The Ensemble Project: Using Fedora to Support the Development of the Semantic Web for Education

Agustina Martinez, Patrick Carmichael and Louise Corti
University of Cambridge & University of Essex
The Ensemble Project

- Semantic Technologies for the Enhancement of Case Based Learning
- 3 Year, £1.5 Million ESRC/EPSRC Project: Research, Development and Implementation (2008-2011)
- “working with teachers and students in undergraduate and postgraduate courses to explore both the nature and role of the cases around which learning is focused, and the part that emerging semantic technologies can play in supporting this learning”
  - A big, happy interdisciplinary and multi-institutional extended family …
  - Website: http://www.ensemble.ac.uk

The Ensemble project (‘Ensemble: Semantic Technologies for the Enhancement of Case Based Learning’) is a three-year research and development project jointly funded by the UK’s Economic and Social Research Council and Engineering and Physical Science Research Council. It’s an interdisciplinary project that brings together educational researchers, computer scientists and teachers and students from various disciplines to explore how semantic technologies can support and enhance teaching in learning in areas of higher education where complexity and change makes case based learning the pedagogical approach of choice.
In the initial stage of the project we differentiate two parallel strands of work that are now converging together: a) the research in the different educational settings and b) the technology analysis and initial prototyping and pilot development. Before describing in detail the system configured to support the different developments I’d like to provide a brief overview of the main technologies being analyzed. The developments cover three main groups of technologies:

- **BACKEND** We are working with large and heterogeneous data sets so we need the support of tools to manage those datasets and provide a common interface that will allow us to expose our data in a wide variety of formats.

- **MIDDLEWARE** Aggregators, converters and tools for mapping between data in different formats and the definition of languages for querying and searching the data in a standardized way. For aggregating the data we can use triplestores which allow us to store the data in RDF format so it can be queried by endpoints or APIs using common languages like SPARQL.

- **FRONTEND** Once we have the data aggregated in a triplestore we can search across it and display the results using different visualization tools like the MIT Simile Toolkit.

The main aim of the demonstrators implemented is to show how semantic technologies can improve current applications. We’ve been experiencing with Fedora not just as a provider of resources and metadata but also as a data source in which convertible formats like plain text, comma-separated values or ‘native’ XML/RDF can be streamed directly to other applications. Now data from multiple sources: repositories, external providers can now be aggregated together into triplestores, queried using standard languages such as SPARQL and finally presented in different formats, user-centred, by using different visualizations toolkits like Exhibit.
We have analysed different digital repositories and its functionalities. We have decided to use Fedora repository because it introduces new semantic functionalities that allow us to work with large and heterogeneous datasets and expose them in a common interface that uses semantic technologies and can be accessed by different applications.

Furthermore, we have developed an extended search interface that adds more capabilities to the semantic-ready tools introduced in Fedora.

Before describing in detail the extensions developed I will explain the generic configurations of Fedora to demonstrate why in the project we needed to extend Fedora semantic tools with new functionalities.

By default, Fedora provides an interface to query the repository. Every object stored in the repository presents a Dublin Core Metadata record to provide a description about the object data contents. At this point we can access the data stored in the repository in two different ways: a) from external providers by exposing the descriptive metadata using the OAI-PMH or b) indexing the metadata in a relational database and the search is performed by using just the description information about the objects. This approach is quite limited because it just uses the metadata information and it is fairly dependant on how well described are the objects stored in the repository.
In this second configuration of Fedora we show the new semantic functionalities. Fedora introduces the CMA, which allows both to classify our data sets and to build collections of resources by defining relationships between objects. At this point we can add to an object a special datastream that expresses, in RDF statements, the relationships between the object and the other objects stored in the repository. Now we have a new search interface based on the Resource index module. The Fedora Resource Index module enable to express the relationships in a machine-readable way so that it can be indexed in a database or aggregator like a triplestore. In order to index this type of information the Resource Index module builds on RDF statements and a simple ontology that defines a set of classes and relationships between objects since RDF is too generic.

The RI search combines the normal searching, using just the metadata and the relation information by querying an internal triplestore instance using a set or query languages: SPARQL or iTQL.
Fedora RI module just indexes in the triplestore the metadata and the relationship information. What happens if we have datasets in RDF format so they are ready to be aggregated into semantic applications? This is the main aim of implementing an extension to the RI approach so that we can now store in the triplestore not only the metadata but also every RDF datastream stored within the objects. To illustrate this approach I will use the following example: a small collection stored in the repository. It comprises one root data object defining the collection and two data objects:

a) data object with a spreadsheet datastream and,
b) one data object containing the contents of the spreadsheet in RDF format.

The extension implements a middleware interface that performs the following process:

1) To obtain the datastreams that contain RDF inline datastreams we need the object identifiers and the datastream names. We obtain this information by querying the RI graph indexed in the Mulgara triplestore.

2) The results obtained from the triplestore are in Sparql-xml format so we need to parse them to construct a list of object-id, datastream-name pairs that will be used to compose REST like queries to the Fedora repository by using the API-A interface. The API-A interface presents a set of methods to perform operations across the repository.

3) The Fedora repository is queried to obtain the contents of the RDF datastreams.

4) Every datastream is now uploaded to a new Mulgara graph by using the Mulgara HTTP SPARQL endpoint.
Web application to query a small dataset using Semantic technologies and visualization tools

- **Plants distribution in the Mediterranean.** The application uses a small dataset of Plants distribution in the Mediterranean. It comprises two different sets: a) plants descriptions and b) plant observations. The data is obtained from the GBIF portal and converted to RDF prior to its store in Fedora.

- **Aggregates data...** One of the key settings of the application is that enables as to aggregate different data together in the triplestore enabling running queries across these data and filter information in a more accurate way.

- The SPARQL endpoint interface provides the following functionalities, on one hand provides a fully functional Sparql query interface to query the datasets stored in the Mulgara triplestore. This is more oriented to experienced users that would like to request complex queries or reason across the data and on the other the application presents a set of predefined queries showing how to query the triplestore and the kinds of questions they can ask based on very specific datasets.

- The Endpoint provides two different visualization on the results: a) HTML data presentation (plain dumping), different semantic formats such as RDF, N3 (triples) and b) faceted browsing using visualization tools like Exhibit toolkit from MIT or Google maps API.

- Combining the triplestore data with rules models and reasoners we can infer new information by exploring the data and applying complex queries.
Since we have a stable configuration of the Fedora repository and the Mulgara triplestore, we are preparing new archives to be stored as collections in the repository. Some of the most important projects are the following:

1) The archive contains documents related to important educational evaluations undertaken by members of the Cambridge Conference on Evaluation. The application aggregates together different data sources enabling to make connections between the data and that can replicate student and researchers questions, like for example: ‘show me researchers interested in data analysis and find publications where this is reported’.

2) The major part of the collection comprises life story interviews originally collected as part of the study The Edwardians: Family Life and Work Experience Before 1918. The collection comprises metadata annotations, interview files and transcriptions. The metadata can be stored in the triplestore and connected with contextual information like for example social and economic events of those years.

3) Web application to display information about earthquakes registered globally and their characteristics. The application brings together information about the earthquakes and global media news published about them (here we have a new example of live aggregation of data). The application uses a customized Simile Map extension to display the information and the data is obtained using a web service of recent seismic events provided by the US Geological Survey.
Summary

- **Fedora Digital Repository** provides a framework to store large and heterogeneous data
  - Not only access to the metadata descriptions but access to the actual data
- **Data structured and defined in semantic-ready format**
  - Triplestores like Mulgara enable to aggregate and reason across different data sources
- **Visualization and presentation tools**
  - Process semantic-ready data and present the information in different formats
Questions