UNIVERSITY OF Southampton

Interoperate With Whom?

Formality, Archaeology and the Semantic Web

Raleigh, NC. April 26-27

WebScience10

Leif Isaksen
Electronics and Computer Science
University of Southampton
leif.isaksen@soton.ac.uk

Kirk Martinez
Electronics and Computer Science
University of Southampton
km@ecs.soton.ac.uk

Nicholas Gibbins
Electronics and Computer Science
University of Southampton
nmg@ecs.soton.ac.uk

Graeme Earl
Archaeology
University of Southampton
graeme.earl@soton.ac.uk

Simon Keay
Archaeology
University of Southampton
sjk1@soton.ac.uk

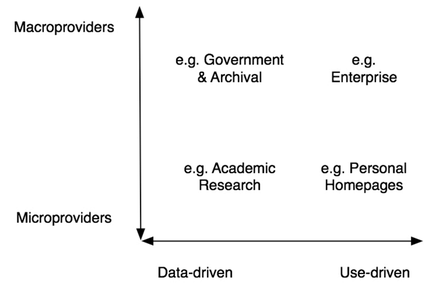
Abstract

'Interoperability' is often cited as a desirable end-goal for information systems, but the highly abstract nature of this apparent benefit sits uneasily with the task-oriented realities of data-management. The approach most frequently advocated is to increase the formality of the system, which facilitates system-integration yet also raises additional barriers to entry that reduce the potential pool of systems to interoperate with. The Semantic Web initiative in particular has faced accusations that difficulties associated with its adoption can outweigh the perceived benefits of data-sharing. This issue will be discussed in reference to current doctoral research being undertaken in Humanities data integration. It will argue that technologies that either heavily front-load or defer dealing with semantic complexity are unlikely to be viable across the producer spectrum. Recourse to altruistic arguments suggests a tacit acceptance that application of such technologies may not be in the immediate interest of the curator. An approach which offers multiple 'pay-off points' is inherently more attractive to potential adopters. In particular, we focus on means by which data-driven microproviders - owners of the small but important datasets that tend to form the 'long tail' of academic data in the Humanities - can participate in semantics-driven datasharing.

We propose that (at least) five escalating levels of semantic formalization can be identified, each with differing requirements and benefits for the implementer: i. Literal Standardization, ii. Instance URI generation, iii. Canonical URI mapping, iv. RDF generation, and v. Database-schema-to-Ontology mapping. We note that Linked Data - hitherto seen as the simplest semantic approach - is relatively advanced in this scheme. We argue that data providers should be encouraged to migrate towards full semantic formalization only as their requirements dictate, rather than all at once. Such an approach acts as both a short and long-term investment in semantic approaches, in turn encouraging increased community engagement.

We also propose that for such processes to be accessible to data-curators with low technical literacy, assistive software must be created to facilitate these steps. We have been developing a prototype package targeted specifically at archaeologists that enables them to produce valid, globally-integrated RDF from unnormalized excavation data with minimal technical knowledge. This takes the form of a Wizard that inspects legacy relational data, and provides predictive mapping and association which users can confirm or amend. A secondary program uses the resulting output in conjunction with the original data in order to produce valid RDF/XML as desired. Using this software we have demonstrated that archaeological excavation data encoded in a variety of formats, languages and schemas can be successfully integrated by its curators.

1. Different Communities



Data providers - that is, anyone who publishes data to the web - can be divided along (at least) two axes, providing a useful heuristic for understanding requirements. These divisions are not absolute but important for understanding why specific formalizing strategies may be inappropriate in some contexts.

Data-driven vs. Use-driven *Data-driven* providers emphasize the value of the dataset itself. This may be because it forms a holistic conceptual entity which would be invalidated by a corruption of its constituents, or because constituent elements may be deemed more important (or sensitive) than their combined value as a resource. For such providers, fidelity to their source is primary. *Use-driven* providers treat datasets as a means to an end - for mining, analysis or attracting customers. Such resources are generated principally for their utility in achieving specific tasks and frequently predicate richness and volume.

Microproviders vs. Macroproviders *Microproviders* generally have a single-purpose dataset that they are willing to contribute, often personal details or research. Structure will derive from the needs of a single or small number of tasks. *Macroproviders* have heterogeneous data, frequently of diverse provenance. They are more likely to benefit directly from approaches that facilitate integration.

2. Choosing Approaches to Data Publishing

Every provider must make a cost-benefit decision based on the realities of their situation - the effort one is willing or able to expend (limited by resources) vs. the potential computational power required to achieve their goals (sufficiency). Effort beyond one's means is impossible and redundant computational power undesirable if achieved at additional cost. Ideally, a provider will expend the least level of effort required to reach their sufficiency point. Different formalization processes have different utility curves in order to suit different user needs. Deferring formalization processes have a high power to effort ratio early on which then rapidly degrades. Front-loading formalization processes require high initial investment in order to improve gain later. (see Figure 1).

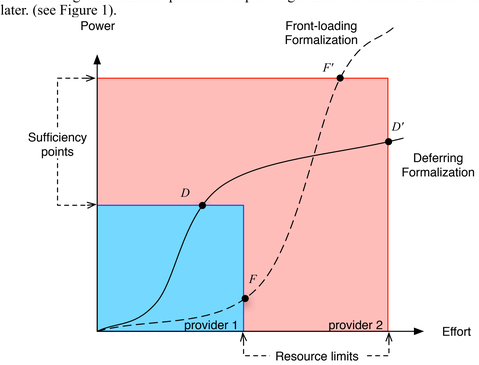


Figure 1: *Formalization benefit variance across users. Provider 1 has few resources and low requirements. A Front-loading formalization process will not meet their sufficiency criteria before their resources are exhausted (F) whereas Deferring formalization processes will meet their needs at low cost (D). Provider 2 has both greater needs and resources. Deferring processes becomes too complex before sufficiency is reached (D') whereas a Front-loading process is both achievable and has longer-term potential (F').*

In a single-use context individual providers can find a formalism which is 'just right' for them. In an open-use context the challenge is to accommodate unexpected outsiders as well. As innate complexity itself cannot be reduced, the only way to do this is to provide a wave-type formalization process, thereby providing multiple pay-off points (see Figure 2). Note that the immediate benefit to each provider is likely to be less than those they would achieve with a custom solution but this may be offset by the network effects generated by a standardized approach.

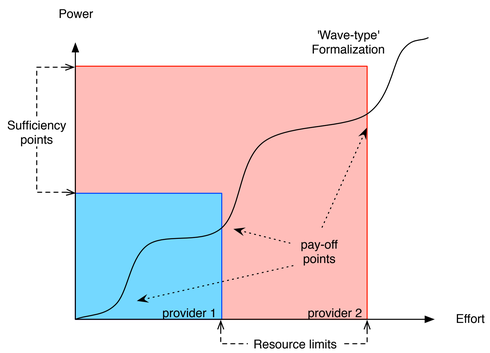
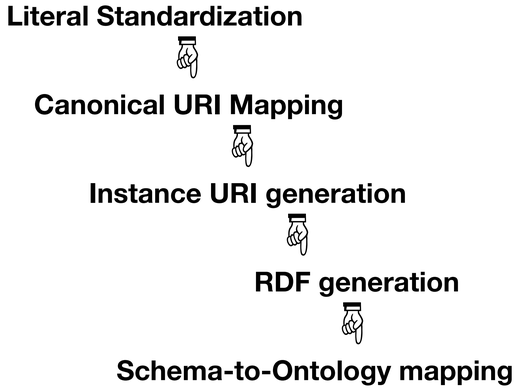


Figure 2: *A wave-type formalization process provides multiple pay-off points in order to maximize adoption (and thus communal benefits) but at the expense of large task-oriented gains to individual providers.*

3. Levels of Semantic Formalization

The well-known Semantic Web 'Layercake' provides an obvious starting point when considering a wave-type formalization process. As a stack of dependent technologies, it makes clear that some semantic technologies cannot be brought into play before others. Nevertheless, it does not outline the benefits or costs of adopting them to any given point, making it hard for potential adopters to evaluate what subset is appropriate for them in achieving their immediate objectives. We identify the following five levels at which a clear user benefit can be gained, each based upon the previous level of formal complexity. These tasks provide a helpful roadmap for resource-poor microproviders providing concrete benefits at each stage that can be evaluated against the cost of implementing them.



Literal Standardization
At the most basic level, users must transform unconstrained 'free' data into a locally-defined controlled vocabulary. This is the level at which the vast majority of data formalization projects start and end. *Benefits*: A higher degree of normalization, permitting stronger internal querying capabilities. Limited federated searching across separate databases is also possible. *Costs*: A number of complex tasks can be involved in this process including the decision of which data to standardize, the level of semantic granularity, the selection of terms, and even character encoding and date formatting.

Canonical URI Mapping
Association with other data requires reference to common concepts (whether abstract concepts or singleton instances) denoted by global identifiers provided elsewhere. *Benefits*: Users can map and compare to related data from external sources. *Costs*: Discovering relevant URIs and choosing between them requires infrastructural vocabulary services provided by the domain community as well as legal or *de facto* agreements on their usage and services for discovery.

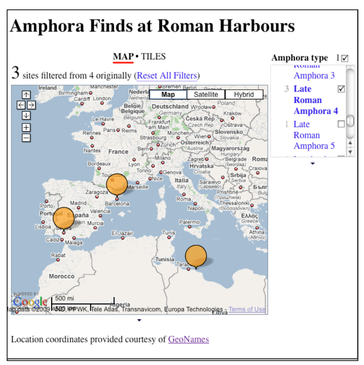
Instance URI generation
The first step to interoperability is the production of global identifiers for concepts unique to the provider. *Benefits*: Individual data instances can be made immediately accessible by web access and uniquely referred to by third parties. Search and discovery by outside parties is greatly enhanced. *Costs*: As the sole benefits are global access and differentiation, control of a web domain and access to web-hosting facilities are additionally required to realize them.

RDF generation
The use of RDF to associate URIs is the level at which many feel it can first be considered Linked Data. *Benefits*: Machine readability starts at this level, although purely in terms of aggregation and resource discovery. *Costs*: Introduction of new - and usually unfamiliar - storage mechanisms. Creation and/or discovery of predicate URIs. Basic grasp of ontological modelling required and tools or scripting needed for RDF generation.

DB-Schema-to-Ontology mapping
Inferencing power can be increased to an arbitrary degree by adoption and development of ontologies. *Benefits*: Sophisticated discovery and querying over heterogeneous and distributed data. *Costs*: Requires a solid understanding of implicit local semantics, external ontology semantics, and mapping vocabularies (e.g. OWL) as a minimum.

4. Case Study - Archaeology

The Roman Port Networks Project [1], is identifying patterns in maritime communications in the Mediterranean during the Roman and Late Antique periods (c. 200 BC - AD 600).



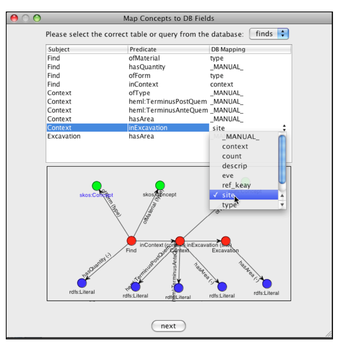
Large quantities of legacy excavation data exist (both published and unpublished) in the form of databases and spreadsheets but are held by separate institutions in different countries. Whilst the datasets all pertain to the same domain, they frequently employ mixed taxonomies and have different schema. Term normalization is rare, uncertainty is frequent and variant spellings are common. Different methodologies also give rise to alternative quantification and dating strategies.

1. http://www.romanportnetworks.org/ (British School at Rome; University of Southampton)

The technical aspect of the project has been to find a means by which to allow domain experts to translate their holdings into a common structure. In order to do so we have created both a procedure and the associated technology to enable data providers to:

1. Generate URIs denoting instance concepts specific to an excavation
2. Align locally-used terminology with canonical URIs
3. Map local RDB schemas to the concepts represented in a domain ontology
4. Export RDF to a triplestore and RDF/XML in order to promote the Linked Data Initiative

Fundamental to this process has been the principle of keeping the domain experts informed as to the nature of the Semantic Web paradigm whilst hiding implementation details where possible. Reliance on overtly technical approaches (such as scripting languages) at best inhibits, and at worst discourages, participation. For this reason we are developing a workflow-style toolkit that uses predictive methods to guide and speed up the mapping procedure.



Acknowledgments: The authors would like to thank Mark Weal, Graham Klyne and the members of the Roman Port Networks project for their valuable input on the themes discussed in this paper.