SIMILE Position Paper

HP Team: Mick Bass, Mark H. Butler, John Erickson, Andy Seaborne, Paul Shabajee, Kevin Smathers, Robert Tansley

This position paper outlines the wider agenda of the SIMILE project as seen from the HP perspective, and relates this to the agenda for the current demo activity.

Agenda for SIMILE

Demonstrate the utility of the Semantic Web tool stack

SIMILE aims to demonstrate it is possible to represent any library metadata in RDF. RDF has a number of features that although not novel in isolation are novel in combination.

First it is semi structured, which is attractive because the real world does not always fit neatly into representations like relational databases. In fact, the library community is still a large user of hierarchical databases, which means semi structured representations are likely to be advantageous in the SIMILE context.

Second RDF is based on directed graphs of URIs, and URIs support namespacing which allows us to distinguish between elements that have the same short name but different meaning. It also means when different fragments of RDF adopt the same URI to refer to the same resource, when we merge those fragments we automatically merge all the instances of the same resource.

Third it can be serialized and read by systems that do not have explicit knowledge of the schema used by the instance data. This has the advantage that it reduces the investment necessary to read in new instance data.

Fourth it explicitly differentiates between classes and properties. This makes it easier for people to understand the data model, as formats such as XML do not make this distinction so they require people to invest extra effort to understand the data model with respect to this.

Fifth unlike XML, where many ways of expressing type have evolved, for example Schema, DTD and using attributes or nodes as type indicators, RDF has a single mechanism for type expression standardized in the RDFS recommendation.

Sixth when RDF is used in association with a schema or ontology language, additional information can be inferred due to explicit relations in the schema e.g. class and property inheritance. This may have the advantage that such languages will simplify the reuse of data, for example when mapping between data using vocabularies.

Finally using RDF as a common format also means we can leverage a common set of tools to interact with that metadata. Many of the tools that build on RDF are declarative, such as RDF Schema or OWL, so a second aim is to demonstrate that

these tools are applicable to the library problem domain and provide an advantage over conventional imperative approaches.

Demonstrate the Semantic Web can help with metadata complexity

One problem that features in SIMILE is we require ways of dealing with the complexity associated with metadata. For example how do we deal with metadata during its lifecycle? When should we track changes and when should we disregard them? How do we handle the schema lifecycle, and how do we deal with metadata when schemas change? How do we support data provenance and integrity? How do we help the user to discover information by modeling the inherent relations? Here the hope is that the Semantic Web tools, as they focus on managing metadata, can provide some identifiable benefit above and beyond existing technologies.

We would like a metric that represents the difficulty or alternatively return on investment that customers currently have due to schema and metadata complexity. At present the main way of solving this complexity is relational databases and related solutions such as data warehouses, so we need to be demonstrate how the Semantic Web can be of benefit here, both quantitatively and qualitatively. Qualitative benefits are also important because failure of new information architectures, standards etc are often caused because of issues related to people, business practices, etc, so solutions that get these issues right may be more compelling than solutions that have only considered quantitative benefits.

Demonstrate the reality of the Semantic Web

HP would like SIMILE to demonstrate a running system, providing practical benefit, based on the Semantic Web architecture, which is fit for the purpose defined in the problem statement. This should be a publicly visible and accessible Web resource that acts as a demonstration of the Semantic Web. This is the concrete goal we share with W3C and MIT libraries.

We observe that one of the barriers to the adoption of the Semantic Web is availability of data in RDF. In the Semantic Web vision, this data is made available using a publishing model for data and vocabularies that adopts a federated rather than centrally controlled approach. Therefore to be a true Semantic Web application, SIMILE will not only be internal infrastructure, i.e. capture and integrate metadata for internal use within a repository, but will also expose the gathered metadata in RDF and associated ontologies to the wider community.

This interest in federation and distribution is shared by the Genesis project group in HP, which is investigating barriers to commercial adoption of RDF. Genesis is particularly concerned with examining the hypothesis that aspects of distribution, security, higher level abstractions and efficiency will be required for the development of real applications of the Semantic Web.

Demonstrate integration of collections

In the SIMILE demonstrator, we have decided to concentrate on integrating collections, and there are several reasons why this is of interest to SIMILE users. We are trying to address the problem encountered in the library community where the library acts as a custodian for resources provided by different communities, and would like to provide a common way of browsing and searching these collections.

Although collections are aimed at specific audiences, they may have resources that are of interest to others. This also enables researchers to move from one axis of interest in a specific collection to another, e.g. from a query about an artist to examples of their work, on to works of a similar type, genre, or period by other individuals. Here integrating collections means the effective corpus from which they draw is much larger and potentially much richer and diverse in data and metadata. Other possible reasons include getting greater value, economic and intellectual, from collections, reducing the effort and overhead on users to locate resources and providing access to surrogates. One of the potential advantages of using Semantic Web tools here is that RDF based schema and ontology languages support mapping between different vocabularies.

Having structured metadata associated with resources adds value by allowing searches to be more specific and less ambiguous, for example allowing us to retrieve information about Madonna the singer rather than Madonna qua Saint Mary, the mother of Jesus of Nazareth. It also simplifies exploring collections by browsing as it can support the hierarchical partitioning of large sets of resources using techniques such as faceted browsing. Whereas it is difficult to construct search queries that overlap multiple collections, faceted browsing allows the exploration along multiple different axes so it should be possible to identify which axes feature prominently in overlap between collections and hence explore these overlaps. There are also advantages in true faceted search e.g. advanced searches that allow a text based query (over the controlled vocabularies, generally combines with a thesaurus) within one facet, as demonstrated by the Fishbase interface [1].

Cooperation between SIMILE and DSpace

Since DSpace has attracted wide interest and adoption within the library community, it has the potential to provide a deployment channel for the SIMILE work. For this to be achieved, it must be possible to deploy the tools and components developed for SIMILE within DSpace, without unduly disrupting existing DSpace services. The components should also be optional so that the added value that the SIMILE components bring can be clearly seen by the community. An additional desirable goal is for the SIMILE components and tools to be re-usable outside of a DSpace context.

Agenda for the demo

Demonstrate a core tool stack

One of the outcomes of the demonstrator should be a prototype core tool stack that demonstrates a possible architecture for a future set of SIMILE components that is deployable on DSpace. Prototyping this stack for the demo should help us understand the functionality of different tools, why they are important and useful, and how they work together. However there are some caveats here: not all of the tools will be in a sufficient state to be reused beyond the demo, although the demo will identify where this is the case. Therefore we do not expect the demo to result in a nicely packaged toolkit but it should start to give us an indication on what would be the requirements to overlap this on DSpace.

Make data corpuses available

As already noted, the Semantic Web depends not just on the existence of RDF data, but also on the public availability of that data. Therefore when we work on data for the demo, we not only need to ensure it is available as RDF, but also that is representative of the SIMILE use case, and ideally in the future we want publicly available data.

One issue here is that data is unlikely to be born RDF native and currently XML is the format widely adopted by industry, so it is necessary for SIMILE to devise ways of taking existing collections written in XML and converting them to RDF. However XML is only one possible format: an important outcome of SIMILE is some guidance about how we turn legacy data into RDF.

SIMILE has already made good headway at tackling this problem for specific datasets, but it is not clear how to explain this in the demo. As already noted one of the advantages of RDF is it makes the data model explicit, so when performing the transformation one crucial step is to understand what was implicit in the author's mind when the original metadata or metadata schema was created. Therefore the translations themselves may not be particularly complicated, it is the process that arrives at that transform that is complicated and that is hard to demonstrate.

The demonstrator also needs to take clear the advantage of representing data in RDF, as described in the first section.

User interaction

We wish to show that once in RDF, the metadata can be presented to and navigated by the user in an intuitive way, and we anticipate the Haystack will play an important role here. As noted, browsing may illustrate the concept of mapping between multiple collections in a more intuitive way than searching. However browsing requires a richer interface than searching, which may just use a single text box with complexity being built into a structured query. Therefore we anticipate the demonstrator will use a rich browse interface, based on Haystack, but ideally presented as web client. This will serve two purposes: first it will make the demonstrator present well, and second it means we are demonstrating a system that can clearly be used by end users.

Other desirables

So to conclude, we have described the relationship between the Semantic Web, technology developed by SIMILE members that embodies this vision, and the SIMILE project. We have discussed some of the benefits from these technologies. We have also discussed how this relates to the problem of integrating collections. We have then discussed how these issues relate to specifics we would like to see in the demo.

Finally we note that the demo does not need to solve all the problems related to SIMILE, only to give a glimpse of how these technologies can be applied to a compelling use case.

---

[1] Fishbase interface, http://www.fishbase.org/search.html