Szostak, R. (2014). Advances in Classification Research Online 2013 Classification, Ontology, and the Semantic Web.
Advances In Classification Research Online, 24(1). doi:10.7152/acro.v24i1.14674

Advances in Classification Research Online 2013 Classification, Ontology, and the Semantic Web

Rick Szostak
University of Alberta
Tory Building 9-18, Edmonton, AB, T6G 2H4, CANADA
rszostak@ualberta.ca

ABSTRACT

The Semantic Web is developing slowly, but arguably surely. Two inter-related sources of delay are network effects and ontologies. The Semantic Web has come over time to rely on formal ontologies but there are many of these and they are each hard to master. The ability to link databases is compromised by the use of incompatible ontologies. But the RDF triplet format at the centre of the Semantic Web insists only on triplets of the form (object)(predicate or property)(subject). This paper explores the potential for a classification system that contains these three types of hierarchies (things, predicates, properties), plus a minimal set of rules on how they can be combined, to serve the needs of the Semantic Web. To this end, it surveys the roles (both the intended roles and side-effects) that formal ontologies play within the Semantic Web. The paper also briefly reviews the challenges faced in applying existing classification systems or thesauri to the Semantic Web.

Keywords

Classification, ontology, semantic web, RDF triplets, basic concepts.

INTRODUCTION

This may be a moment in time – much like the late nineteenth century – when approaches to classification are developed for a particular environment but have long-lasting impacts. Work on the Semantic Web has been dominated by IT professionals. Yet there appears to be an important role for input from experts on classification *at precisely this point in time* to ensure that the Semantic Web evolves in a manner that reflects our understanding of how best to classify. Classifications developed for the Semantic Web are likely to have a pervasive influence in the future, even in areas dominated at present by information scientists.

We should in particular want the Semantic Web to be open to diversity of all sorts. The web-of-relations approach advocated by Olson (2007) in the interests of social diversity involves fortuitously precisely the components that we shall find necessary for the Semantic Web (Szostak 2013a). The purpose of the Semantic Web, after all, is to find new connections.

The paper begins by discussing the RDF triplets that lie at the heart of the Semantic Web. It is argued that these are best facilitated by a particular approach to classification. The next section reviews the roles that ontologies must play on the Semantic Web, and argues that these might be accomplished by adding some straightforward syntactic rules to the sort of classification recommended in the first section. The third section briefly reviews some of the problems associated with ontology at present, and notes that these would be obviated by the approach recommended here. A brief concluding section follows.

RDF TRIPLETS ARE THE KEY

The purpose of the Semantic Web is to allow computers to make connections across different databases. If one website says that swans are black, and another that Fred is a swan, the computer can deduce that Fred is black. For this sort of computer inference to be possible, all databases need to be coded in terms of what are called RDF triplets. These take the form (object) (predicate or property)(subject). Thus we can code (swans)(are)(black) and (Fred)(is a)(swan).

The first critical point to appreciate is that databases have to be coded. The classification community has worried for decades that full-text searching might supplant classification as the key approach to information retrieval. The Semantic Web does not rely on full-text searching but rather purposeful coding. Indeed much of the original impetus for the Semantic Web reflected concerns that full-text searching and internet search algorithms focusing on links between websites failed to identify many valuable information sources. [There are programs for extracting RDF triplets from a text but these are imprecise.]

The second critical point follows from the first. The ability of computers to draws inferences across databases will depend entirely on controlled vocabulary. If the two databases in our example had used different terms for swan (perhaps one employed the scientific name), the computer could only draw the inference that Fred is black if it knew that the two terms were equivalent. The success of the Semantic Web, then, depends entirely on the acceptance of a unique controlled vocabulary and/or the development of a thesaurus that can seamlessly translate across different controlled vocabularies. If different databases employ incompatible vocabularies, computer inference will be impossible. The development of the Semantic Web thus heralds a renaissance in the role of controlled vocabulary. Information scientists have displayed much interest in putting bibliographic information in RDF format, but have been concerned by the "messiness" of terminology on the Semantic Web (Pattuelli and Rubinow, 2013).

A third critical point follows from the first two, but seems little appreciated in either the Semantic Web or classification literatures. Due to the form that RDF triplets take, the controlled vocabulary required by the Semantic Web involves three types of concept: the things that can serve as object or subject in RDF triplets; predicates (relationships) that can link subjects and objects; and properties that can be ascribed to things. If the classification research community wishes to contribute to the development of the Semantic Web it should develop classifications of these three types of concept.

It should be noted that the "properties" referred to in the Semantic Web literature often mean what information scientists would consider a relationship. The "is a" relationship is conceived as a property. But it is also true that the objects in an RDF triplet are often adjectival properties. And thus it remains the case that RDF triplets call for separate classifications of things, relationships, and properties.

A fourth critical point also follows. The Semantic Web will depend on the free combination of things, properties, and relationships. To be sure, we will want to rule out false statements: if all swans really were black we would not allow (swans)(are)(white). But the point here is that we would want such limitations to be inspired by our understanding of the world, not artificially constrained by some classificatory structure.

The author should declare his bias here. He has for some time been developing a classification system for use in libraries that involves precisely the free combination of things, relationships, and properties (Szostak 2013b). It is his good fortune, perhaps, that a classification developed for library use seems admirably suited to the needs of the Semantic Web. But its suitability for that latter purpose flows naturally from the very nature of RDF triplets. And it is noteworthy that the sort of classification that facilitates the Semantic Web is also suitable to library classification.

There are reasons why this coincidence should not be a surprise. First, I have long appreciated that my classification might be particularly well suited to digital libraries, and that it is much more feasible in the digital age than it would have been in an era of card catalogues. I have also noted that it is well suited to the classification of ideas as well as documents. The Semantic Web aspires to link ideas. Second we should want library classifications, like any other database, to be computer searchable (DeRidder, 2007). And we should want bibliographic information to be integrated with other sorts of information on the Semantic Web. Third, the desire to facilitate interdisciplinary connections is a powerful motive behind both my classification and the Semantic Web.

A fifth critical point deserves mention. Existing classification systems such as Dewey (DDC) or Library of Congress (LCC) are enumerative, and thus contain mostly compound terms as entries. To use such classifications as controlled vocabularies for the Semantic Web would require that each entry be individually coded. That is, they would have to be broken into their constituent things, properties, and relationships. For the purposes of the Semantic Web a classification that already breaks complex concepts into basic concepts (a strategy advocated in Szostak 2011) is a far superior source of controlled vocabulary. [SKOS (Simple Knowledge Organization System) allows such classifications as LCC and DDC to be accessed on the semantic web, but hardly facilitates their application there. For example, SKOS does not distinguish "type of" from "part of" relationships, a distinction that we will find is of critical importance.]

ONTOLOGIES AND THE SEMANTIC WEB

As noted above, the Semantic Web requires not just a controlled vocabulary. If computers are to draw inferences across databases, they must be instructed regarding how concepts can be related. Hart and Dolbear (2013) note, for example, that it might be useful to tell the computer that rivers generally run into lakes or seas. Computers can then connect statements made about rivers and about lakes. It has been hoped that formal ontologies could provide both the controlled vocabulary and inferential rules required by the Semantic Web. But there are multiple formal ontologies grounded in different assumptions. At this point in time, there is little likelihood either that one (family of) ontology will become dominant, or that it will be easy to translate terminology across ontologies. Research in the Semantic Web field has indeed turned away from exploring ontologies (Hart and Dolbear 2013), though it is far from clear what if anything can replace them.

It therefore seems worthwhile to explore the possibility that the sort of classification system urged in the previous section – where classifications of things, relationships, and properties can be freely combined – could serve the functions that it was once hoped formal ontologies could serve for the Semantic Web. We have already argued that such a classification could potentially serve the controlled vocabulary needs of the Semantic Web. We should explore this argument in more detail, and can then proceed to examine the other purposes that ontologies were expected to fulfill.

In other words we should explore the possibility that a "building up" strategy can succeed where a "top down" strategy has not. Formal ontologies embed a host of rules and definitions in one complex structure. Any flaws in that complex structure limit the effectiveness of the Semantic Web. The alternative explored here is to start with a straightforward approach to classification grounded in the nature of RDF triplets, and then build onto this the minimum set of restrictions necessary for the Semantic Web,

Controlled Vocabulary

Ontologies, it was hoped, would give very precise definitions of all concepts employed. They would do so in large part by carefully describing hierarchical and other relationships between concepts. A logical hierarchy of things serves an important definitional role: it establishes precisely what sort of thing something is and what sort of thing it is not. But extant classifications often abuse hierarchy: recycling is treated as a subclass of garbage when it is rather something that we do to garbage (Mazzocchi et al., 2007). Freely combining things, relationships, and properties frees the classificationist from the temptation to abuse hierarchy. Computers can correctly infer that a subclass within a logical hierarchy where subdivision occurs in terms of “type of” or “example of” has the same characteristics as the broader class. Swans thus will have all the characteristics ascribed generally to birds or animals. Computers can be told when subdivision instead occurs with respect to “part of,” and can then make appropriate inferences there as well.

The role of hierarchy with respect to things can arguably be served by combinations with respect to relationships. Szostak (2012) showed how an exhaustive set of relationship concepts could be obtained by combining some 100 basic relationship concepts with each other or with things or properties. A computer could easily be programmed to appreciate these combinations. It could then infer that “run” is equivalent to “walk fast” and draw appropriate connections between statements about walking and statements about running.

Information scientists (including me) have devoted much less attention to properties than to things or relationships. The Basic Concepts Classification (Szostak 2013b) contains a class of properties. But these have been identified inductively, and only the most primitive attempts have been made to organize these concepts logically.

Though more research is called for, especially with respect to properties, it must seem at least possible that the controlled vocabulary needs of the Semantic Web can be served by logical hierarchies of things, well-defined combinations of relationships, and some logical treatment of properties.

The classification itself might then be supplemented by an exhaustive thesaurus that translated all concepts into controlled vocabulary (or perhaps allowed equivalent terms to be employed as controlled vocabulary). Note that identifying equivalent terms could be very useful but that vaguely identifying “related terms” is not.

Syntactic Relations

It should first be stressed that there is a cost associated with placing unwarranted restrictions on how concepts can be combined. If some swans are white then the inference that Fred is black will be mistaken. If some rivers disappear underground or die in the desert, assuming a connection with lakes or seas will be misleading. It follows that formal ontologies may inadvertently lead to mistaken inferences if they impose restrictions that do not precisely reflect restrictions in the real world.

Nor is this a trivial concern. The literatures on undiscovered public knowledge, literature-based discovery, and serendipity extol the advantages of combining pieces of information that have never previously been juxtaposed. And the Semantic Web itself is expected to achieve this sort of connection (though these three literatures are rarely referenced by writers on the Semantic Web). Falsely limiting the possibility that two concepts can be connected will prevent a subset of valuable juxtapositions from being discovered.

There are also, of course, costs of under-constraining connections. Information scientists have long appreciated that getting numerous false hits is a problem, though perhaps less problematic than missing important sources of information. A prudent strategy for the Semantic Web would seem to involve building up individual restrictions one-by-one, taking care that each restriction accurately reflects the way the world works. The formal ontology approach imposes a set of restrictions at the outset. If we start from a classification, and add restrictions as necessary, we take the prudent approach.

So what sort of inferential rules are necessary for the semantic web?

Hierarchy

Hierarchy is stressed in the Semantic Web literature. We want the computer to infer that all characteristics associated with animals in general are applied also to subclasses of animal. As noted above, we need then to insist rigidly on logical hierarchy. And we need also to distinguish “type of” subdivision from “parts of” subdivision (which is a property rather than a hierarchic relation on the Semantic Web). The Semantic Web stresses “type of” hierarchy.

A classification that employs a strictly logical "type of" approach to classifying things will thus admirably serve the inferential as well as definitional needs of the Semantic Web. Cases of "part of" subdivision can be clearly distinguished.

The combinatory approach to relationships suggested above will likewise serve both inferential and definitional purposes. A computer told that walking involves moving ones legs can infer that running likewise involves moving ones legs, albeit faster.

Class Distinctions

It may be useful to identify the difference between subclasses (say, creek and river). This a classification alone cannot do. But it may prove relatively straightforward in many cases to identify class distinctions (creeks have less water flow than rivers).

It is harder to identify the differences between, say, cats and dogs. But many of these differences will be signaled by RDF triplets themselves. The computer may need to know little at the outset beyond the fact that they are different kinds of animal. And the classification itself tells the computer that dogs and cats are different kinds of animal.

Causal Connections

Hart and Dolbear (2013) give the example "A weir is a form of flood defence." Such information allows the computer to infer something about flood defences from data on weirs. They appreciate that weirs are not the only form of flood defence. They likely also appreciate, but do not state, that weirs can serve other purposes. Care would have to be taken to ensure that computers were not inadvertently programmed to ignore these other purposes. It is certainly possibly to employ RDF triplets to express "Weirs can serve as flood defence" and also "Weirs can create reservoirs."

One question that arises here is how much of this sort of information needs to be explicitly programmed at the outset. A computer trawling the internet will presumably find many references to weirs preventing floods and also doing other things. These will be captured by the RDF triplets associated with various databases. As long as we have solved controlled vocabulary challenges, the computer may be able to identify causal relationships unaided.

And this is critical for the process of discovery. There may be other physical features out there that serve an important flood-control role but indirectly. Computers are well-suited to appreciating that an argument in one database that A influences B can be connected to an argument elsewhere that B influences C in order to generate an appreciation that A exerts an important but indirect influence on C.

It is an open question whether we want to effectively prioritize certain causal relations by programming these into computers before they search databases. If so it is certainly possible to do so. The alternative is to set computers with a certain research task (what affects C?) and let the RDF triplets out there in the world guide them to answers.

Likewise we might wonder how much it is necessary to include restrictions at the outset. We know that dogs cannot breathe underwater. But if no set of RDF triplets would imply such a thing, there is no value in forbidding the connection from being made.

Of course, in the real world, many websites do say things that are untrue. We might need to use some probabilistic algorithm to dismiss connections posited by a small minority of sites. But we would then risk losing some important insights that are only rarely appreciated.

Properties

Which properties can a particular thing possess? If we are able to achieve small schedules of both things and properties (and Szostak, 2013b, suggests that this is the case, at least for human science), it would be quite feasible to identify which properties can be attached to which things. We would want to be very careful that we did not accidently prohibit a combination that exists in the world. And again we have to wonder if computers can infer which combinations are feasible from RDF triplets themselves.

Definitions

As noted above, much effort in formal ontologies is devoted to providing precise definitions of each term. This effort could be derided by those who, following Wittgenstein, appreciate that the sort of precision being sought is in fact impossible. There is nevertheless some advantage in defining terms. The computer can only draw correct inferences if all databases are employing concepts in a similar manner, and thus those ascribing RDF triplets to diverse databases need a shared understanding of the meaning of concepts. One advantage of classifying basic concepts – the things, relationships, and properties that we perceive in the world around us – is that it is much easier to achieve broadly shared understandings of what each concept means. And if we insist on logical hierarchy for things, and combinations for relationships, and develop some logical approach for properties, the definitional challenge is further limited: many terms can be defined well enough as combinations of or types of other well-defined terms. As noted above, subclasses are defined in important ways simply by classifying: we know what kind of thing they are and what kind they are not.

Though the people coding RDF triplets need some idea of what terms in the controlled vocabulary mean (and we can note that it is quite possible to add scope notes within the RDF approach), it is not clear how much definition the computers trolling the Semantic Web need. In our example above it was quite possible to deduce that Fred is black without knowing what a swan is.

Indeed the Semantic Web has often been criticized for not really being about semantics, which (in philosophy at least) refers to how linguistic units relate to the real world. It might better be termed the "Syntactic Web" for it focuses on how linguistic units relate to each other [as we have in this paper]. Though the impetus for the Semantic Web (and thus its name) may have reflected a sense that computers needed to understand semantics in order to be able to draw inferences (especially from natural language), it has evolved in a manner that emphasizes instead identifying different types of links between concepts (Guns 2013).

Inverses, Symmetry, Transitivity

It is useful to program inverses: "own" is the "inverse of owned by." This is easily done. Indeed the Basic Concepts Classification (Szostak 2013b) already codes for inverses, and for the same reason: so that "Bill owns that truck" is treated identically to "That truck is owned by Bill." The same holds for symmetry: "Bill is next to the truck" should be and is treated identically to "The truck is next to Bill." As for transitivity, we want the computer to appreciate that if A is bigger than B and B is bigger than C that A must be bigger than C. This requires only that we designate which properties or predicates are transitive.

Summary

It seems quite feasible to add the few syntactic rules necessary for the Semantic Web to a classification that provides the necessary controlled vocabulary of things, relationships, and properties. This will be especially the case if we are able to allow the computer to infer some of these from the universe of RDF triplets itself.

CHALLENGES IN EMPLOYING ONTOLOGIES

What challenges are faced at present in the application of ontologies on the Semantic Web? We have already addressed the most important challenge above: that there are a host of ontologies to choose from, and that since these employ different starting assumptions it is not easy to translate across these. The negative implication for the Semantic Web is severe. It is impossible for a computer to draw connections across databases employing incompatible ontologies. The approach recommended in this paper, of adding as few syntactic rules as possible onto an easily-understood classification, holds out hope of allowing all databases to be connected.

The second challenge is that the terminology employed in especially upper-level (that is, general) ontologies is often frustratingly vague. "There are, however, some drawbacks to using upper ontologies, not least because it can be very difficult for an expert in a particular domain such as GI [geographic information] to understand exactly which of the oddly termed classifications to assign to their concepts. Should a County be classed as a Physical Region or a Political Geographic Object? Is a flood an endurant or a perdurant? It depends on your point of view. These quandaries become even more apparent when confronted with terms like 'Non-Agentive Social Object' or 'Abstract'." (Hart and Dolbear 2013, 13-4). The classification recommended in this paper is grounded in the things, relationships, and properties that we perceive in the world around us. It is simply not necessary to resort to vague terminology.

Hart and Dolbear also note that every constraint imposed both slows the inferential process of the computer and increases the chances of programming error. These concerns reinforce the argument made above that we should be prudent in imposing restrictions. And Hart and Dolbear stress that different constraints may be useful for different queries. This indicates that we should limit the constraints imposed on the Semantic Web as a whole.

CONCLUDING REMARKS

The development of the Semantic Web is hobbled at present by the absence of an agreed-upon controlled vocabulary and set of syntactic rules. Yet the potential value of the Semantic Web, and the number of researchers pursuing it, is so large as to suggest that this hurdle will one day be overcome. The question is how. This paper has argued that classification researchers have an opportunity at this point in time to shape the structure of the Semantic Web. And the Semantic Web itself is likely to shape approaches to classification far into the future. It is critically important that developers of the Semantic Web be guided by the expertise of the information science community.

Classification for the Semantic Web must accord with the format of RDF triplets. This means the separate classification of things, relationships, and properties. These can then be freely combined in RDF triplets, with the imposition of a (hopefully limited) set of syntactic constraints. This paper has made the bold suggestion that the present "top-down" strategy of imposing formal ontologies on the Semantic Web, which has proven highly problematic, be replaced by a "building-up" strategy of developing classifications of things, relationships, and properties, and then adding constraints as necessary.

REFERENCES

DeRidder, J.L. (2007) The immediate prospects for the application of ontologies in digital libraries, *Knowledge Organization* 34:4, 227-46.

Guns, R. (2013) Tracing the Origins of the Semantic Web. *Journal of the American Society for Information Science and Technology* 64:10, 2173-81.

Hart, G., and C. Dolbear (2013) *Linked Data: A Geographic Perspective.* CRC Press.

Mazzocchi, F., Tiberi, M., De Santis, B., & Plini, P. (2007) Relational semantics in thesauri: some remarks at theoretical and practical levels. *Knowledge Organization 34:4,* 197-214.

Olson, H. (2007) **How we construct subjects: A feminist analysis,** *Library Trends* 56:2, 509-41.

Pattuelli, M.C., and S. Rubinow (2013) The knowledge organization of DBpedia: A case study, *Journal of Documentation* 69:6.

SKOS (Simple Knowledge Organization System) Retrieved September 18, 2013 from http://www.w3.org/TR/skos-reference/#collections

Szostak, R. (2013a) Classifying for diversity. Paper prepared for NASKO conference, June. Under review.

Szostak, R. (2013b) *Basic Concepts Classification*. Retrieved May 30, 2013 from https://sites.google.com/a/ualberta.ca/rick-szostak/research/basic-concepts-classification-web-version-2013

Szostak, R. (2012) Classifying relationships, *Knowledge Organization* 39:3, 165-78.

Szostak, R. (2011) Complex concepts into basic concepts, *Journal of the American Society for Information Science & Technology* 62:11, 2247-65.