THEME ISSUE PAPER

Requirements for vocabulary preservation and governance

Thomas Baker
Dublin Core Metadata Initiative, Washington, District of Columbia, USA
Pierre-Yves Vandenbussche
Fujitsu (Ireland) Limited, Swords, Ireland, and
Bernard Vatant
Mondeca, Paris, France

Received 4 March 2013
Revised 4 July 2013
Accepted 23 July 2013

Abstract

Purpose – The paper seeks to analyze the health of the vocabulary ecosystem in terms of requirements, addressing its various stakeholders such as maintainers of linked open vocabularies, linked data providers who use those vocabularies in their data and memory institutions which, it is hoped, will eventually provide for the long-term preservation of vocabularies.

Design/methodology/approach – This paper builds on requirements formulated more tersely in the DCMI generic namespace policy for RDF vocabularies. The examination of requirements for linked open vocabularies focuses primarily on property-and-class vocabularies in RDFS or OWL (sometimes called metadata element sets), with some consideration of SKOS concept schemes and Dublin Core application profiles. It also discusses lessons learned through two years of development of the linked open vocabularies (LOV), of which main features and key findings are described.

Findings – Key findings about the current practices of vocabulary managers regarding metadata, policy and versioning are presented, as well as how such practices can be improved to ensure better discoverability and usability of RDF vocabularies. The paper presents new ways to assess the links and dependencies between vocabularies. It also stresses the necessity and importance of a global governance of the ecosystem in which vocabulary managers, standard bodies, and memory institutions should engage.

Research limitations/implications – The current paper is focused on requirements related to a single type of vocabulary but could and should be extended to other types such as thesauri, classifications, and other semantic assets.

Practical implications – Practical technical guidelines and social good practices are proposed for promotion in the Vocabulary Ecosystem (for example, by vocabulary managers).

Originality/value – This paper brings together the research and action of several important actors in the vocabulary management and governance field, and is intended to be the basis of a roadmap for action presented at the Dublin Core conference of September 2013 in Lisbon (DC 2013).

Keywords Governance, Sustainability, Linked Data, Vocabularies

Paper type Research paper

Foreword

One must be cautious when speaking, writing, or reading about vocabularies. Treacherous paradoxes lurk where language is used recursively to speak about itself –

An earlier version of this paper was presented at the LOV symposium, held in Madrid, Spain, on 18 June, 2012.

Library Hi Tech
Vol. 31 No. 4, 2013
pp. 657-668
© Emerald Group Publishing Limited
0737-8831
DOI 10.1108/LHT-03-2013-0027

where terminology is overloaded by the inconsistent meanings of so many communities and viewpoints. We therefore clarify, upfront, what we mean by critical words such as "vocabulary" and "term." The reader is asked to pay careful attention to those definitions, even if they seem unfamiliar.

Introduction

Linked open data (www.w3.org/standards/semanticweb/data) is structured data, published on the web, which uses standard formats to support the creation of meaningful "links" among a diversity of datasets, at a high degree of granularity, without requiring pre-coordinated arrangements among data providers (Berners-Lee, 2006). Linked Data bridges the barriers among separately maintained data silos. Since 2006, a "cloud" of interlinked data sources has grown to encompass a wide range of resources, from census data to bibliographies and biomedical databases. Publishing data on the web as linked data makes it easy for other organizations and data providers to create detailed links to your data (and vice-versa) and to integrate your data interoperably into other contexts, making data more visible and more re-usable and potentially contributing to the creation of data aggregates that are more than the sum of their parts.

Linked open data is expressed using a simple three-part data structure: the "RDF Triple," as defined by a W3C standard, Resource Description Framework (RDF) (www.w3.org/TR/rdf-primer/). An individual RDF Triple makes a simple statement of the form "A is related to B," as in "Book X was authored by Person Y" or "Book X has the title *Being There*." Some elements of RDF statements are string values, such as "Being There." For all other elements, RDF leverages the globally managed address space of uniform resource identifiers (URIs) on the world wide web – for example, the ubiquitous http:// URL – as a source of identifiers to denote not just specific web pages, but also non-digital entities such as people, books, and concepts. URIs make people, books, and concepts globally citable.

RDF triples use "vocabularies" of properties (to express relationships such as "was authored by") and classes (to say that Resource X is a "book"). Vocabularies bring meaning to data. RDF vocabularies are themselves expressed as linked data – i.e. they are defined using RDF triples and published on the web, using URIs as identifiers for their terms. In today's web environment, RDF vocabularies are created by a wide range of people and institutions, from individual researchers to national libraries and for-profit corporations, for a wide range of descriptive requirements.

Like other forms of linked data, vocabularies are linked among themselves, for example to say that Term X in Vocabulary A means the same as Term Y in Vocabulary B. The set of linked open vocabularies may be thought of as a vocabulary ecosystem. The usability of linked data – i.e. the ability to interpret what the data means – depends on the availability of the RDF vocabularies used in the data. The health of the linked open data cloud, in other words, depends on the health of the vocabulary ecosystem.

This paper builds on requirements and recommendations formulated in several previous documents. The general principles of linked data have been defined in the reference "5-star" document (Berners-Lee, 2006). Best practice recipes for publishing vocabularies have been defined (Berrueta and Phipps, 2008). Various papers stress the importance of reuse in ontology engineering (e.g. Suárez-Figueroa *et al.*, 2011; Sure *et al.*, 2009). Such documents provide precious guidelines for individual vocabulary engineering and publication but barely address the issues of global governance and long-term preservation. Some of the relevant requirements proposed here have been previously introduced in the DCMI Generic Namespace Policy for RDF vocabularies (Baker *et al.*, 2011).

Beyond the above literature, we ground our proposal on lessons learned during two years of work in the linked open vocabularies (LOV) project, during which we have studied in depth the state of the vocabularies ecosystem at both technical and social level. Through hundreds of exchanges with vocabulary managers, we have identified the critical issues the ecosystem faces, due to the lack of global vision of its various actors.

The requirements presented here are directed to all stakeholders of the vocabulary ecosystem: linked data providers who use those vocabularies in their data and, ultimately, memory institutions which, it is hoped, will take measures to provide for the long-term preservation of the vocabularies.

Scope of this paper

"Vocabulary" means different things to linguists, library scientists, and the semantic web community. This paper follows a definition of World Wide Web Consortium (W3C) whereby semantic web vocabularies "define the concepts and relationships (also referred to as 'terms') used to describe and represent an area of concern"(www.w3.org/standards/semanticweb/ontology). A semantic web vocabulary classifies its terms into categories of a grammatical nature, such as properties or classes. The vocabulary may also specify relationships among terms, such as sub-property or sub-class relationships, and it may define formal constraints on their meanings. Although there is no clear formal difference between vocabularies and "ontologies," the two types of vocabulary are commonly distinguished by their usage: "The trend is to use the word "ontology" for more complex, and possibly quite formal collection of terms, whereas "vocabulary" is used when such strict formalism is not necessarily used or only in a very loose sense."

Translated into terms more familiar to library scientists, RDF vocabularies may be seen as constituting metadata element sets (vocabularies of properties and classes) or value vocabularies (vocabularies of concepts typically used as values such as subject descriptors, which are often described using the W3C specification Simple Knowledge Organization System, or SKOS) (Isaac *et al.*, 2011; Miles and Bechhofer, 2009). Moreover, some communities distinguish between vocabularies as they are originally declared and vocabularies comprised of terms originally declared elsewhere, as in the Dublin Core community's notion of an application profile. Element sets, value vocabularies, OWL ontologies, SKOS concept schemes, and Dublin Core application profiles can all be considered linked open vocabularies. An important vocabulary project of several major search engines, Schema.org, is based (though somewhat loosely) on RDF.

The examination of requirements for linked open vocabularies in this paper focuses primarily on property-and-class vocabularies in RDF or OWL, with some consideration of SKOS concept schemes and Dublin Core application profiles.

Services which monitor the ecosystem

Analogously to natural-language vocabularies, machine-readable vocabularies are created for uses which range from the very generic to the very specialized and domain-specific. They are used in contexts which range from local, time-limited projects to stable institutions of global reach. As in agriculture, this diversity of vocabularies is arguably healthier than monoculture, but the decentralized, uncoordinated nature of the vocabulary ecosystem poses challenges for its users. Data providers and vocabulary designers cannot follow the principle of re-use unless they can find what they need to re-use. Existing services such as the following scan or catalog parts of the vocabulary ecosystem, from various perspectives, to help users discover vocabularies, find specific terms and their uses, and understand relationships among vocabularies:

- LOV (http://lov.okfn.org/dataset/lov/) – an observatory focused on providing an inventory of, and analytical metadata about, property-and-class vocabularies used in linked data. Started in 2011 as part of a research project, DataLift (Scharffe *et al.* 2012), LOV has operated since 2012 under the umbrella of the Open Knowledge Foundation. Its crawling tool, LOV-bot, scans the vocabulary ecosystem each day to update metadata about terms and their interconnections.
- Swoogle (http://swoogle.umbc.edu/) (Ding *et al.*, 2004), Falcon-S (Cheng *et al.*, 2008) and Watson (d'Aquin and Motta, 2011) – search engines that index ontologies for querying.
- Sindice (http://sindice.com/) (Tummarello *et al.*, 2007) – a project which crawls web documents for semantic data and can therefore help in discovering and quantifying vocabulary usage.
- Open Metadata Registry (http://metadataregistry.org/about.html) – an online environment for creating and maintaining property-and-class vocabularies and SKOS concept scheme.
- Meta-Bridge (www.metabridge.jp/infolib/metabridge/menu/?lang = en) – an online metadata schema registry in Japan focused on RDF property-and-class vocabularies and application profiles.
- Joinup (https://joinup.ec.europa.eu/catalogue/all) – a portal for "semantic assets" used in eGovernment applications, managed by Federation of Semantic Asset Repositories of European Commission's ISA Programme.
- CKAN Datahub (http://datahub.io/) – a community project for cataloging Linked Data datasets of all types and therefore also of vocabularies.
- prefix.cc (http://prefix.cc) – a service that tracks how vocabulary URIs are abbreviated in data using conventional prefixes.

Act locally

Since we deal here with concepts close to those of sustainable development, the requirements are sorted in two sections, inspired by a famous tagline borrowed to this field. Under "Act locally" we gather requirements for individual vocabulary managers, and under the subsequent section "Think globally" we review requirements for a sensible governance by the community.

Leverage the linked data and web architecture

To be usable in linked data, vocabularies must meet several criteria considered by the authors as more-or-less "hard" requirements. They represent the minimal technical requirements for a vocabulary to be considered part of the global linked data ecosystem:

- Each term in a vocabulary must be citable by its own URI (here: "term URI"). Vocabularies should be citable as a whole by their "namespace URI" – the base substring used to form the term URIs of the vocabulary. Note that the W3C standards community is moving to IRIs, or Internationalized Resource Identifiers, which are like URIs but allow the use of characters from the Universal Character Set, Unicode/ISO 10646, which includes Korean hangul, Japanese kanji, Russian Cyrillic characters, and the like.
- Each term URI must be resolvable ("dereferencable") to a formal, machine-readable expression of its meaning. The formal expressions of vocabularies of properties and classes typically use one of the semantic web languages, RDF Schema or OWL. The formal expression may be embedded in web pages using a special syntax, RDFa, or published on the web in one of several alternative RDF serialization syntaxes. It is good practice to offer a choice of representations, including human-readable web pages, through browser- or software-mediated content negotiation (Berrueta and Phipps, 2008).

Provide metadata

One key finding from the first year of LOV project has been the widespread lack of the most basic metadata on vocabularies, in particular for vocabularies published in the early years of the semantic web (roughly, before launch of the linked open data project in 2006). For example, little more than 20 percent of the vocabularies added to the LOV database in the first year explicitly declared a date of last modification. Similarly few supply any information about creators, contributors, and editors. Early vocabulary publishers apparently saw little point in providing such metadata. Using contextual information, LOV can reconstruct such information for 90 percent of the vocabularies listed. Actively seeking input and information from vocabulary maintainers and promoting good practices has yielded tangible improvements.

- Each vocabulary should describe its maintainers and the policies by which it is maintained. In the recursive manner characteristic of RDF, vocabularies should provide descriptions of themselves, preferably using one of several emerging standards, such as Dublin Core, ADMS, VANN, and VOAF (Vandenbussche and Vatant, 2012). Given that any living vocabulary can and should evolve over time, the vocabulary should point to policies about the management of changes. The DCMI Namespace Policy, for example, articulates four categories of change, from minor editorial errata to substantive semantic changes, to specify allowable limits within which the definition of a term may be changed before triggering the creation of a separate term and term URI (Powell *et al.*, 2007).
- Where possible, vocabularies should provide labels and definitions in one or more explicitly declared natural languages. Most vocabularies provide labels and definitions for their terms in English only, but internationalization is underway. The vocabularies listed in the LOV database use more than 20 different languages. Ideally, information should be specified in multiple languages both at the level of the vocabulary as a whole and for each term label and definition.
- Vocabularies should be published under copyright terms that explicitly allow and encourage re-use. In the case of actively maintained vocabularies, ideally,

maintainers should clarify how users may provide feedback regarding errata and proposed changes or extensions to their vocabulary.

- Successive versions of a vocabulary should be discoverable and accessible. Many vocabularies have no clear policy regarding versioning and history. A namespace URI should always dereference to the latest version, from which it should be possible to access previous versions. Descriptions of vocabularies in the LOV database provide a visualization of the timeline of versions. Ideally, version histories are explicitly defined by the maintainers of a vocabulary. In the absence of such information, the LOV software can detect and track changes by monitoring URIs on a daily basis.

Reuse, reduce, recycle

Inasmuch all terms in the semantic web environment are, in principle, globally citable by URI, it is generally considered good practice to use available vocabularies when describing resources: "If suitable terms can be found in existing vocabularies, these should be reused to describe data wherever possible, rather than reinvented. Reuse of existing terms is highly desirable as it maximises the probability that data can be consumed by applications that may be tuned to well-known vocabularies, without requiring further pre- processing of the data or modification of the application." (Heath and Bizer, 2011) Likewise, methodologies for "ontology engineering" promote re-use where possible (Suárez-Figueroa *et al.*, 2011; Sure *et al.*, 2009). This suggests the following soft requirements for vocabularies:

- *Where appropriate, design your vocabulary to re-use other, well-known vocabularies*. It is good practice to avoid creating exact equivalents of terms that exist elsewhere. Where a term with a meaning broader or more specific than that of an existing term is needed, it is good practice, when coining the new term, to specify how it relates to near-equivalents elsewhere by explicitly declaring relationships such as sub-property or sub-class. The self-description of a vocabulary should use well-known vocabularies created for such purposes. The 350 (as of July 2013) property-and-class vocabularies indexed by the LOV Project link in such ways to, on average, four other vocabularies.
- *When re-using terms from other vocabularies, respect the formal definitions and constraints declared by their maintainers*. In a semantic web context there is a strong social convention by which it is the owner of a vocabulary – or to be more precise, the owner of the URI domain under which its terms are coined – who declares the meaning of a given term, and anyone re-using that term should respect its declared meaning. Of course, terms are subject to redefinition through actual use – for example, the term owl:sameAs has been misused on such a massive scale that its original meaning has arguably been compromised – but URIs can only serve to anchor meaning if this principle is followed.
- *Where appropriate, map your terms to related terms in other vocabularies*. Asserting relationships of sub-property, sub-class, equivalence, or near-equivalence between terms in different vocabularies mitigates the effects of redundancy by making it easier to align datasets based on different vocabularies. Providing labels for terms in multiple languages also constitutes a

form of mapping inasmuch it makes a vocabulary accessible to, and thus usable by, speakers of other languages.

That some organizations re-invent existing terms instead of re-using them from elsewhere is sometimes due to a desire to avoid making their vocabularies, and thus their data, dependent on multiple external specifications beyond their control and with unknown futures – an issue the requirements for vocabulary preservation, formulated below, are intended to address.

Think globally

To trust that independent application of the above requirements by vocabulary creators and publishers will be enough to ensure the healthy development of the ecosystem as a whole seems overly optimistic. We now consider what should to be done at a more global level for a better monitoring, governance, and long-term preservation of the ecosystem.

Improved monitoring tools

Meeting requirements at the global level requires efficient tools for monitoring the ecosystem. Data publishers should find it easy to discover and understand vocabularies appropriate for their needs, while vocabulary creators should be able to get an overview of the vocabularies they reuse or could reuse, as well as how their own vocabulary is itself reused. In the absence of dedicated services, it is not obvious how vocabulary curators can learn of changes in the vocabularies on which they rely or assess the impact of such changes in their own vocabularies. Services exist (see above), but they need to be consolidated and improved to meet the following requirements.

- *Provide web analytics with empirical data about the usage of vocabularies in the wild*. This requirement addresses a basic need both of data providers and of vocabulary publishers. A "good" vocabulary, seen in terms of semantic interoperability, is one that is widely used. Before using a vocabulary, however it may have been discovered, data providers need to know if the vocabulary is widely used. There is a bootstrapping issue here, since publishers of vocabularies of obvious quality often complain that similar vocabularies of poorer quality are used more widely. Early users of a vocabulary should be able to promote further use of a new vocabulary which they have found relevant to their needs.
- *Enable data providers to discover which vocabularies are used, and with what patterns, in similar datasets*. The vocabularies used by datasets can be indicated in VoID (http://vocab.deri.ie/void) descriptions, such as those provided by CKAN. But such descriptions do not detail, for example, which terms are most frequently used or with which patterns. The definition of Application Profiles is an important way forward, but this approach has hitherto been used with small set of vocabularies, mainly around Dublin Core.
- Enable vocabulary designers to find terms related to terms in their own vocabulary. Enabling such discovery has been the main focus of the LOV project, both by making vocabulary-level links explicit using the VOAF (http://purl.org/vocommons/voaf) vocabulary (e.g. "relies on," "specifies," "extends") and by enabling search across the aggregated database of hundreds of vocabularies.

One component of the Datalift platform enables discovery of potential mappings from a local vocabulary to equivalent terms in the LOV cloud.

- *Helping vocabulary maintainers describe their own vocabularies*. Most vocabularies now submitted to LOV provide more detailed metadata, examples of good practices are spreading, and the overall quality of metadata in the ecosystem is improving. The next step should be to develop and promote an application profile for describing vocabularies. The profile should use terms from general vocabularies with more specific terms from vocabularies such as VOAF, as recommended by the LOV Project.
- *Understanding relationships between vocabularies*. The LOD Project uses the Vocabulary of Interlinked Datasets (VoID) to describe relationships between datasets and the vocabularies they use (Alexander and Hausenblas, 2009). Similarly, VOAF has ways to express relationships among vocabularies, such as specialization, generalization, extension, equivalence, and disjunction. These relationships are computed using SPARQL (www.w3.org/TR/rdf-sparql-query/) queries over the content of vocabularies to evaluate links between individual terms. This approach allows one to visualize dependencies, and the importance of a given vocabulary in the ecosystem can be assessed from the number and type of incoming and outgoing links.

A sustainable business model for those monitoring services has yet to be found. Most existing services have been developed in the framework of projects of limited duration compared to the long-term perspective adopted here. Transforming those projects into sustainable services will be a challenge for the semantic web community.

Sustainable governance for long-term preservation

As a foundation for data sources meant to be usable in the long term, the value of any given vocabulary depends on the perceived certainty that the vocabulary – in both its machine-readable and human-readable forms – will remain reliably accessible over time and that its URIs will not be sold, re-purposed, or simply forgotten. Vocabulary maintainers move on to other projects or retire. Resources owned by institutions may be neglected or become unavailable. As the givers of meaning to datasets, vocabularies are of vital importance to the scholarly record and cultural memory. However, their preservation will not happen automatically; it must be planned. The requirements for long-term preservation must consider a timeframe that stretches beyond the planning horizon of any institution that exists today.

- *Institutional guarantees for the persistence of URIs*. One good first step is for owners of URI domains to publish a commitment that any URI coined for a term in a vocabulary will be used to refer to the same term in perpetuity and will not be repurposed. This is related to, but transcends, the requirement for semantic change policies articulated above.
- *Persistence of documentation*. Each term URI should remain resolvable to "namespace documents" – descriptive documentation in HTML and/or machine-readable representations such as RDF schemas. Note that "persistent" URIs may redirect to documentation held at non-persistent locations. URIs and

the associated documentation remain "persistent" to the extent that the link between the two is maintained as documentation is moved between servers.

- *Persistence of institutional support*. No institution on the planet can possibly guarantee, today, that it will able to honor a preservation commitment of decades – even in the case of the most well-endowed and politically secure national libraries. If so, then the problem of institutional commitments becomes one of creating commitments among vocabulary maintainers. The Friend of a Friend (FOAF) Project, for example – a vocabulary maintained by two private individuals – has a cooperation agreement with the Dublin Core Metadata Initiative (DCMI) whereby DCMI maintains an up-to-date snapshot of the FOAF vocabulary, may temporarily host the vocabulary by redirecting FOAF URIs to the DCMI website if needed, and could assume maintenance responsibility for the vocabulary if the FOAF Project should cease its normal activity (Brickley *et al.*, 2011a). The intent of this agreement is in part to promote the idea of long-term planning in the vocabulary maintenance community and to affirm best-practice principles and policies for RDF Vocabularies (Baker *et al.*, 2011).
- *Standard, well-understood bundles of commitments*. Creative commons took off, as an idea, when it provided a range of menu of standard, well-understood contracts detailing bundles of legal choices related to copyright. Imagine a menu of commitments about rights and duties regarding rapid interventions (e.g., redirecting URIs if disaster strikes) and the transfer of internet domain names and of maintenance responsibility in the long-term. The involvement of major players in the vocabulary ecosystem could help ensure, at a minimum, that the popular vocabularies have such mechanisms in place.
- *The principle of "safety through redundancy"*. The general principle is summarized in the acronym for the LOCKSS service: "Lots of copies keep stuff safe" (Maniatis *et al.*, 2005). This can be achieved by mirroring information caches among multiple repositories. This is analogous to how the internet's domain name system itself is cached, or indeed how living organisms ensure the survival of their genes by copying their genetic information into new carriers rather than by defending a single, ultimately mortal, cache of information.
- *Redundant caching of vocabularies*. Because the software used for LOCKSS provides an automated system for sharing caches of digital content within a secure, closed peer-to-peer network – an idea that has already been implemented successfully to preserve digitally curated journal holdings – it has been suggested that this software, or something like it, could be used to provide mirrored caches of vocabularies (Halpin and Baker, 2010). The LOCKSS system monitors the integrity of a local information cache by continually comparing the local cache to exact copies of that cache held within a closed system of partner institutions and sounding an alarm if discrepancies arise. It is mathematically very improbable that a even a fairly small number of independently maintained information caches, e.g. seven, should be compromised by server failures or deliberate manipulation in a way that precludes diagnosis and repair.
- *Flexible cooperation among memory institutions*. Implementing a redundancy strategy for the long-term preservation of vocabularies implies a flexible form of governance among memory institutions. A robust preservation strategy would

embrace memory institutions large and small, in countries rich and poor, from the politically guaranteed to the politically more fragile and provisional.

Conclusion

As a foundation for linked open data, the value of any given linked open vocabulary depends on the perceived certainty that the vocabulary and the URIs used to identify its terms, will remain reliably accessible over time. Drawing on long-standing discussions in the W3C and Dublin Core communities, the academic literature on ontology engineering, and two years of experience in the LOV Project, this paper proposes a set of requirements for vocabulary preservation and governance.

Under the heading "Act locally", some of these requirements are addressed to individual vocabulary maintainers: that each term in a vocabulary be citable by a URI and resolvable to a formal, machine-readable representation of its meaning, and that information ("metadata") be made available about the maintainers of a vocabulary, related maintenance, copyright, and versioning policies, and the definitions of its terms be made available in different natural languages. Vocabulary maintainers are encouraged to "reuse, reduce, and recycle" – where possible, to re-use existing, well-known vocabularies, respecting the definitions declared by their maintainers, and to map the terms of their vocabulary to related terms in other vocabularies. To meet these requirements, software developers and institutional projects should focus on tools and methods for their validation. This can be easily achieved at a technical level by defining incremental targets for each requirement (similarly to the 5-star levels for Linked Data) and by writing queries to verify compliance with a requirement (e.g. detecting the presence of metadata or assessing the connectivity of a vocabulary).

Under the heading "Think globally," other requirements relate to the governance of the vocabulary ecosystem as a whole: to the availability of maintainer-provided descriptions, of empirical analytics about the use of vocabularies in the wild, and of services for exploring usage patterns, relationships between terms, and relationships between vocabularies. The paper proposes an expanded role for memory institutions in guaranteeing the persistence of vocabularies with transparent commitments and cooperative strategies for data redundancy. Meeting these requirements requires the vocabulary community to collaborate on vocabulary governance policy including cooperation agreements among maintainers and memory institutions (following the DCMI-FOAF example), and to support existing tools for monitoring and maintaining the vocabulary ecosystem. Evaluating the set of requirements and recommendations formulated in this section is difficult as it involves a large and diverse set of stakeholders. Monitoring the vocabulary ecosystem is of prime importance but further methods to assessing sustainability (e.g. dedicated metadata vocabulary pointing to sustainable policy documents) are needed.

Many aspects of the strategy proposed here are not unique to the problem of preserving RDF vocabularies but hold for many other types of information. However, compared to the much larger problem of preserving the human record as a whole, the problem of preserving several thousand RDF property-and-class vocabularies is comparatively more tractable. The goal is worthwhile and practical both in its own right and potentially as a first step towards addressing preservation goals more generally.

References

Alexander, K. and Hausenblas, M. (2009), "Describing linked datasets-on the design and usage of void, the vocabulary of interlinked datasets", paper presented at Linked Data on the Web Workshop (LDOW 09), in conjunction with 18th International World Wide Web Conference (WWW 09).

Baker, T., Brickley, D. and Miller, L. (2011), "DCMI generic namespace policy for RDF vocabularies", available at: http://dublincore.org/documents/dcmi-namespace-generic/

Berners-Lee, T. (2006), "Linked data", available at: www.w3.org/DesignIssues/LinkedData.html

Berrueta, D. and Phipps, J. (2008), "Best practice recipes for publishing RDF vocabularies", available at: www.w3.org/TR/2008/NOTE-swbp-vocab-pub-20080828/

Brickley, D., Miller, L. and Baker, T. (2011a), "Agreement between DCMI and the FOAF Project", available at: http://dublincore.org/documents/2011/05/02/dcmi-foaf/

Brickley, D., Miller, L. and Baker, T. (2011b), "Agreement between DCMI Generic Namespace Policies for RDF Vocabularies", available at: http://dublincore.org/documents/2011/05/02/dcmi-namespace-generic/

Cheng, G., Ge, W. and Qu, Y. (2008), "Falcons: searching and browsing entities on the semantic web", *Proceedings of the 17th International Conference on World Wide Web*, ACM, New York, NY, pp. 1101-1102.

d'Aquin, M. and Motta, E. (2011), "Watson, more than a semantic web search engine", *Semantic Web*, Vol. 2 No. 1, pp. 55-63.

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y. and Sachs, J. (2004), "Swoogle: a search and metadata engine for the semantic web", *Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management*, ACM, New York, NY, pp. 652-659).

Halpin, H. and Baker, T. (2010), "Vocabulary hosting: a modest proposal", paper presented at AAAI Spring Symposium on Linked Data Meets Artificial Intelligence, Stanford University, Stanford, CA, available at: www.aaai.org/ocs/index.php/SSS/SSS10/paper/view/1140/1450

Heath, T. and Bizer, C. (2011), "Linked data: evolving the web into a global data space", *Synthesis Lectures on the Semantic Web: Theory and Technology*, Vol. 1 No. 1, pp. 1-136.

Isaac, A., Waites, W., Young, J. and Zeng, M. (2011), "Library Linked Data Incubator Group: datasets, value vocabularies, and metadata element sets", W3C Incubator Group Report, 25 October, available at: www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/

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.*, Vol. 23 No. 1, pp. 2-50.

Miles, A. and Bechhofer, S. (2009), *SKOS Simple Knowledge Organization System Reference*, available at: www.w3.org/TR/skos-reference/

Powell, A., Wagner, H., Weibel, S., Baker, T., Matola, T., Miller, E. and Johnston, P. (2007), "Namespace policy for the Dublin Core Metadata Initiative (DCMI)", available at: http://dublincore.org/documents/dcmi-namespace/

Scharffe, F., Atemezing, G., Troncy, R., Gandon, F., Villata, S., Bucher, B., Vandenbussche, P-Y. and Vatant, B. (2012), "Enabling linked-data publication with the datalift platform", *Proc. AAAI Workshop on Semantic Cities, July*.

Suárez-Figueroa, M.C., García-Castro, R., Villazón-Terrazas, B. and Gómez-Pérez, A. (2011), "Essentials in ontology engineering: methodologies, languages, and tools", *Session: Ontological Engineering State of the Art*, p. 9-21.

Sure, Y., Staab, S. and Studer, R. (2009), "Ontology engineering methodology", *Handbook on Ontologies*, Springer, New York, NY, pp. 135-152.

Tummarello, G., Delbru, R. and Oren, E. (2007), "Sindice.com: weaving the open linked data", *The Semantic Web, Lecture Notes in Computer Science*, Vol. 4825, pp. 552-565.

Vandenbussche, P-Y. and Vatant, B. (2012), "Metadata recommendations for linked open data vocabularies", available at: http://lov.okfn.org/dataset/lov/Recommendations_Vocabulary_Design.pdf

Further reading

Nagamori, M., Kanzaki, M., Torigoshi, N. and Sugimoto, S. (2011), "Meta-bridge: a development of metadata information infrastructure in Japan", *Proceedings of the DCMI International Conference on Dublin Core and Metadata Applications*, available at: http://dcpapers.dublincore.org/pubs/article/view/3632?show

Appendix. DCMI generic namespace policy for RDF vocabularies

RDF vocabularies require:

- *Use of URIs*. Each term (e.g., property or class) in an RDF vocabulary must be identified with a URI.
- *Stable identifiers*. Each term URI should refer to the same term in perpetuity and should not be repurposed, sold, or forfeited through neglect (e.g., by non-payment of domain name fees). This commitment should be backed by institutional guarantees. Note that stable URIs (such as PURLs) do not automatically refer to stable documentation (such as content to which PURLs are redirected).
- *Machine-processable documentation*. Each term URI should remain resolvable to a machine-processable expression of its semantics in accordance with principles of web architecture. As of 2011, such expressions include RDF schemas and OWL ontologies and are documented in forms ranging from stand-alone schema files to formal representations embedded in web pages.
- *Change policies*. 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 individuals or organizations. Change histories should also be published so that a vocabulary's evolution over time is a matter of public record.
- *Open access provisions*. Vocabularies should be made available for public access under the terms of copyright models that encourage re-use and collaboration and with well-defined mechanisms for community feedback.
- *Preservation provisions*. As for any other artifact of cultural and historical significance, arrangements should be made for the long-term preservation of a vocabulary.

Subscribers to these principles:

- FOAF Project.
- Dublin Core Metadata Initiative.

Source: Brickley *et al.* (2011b).

Corresponding author

Pierre-Yves Vandenbussche can be contacted at: py.vandenbussche@gmail.com