teacher Education (TE) content. The heart of the Share.TEC system is the central repository, storing metadata about TE resources. In this paper we describe the design of the digital Share.TEC repository, providing the more flexible and powerful ways for representing Common Metadata Model (CMM) metadata records and objects from the Teacher Education Ontology (TEO), and ensuring the most efficient and comprehensive search and reasoning abilities, as the key factors for the success of the Share.TEC project. We describe the data models for representing CMM and TEO, as well as the processes ensuring their correct coexistence.

Keywords: Teacher Education, Ontology, User Functionality, User Interface

1 Introduction

Digital research repositories are already well established throughout many countries in the European Union [16]. Recent surveys in the US show similar results. Digital repositories are on their way to become a permanent part of the scholarly communication and documentation research infrastructure. Most of the European digital repositories (95%) support the Open Access [17].

The Share.TEC system has the main goal to establish a highly visible and functional portal with advanced brokerage services that will provide personalised access to a wide-range of Teacher Education (TE) content. The heart of the Share.TEC system is the central repository, storing metadata about TE resources. All metadata stored in the repository follow the Common Metadata Model (CMM) metadata format [1], which is an extension of the Learning Object Metadata (LOM) format [18]. For providing more robust, flexible and powerful way for classifying TE resources, the Share.TEC project develop specific Ontology, called TEO (Teacher Education Ontology) [19]. So, the main functions of the Share.TEC repository are to provide the most useful and convenient support for all operations related to CMM and TEO. We choose to use Fedora Commons Repository system as a central repository (cache). We also recommend Fedora to be used as a local repository for partners that do not yet have a repository system. The main reasons for choosing Fedora for the central repository cache are:

- Fedora was recognized as the best repository system to be used as a central hub by various independent research surveys (see for example [2, 3]).

- Fedora provides extensive support for representing and using ontologies, which feature is difficult to find in any other existing repository software.
- Fedora supports many different search engines, and their combined use has no match within existing repository systems.
- Fedora is open source system, but it is also supported by many strong research organizations and is implemented in many big Universities.
- Fedora supports very powerful and flexible data model representations, which will enable us to combine in an easy way and use together both CMM and TEO.
- Fedora supports all the main protocols enabling the development and use of software services: Simple Object Access Protocol (SOAP), REpresentational State Transfer (REST), Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH), Resource Description Framework (RDF) and Web Ontology Language (OWL) for ontology representation.

As providing the more flexible and powerful ways for representing CMM and TEO, and ensuring the most efficient and comprehensive search abilities are the key factors for the success of the Share.TEC project, the choice of the Fedora system seems to be the best one, ensuring the ability to create the first project prototype on time and with almost all the features planned.

2 Implementation of CMM and interconnections between CMM and TEO data models

The Fedora object model [4, 5, 6] supports the expression of many kinds of complex objects, including documents, images, electronic books, multimedia learning objects, data sets, computer programs, and other compound information entities.

The Digital Object is the basic unit for information aggregation in Fedora. At a minimum a digital object has a persistent identifier (PID) and Dublin Core metadata [20] that provide a basic description of the digital object.

A Datastream is a component of a digital object that represents a data source. A digital object may have just the basic Dublin Core datastream, or any number of additional datastreams. Each datastream can be any mime-typed data or metadata, and can either be content managed locally in the Fedora repository or by some external data source (and referenced by a URL).

The architectural view of Fedora digital object model is shown below [5].

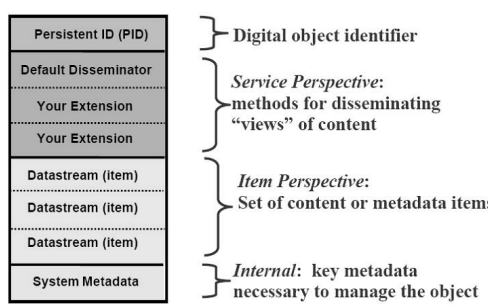

Fig. 1. Fedora Digital Object Model [5].

The Fedora object model [4] allows the definition of virtual representations of a digital object. Such a virtual representation, known as *dissemination*, is a view of an object that is produced by a service operation (a method invocation) that can take as input one or more of the datastreams of the respective digital object.

Starting with version 3.0 Fedora supports a new Content Model Architecture (CMA) that:

- Establishes a uniform way to classify objects;
- Provides a uniform way to access the model;
- Includes a simple content modeling language;
- Separates "Architecture" from "Model" concerns;
- Enables sharing content and service designs, validating objects;
- Enables adding customized functionality to content and sharing services.

The method in [7] for implementing OWL LITE in Fedora objects would be sufficient if we aim at having a Fedora-based representation of TEO. However, TEO internal structure should also support other functionalities like faster search and navigation (in respect to criteria provided by the use cases), on-the-fly mapping of incomming data for building references from these data to TEO nodes, and bi-directional translations of TEO entities.

Furthermore, the approach described in [7] has some limitations: OWL properties are defined locally for a class; "rdf:range" and "rdf:domain" are not allowed on any properties, the maximum cardinality of a property is one, while we need it to be greater than one for the multilanguage support of TEO, etc.

These additional requirements from TEO as well as the special tags in the OWL representation of TEO require a customized Fedora objects' structure and additional processing from OWL to Fedora CMA.

This is the reason for providing a customized Share.Tec-aware Fedora representation of the ontology.

The TEO ontology consists of 3 types of objects – classes, properties and individuals. We have defined a content model for each type:

- teo-CM:Class – a content model for classes;
- teo-CM:Property – a content model for properties;
- teo-CM:Object – a content model for individuals (objects).

All these content models define that each ontology object in Fedora has the following datastreams:

- DC – Dublin Core metadata that describes the objects. Typically, the following DC fields can be used: identifier, title and description. DC fields are automatically indexed and can be used in Resource Index Search.
- RELS-EXT – defines the relations for the object. The relations in RELS-EXT are also automatically indexed and can be used for Resource Index Search.
- ONTOLOGY – defines the part of the TEO ontology (in OWL) that is represented by the corresponding Fedora object.

For the implementation of the TEO ontology in Fedora (we will call it TEO-Fedora for short) we will restrict to OWL Lite [9, 10] but with some extensions (properties can have cardinality greater than one, multilanguage support) and some further restrictions (there are no equivalent classes, no 'allValuesFrom' and 'someValuesFrom' property restrictions).

The structure of digital objects in Fedora for representation of OWL classes is shown on Fig. 2. All objects have content model teo-CM:Class which defines the datastreams in the objects.

Fig. 2. Digital object structure for OWL classes.

PID is the persistent identifier of the digital object in Fedora. For the OWL classes in TEO-Fedora the PID has a prefix teo-class.

ONTOLOGY, DC and RELS-EXT are datastreams with internally managed XML content. The ONTOLOGY datastream contains the OWL description of the class. The DC datastream contains the Dublin Core metadata for the object. The RELS-EXT datastream defines the relations of the object. It should contain relations hasModel and subClassOf.

The structure of digital objects in Fedora for representation of OWL properties is the same as for OWL classes (see Figure 2). All objects have content model teo-CM:Property which defines the datastreams in the objects.

The PID for the OWL properties in TEO-Fedora has a prefix teo-property. The ONTOLOGY datastream contains the OWL description of the property. The DC datastream contains the Dublin Core metadata for the object. The RELS-EXT datastream defines the relations of the object. It should contain relations hasModel, definesObjectProperty (for Object Properties), definesDataProperty (for Datatype properties), domain, range and subPropertyOf (optional).

The structure of digital objects in Fedora for representation of OWL individuals is the same as for OWL classes (see Figure 2). All objects have content model teo-CM:Object which defines the datastreams in the objects.

The PID for the OWL individuals in TEO-Fedora has a prefix teo-object. The ONTOLOGY datastream contains the OWL description of the property. The DC datastream contains the Dublin Core metadata for the object. The RELS-EXT datastream defines the relations of the object. It should contain relations hasModel, type and the properties of the individual.

The TEO ontology is represented by a large number of digital objects in Fedora. So the consistency and validity of the ontology is very critical.

The basic assumption is that if we merge all parts of the ontology from the ONTOLOGY datastream of all digital objects (classes, properties and

individuals), the resulting file should be a consistent valid ontology in OWL that can be opened and processed in Protégé [8].

Note that the ONTOLOGY datastreams for the properties and the individuals should not contain the ontology description explicitly as described in the above sections. Since all the needed information is stored also in the DC and RELS-EXT datastreams we will develop dissemination methods for automatic generation of the content of the ONTOLOGY datastream from DC and RELS-EXT through appropriate XSLT transformations. If we restrict the ontology not to use property restrictions in the definitions of the classes, the content of the ONTOLOGY datastream for the classes can also be generated automatically.

Since the digital objects for the TEO ontology will be created, searched, edited and deleted asynchronously, validation services (methods) will be developed that for each type of object (class, property or individual) will check:

- whether a class references existing classes and properties;
- whether a property references existing classes and properties;
- whether an individual has the required properties, references existing individuals and the values of the datatype properties are valid.

The structure of harvested objects with CMM is shown on Fig. 3. All objects have content model sharetec:CM_Digital_Objects which defines the datastreams in the objects.

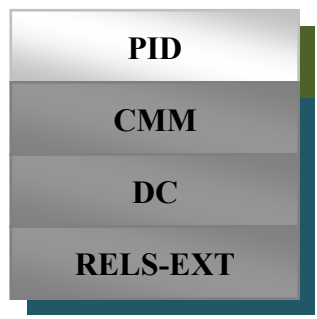

Fig. 3. Digital object structure for CMM objects.

The PID for the harvested objects with CMM has a prefix sharetec.

The datastream CMM contains the CMM metadata record of the harvested objects. The GSearch component of Fedora is configured to automatically index this datastream so faceted search on CMM metadata can be performed.

The datastream DC contains the Dublin Core metadata for the object. The values for the DC (dc:title, dc:description, etc.) will be extracted from the original CMM record via appropriate XSLT transformation before ingesting the object in Fedora.

The RELS-EXT datastream defines the relations of the object. It should contain the following relations:

- hasModel – defines the content model for the object (sharetec:CM_Digital_Objects).
- isMemberOf – defines the collection for the object. For each repository

we will have a separate collection that contains the harvested objects. The collections are represented also as Fedora objects.
- itemID – defines the original Id (OAI identifier) of the harvested object.

3 Importing the TEO to the central repository

The original Teacher Education Ontology (TEO) is defined in Protégé and is available as OWL [9, 10] file. More recent versions of TEO will also be available as OWL files. TEO is represented in Share.TEC repository and will be used in real-time for:

- Constructing relations when data are being stored in the repository.
- Facilitating hierarchal search and navigation.
- Retrieving language-dependent translations of ontology concepts.
- Answering specific non-trivial queries regarding CMM and TEO objects and relations.

The current OWL implementation of TEO cannot be directly imported into Fedora, because the OWL file format as exported from Protégé is not directly supported by Fedora. A simpler OWL representation is needed in order to be able to automatically import it to Fedora. This problem leads to the following decision: OWL being an XML [11] file could be transformed by an XSLT [12] and XPath [13] into a set of XML files that can be ingested directly into Fedora. In the rest of this section the Fedora representation of TEO which is derived from its OWL representation will be referred as TEOWL – TEO+OWL.

TEOWL is generated in a process of transforming OWL file through an XSLT script. The overall transformation process is shown in Figure 4.

The transfer of OWL into Fedora makes several assumptions about how data are represented in the OWL file. This section describes some requirements, which must be considered during the future modifications of TEO.

- Using <DEF>…</DEF> and <TERM>…</TERM> tags in translations of concepts and their descriptions are not used by the script. They are also not processed by Protégé as XML tags, but are encoded as plain text. If these tags are used, they will also appear in the final translations in Fedora digital objects. The easiest way to handle this is to remove all these tags from the OWL file (by opening the file in simple text editor, removing the tags, and saving it again as text file with OWL extension)
- The languages must be encoded with @xml:lang.
- Fedora cannot accept long PIDs, so the names in OWL must be as short as possible. Currently a few instances cannot be ingested in Fedora because their names are too long. They require manual processing.
- All entities which are introduced for test purposes must be removed from the OWL file.

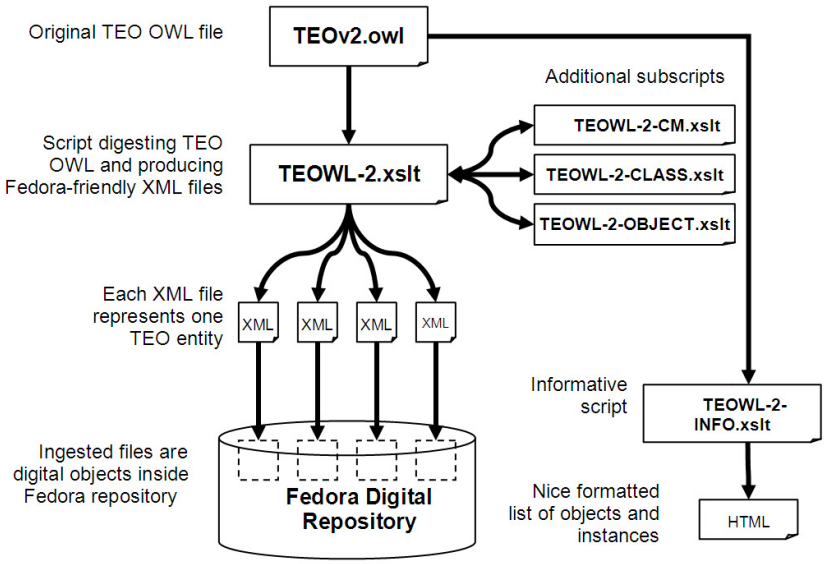

Fig. 4. Transferring TEO from OWL to Fedora.

4 Conclusion

In this paper we described the Share.TEC repository system prototype. The central repository cache contains metadata records of teacher education resources annotated following the CMM metadata format proposed in the Share.TEC project. For more precise and flexible description and classification of TE resources, the TEO ontology is used. For the representation of the teacher education resources both CMM and TEO are used, as well as the appropriate mappings between TEO and CMM are available.

Two search engines are employed in the searching service of the Share.TEC repository prototype. On the one hand, Lucene [14] is proposed for resolving the queries corresponding to CMM fields. On the other hand, the Resource Index of Fedora [15] implemented as a Mulgara RDF triple store is the search engine in charge of processing those queries that include links among teacher education resources and TEO elements such as knowledge areas.

The user interface in the searching service for the first Share.TEC system prototype includes basic search, advanced search and browsing.

Acknowledgments. This work is supported by EC project Share.TEC - SHAring Digital REsources in the Teaching Education Community, eContentplus programme (ECP 2007 EDU 427015); http://www.sharetecproject.eu/.

References

1. Share.TEC Project Deliverable D2.2 Common Metadata Model (CMM): version 1, http://www.share-tec.eu/content/1/c6/04/41/02/D2_2_Common_Metadata_Model.pdf
2. Repository Software Survey, March 2009, http://www.rsp.ac.uk/software/surveyresults

3. Technical Evaluation of Research Repositories, https://eduforge.org/docman/view.php/131/1062/Repository Evaluation Document.pdf
4. Lagoze C., Payette S., Shin E., Wilper C. Fedora: An Architecture for Complex Objects and their Relationships, Journal of Digital Libraries, Special Issue on Complex Objects, Springer, 2006, pp. 124-138.
5. Fedora Commons Repository Software, http://www.fedora.info/
6. Fedora Digital Object Model, Fedora Repository 3 Documentation, http://www.fedoracommons.org/confluence/display/FCR30/Fedora+Digital+Object+Model
7. Blekinge-Rasmussen, A. Enhanced Content Models for Fedora. Open Repositories, North America, 10-04-2009.
8. The Protégé; Ontology Editor and Knowledge Acquisition System, http://protege.stanford.edu/
9. OWL Web Ontology Language Guide, http://www.w3.org/TR/2004/REC-owl-guide-20040210/
10. OWL Web Ontology Language Reference, http://www.w3.org/TR/2004/REC-owl-ref-20040210/
11. XML, http://en.wikipedia.org/wiki/XML
12. XSL Transformations, http://en.wikipedia.org/wiki/XSL_Transformations
13. XML Path Language (XPath) 2.0, http://en.wikipedia.org/wiki/XSL_Transformations
14. Apache Foundation. Apache Lucene Website, http://lucene.apache.org/
15. Fedora Commons. Resource Index. http://fedoracommons.org/confluence/display/FCR30/Resource+Index
16. M. Feijen, W. Horstmann, P. Manghi, M. Robinson and R. Russell, DRIVER: Building the Network for Accessing Digital Repositories across Europe. Ariadne Issue 53 October 2007.
17. Open Archives Initiative, http://www.openarchives.org/
18. IEEE Standard for Learning Object Metadata 1484.12.1-2002. IEEE LTSC, http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf
19. S. Alvino, S. Bocconi, J. Earp, L. Sarti (2008) A Teacher Education Ontology for Sharing Digital Resources across Europe, in Proceedings of the 5th International TENCompetence Open Workshop "Stimulating Personal Development and Knowledge Sharing", ed. R. Koper, K. Stefanov and D. Dicheva, Sofia, Bulgaria, 30-31 October 2008, ISBN: 978-954-92146-5-9, pp. 26-29
20. Dublin Core Metadata Initiative, http://dublincore.org/