

Current issues with Social Network Representations

Peter Mika (Yahoo! Research)

November 22, 2008

In the social network interoperability space, several organizations have made great strides in forging agreements on interaction protocols among applications and between applications and their host containers. Less attention has been paid to another critical aspect of application interoperability, i.e. the shared representations of social network data. The importance of a shared syntax and semantics can not be overestimated: in practice the mappings among (versions of) data formats and vocabularies are too difficult to infer automatically and result in a serious loss of information.

Semantic technologies are likely to offer a way out, but do not provide an adequate solution at the current state of the art, largely because of a lack of coordination among related efforts. In this paper, we will illustrate some of the practical problems we have encountered when helping large publishers to annotate their content with microformats and existing RDF vocabularies (ontologies) in the context of SearchMonkey, Yahoo's Semantic Web product. We will also discuss potential solutions. While some of these issues may seem trivial at first, a lack of agreement on their solution still means that often applications developed by our users cannot fully rely on the data they receive.

1 The dichotomy of microformats and RDF vocabularies

Currently, microformats simplify the annotation process for publishers in that they provide a customized syntax and a vocabulary for a given task. Still, often publishers find that they cannot mark up all the information they would like to using microformats, and because of the inherent lack of extensibility of microformats, they eventually migrate to RDFa. At that point, they are forced to find an RDF vocabulary that is a direct mapping of a particular microformat. An example is moving from hCard to annotating with the VCard ontology using RDFa syntax. Unfortunately, not all microformats have an equivalent RDF ontology. In some cases, such as for VCard, there are multiple incompatible RDF vocabularies.

In the future, moving from microformats to the RDF world should be a seamless process. There should be a one-to-one mapping from microformat data to its RDF equivalent. In terms of social network data this means an agreed-upon RDF representation for hCard, XFN and hResume.

2 FOAF vs. VCard

Currently, once publishers are in the RDF space, they are faced with the problem of combining existing vocabularies to fit their need. In the social space, this most often boils down to combining terms from FOAF and one of the RDF versions of VCard. However, the relationship between FOAF and VCard is unclear: on the one hand there is some overlap (name, email, homepage, geo-location), but also different functionality (e.g. address, friends). In our daily practice, we suggest them to create instances that are typed both foaf:Person and vcard:VCard and use properties from both vocabularies in combination. It is not clear whether this is good practice.

In the future, the relationship between FOAF and VCard should be clarified, illustrated by examples on how to use them in combination. In general, overlaps among ontologies are not completely avoidable, in fact decentralized ontology development is one of the key values of the Semantic Web. However, a balance need to be found between the efficiencies of centralized ontology development and the reach of decentralized methods.

3 Vocabularies for annotation

Currently, in our practice, we see the need to adapt existing ontologies to be useful for annotation. Before the advent of RDFa, Semantic Web vocabularies were designed solely from a consumer perspective, not from a producer perspective and therefore lack some of the terms required to fully annotate social sites. For example, in social networking sites it is common to expose only the user's age and number of friends (for which FOAF has no properties), not the birthday and the list of friends (for which it does). Principles of ontology engineering may dictate that one should not introduce time-bound properties or properties for counts of things, or multiple properties for saying the same things in different ways. However, in the interest of growing the semantic ecosystem we need to accept whatever information the publisher is willing to expose. Another example from the social space is the lack of a foaf:description property, even though many social sites have a free text field where the user can introduce himself or herself.

In the future, new ontologies should be designed from the start up by taking into account what information publishers are willing to expose. In the microformat community, this is achieved by visiting the websites of major publishers in a given domain, and taking count of what is exposed on the web interface. The idea behind this method is that whatever a publisher is exposing on the HTML

interface, it may be willing to expose in a structured form as well. Existing ontologies should be adapted in a similar fashion.

4 Relaxing constraints

Currently, our practice reflects that many of users find it difficult to deal with some of the complexity of RDF such as the distinction between resources and literals, or understanding the difference between classes (a property of resources) and datatypes (a property of literals). This leads to a lot of 'abuse' where, for example, homepages and email addresses are represented as literals, not resources.

In the future, there is some role for education or alternatively, devising 'lightweight' representations that do not force a distinction between resources and literals, classes and datatypes.

5 Mobile readiness

Currently, on occasion we find the need to display some information differently in a mobile environment than in an online environment. The lack of space means that we would often prefer, for example, a graphical representation of a rating or the type of a phone number. It is unclear how to indicate that an image is a graphical representation of a value.

In the future, agreements should be made about how to treat multiple representations of the same value.

6 Conclusion

In conclusion, we see a number of issues with current methods for representing social network data on the Web, in particular when it comes to exposing data using semantic annotation (microformats, RDFa). These problems do not seem to require new efforts in standardization, but rather in forging agreements about the usage of existing Web standards. Once these agreements have been reached, we also see a significant role for educating the developer community about the best practices in social network data representation.