The Basic Concepts Classification as a Bottom-Up Strategy for the Semantic Web

Rick Szostak[*]

ARTICLE INFO

Article history:
Received 18 March 2014
Revised 26 May 2014
Accepted 1 June 2014

Keywords:
Classification, Ontology, Semantic Web, RDF Triples, Basic Concepts

ABSTRACT

The paper proposes that the Basic Concepts Classification (BCC) could serve as the controlled vocabulary for the Semantic Web. The BCC uses a synthetic approach among classes of things, relators, and properties. These are precisely the sort of concepts required by RDF triples. The BCC also addresses some of the syntactic needs of the Semantic Web. Others could be added to the BCC in a bottom-up process that carefully evaluates the costs, benefits, and best format for each rule considered.

1. Introduction

The Semantic Web requires a controlled vocabulary – that is, some well-defined set of concepts to be employed in RDF triples – and a set of syntactic rules that enhance the ability of computers to draw inferences across databases. As in any network situation, the value of coding any database in terms of a particular controlled vocabulary and set of syntactic rules depends critically on the number of other databases using the same controlled vocabulary and syntactic rules in coding RDF triples. But the Semantic Web community is far from consensus. There are a host of ontologies to choose from. 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. While existing ontologies serve valuable purposes, it would be advantageous if it were possible to develop an ontology that was more widely applied.

Existing ontologies have generally been developed in a 'top-down' manner: core axioms are postulated from which a set of classes and syntactic rules governing relations among classes are derived. An alternative 'bottom-up' strategy is possible in which we start from a classification of the elements required for RDF triples and then add syntactic rules as necessary. A bottom-up strategy has the advantage that consensus can be sought one step at a time rather than all at once. This

[*] Department of Economics, University of Alberta, CANADA (rszostak@ualberta.ca)
International Journal of Knowledge Content Development & Technology, 4(1): 39-51, 2014.
http://dx.doi.org/10.5865/IJKCT.2014.4.1.039

may significantly increase the likelihood of achieving widespread ontological consensus.

A bottom-up strategy may have further advantages. The main classes in extant ontologies are often vague constructs reflecting core axioms rather than general understandings of the world (Hart and Dolbear, 2013). A bottom-up strategy can strive to ground all classes in shared experience of the world. Differences across ontologies generally reflect the fact that the core axioms of particular ontologies are controversial (Masolo et al., 2003). A bottom-up approach does not require such axioms. Last but not least, a bottom-up ontology grounded in the shared human experience of the world could prove much easier to master and apply to diverse databases than extant ontologies.

This paper outlines a particular bottom-up strategy. The next section describes the Basic Concepts Classification (hereafter BCC; Szostak, 2013a). This classification has been developed with the original purpose of classifying documents (for libraries), or materials in museums and archives. But from the start it was also hoped that the classification could also serve to classify ideas. Happily, two key characteristics of the BCC make it extremely well-suited to the needs of the Semantic Web. Since this paper employs a bottom-up strategy, we will in this paper outline the advantages of this approach solely in terms of RDF triples, the basic building block of the Semantic Web:

- The BCC involves separate classifications of the things that we observe in the world, the relationships that exist among things, and the properties of things and relationships. These can then be freely linked to identify any work, object or idea. It will be argued that this approach is admirably suited to providing the controlled vocabulary for RDF triples.
- These classifications of things, relationships, and properties are each performed in terms of 'basic concepts.' As argued in Szostak (2011) the complex concepts that are understood differently across groups and individuals can be broken into basic concepts for which a broadly shared understanding is possible. It is thus much easier to achieve consensus around a shared controlled vocabulary at the level of basic concepts than complex concepts. But complex concepts can be conveyed by combining these basic concepts. It will be argued that this approach also is well-suited to RDF triples.

The subsequent section then explores how the syntactic rules necessary for the Semantic Web could then be added to the BCC. The bottom-up approach allows each type of syntactic rule to be addressed in turn. What is the purpose of particular syntactic rules?; how can they best be associated with the BCC?; and what are the advantages of this particular approach? We will explore in turn hierarchy, class distinctions, causal relations, properties, definitions, inverses, symmetry, and transitivity. This list is not exhaustive, but does reflect the types of syntactic rules stressed in the literature. One question that will be raised more than once is whether particular rules need to be imposed on the Semantic Web or could in fact be derived from the universe of RDF triples itself.

A brief concluding section follows.

2. The Basic Concepts Classification and RDF Triples

The BCC has been developed over the last decade. Over 20 books and articles that provide a philosophical justification for the BCC, evaluations of its desirability and feasibility, and outlines of its structure are described in Szostak (2013a). The original motivation for the BCC was interdisciplinarity: existing systems of classification in the world place many unnecessary barriers in the way of interdisciplinary scholarship. They likewise hobble the general user who wishes to follow their curiosity from one topic to a related topic without needing to master an arcane set of subject hierarchies. An interdisciplinary mindset also motivates the Semantic Web: the hope is that inferences can be drawn across databases developed for quite different purposes by individuals or groups with quite different goals, expertise, and worldviews. It should thus not be surprising that the BCC is well suited to the needs of the Semantic Web.

Moreover, the BCC was designed from the start for the digital age. It reflects the belief that any classification should be susceptible to computer searching (DeRidder, 2007). A disciplinary expert does not need much from a classification system, for they are familiar with the terminology and key journals of their field. An interdisciplinary scholar needs more: they may want to know not just the multiple things, relationships, and properties addressed in a work − and importantly how these are connected to each other − but the theories, methods, and perspectives employed, the disciplinary affiliation of authors, and perhaps more. Classifications developed for an age of card catalogues could hardly provide all of this information, but a classification designed for the digital age can. There is again a broad similarity with the goals of the Semantic Web: it is hoped that RDF triples can capture the unique insights, ideas, or information within any database, such that these can be related to the unique elements of any other databases.

RDF triples take the form (subject)(predicate or property)(object). They can thus make two broad types of statements: "X has property Y"; or "X is related in manner N to Z." X and Z in these statements refer to things in the world. N refers to a relationship that exists between things. Y refers to properties of things. We might also wish to allow statements of the form "Relationship N has property Y." The controlled vocabulary needs of the Semantic Web are thus threefold:

- Things in the world.
- Relationships among things
- Properties of things (and perhaps relationships)

The challenge is to develop classifications of each of these that allow the full range of RDF triples to be expressed in a manner that can be understood by both human users and computers.

The BCC itself is also outlined in Szostak (2013a). As noted above, it contains separate classifications of things, relationships, and properties. The classification of things is grounded in Szostak (2003), a book that showed how scholarly works in the social sciences and humanities could be understood as reflecting relationships among some thousand things (phenomena). These things were organized hierarchically within ten broad categories. This original classification of things was developed using a mix of deduction and induction: a logical structure was developed and this was adjusted and expanded through referencing hundreds of works from across the social sciences and humanities. The classification has been further expanded and adjusted since then (and is being loaded onto the Protegé ontology editor). In particular the natural sciences are now embraced, though work remains to be done there. Though some elements of the classification must be tentative – psychologists still debate the best way to classify some personality traits – most of the hierarchies represent consensus among scholars both within and across fields of study.

Texts and ideas in the social sciences and humanities can be classified with reference to very compact hierarchical schedules of the things that are perceived or studied in these fields. The natural sciences present a greater challenge here. There are thousands upon thousands of chemical compounds. Potentially these can each be signified through reference to chemical formulae and other notations developed by chemists. The millions of species create even greater challenges (especially as millions may remain to be discovered), but advances in genetic analysis suggest that within the next years there may be broad agreement on a hierarchical classification of many of these in terms primarily of genetic descent.

The classification of relationships was developed in Szostak (2012a, b). Again a mix of deduction and induction were used. The SUMO upper ontology, the Art and Architecture Thesaurus, and Wordnet were among the sources consulted. It was argued that a basic classification of some one hundred relators (both causal and non-causal) could generate, through combinations with other relators, things, or properties, the thousands of relators that are employed in the world. The hundred core relators are themselves grouped into about a dozen flat hierarchies. The BCC website (Szostak, 2013a) provides hundreds of examples of how these can be combined to generate further relators ("Some Compounds of Basic Concepts").

The classification of properties has largely occurred inductively. As (some 200) properties have been encountered they have been grouped into a couple of dozen classes.

The approach of the BCC was justified in Szostak (2011) in terms of the key philosophical concept theories. In particular the theory of conceptual atomism argues that we will have the greatest degree of shared understanding of concepts that refer to things and relationships (and also likely properties) that we observe in the world. The BCC thus stresses such concepts, and strives to capture more complex concepts through combinations of these.

Three key differences between the BCC and existing faceted classifications such as Colon, Bliss, or the Universal Decimal Classification deserve mention here. First, the BCC is not organized around disciplines. This characteristic greatly facilitates interdisciplinary exploration: all works about a particular thing or relator are classified in the same way. As noted above, an interdisciplinary impulse motivates the Semantic Web, and we will be better able to achieve cross-database inferences if different databases are not coded in different ways depending on the disciplinary inclinations of authors and managers. A second difference follows: the BCC allows any concepts to be freely combined whereas Colon, UDC, and Bliss each provide quite different rules for combinations within and across disciplinary classes. This should make the BCC easier for a computer to navigate, and does not arbitrarily privilege some combinations over others. A third difference involves facet indicators: while the BCC follows within the facet tradition – it appreciates that we wish to capture a set of possible facets of a work or idea – it eschews the use of facet indicators but relies instead on the logical structure of concept strings: the string (X)(causes in manner N)(Y) indicates clearly which is the 'agent', 'operator', and 'product'. Szostak (2013a) describes in detail how the 13 facets recognized in the Bliss Classification are each treated in BCC. This characteristic may be especially important for application to the Semantic Web for it allows works and ideas to be captured entirely through combinations of things, relators, and properties as in RDF triples (It also distinguishes the BCC from the Integrative Levels Classification; see ILC, 2014).

Most scholarly works, and likely most general works, investigate how one thing or set of things influence in a particular way a different thing or set of things. The best way to classify such works is thus to synthetically link things and relators: (chemical)(reduces)(blood pressure) or (dogs)(bite)(mail carriers). A minority of works describe the properties of a thing: (steel)(is)(strong). Such works are also best captured synthetically, this time by linking a thing and a property. Note that works are thus classified in terms of the key ideas or insights that they contain. The hope for the Semantic Web is that RDF triples can capture the key ideas or insights within diverse databases. It is no surprise that the best way of doing so is to link things with relators or properties. It should also be no surprise that a classification designed to do precisely that, albeit in a different environment, is well-suited to the needs of the Semantic Web.

The BCC thus has the potential to serve the controlled vocabulary needs of the Semantic Web. But are the things, relators, and properties in the BCC defined precisely enough? Much of the effort in developing formal ontologies has been devoted to providing very precise definitions of each concept employed. Computer inference, it is feared, depends on very precise definitions of terminology. The BCC is grounded in philosophical concept theory. Philosophers are far from consensus on the nature of concepts. In Szostak (2011) it was nevertheless shown that the BCC approach of breaking the complex concepts that are interpreted differently across disciplines or groups or individuals into basic concepts for which there are broadly shared understandings could be justified in terms of a broad range of concept theory. The theory of conceptual atomism suggests, in particular, that we will have the greatest sense of shared understanding of those concepts that signify the things and relationships [and properties] that we perceive in the world around us.

The concepts within the BCC are thus mostly if not entirely concepts for which humans will have a broadly shared understanding. The placement of these concepts within simple and logical hierarchies (an important characteristic in the next section of the paper) serves to further clarify meaning. Information scientists have worried that a choice must be made (in naming classes) between vague natural language and precisely defined artificial language (see Svenonius, 2004). It is argued in Szostak (2015) that we can have the best of both worlds with basic concepts: these are the concepts for which natural language is least ambiguous. We can thus anticipate that different humans coding RDF triples for different databases will employ these concepts in a similar fashion. Is this degree of similarity sufficient for drawing computer inferences across databases? While we should always be hesitant to draw empirical conclusions from theoretical conjectures, it seems likely that this would be the case. But we cannot reach a judgment on this matter without also exploring the syntactic rules necessary for computer inference.

Complex concepts are then communicated as necessary through combining basic concepts. This forces clarity. "Globalization" is a concept with diverse meanings. It might refer to how (trade) (increases)(incomes) or how (watching)(American)(movies) affects (French)(cultural attitudes). The terms in parentheses are much less ambiguous than globalization itself. And further clarification comes from identifying hundreds of particular cultural attitudes (and employing these rather than the broader term whenever a document or database is referring to a particular attitude or attitudes). The BCC approach would encourage RDF coding in terms of terminology for which there is shared understanding, and require that complex (and thus contested) terminology be built up through a combination of RDF triples.

The BCC has been tested in one important way. In Szostak (2013b) several thousand entries in the Dewey Decimal Classification (DDC) – as well as all classes in ICONCLASS – were translated into BCC terminology. This could always be done with a manageable set of basic concepts. When there was ambiguity in translation this could always be traced to vague DDC terminology. Quite often the BCC translation was far more precise than the DDC entry. Often DDC entries needed to be translated into multiple BCC entries.

This empirical exercise, while important (in particular for establishing that the BCC can handle the full range of works that the DDC strives to cope with), was biased against the BCC. As noted above, works are generally best handled through a synthetic approach to classification. The DDC, like other major classifications in widespread use, tries to identify a set of complex headings that can substitute for synthetic headings. The true advantage of the BCC is not captured by trying to translate DDC terms into BCC, but will only be established by showing that individual works are handled much better by BCC than other classifications.

The next tests of the BCC will thus focus on application to particular works or objects or ideas. One key question here would be how well the BCC lends itself to the coding of RDF triples. Though there are theoretical reasons to anticipate success, it could well be that adjustments need to be made.

Information scientists have worried a great deal in recent years about how to allow users to seamlessly navigate library classifications, archive catalogues, museum inventories, and websites of various types [This is the theme for the 2014 Dublin Core Metadata conference, and to some extent also of the 2014 ASIST conference.]. Users may find valuable information in each, but need at present to master different classification systems in order to access these different information sources. It is recognized that managers of these other information resources are unlikely to master the intricacies of extant library classifications. Information scientists have also 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). The BCC, with its synthetic approach, basic concepts, and compact hierarchies, might lend itself to employment across diverse settings. And if it can allow humans to easily code and search across diverse databases, it can be hoped that it will also facilitate computer navigation.

Ideally, the BCC would be supplemented by a thesaurus that translates other terms into BCC terminology. But this thesaurus would need to be different from the thesauri commonly developed within information science in important ways. Existing thesauri identify hierarchy (broader terms and narrower terms), but do not distinguish different types of hierarchical relationship as the Semantic Web requires. The other common indicator in thesauri is "related term." This vague descriptor will be of limited use to a human coding RDF triples and of no use whatsoever to a computer seeking inferences. We need to identify equivalent terminology, and it may be desirable to identify degrees of equivalency. Perhaps most importantly, we will often wish to translate other terms into combinations of BCC concepts.

3. Syntactic Relations

If computers are to draw inferences across RDF triples, they will need some instructions on what connections it is desirable to draw. It should first be stressed that there is a cost associated with placing unwarranted restrictions on how concepts can be combined. The most obvious cost occurs if we inadvertently place false restrictions. Ideally the Semantic Web will allow computers to draw inferences that had not previously occurred to any human actor. Computers will do this by juxtaposing insights that have never previously been juxtaposed. Within the information science community, the value of drawing novel connections between extant ideas is stressed in the literatures on “undiscovered public knowledge,” “literature-based discovery,” and “serendipity.” Information scientists thus encourage the development of classification systems that can guide users not just to the information that they know to seek but also to related information that they would not have known to look for (Davies, 1989). It turns out that a synthetic approach to classification is beneficial in this respect, for a user interested in how X affects Y in manner Z can then easily explore information about X, Y, and Z in other contexts. The Semantic Web may greatly accelerate the rate of “serendipitous” discoveries. But if syntactic rules arbitrarily prevent certain connections from being drawn then some such discoveries will not occur.

A second cost occurs in terms of computing time and cost. Though syntactic restrictions could potentially reduce computing time and cost (by, say, limiting the set of RDF triples surveyed), the more common effect seems to be an increase (Hart and Dolbear, 2013). Moreover, Hart and Dolbear (2013) worry about the possibility of programming error with every additional rule imposed.

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.

This strategy is not, though, the one at present pursued in ontological development. Ontologies are expected to specify which properties and relationships can be associated with which things. They are not expected to allow properties or relators to be freely combined with things, unless particular combinations are expressly prohibited. Ontologies in widespread use all start by listing a constrained set of possible combinations. To be sure, the Open World Assumption (see Sequeda, 2012), which guides the development of ontologies, asserts that we should not assume that a statement is false because it does not exist in the database(s) (it may simply be unknown), and implies among other things that one can always add new properties to existing concepts. The point to stress here is that each new property has to be specifically authorized for use in combination with a particular thing. An alternative approach would allow all properties to be associated with all things unless expressly prohibited. There does not appear to have been extended reflection of the costs of this "traditional" approach to ontology development, both with respect to constraining the set of possible inferences that might be drawn, and making it much harder to achieve consensus on an ontology to employ across diverse databases. The approach recommended in this paper instead provides for a very large set of possible combinations, and then asks what limits might need to be placed on these.

This paper thus takes a minimalist approach: we should seek the minimum set of syntactic rules. This sensible guideline is likely easier to achieve within a bottom-up approach.

So what sort of inferential rules are necessary for the Semantic Web?

3.1. Hierarchy

Hierarchy is stressed in the Semantic Web literature (Hart and Dolbear, 2013). We want the computer to infer that all characteristics associated with animals in general are applied also to subclasses of animal. Otherwise we need to indicate these characteristics for each animal individually. We need then to insist rigidly on logical hierarchy. This is not done within most classifications developed historically within the information science community. "Type of" hierarchies are not always clearly distinguished from "part of" hierarchies. The latter are treated as properties within the Semantic Web literature, in order to ensure that false inferences are not drawn: wheels do not possess all of the characteristics of automobiles. Non-synthetic classifications also regularly abuse hierarchy (Mazzocchi et al., 2007): there is no other place to put recycling and so it is treated as a subclass of garbage when it is rather something done to garbage.

A classification that employs a strictly logical "type of" approach to classifying things will thus admirably serve the inferential as well as controlled vocabulary needs of the Semantic Web. The BCC generally holds to a strictly logical approach to subdivision of classes. This almost always occurs in terms of "type of" subdivision. Cases of "part of" subdivision are clearly distinguished. This logical approach is possible because of the two characteristics of the BCC stressed above. A synthetic approach allows recycling to be treated properly as a relator rather than improperly as a misplaced thing. Freely combining things, relationships, and properties frees the BCC from the temptation to abuse hierarchy. The approach of breaking complex concepts into basic concepts allows us to classify only real (generally) observable things in the world, rather than aspire to logically subdivide within a hierarchy of things vague terminology such as globalization.

The combinatory approach to relationships outlined 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.

The inductive approach taken to date in the classification of properties does not lend itself to drawing inferences across properties. Some rules regarding properties will nevertheless be suggested below. And further research may suggest logical hierarchies of BCC properties as well.

3.2. 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 triples 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.

Empirical research is called for regarding the costs and benefits of identifying particular class distinctions. In particular, it must be asked whether these can generally themselves be correctly inferred.

3.3. Causal Connections

The example, "A weir is a form of flood defence," is given in Hart and Dolbear (2013). 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 triples 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 triples 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.

Note that it is quite possible that the influence of A upon B is not widely appreciated. It may not be one of the main influences associated with A. The standard approach to building ontologies, which insists on identifying precisely which properties or predicates can be associated with each thing might easily exclude the particular effect that A has on B.

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 triples 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 triples 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. An alternative involves employing software that looks for certain tell-tale signs of deliberate falsehood or exaggeration (see Lukoianova and Rubin, 2013), but of course such software is not flawless either.

3.4. Properties

Which properties can a particular thing possess? If we are able to achieve small schedules of both things and properties (and Szostak (2013a) 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. But it would hardly be a trivial task: there are hundreds of possible properties, and one might need to explore deep into several distinct hierarchies of things to evaluate which properties might be associated with which things. Tens of thousands of possible combinations might need to be evaluated. We would want to be very careful that we did not accidently prohibit a combination that exists in the world. Again we have to wonder if computers can infer which combinations are feasible from RDF triples themselves. If so, we could then allow any property to be associated with any thing. Only if a particular combination were found to be problematic in practice would we − after careful evaluation − prohibit it.

3.5. 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 unattainable. 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 triples 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.

The terminology employed in especially upper-level 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, p. 13-4). The classification recommended in this paper is grounded in basic concepts: the things, relationships, and properties that we perceive in the world around us. It is simply not necessary to resort to vague terminology.

The approach of logical classification further clarifies meaning. Placing a concept within a strictly logical hierarchy tells us what sort of thing it is, and also what sort of thing it is not. And if we insist on logical hierarchy for things, and combinations for relationships, and perhaps 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.

Though the people coding RDF triples 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. If told that “Fred is a swan,” and that “swans are white,” the computer needs no definition of swan in order to infer that Fred is white.

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).

3.6. 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 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.

3.7. 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 triples itself.

4. Concluding Remarks

The development of the Semantic Web is limited at present by the absence of an agreed-upon controlled (and accessible) vocabulary and set of syntactic rules. The question is how, and how well, this limitation will be overcome. At present it seems likely that the Semantic Web will develop in a fractured manner with different sets of databases coded in terms of incompatible ontologies. This paper has argued that the Basic Concepts Classification (BCC) can serve the controlled vocabulary needs of the Semantic Web. The BCC also addresses some of the syntactic needs; other syntactic rules can be built onto the BCC as necessary.

The bottom-up approach using the BCC has two key advantages. First, it greatly enhances the probability of achieving widespread consensus, since consensus can be sought step-by-step rather than after the fact when faced with complete, complex, and incommensurate ontologies. Second, and equally important, this approach does not arbitrarily constrain the set of possible combinations. This is easily done if we start from an exhaustive set of combinations, and place a minimal set of limitations. It is virtually impossible if we strive to identify each possible combination in advance. Serendipitous discovery in particular will be limited in such an approach.

Classification for the Semantic Web must accord with the format of RDF triples. This means the separate classification of things, relationships, and properties. These can then be freely combined in RDF triples, with the imposition of a (hopefully limited) set of syntactic constraints.

One key question raised in the paper is the degree to which many possible syntactic rules can themselves be inferred from the universe of RDF triples. The bottom-up approach grounded in the BCC is quite feasible even should this not prove possible. But if it is possible to infer rather than impose many syntactic rules, then the bottom-up approach is even more advantageous.

This paper has made a theoretical case for the BCC as the basis of a bottom-up ontology for the Semantic Web. The next step is to practically illustrate this theoretical possibility. As noted above, the BCC is being loaded onto the Protegé ontology editor. Yet this and other ontology editors were naturally designed to facilitate top-down ontologies. In particular, it is generally expected that there will be a limited set of properties and predicates associated with any one class of things. It may well prove that subtle but important changes in programming may be required to facilitate the use of the BCC as an ontology for the Semantic Web. Such a development may best be achieved through interdisciplinary collaboration.

References

Davies, R. (1989). The Creation of New Knowledge by Information Retrieval and Classification. *Journal of Documentation*, 45(4), 273-301.

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., & Dolbear, C. (2013). *Linked Data: A Geographic Perspective*. Boca Raton, FL: CRC Press.

Integrative Levels Classification (ILC). (2014). Available at <www.iskoi.org/ilc> Retrieved 2014.05.15.

Lukoianova, Tatiana, & Rubin, Victoria L. (2013). Veracity roadmap: Is big data objective, truthful, and credible? *Advances in Classification Research Online*. Available at <https://journals.lib.washington.edu/index.php/acro/article/view/14671> Retrieved 2014.05.15.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., & Oltramari, A. (2003). Ontology Library. Laboratory for Applied Ontology - ISTC-CNR. Available at <http://wonderweb.man.ac.uk/deliverables/documents/D18.pdf> Retrieved 2014.05.15.

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.

Pattuelli, M.C., & Rubinow, S. (2013). The knowledge organization of DBpedia: A case study. *Journal of Documentation*, 69(6), 762-72.

Sequeda, Juan. (2012). The Open World Assumption versus Closed World Assumption. Semantic web.com. Available at <http://semanticweb.com/introduction-to-open-world-assumption-vs-closed-world-assumption_b33688> Retrieved 2014.05.15.

Svenonius, E. (2004). The Epistemological Foundations of Knowledge Representations. *Library Trends*, 52(3), 571-87.

Szostak, R. (2003). *A Schema for Unifying Human Science: Interdisciplinary Perspectives on Culture*. Selinsgrove PA: Susquehanna University Press.

Szostak, R. (2011). Complex concepts into basic concepts. *Journal of the American Society for Information Science & Technology*, 62(1), 2247-65.

Szostak, R. (2012a). Toward a Classification of relationships. *Knowledge Organization*, 39(2), 83-94.

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

Szostak, R. (2013a). Basic Concepts Classification. Available at <https://sites.google.com/a/ualberta.ca/rick-szostak/research/basic-concepts-classification-web-version-2013> Retrieved 2014.05.15.

Szostak, R. (2013b). Translation table: DDC [Dewey Decimal Classification] to Basic Concepts Classification. Available at <http://www.economics.ualberta.ca/en/FacultyandStaff/~/media/economics/FacultyAndStaff/Szostak/Szostak-Dewey-Conversion-Table.pdf> Retrieved 2014.05.15.

Szostak, R. (2015). “A Pluralistic Approach to the Philosophy of Information Science” Invited for a special issue of Library Trends.