Vocabulary Hosting: A Modest Proposal

Harry Halpin
World Wide Web Consortium and University of Edinburgh
10 Crichton St.
Edinburgh, United Kingdom EH8 9AB

Tom Baker
Dublin Core Metadata Initiative Ltd.
50 Geylang East Avenue 1
Singapore 389777

Abstract

Many of the benefits of structured data come about when users can re-use existing vocabularies rather than create new ones, but it is currently difficult for users to find, create, and host new vocabularies. Moreover as a foundation for applications, the value of any given vocabulary depends on the perceived certainty that the vocabulary – both its machine-readable schemas and human-readable specification documents – will remain reliably accessible over time and that its URIs will not be sold, re-purposed, or simply forgotten. This note proposes two approaches for solving these problems: one for multiple Vocabulary Hosting Services and a Vocabulary Preservation System to keep them linked together.

Many of the benefits of structured data come about when users can re-use existing vocabularies rather than create new ones, but it is currently difficult for users to find, create, and host new vocabularies. As a foundation for applications, the value of any given vocabulary depends on the perceived certainty that the vocabulary – both its machine-readable schemas and human-readable specification documents – will remain reliably accessible over time and that its URIs will not be sold, re-purposed, or simply forgotten. This note proposes a Vocabulary Hosting Service (VHS) for meeting immediate needs and outlines a more comprehensive Vocabulary Preservation System (VPS) of the sort needed as a solution for the long term.

A **Vocabulary Hosting Service** (VHS) needs to be made to meet immediate needs. Ideally, these sort of services would fulfill the following characteristics.

- Addresses an immediate need by application developers for a managed persistence context in which to declare vocabularies.
- The service would own a domain under which vocabularies can be created (in principle by anyone) with the certainty that the vocabulary URIs and related documentation will remain accessible.
- Stewardship of such a service would lie with a group of stakeholders involving (potentially) groups like the W3C, DCMI, W3F, Library of Congress that are known long-term entities with an interest in stable vocabularies and commercial companies such as Google, Yahoo, and Talis whose business model does or may involve vocabularies.

A **Vocabulary Preservation System** can then link the various vocabulary hosting services together. This more comprehensive Vocabulary Preservation System (VPS) is less pressingly needed now, but is needed as a solution for the long term. This system would fulfill the following characteristics.

- Addresses a longer-term need to put vocabularies created in a wide range of institutional and commercial contexts into a well-understood preservation context that will guarantee the resolvability of vocabulary URIs to namespace documentation over the long term.
- Memory institutions (such as libraries) maintain mirrored caches of crucial namespace documentation, step in to maintain uninterrupted service in response to well-defined trigger events, and inherit stewardship of the vocabularies in the long term.

In the end, pragmatic concerns will dominate. Users who want to create a vocabulary need to be able to host the vocabulary somewhere easily, possibly without purchasing their own domain name or webserver. Experiences at recent 'open data' events like Vocamps[1] show that this type of service is needed, as much of the work done at these events is lost. At present, projects or organizations that wish to publish vocabularies and make them available for use by others are rarely in a position to guarantee the long-term persistence of their vocabularies' URIs or continued availability of associated documentation. End-users face a bewildering number of options for actually finding and using vocabularies, and few reliable third-party hosts provide any technical or legal assurances about the use or persistence of their vocabularies. Furthermore, companies that benefit from the deployment of structured data by end-users would be best served by keeping the vocabularies of structured data of the Web shared and not controlled by any single company or application, much like HTML is shared amongst various browser vendors. This makes it easier for users (including other companies) to create structured data, providing more likelihood that a critical mass of structured data can be created that can

[1] A sort of 'barcamp-style' event for creating vocabularies, see `http://vocamp.org`.

then serve as the foundation for new, exciting, and potentially profitable applications. This is true for producers of Linked Data in general, and as more and more people release data as Linked Data, the need for vocubulary hosting will only increase. However, first we will look at the use of vocabularies in consumers in the case of structured search.

The Use Case of Common Vocabularies in Structured Search

Currently, there is increasing interest in the ability to add structure to the Web, as exhibited by the recent move of search engines to index and handle vocabularies in RDFa (Adida et al. 2008). RDFa fulfills a use-case that is not currently covered by microformats (Suda 2006), in particular, the use-case of a user wanting to mark up some structure in HTML that does not have already valid microformat. Note that using technologies like GRDDL, microformats and RDFa can be made compatible (Connolly 2007).

As more and more structured data gets on the Web, users will want to be able to deploy a 'long-tail' of vocabularies, and many niche markets will want to take advantage of these vocabularies. For example, while microformats like hCard and XFN cover social relationships, in a scientific domain like medicine it is likely useful to large and detailed specialist vocabulary for health care, and to mark up health reports using this vocabulary. Yet such a vocabulary can not easily be created as a microformat, since it appeals only to a small community and its creators may not have the time nor wherewithal to engage in the process around creating a microformat for their domain-specific knowledge.

Equally, for the discovery of information about products like MP3 players, some vocabulary listing of MP3 player brands, options, and prices would be needed. For example, the GoodRelations vocabulary currently allows one to describe products (Hepp 2008), and while this is supported by Yahoo, it is not supported by Google, who instead created their own vocabulary for products.[2] How can the owner of a store that wants to list MP3 player brands easily find the vocabulary to put that sort of information into his web-page using RDFa? Currently, it is difficult to even discover vocabularies like GoodRelations using Semantic Web search engines like Sindice (Oren et al. 2008). The results of these Semantic Web search engines, while useful for a Semantic Web expert, do not provide the guidance an ordinary user needs. Furthermore, if the store owner tries to optimize for either vocabulary, they risk alienating a portion of their audience or applications that use a particular vocabulary. While there are clear benefits for a search engine to find and use this structure from a web-page, the costs currently outweigh the benefits for the end-user, as there is no process for creating a new RDFa vocabulary that can be easily found, re-used by other users, and indexed by search engines is needlessly difficult.

The same issue holds not just for product reviews and prices, but for almost all Semantic Web vocabularies, which are already often fractured into incompatible versions despite a minimum of uptake. So, currently if a user wants to use RDFa to mark up something as simple as some personal information attached to a review, how should they do it? There are a number of independent personal information vocabularies FOAF,[3] W3C RDF mappings of vCard,[4] and Google's vocabulary.[5] Obviously, the same user or developer that wants to add structured data to their web-page should not have to choose whether they want to optimize their RDFa for either Google or Yahoo!. They should just be able to find the 'best' vocabulary for their needs and mark up their web-page with minimal concern, and if they can't find a vocabulary, they should be able to create one and host the vocabulary someplace neutral with a guarantee that their vocabulary can be re-used without any charge, and the assurance that the documentation and URIs used by the vocabulary will be persistent.

Currently, the two main search engines that deploy structured search are Google and Yahoo. Yahoo's SearchMonkey, based on the earlier work of Microsearch (Mika 2008), tries to point users to relevant specifications, such as FOAF, but does provide any sort of infrastructure for users to find and create vocabularies. Google supports and offers examples in RDFa to vocabularies for reviews, people, products, and organizations. The vocabularies used by Google are often straightforward vocabulary mappings from microformats. Google also does not currently seem to support open-ended use of RDFa or the ability for users to create vocabularies, although it supports and hosts all its supported RDF vocabularies at Google's own URI. There is a powerful logic behind Google's choice: If Google does not feel that other URIs are likely to be trusted, maintained, preserved, or provide a decent user-experience, then of course Google will host the vocabulary at its own URI, as at least Google can then guarantee as a company the persistence and maintenance of those vocabularies. Ideally, what is needed is an infrastructure, usable and supported by both major companies such as Google and Yahoo, that is hosted at a neutral third-party. This neutral vocabulary hosting service would then make life easier for users of these structured vocabularies.

Requirements for Vocabularies

However, what would the requirements be for vocabularies for structured data? There are many types of vocabularies. In general, vocabularies are used across a vast variety of systems, ranging from strings used in everything from SQL databases microformats, to more complex approaches such as to the QNames typically used in XML and the use of URIs in RDF. In particular, we will in our proposed system will map these all to URIs so that vocabularies in different domains that re-use the same text string can remain independent of each other. In this vein, vocabularies require:

- **Use of URIs**: Each term in a vocabulary has a URI.
- **Persistent URIs**: Each term URI will be used to refer to the same term in perpetuity and will not be repurposed.

[2] See `http://rdf.data-vocabulary.org/`.

[3] See `http://xmlns.com/foaf/spec/`

[4] See `http://www.w3.org/TR/vcard-rdf`

[5] See `http://rdf.data-vocabulary.org/`

Institutional guarantees are key; note that stable URIs (e.g., using redirection services) do not automatically refer to stable documentation.

- **Persistent documentation**: Each term URI should remain resolvable to 'namespace documents' – descriptive documentation in HTML and/or machine-readable representations such as RDF schemas.
- **Change policy**: The stability of the meaning of the terms should be determinable – i.e., the meanings of terms should evolve according to known change management policies and with responsibility for changes traceable to either individuals or organizations.
- **Preservation**: Vocabularies should be preserved, like any other business-critical component or component of cultural heritage, ideally not in dependence on a single vocabulary hosting service.
- **Analytics**: Vocabularies should be plugged into actual empirical data about their usage in the wild, allowing vocabulary producers and consumers to see how the vocabulary is actually used. This is invaluable for future versioning of the vocabulary and for discovering valuable patterns in the vocabulary usage.
- **Healthy ecosystem**: Like any other aspect of a machine or human language, vocabularies are needed in contexts ranging from general to the very specific domains, from the informal to the highly institutionalized. A diversity of vocabulary maintainers using a multiplicity of domain names is healthier than a vocabulary monoculture.

A few notes on URIs and persistence are in order. A URI can always be redirected to persistent documentation, as with `purl.org`. The problems of persistent URIs can therefore be separated from the goals of persistent documentation and change policy. One organization host might provide persistent URIs while another provides a hosting service for documentation. For example, an organization (such as DERI) might host URIs under its domain (like `semanticweb.org`) while the documentation is hosted using resources elsewhere (even cloud resources like `Amazon.com`); this can be easily done.

The persistent URI problem is not trivial, and may require clever optimization of caching if the URI is frequently retrieved. However, as URIs are normally just used for disambiguation in vocabularies without retrieval, this problem is not as endemic as it might seem.

Vocabulary Hosting Service

A vocabulary hosting service would own a domain, let people create new vocabularies under that domain, and publish human-readable documentation about the vocabularies. Once published, documentation would fall under the persistence guarantee of the service. Vocabulary maintainers would have the right to install updated documents, but all historical versions of the vocabulary would be preserved and remain accessible.

The service would promise to keep the latest version of the documentation available at the vocabulary URIs. Just as `sourceforge.net` guarantees the preservation of old versions of software and provides access to the latest version ('latest version persists') while making no guarantees that the software posted on its servers will work as described, the vocabulary hosting service does not have to make guarantees regarding the soundness of hosted vocabularies, nor would the service itself be in a position to make any promises regarding the further development of any hosted vocabulary. Social systems that provide user-driven ratings could even allow vocabulary users themselves to ascertain soundness and provide explicit feedback.

A vocabulary hosting service could do for structured data what `programmableweb.com` does for APIs and mashups. Such a hosting service might also offer easy-to-use examples for users, tutorials, recipes, best-practice documents on managing vocabularies, a FAQ, and mailing lists; a staging area for testing documentation (human and machine-readable) before publication; a publication wizard for uploading namespace documents and setting MIME types and 303 redirects; and a Web-based authoring environment for vocabularies. Options provided by the hosting service could range from simple uploading of schemas and documents (with automated configuration and testing before final publication) to rich editing environments and command-line access to a server.

The service would require explicit permission from vocabulary registrants, such as a license to distribute content in perpetuity, and it would need to provide disclaimers regarding copyrights, trademarks or patent violations. It should ideally guarantee some type of royalty-free licensing for the use of a vocabulary. A hosting service may by default acquire the right to manage (i.e., make changes) to the vocabulary in the long term if or when its maintainers disappear.

Vocabulary hosting services may have different policies about vocabulary creation. Some might let anyone register a vocabulary with the service free-of-charge with instant deployment. Others may require a longer community-driven process of review before deploying the vocabulary. A vocabulary hosting service may allow URIs to point to documentation or schemas outside the vocabulary host itself.

Vocabulary hosting services could become domain-specific. Vocabulary hosting services may restrict people to hosting abstract vocabularies, while others may allow URIs to be minted for physical objects, people, places, or ideas. For example, one might want a taxonomy of automobile types and options, but also a single URI to describe in detail a particular automobile (the 'Volkswagen Bug') that fits within the taxonomy of more abstract automobiles ('convertibles'). Thus, one could restrict users to RDF properties and classes about automobiles, but some hosts might allow objects, people, places, or ideas, or even entities of an undeclared type.

Note that the problem of vocabulary hosting can be separated from the problem of guaranteeing the availability of vocabularies in a distributed manner over a longer term ('vocabulary preservation'). A Vocabulary Hosting Service might be part of a larger Vocabulary Preservation System which maintains up-to-date caches of all vocabulary-related materials multiple institutions.

One possible objection is that redirection services, such as the use of HTTP 303 or `purl.org` are enough. However, while redirection services such as `purl.org` can guarantee that URIs will remain resolvable, they cannot guarantee that schemas and documentation will remain available when a vocabulary is no longer actively maintained. Furthermore, even `purl.org` has gone down on occasion. What few vocabulary hosting sites exist currently are not regularly updated and do not allow use by the general public. Dan Brickley's `xmlns.com` and Ian Davis's `vocab.org` may go the way of the currently non-maintained `schemaweb.info`. Other organizations that maintain vocabulary hosting sites, such as DERI's `rdfs.org` and Revelytix's `knoodl.com`, may also suffer the same fate if either organization has problems.

There exists a number of codebases that, while each has a large number of problems, could with enough work serve as the foundation for such a system. It should be a requirement that such a system should allow one to edit and maintain a vocabulary purely through a Web-based interface and so without downloading any software. This immediately makes unusable most traditional ontology editing environments. Rather unfortunately, while work like the Ontolingua system are the inspiration for this modest proposal for Linked Data (Gruber 1993), most work from the ontology engineering community like the Neon Toolkit is aimed at specialists and does not have a Web-based interface (Gómez-Pérez and del Carmen Suárez-Figueroa 2009). However, Semantic Mediawiki, could serve as a basis in combination with WebProtege (Völkel et al. 2006). A more light-weight client like Neologism[6] could also serve as a foundation in combination with a content management system. Since vocabulary creation and maintenance is social process, the re-purposing of a content management system like Drupal that already has a strong social component (with ratings, tags, social networks, and blogs) would likely be a necessary component of a larger vocabulary hosting platform.

Vocabulary Preservation System

Regardless of the intent, the best persistent URIs and Vocabulary Hosting Services may still at some point go down. The only way to counter this to have some loose cooperation among institutions to cache and preserve vocabularies, much as the domain name system itself is cached. The preservation approach should provide a context for the long-term preservation of vocabularies in a distributed manner and in a wide range of social contexts. Indeed, as they may be important to future applications or scholarship, it is important to consider these vocabularies as an artifact to be preserved like any other artifact important for cultural memory. To summarize, a Vocabulary Preservation System (VPS) would:

- **Contracts**: Provide a well-understood framework for contracts between Vocabulary Hosting Services and the Vocabulary Preservation System, defining processes and governance for handling rapid interventions (e.g., redirecting URIs if disaster strikes) and for transferring ownership and maintenance responsibility in the long-term.
- **Open:** Offer itself as a partner to any Vocabulary Hosting Service that tries to aim for best practices in vocabulary hosting, and so serve as a further 'seal-of-approval' for Vocabulary Hosting Services.
- **Reliable:** Aim at ensuring that at least all large-scale (popular) vocabularies have such mechanisms in place.
- **Meta-review:** Provide regular reports on the status of maintenance activities for covered vocabularies, possibly as a group with other Vocabulary Preservation Systems. As a by-product of this reporting, the VPS or coalition of VPS systems would function as a clearinghouse for information about vocabularies.

Note that this system may not be difficult to implement. The LOCKSS (Lots of Copies Keep Stuff Safe) of Stanford would provide an ideal open source code-based for distributing copies across the world (Maniatis et al. 2005). The LOCKSS system achieves this redundancy by providing an automated system for sharing caches of digital content within a secure, closed peer-to-peer network, and has already been implemented successfully to preserve digitally curated journal articles (Maniatis et al. 2005). Due to the mathematical probability of unrelated server failures, a fairly small number of vocabulary caches should be sufficient in order to guarantee vocabulary preservation in the long-term.

However, at the moment we need at least one long-term, neutral, non-profit vocabulary hosting site with a clearly established social process for adding and editing vocabularies. Once this is in place, more Vocabulary Hosting Services will undoubtedly arise.

A Modest Proposal

There is a clear need today for at least one neutral and non-profit Vocabulary Hosting Service. This hosting space should ideally be managed by organizations and companies that have a long-term interest in the area and are actively deploying these technologies. This should be created as soon as possible. The elements of a solution are:

- **Domain-name**: A neutral domain name should be used – one that belongs to a non-profit organization or could be transferred to one. Suggestions have been `sharednames.org`, `semanticweb.org`, a new domain name, or another domain name of an existing body.
- **Legal Body**: Some sort of legal body needs to be host the domain name or purchase. Another option could be an existing organization, and the W3C or IANA would be possibilities. A legal policy that provides the proper disclaimers and royalty-free status of the vocabularies would need to be crafted; the W3C Royalty-Free Patent Policy could serve as a starting point, as well as the Open Data Commons' Open Database License (ODbL).[7]
- **Technical Infrastructure**: The site would need a strong policy on service and maintenance, with a robust and extensible infrastructure. Some combination of Amazon,

[6] See `http://neologism.deri.ie/`

[7] http://www.opendatacommons.org/licenses/odbl/

Google, and Yahoo! could help provide this infrastructure. Organizations like IANA or the W3C do not have at the moment the technical resources to maintain such a vocabulary hosting service by itself, although they might if there was some funding by external entities. The baseline infrastructure should allow users that agree to the legal policy to upload their schema and human readable-documentation, generating this via simple transformations. This could later be upgraded with a more comprehensive and interactive user interface.

- **Long-term financial stability**: Since managing persistent URIs and hosting documentation takes resources, there needs to be some baseline financial model. This can initially be supported by work from interested parties, and if enough interested parties sign up, the domain name can be guaranteed for a long period of time. Hosting is more difficult, but suitable arrangements could be made. A simple corporate and organizational sponsor revenue stream may work, but in the worst case some sort of fee for hosting and maintaining vocabularies could be deployed.
- **Clear Social Process**: There are two main options as regards social process for vocabulary creation: the completely open model and the working group model. In the open model, similar to a Wikipedia for vocabularies, any person could easily submit, update, and comment upon a vocabulary, with a simple line of responsibility going to a single individual or the VHS itself. In this way, vocabularies could be allowed to be assigned on a 'first-come first-serve' basis. In the working group model, a group of interested parties is chaired for some period of time to come to consensus on a vocabulary, like in any other standardization process, and vocabularies are then assigned based on consensus.

The vision of this modest proposal is at least one neutral vocabulary hosting system that would serve the various vocabularies for the 'links' in Linked Data and structured search. Far from an 'Academie Francaise' that dictates how vocabularies should be used, a bottom-up approach that allows both producers and consumers of data to contribute to these vocabularies would be ideal: A Wikipedia for the vocabularies that give data their meaning.

References

Adida, B.; Birbeck, M.; McCarron, S.; and Pemberton, S. 2008. RDFa in XHTML: Syntax and Processing. W3C Recommendation, W3C. http://www.w3.org/TR/rdfa-syntax/.

Connolly, D. 2007. Gleaning Resource Descriptions from Dialects of Languages (GRDDL). W3C Recommendation, W3C. http://www.w3.org/TR/grddl/.

Gómez-Pérez, A., and del Carmen Suárez-Figueroa, M. 2009. Scenarios for building ontology networks within the neon methodology. In Gil, Y., and Noy, N. F., eds., *K-CAP*, 183–184. ACM.

Gruber, T. R. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5(2):199–220.

Hepp, M. 2008. Goodrelations: An ontology for describing products and services offers on the web. In *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management*. 332–247.

Maniatis, P.; Roussopoulos, M.; Giuli, T. J.; Rosenthal, D. S. H.; and Baker, M. 2005. The lockss peer-to-peer digital preservation system. *ACM Trans. Comput. Syst.* 23(1):2–50.

Mika, P. 2008. Microsearch: An Interface for Semantic Search. In *Proceedings of Semantic Search Workshop at the European Semantic Web Conference*.

Oren, E.; Delbru, R.; Catasta, M.; Cyganiak, R.; Stenzhorn, H.; and Tummarello, G. 2008. Sindice.com: A document-oriented lookup index for open linked data. *International Journal of Metadata, Semantics, and Ontologies* 3(1):37–52.

Suda, B. 2006. *Using Microformats*. O'Reilly. http://safari.oreilly.com/0596528213 (Last accessed Oct 12th 2008).

Völkel, M.; Krötzsch, M.; Vrandecic, D.; Haller, H.; and Studer, R. 2006. Semantic Wikipedia. In *Proceedings of the International Conference on World Wide Web (WWW)*, 585–594.