LA-UR- 09-003760

Approved for public release;
distribution is unlimited.

Title:	Using Architectures for Semantic Interoperability to Create Journal Clubs for Emergency Response
Author(s):	James E. Powell Linn Marks Collins Mark L. B. Martinez
Intended for:	6th International Conference on Information Systems for Crisis Response and Management Special Session on Architectures to Support Interoperability May 10-13, 2009 Göteborg, Sweden

Form 836 (7/06)

Using Architectures for Semantic Interoperability to Create Journal Clubs for Emergency Response

James E. Powell, Linn Marks Collins, and Mark L.B. Martinez
Los Alamos National Laboratory
Los Alamos, New Mexico USA 87545-1362
{jepowell, linn, mlbm}@lanl.gov

ABSTRACT

In certain types of "slow burn" emergencies, careful accumulation and evaluation of information can offer a crucial advantage. The SARS outbreak in the first decade of the 21st century was such an event, and ad hoc journal clubs played a critical role in assisting scientific and technical responders in identifying and developing various strategies for halting what could have become a dangerous pandemic. This paper describes a process for leveraging emerging semantic web and digital library architectures and standards to (1) create a focused collection of bibliographic metadata, (2) extract semantic information, (3) convert it to the Resource Description Framework /Extensible Markup Language (RDF/XML), and (4) integrate it so that scientific and technical responders can share and explore critical information in the collections.

INTRODUCTION

In late 2002 and early 2003, World Health Organization officials were called upon to investigate reports of a mysterious, highly contagious, and often fatal respiratory ailment that first appeared in southeast Asia. During the outbreak, over 8400 people were infected and over 800 died. The mystery illness spread among residents of a Hong Kong housing complex, among guests attending a conference, among passengers on several domestic and international airline flights, and managed to travel thousands of miles from Asia to Canada. The disease that came to be known as SARS (Severe Acute Respiratory Syndrome) was perilously close to exploding into a worldwide pandemic.

One factor that prevented this was the information-centric response by the World Health Organization (WHO) and by the health officials of many countries affected by the outbreak. Indeed, one of the most intriguing aspects of the outbreak was the way in which researchers tackled the problem of sharing information. In order to keep pace with relevant technical reading, WHO researchers resorted to forming an ad hoc "journal club group" (World Health Organization, 2006). The "journal club" has its roots in the editorial practices of peer-reviewed journals, and involves practitioners reviewing, discussing and debating various aspects of a set of published journal articles. In the SARS outbreak, the journal club played an important role in identifying and preventing this newly emergent zoonotic virus from becoming a worldwide pandemic.

In December 2008, the United States released a new bipartisan National Intelligence Estimate (NIE) that asserts a "large-scale bioterrorism attack is likely by 2013" (http://www.dni.gov/press_releases/20070717_release.pdf). Some observers have noted that the SARS outbreak may have been a dry-run for future outbreaks, whether natural or the results of terrorist activities. The information required by responders to outbreaks of highly infectious diseases with a high rate of mortality are diverse, deep, and critical. For example, responders may need to know under what circumstances quarantine might be appropriate and what practices make for effective quarantines given the characteristics of an infectious agent, what diseases exhibit similar symptoms, what treatments are available or appropriate, and many other facts and opinions. In such instances, content from digital libraries—such as bibliographic data describing the contents of peer-reviewed scientific and technical journal articles—emerges as particularly advantageous to scientific and technical responders, especially if it can be combined with the ability to interact, share, discuss, and recommend content in near real-time within a collaboration environment.

This paper discusses our approach in support of the mitigation of such life-threatening events and is based upon digital library and semantic web technologies. It involves rapid harvesting, normalizing, augmenting, and exposing of custom collections of metadata from heterogeneous, distributed digital libraries, with the ultimate goal of integrating these collections with a collaboration framework.

INFORMATION COLLECTION, EXTRACTION, AND INTEGRATION

Custom collections may be based upon simple or complex queries against digital library search services, or consist of a large set of data harvested from one or more repositories, or some combination thereof. These

highly focused, topic-specific collections may then be explored using various awareness tools within the context of a collaboration space (Collins, Powell, Dunford, Mane, Martinez, 2008), and shared with other users regardless of their physical location. The quality of content harvested from digital libraries tends to be high, since the material has been vetted by subject librarians prior to inclusion in their collection.

Web 2.0 technologies have enabled sophisticated methods of content sharing and annotation by users. As part of the E-SOS project (Collins et al., 2008), we have explored some techniques for assembling, normalizing, augmenting and exposing custom collections for emergency responders. Furthermore, we provide several Just-in-Time-Information-Retrieval (JITIR) awareness tools that allow users to explore these collections in various ways, since the number of results of interest and relevant to the topic at hand can number in the hundreds or thousands. Users of such collections may be motivated to explore the information more deeply than the typical user of a search engine, and the recall and precision of queries into a topic-specific data set will be better than similar queries against a search service exposing content on a wide range of topics. By essentially prefiltering the content, and offering custom ways of exploring it through context-aware writing tools, interacting with larger results sets (hundreds or thousands of items rather than tens of items) becomes a more manageable and realistic task.

We have developed tools that harvest data from OAI-PMH (Open Archives Initiative-Protocol for Metadata Harvesting) (http://www.openarchives.org/OAI/openarchivesprotocol.html) repositories, or from our own Apache SOLR-based internal search service (http://lucene.apache.org/solr/), which indexes and exposes 94 million bibliographic records describing academic research papers on science and technology topics spanning over a century. Moreover, our tools can convert Dublin Core metadata (http://dublincore.org/), or MARC-XML records (http://www.loc.gov/standards/marcxml/), into RDF triples that describe various key aspects of the publications. Mapping is facilitated by borrowing elements from existing XML specifications and using them as properties in triples, as is a common practice in the semantic web. For example, bibliographic metadata may be expressed using Dublin Core elements as properties, such as this triple describing the title of a publication in a collection of about 1600 results harvested from our local collection in response to the query ("humanitarian assistance" or "disaster relief"):

```
<uuid:a65564c6-1869-4f89-a3b4-254f9b526101> <dc:title> "Negotiation issues in
multinational Humanitarian Assistance/Disaster Relief".
```

When possible we also used existing defined ontologies, such as Friend of a Friend (FOAF) (http://www.foaf-project.org/) for describing people and the relationships between them, and the geonames ontology for describing characteristics of geographic locations, here represented as an RDF/XML code snippet:

```
<geo:name>Zaire</geo:name>

<geo:lat>-7.0</geo:lat>
<geo:long>13.8333333</geo:long>
```

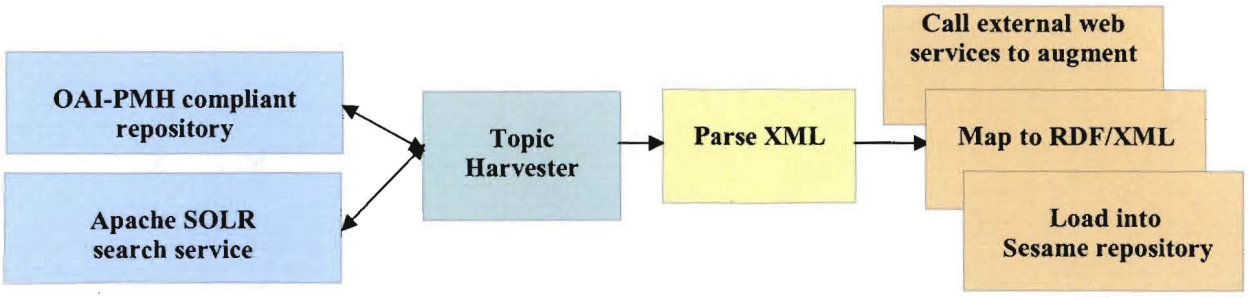

Figure 1: Converting metadata to RDF/XML

As we map this bibliographic data into triples, we also perform some augmentation of the data during the mapping process. For example, our tool looks for place names which may occur in certain MARC bibliographic record fields, and if a place name is found and extracted, the mapping tool will request georeference coordinates from a service such as http://ws.geonames.org/search?name=<someplacename> and incorporate georeference triples for the record into the triplestore. Implicit social networks represented in the record, in the form of authorship, are mapped into explicit foaf:knows triples, which enables exploration of the social networks of authors, as this RDF/XML illustrates:

```
<dc:creator>
  <foaf:Person rdf:about="uuid:0989d3f7-5791-4cde-93ba-dfcdf6675123">
    <foaf:name>BUISSON, L</foaf:name>
  </foaf:Person>
</dc:creator>
<dc:creator>
  <foaf:Person rdf:about="uuid:ddc93e8a-8f43-4f19-aa4e-9f0a67f85300">
    <foaf:name>Sullivan, JD</foaf:name>
    <foaf:knows rdf:resource="uuid:0989d3f7-5791-4cde-93ba-dfcdf6675123"/>
  </foaf:Person>
</dc:creator>
```

All of the harvested data of interest (e.g. title, author(s), abstract, publication information), plus the additional data provided by calls to external web services (e.g. georeference data) is mapped to RDF/XML, using the aforementioned XML schemas and ontologies to express properties and relationships. The data is then loaded into a Sesame triplestore (Figure 2) based on the OpenRDF open source application suite (http://www.openrdf.org/doc/sesame/users/index.html).

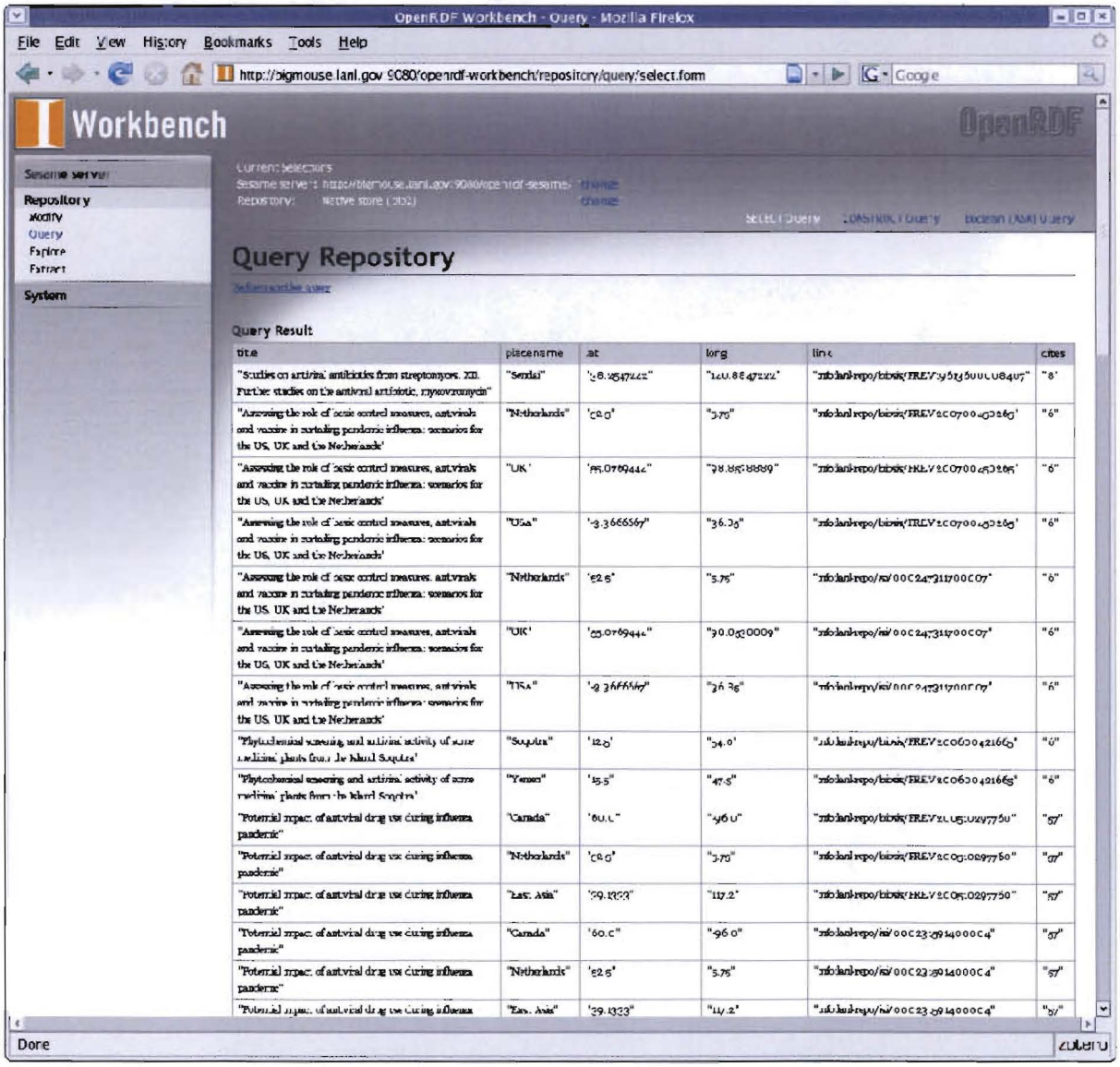

Figure 2. Open RDF Sesame Workbench with results from collection

There are many benefits to building a semantic web repository of technical articles. Unlike a relational database, a triplestore may be extended to incorporate new facts at any time by simply adding new triples. Since there is no database schema to contend with, older entries will simply lack the new properties. Ontologies may be applied in numerous ways to add value and usability to the data, and to support inferencing across collections of triples which may reveal new knowledge or non-obvious relationships within the data. RDF linking makes it possible to relate locally stored data to other semantic web repositories, such as DBPedia (http://wiki.dbpedia.org/About), which is a semantic web version of Wikipedia.

While a custom collection may start out as a set of triples representing results from a digital library query, similar data from a collaboration or content management tool may also be incorporated into a repository. In fact, Drupal, which is at the core of E-SOS, is capable of exporting metadata about content, authorship, and annotation performed within the framework, as RDF triples. Real-time sensor data might be incorporated as well, using Open Geospatial Consortium (http://www.opengeospatial.org/) standards such as SensorML (http://www.opengeospatial.org/standards/sensorml) as a source of properties about sensors. And finally, the inferencing capabilities enabled by semantic web technologies allow applications and users to discover new facts within the data.

With normalized data, it is easier to build search and visualization tools to explore the triples. Since the data has been augmented with georeference coordinates when possible, bibliographic metadata represented by the triples can be overlayed onto a map to rapidly expose relationships among data not otherwise apparent from a textual view. Since relationships between co-authors are captured during the mapping from metadata to triples, SPARQL (http://www.w3.org/TR/rdf-sparql-query/) queries against a repository representing a custom collection can return FOAF triples representing those co-authorship networks.

With this in mind we developed a Social Awareness Tool consisting of a pair of REST-based web services (Fielding, 2000) used in conjunction with a visualization applet provided with the GUESS (Graph Exploration System) toolkit (http://graphexploration.cond.org/). (Figure 3) The first web service uses the Open RDF libraries to connect to and query a Sesame repository. It uses a SPARQL query template to formulate a query against all triples which contain dc:title as a property, as this example illustrates:

```
SELECT ?title ?name ?selfuuid ?knowsuuid
WHERE {
  ?y <http://purl.org/dc/elements/1.1/#title> ?title.
  ?y <http://purl.org/dc/elements/1.1/#creator> ?selfuuid.
  ?selfuuid <http://xmlns.com/foaf/0.1/#name> ?name.
OPTIONAL
  { ?selfuuid <http://xmlns.com/foaf/0.1/#knows> ?knowsuuid. }
  FILTER regex(?title, "influenza", "i") }
```

The results of this query are returned as GraphML (Graph Markup Language) (http://graphml.graphdrawing.org/) or GDF (GUESS Data Format), which represents the edges and nodes of the authorship networks. This output serves as the basis of an authorship network visualization tool.

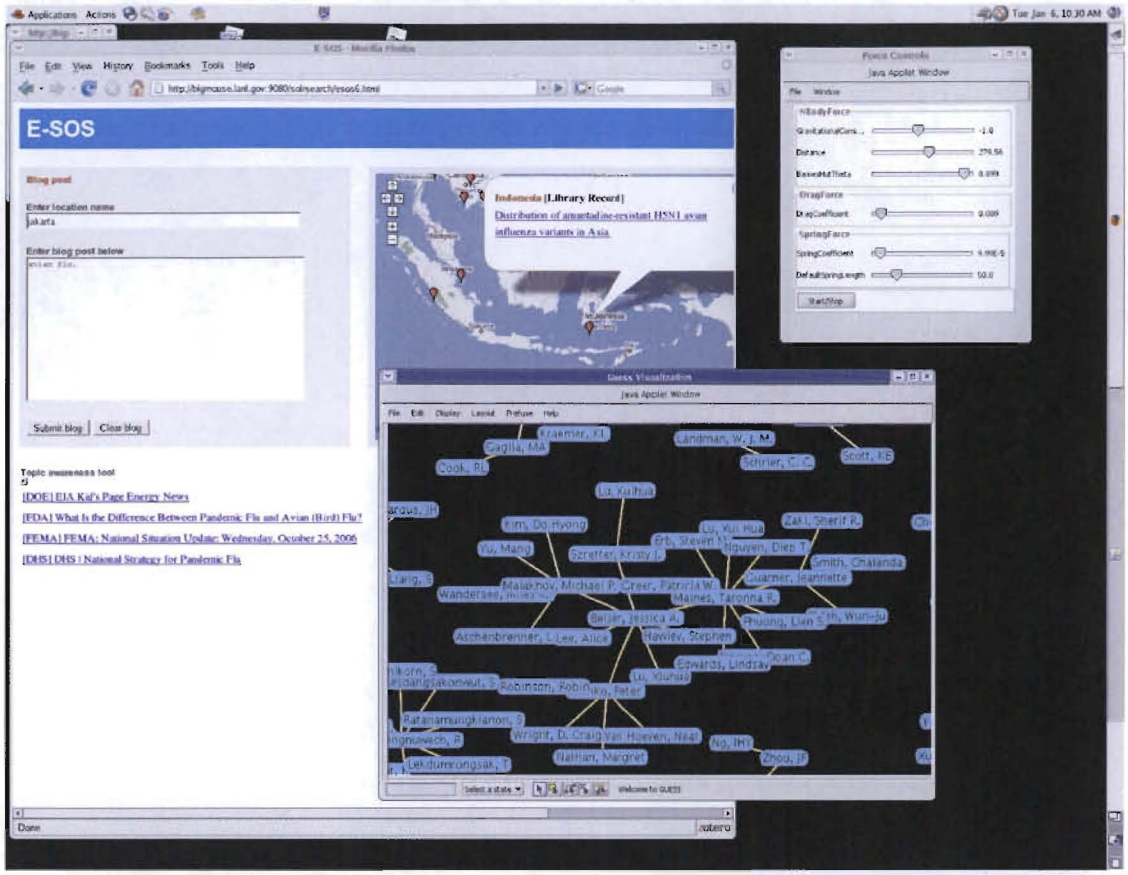

Figure 3: Tools for exploring and finding journal club content and members

We have built nearly a dozen custom collections using the harvesting, augmentation, and mapping technique described above. Four of the team leader's alerts (stored queries) were converted into custom collections , which ranged from 1,600 to over 6,000 results. We also harvested the entire contents of several small OAI-compliant repositories, as well as a subset of Pubmed (http://www.ncbi.nlm.nih.gov/pubmed/) data during testing. But perhaps the most interesting test of the system involved a collection which was based upon an extremely complex query formulated by a group of bioscientists. This particular query included 32 phrases linked by a boolean OR operator, and yielded over 57,000 results, which were harvested, augmented, and mapped to RDF/XML. Such a large number of results would normally be unwieldy and undesirable, but with our conversion and augmentation process, the suite of awareness tools offer ways to explore collections of intermediate sizes that are difficult to explore using the common (as exemplified by Google) results list paradigm provided by most Internet search engines.

CONCLUSION

What can we learn from the global response to the SARS epidemic and how can we better leverage knowledge, information, and technology to mitigate similar future events? We believe that high quality custom collections of content from digital libraries, and the ability to explore it, can be critically important to decision makers and first responders dealing with crises. These collections become even more valuable when offered with tools enabled by semantic technologies. These tools can facilitate visual and task-based exploration of the collection, and provide Web 2.0 collaboration capabilities such as sharing, commenting, rating, and tagging, which are typical of online journal clubs. It is also important that it be possible to generate and extend those collections quickly and in a dynamic, and automated fashion. Although four months elapsed between the first documented case and the preliminary identification of a new coronavirus as the possible cause of the deadly respiratory ailment (World Health Organization, 2006), health officials were urgently grappling with strategies to prevent new infections and find effective treatments for patients. The SARS outbreak is representative of how a bioterrorist incident may unfold. Such an incident may start slowly, and during the earliest stages of the outbreak, officials and specialists will be actively and urgently seeking high quality information to address concerns ranging from quarantine, limiting travel, informing the public, and identifying and treating the culprit.

Access to topic-specific information can be extremely valuable in a number of ways: it can be used in

informational campaigns to help the public understand the nature an outbreak and how they can protect themselves, it can help public health officials develop strategies for protecting the public and health care providers, and it can help researchers as they rush to identify the cause of an outbreak, develop treatments, and explore possible ways to mitigate it. Quality and topical information can be invaluable to medical staff treating infected patients, and researchers evaluating antiviral medications which may reduce the severity of the disease, or who may be exploring possible vaccines, which throughout human history are among the most effective means of stopping and even eradicating an infectious agent. In the 21st century, information systems will play a crucial role in preventing the next catastrophic disease pandemic.

ACKNOWLEDGMENTS

The authors wish to thank Kim Thomas and Miriam Blake for their support. (Ask Helen Cui about using her name.)

REFERENCES

1. Apache SOLR: http://lucene.apache.org/solr/
2. Collins, L.M., Powell, J.E., Jr., Dunford, C.E., Mane, K.K., and Martinez, M.L.B. (2008) Emergency Information Synthesis and Awareness Using E-SOS. *Proceedings of the 5th International ISCRAM Conference*, Washington, DC, USA.
3. Dayton, A. I. (2006) "Beyond Open Access: Open Discourse, the next great equalizer". *Retrovirology*, 3, 55. doi:10.1186/1742-4690-3-55. http://www.retrovirology.com/content/3/1/55
4. DBpedia: http://wiki.dbpedia.org/About
5. Dublin Core: http://dublincore.org/
6. Friend of a Friend (FOAF): http://www.foaf-project.org/
7. Graph Exploration System (GUESS): http://graphexploration.cond.org/
8. GraphM: http://graphml.graphdrawing.org/
9. GUESS Data Format (GDF). GUESS manual: http://graphexploration.cond.org/manual.html
10. National Intelligence Estimate. (2008) http://www.dni.gov/press_releases/20070717_release.pdf
11. MARC-XML: http://www.loc.gov/standards/marcxml/
12. Open Geospatial Consortium: http://www.opengeospatial.org/
13. PubMed: http://www.ncbi.nlm.nih.gov/pubmed/
14. Fielding, R.T. (2000) Architectural Styles and the Design of Network-based Software Architectures. Doctoral dissertation, University of California, Irvine. http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm
15. SensorML: http://www.opengeospatial.org/standards/sensorml
16. Sesame User Guide: http://www.openrdf.org/doc/sesame/users/index.html
17. SPARQL: http://www.w3.org/TR/rdf-sparql-query/
18. Twidale, M. B., Gruzd, A. A., and Nichols, D. M. (2008) Writing in the library: Exploring tighter integration of digital library use with the writing process. *Inf. Process. Manage.* 44, 2, 558-580. http://dx.doi.org/10.1016/j.ipm.2007.05.010
19. Open Archives Initiative-Protocol for Metadata Harvesting: http://www.openarchives.org/OAI/openarchivesprotocol.html
20. OpenRDF: http://www.openrdf.org/
21. World Health Organization. (2006) SARS: How a global epidemic was stopped. World Health Organization.