# Inference and Linking on the Humanist's Semantic Web

**Scholarly and Research Communication**
VOLUME 5 / ISSUE 4 / 2014

John Edward Simpson
*University of Alberta*

## Abstract

The Semantic Web promises that the pools of semantic data it interweaves together will enable people to find information that they could not otherwise find by revealing knowledge not explicitly visible in the distributed source data. In order for this promise to be fulfilled within the humanities, the Semantic Web data being created must have certain features, but what are they? This article provides some background on Semantic Web inferencing and then argues that there are three things that humanists can do to prepare their data to participant in this sort of inference generation: add more data, reciprocate links across repositories, and add metadata specifically to support inferencing.

**John Edward Simpson** is a postdoctoral fellow at the University of Alberta, Calgary, where he splits his time between INKE Interface Design, a project funded by the SSHRC Major Collaborative Research Initiatives Program, and the Text Mining & Visualization for Literary History Project. Email: john.simpson @ualberta.ca

## Keywords

Semantic Web; Inferencing; Reasoning; Humanities

CCSP Press
*Scholarly and Research Communication*
Volume 5, Issue 4, Article ID 0401182, 10 pages
Journal URL: www.src-online.ca
Received July 31, 2014, Accepted August 25, 2014, Published December 17, 2014

Simpson, John Edward. (2014). Inference and Linking on the Humanist's Semantic Web. *Scholarly Research and Communication, 5(4)*: 0401182, 10 pp.

To begin to understand the potential power of inferencing with the Semantic Web, consider the following three (hypothetical) examples.

Example 1. There are 129 towns, villages, and the like in the United States with "Springfield" in the name (Geonames, 2014), if you're willing to include Springfield Downs, Springfield Acres, and one abandoned settlement in Early County, Georgia, that was once called Springfield. In a paper titled "The Springfield Contagion," Rachel Morris links together the popularity of this name with the publishing of a popular work of poetry that makes reference to a fictional Springfield that functions as a practical utopia of sorts. A core part of her argument is drawn from a Semantic Web–enabled search that was able to reveal journal entries of early settlers referencing the poem as they moved west. With this information she was able to correlate the movement of the poem with the general spread of the name as new towns and villages were being founded.

Example 2. Benjamin Bradley was a little-known poet and author living in England during the Industrial Revolution whose writing has long been seen to have important resemblances to the work of Charles Dickens. Bradley is the subject of Morgan Fairchild's dissertation, and just like previous Bradley scholars, Morgan was having trouble actually showing a physical connection between the two authors. This changed when a Semantic Web search connected together two previously separate data sets that allowed for the identification of a mutual friend who hosted parties that both regularly attended.

Example 3. As in the previous example, building the case for influence has historically required the person under investigation to either directly cite a resource (for it to be a candidate for influence) or required a historian to cobble together an argument that a particular work or person influenced a historical figure based on what is known about their acquaintances and the context in which they lived. Not satisfied with the narrow field of view enabled by these traditional methods, Charlie Crampton wrote a set of Semantic Web rules to extract all the possible unidentified influences on the authors listed in the Orlando data set. The rules returned authors likely to have been influenced by works that at least three or more of their friends and acquaintances are known to have read that have not previously been considered influential.

Although these cases are entirely fictitious (with the exception of the 129 Springfields), the promise of allowing scholars to avail themselves of automated inferences that they would not normally have access to, let alone content at hand to support it, is entirely real. This promise comes in three parts, each corresponding to the respective example. The first is targeted harvesting of data through semantic tagging that would be difficult to find otherwise. The second is the ability to extend simple inferencing within and across data sets so that scholars are no longer left to cobble together information across repositories on their own and "by hand." The third is the ability to write inferencing rules that allow for knowledge that is not already embedded explicitly in the data to be extracted, ultimately producing new knowledge that would not otherwise be had.

Clearly these are attractive features in an academic research tool, and while the promise of their delivery will be made good in the future, it will not be the immediate future.

The space between fulfilling the promise of a fully inference-enabled Semantic Web capable of the sorts of feats suggested above and what it is capable of now will be filled with a fair bit of work getting the data in order and endowing people with the skills to use it, especially in case 3. Before I expand on the work required to get the Semantic Web in shape for serious inferencing, it is important to be clear about exactly what the Semantic Web is and what inferencing on it means, and so we will turn our attention there now.

## The Semantic Web and inferencing

The Semantic Web, first described to the public by Berners-Lee, Hendler, and Lassila in a 2001 issue of *Scientific American*, is the result of using a particular method of adding, marking up, and tagging information on the Web so that it becomes more than a collection of pages and instead operates as a knowledge network, a collection of relationships across information within data sets. The markup tool is known as RDF (Resource Description Framework), and it is used to provide semantics to content that would otherwise just be a collection of words. Semantics are added by introducing structures known as "triples" by asserting information as tripartite subject-predicate-object statements. Such a statement might assert that Prince Charles (the subject) is married to (the predicate) Princess Diana (the object) with a structure like the following: Charles marriedTo Diana. These triples, often used in combination with one or more ontologies that allow for the terms used to be interpreted in useful ways by a machine (e.g., an ontology could assert that the predicate "married to" is irreflexive, preventing people from being married to themselves), makes for content of a desired type to be specifically targeted and retrieved.

For example, a search on the standard Internet that asks who Prince Charles was married to in 1994 will at best return pages likely to contain the answer, leaving it up to the user to sift through the pages and determine what the actual answer is.[1] This state of affairs arises because the tools carrying out the search do not know what is meant by marriage, let alone who Prince Charles, Camilla, and Princess Diana are, and so they are consequently performing what amounts to clever word associations and intelligent guessing. The same search carried out on a properly enabled Semantic Web would be able to return that information directly because the information in question would have been tagged specifically, allowing the search tools to "know" exactly what is being asked and to provide an answer instead of pages that contain the relevant words but which might not actually contain an answer.

Targeting information with this level of precision has obvious benefits, but what it amounts to is treating Web repositories as "information jukeboxes," a term introduced by Dominic Oldman (2012, para. 3) – tools that simply return results that are directly associated with the query terms. Although useful in enabling researchers to do more of the same, rapidly finding information is only just a hint of what is possible as far as the information and knowledge the Semantic Web promises to make available.

The next level of sophistication comes as the growth of the Semantic Web and associated tools enables searching for more than only target phrases and keywords with something that resembles the sort of reasoning that humans undertake. It is one thing to ask a Semantic Web repository who all the friends of a particular author were known to be and quite another to ask it to suggest who the undocumented friends might have been. The former question amounts to a jukebox search: just choose all the individuals connected to the author in question with something that amounts to a friendOf predicate. The latter case could be done similarly, possibly by providing a list of predicates that might indicate a friendship. More powerfully, computers can be programmed or trained to recognize what a friendship looks like in terms of the connections that exist between individuals – such as shared acquaintances, geographical proximity, and shared interests – enabling them to suggest friendships that might otherwise have gone unnoticed. This is already done to generate connection suggestions on social networking sites like Facebook, Twitter, and LinkedIn. The shift from the first case to the second amounts to a shift from working with information to working with knowledge.

What has just been described is an example of inferencing in its fullest form, but even 10 years into the production of the Semantic Web, this is still not generally possible. This failure to return on the promise has led many to question the viability of the Semantic Web at all and caused its proponents to remind the naysayers that a great deal of infrastructure is required, and we are only now approaching viability (Hendler, 2007; Shadbolt, Hall, & Berners-Lee, 2006). The tools to do this work are still under active development, and the majority of work on infrastructure is directed toward opening data sets to Semantic Web access and closing gaps across these sets by asserting connections that are implied by the logic of the ontologies used when applied to the underlying RDF. Although the development of both inference engines and rules to apply these inference engines can be technically challenging tasks involving a healthy background in logic (particularly description logic), the Semantic Web, and/or artificial intelligence, there is work that can be done by the owners of the repositories that make up the Semantic Web to build the base on which these inference engines will be able to operate.

## Current support for inferencing on the humanist's Semantic Web

Currently, Semantic Web technologies are predominantly being used to do two things relevant to humanities research: extend the size of information networks and add certainty to search results. The benefits of extending the size of information networks are borne out by the interest of scholars in Web resources such as Europeana, Pelagios, and the members of the Advanced Research Consortium (ARC) and its hubs (NINES, 18th Connect, etc.). Each of these resources federates many smaller data sets via RDF standards and exposes them through a single search interface, enabling researchers to save time by avoiding the repository-by-repository sort of search that would otherwise be needed.

The addition of certainty to search results is most easily experienced by carrying out a search for a person in a modern search engine. Taking Google as a case in point, the result of a search is now less often a list of pages that might possibly contain information about the person being searched, but will frequently include a brief bio of the most likely match, including a picture, date of birth, and other sorts of specific information. This additional information is enabled by a tool that Google is calling the Knowledge Graph, and while the specifics of its implementation are not public knowledge, in principle it is definitely leveraging at least the principles of the Semantic

Web. The graph structure that the Semantic Web enables makes it possible to do more than simply return pages based on word associations and instead target specific information within those pages that is relevant to the subject of the search. In this way the RDF markup of the Semantic Web allows information to be specifically associated with relevant individuals, places, and things, and allows this information to be exposed in a search. Federating data sets or leveraging the massive Web crawls of Google only amounts to taking the jukebox approach mentioned earlier and replacing it with super jukeboxes. Again, jukeboxes, regular or super, are really useful and good to have, but the Semantic Web stands to be much more than this when its inferencing ability is taken advantage of more directly.

Inferencing is generally being used to improve the search capabilities of the Semantic Web in two ways. The first of these is to complete network graphs and check them for consistency in conjunction with an ontology. In such cases an RDF data repository may have thousands, millions, or even billions of triples (Semantic Web Challenge, n.d.), each of which is making an assertion. The associated ontologies change the nature of this repository, shifting it from a mere collection of links by adding rules describing how those links are made and interpreted.  This application of ontologies adds meaning to the resulting graph, changing how it can and will be used by both humans and software agents. The rules added by an ontology to a collection of RDF statements often allow for inferences about connections that should be made between nodes on the basis of current connections. For example, if "Mary siblingOf Mark" is declared as a triple and one of the properties of siblingOf in the ontology that defines the use of this predicate is symmetry, then the inverse also holds and we may assert "Mark siblingOf Mary."

The importance of declaring logical properties such as symmetry should not be underestimated since Semantic Web triples are assumed to be as bare as possible by the tools that process them unless an ontology or other set of interpretive rules tells them otherwise.[2] The inference capabilities that reasoning tools such as Pellet, FaCT++, and HermiT provide out of the box amounts to graph completion; based on the logic expressed in the ontology and the existing connections on the graph, the reasoners will complete the graph. Consequently, tools processing Semantic Web statements start out assuming that each assertion holds only in the stated direction and only between the stated objects, and continue to make this assumption until given reason to do otherwise.

The reasoners that apply the logic embedded in an ontology might actually fill in missing triples in advance of any query, creating a new triple in each case that it is warranted, or simply assume them on the fly during a query. The former approach increases the speed of future searches at the cost of increasing the size of the repository, while the latter keeps repository size down at the cost of slowing future queries. Such reasoning is a relatively straightforward task, while the list of properties attached to predicates is fairly light. With more complicated combinations of Semantic Web collections and ontologies, conflicts may be introduced, and the reasoners can assist in detecting these so that they can be corrected.

Of course, the inferencing that we really want is hidden knowledge detection/suggestion, such as suggested in the third example that began this article. This sort of inferencing requires writing special rules, but to do this we need both fecund RDF data repositories and the technical capabilities to write such rules. The technical capabilities and skills will be acquired by people within the humanities when it is possible to put them to good use, and the possibility of putting them to good use amounts to adopting some best practices around how we manage our RDF data repositories, which we will look at next.

## How to improve inferencing on a data set

Given the limited reasoning support and capabilities currently available on components of the Semantic Web relevant to the humanities, there is room for improvement both immediately and in the future. A great deal of attention is being paid to inferencing on the Semantic Web in terms of the technicalities of the description logics and artificial reasoning routines that it embodies and how to produce technical tools capable of making use of this (Akerkar & Lingras, 2008; Cheng, Yang, & Cheng, 2011; Golbreich, 2004; Kopena & Regli, 2003; Stoilos, Grau, & Horrocks, 2010). However, much of this intricate work requires either a substantial background in logic, or in programming, or quite often in both, to understand the work and make a contribution. These are not backgrounds that most people involved in the humanities have, nor are they likely to acquire them in the immediate future. What then can the average humanist with a data set they would like to expose to the Semantic Web do to help lay the foundation for a Semantic Web that the humanities can take advantage of? In what follows, three easily achievable and understandable steps to begin this contribution are outlined: adding elements, reciprocating links, and adding metadata.

### Add as many elements as possible

As we are seeing from the success of Semantic Web projects that focus on aggregating smaller data sources, such as Europeana and NINES, much of the value of the Semantic Web lies in the volume of information that it exposes. Beyond simply extending the network, new content is useful in assisting reasoning tools to draw the fullest possible answers out of a collection of Semantic Web information. It is important to include as many relevant entities in the knowledge store as possible, lest they be ignored outright. Adding new entities also adds the possibilities that new connections between pieces of information already known will be introduced, further tightening the web of knowledge.

There are effectively two ways to increase the size of a data set: add the elements yourself or connect to another repository. Adding original data oneself by writing RDF statements is an important task that all data repositories must engage in at some point or another, but as soon as there are volumes of information to add, doing this by hand becomes a time-consuming and tedious task, and so scripts and translation tools become the most efficient way forward (Simpson & Brown, 2013).

The value of original creation and curation aside, data sets can be expanded much faster by drawing on the work already done by others. With data sets that are not part of the Semantic Web, the option for expanding a data set at hand with information from another data set is to merge the two data sets. This merging could take the form of simple links between database tables or a complete integration of the data repositories themselves, such as would be achieved by adding rows from one data set to rows of another and navigating data format variations, content concerns, and permissions that almost inevitably arise with these tasks. Despite the hassle that such direct merging can give rise to, it often provides a speed boost that can make it an attractive choice. Although some form of traditional data merging is an available option on the Semantic Web, it is neither the only nor the primary means of connecting data together. By its very nature the Semantic Web allows for links to be made to other data repositories and information stores both on the Web and the Semantic Web. These links are easily made to any available data source, as protocols setting permissions and data formats are already handled by the standardizations built into the Semantic Web.

The downside of this second approach is the potential for a loss of speed; if the server holding linked data is not particularly fast, then it will add time to the search in addition to the time taken to establish the connection. Of course, if the server linked to is relatively fast, then the lag may be inconsequential or even produce a speed boost overall.

### Reciprocate links across repositories

Links between resources on the Internet and the World Wide Web are directional because they can only be followed from the page holding the link to the page that the link points to. Put another way, from the perspective of any given webpage, there is no way to know what other webpages point at it without some sort of additional effort. Even in cases where very large crawls have been done it is not possible to say "all and only these pages point to this one" with the same certainty as "this page points to all and only these pages." While unidirectional links are a commonly accepted feature of the current Web, the situation is by no means ideal. Ted Nelson (1999), in particular, is strongly critical of this practice because it reduces the degree to which the network can be considered connected by limiting what people can connect to. Only being able to follow links forward but never backwards is mitigated to some degree by deep Web crawls, such as performed by Google. This is not the same thing as being able to follow the links directly and knowing first-hand what the true structure of the network is, and deep Web crawls are not as useful as some might like, given that so few people actually have access to a large snapshot of the Web, as a company like Google does.

Since the Semantic Web in being built over and through the current Web, it has inherited unidirectional links. This creates an odd situation in terms of deriving knowledge through inference, because the repositories that everyone points to are not necessarily the best place to begin inferencing from. If these central repositories are missing potentially important information and do not point to that information or to information that points to that information, then beginning an inference-based activity in these central repositories will never reveal this missing information. Starting the inferencing activity at some more peripheral data set that points to the same central data set as well as other relevant ones stands to reveal much more information. This is an inversion of importance compared to standard Web-related thinking where the value of a site is measured in large part by how many other sites point to it. Here a strong case stands to be made for value residing in which resources and how many other resources a given resource points to.

There will, of course, be questions about whether the sites pointed to are relevant, accurate, or the like, but the sentiment remains: links must be made unless you are in the position of holding all the relevant information in one spot, an increasingly unlikely event. This need to link to all relevant data sources is unfortunate because finding these repositories at the edge of the Web is not as easy as finding those in the centre, a reflection of the reality that a page is not really a part of the Web until something else on the Web points to it. This system of selective linking can only be fully overcome by making reciprocal linking a practice to the point where it becomes a specification of the system, a state of affairs that is unlikely to be fully realized.

One of many barriers to making bidirectional linking a standard practice is that it stands to upset current hierarchies on the Web by lessening the perceived importance of the central nodes with large collections. Volume will become less of an attractor (it can be had from any site with the appropriate links), while the quality and accessibility of linked-to content become more important differentiators. This is not to say that volume will not still have clout – holding data on your own server will still stand to be faster than accessing another server, with all else being equal – just that volume will have less clout than it currently has.

### Add inference-assisting metadata

As important as increasing the volume of content and introducing reciprocal linking is, adding metadata to data within repositories stands at least on par. Such metadata can take one of two forms: certainties and provenance. Certainty information changes the topology of the Semantic Web by modifying the attractiveness of following certain connections in comparison to others that are available, since connections that are more certain are to be preferred to those that are not, all else being equal. Without these certainties, as is currently the case for most of the Semantic Web, all links are taken to be of equal strength, which can make it difficult for a reasoning tool with limited understanding of actual semantics to assess how to proceed. Frustrating for users, a lack of information about the quality or certainty of each result returned can also lead to more inferences than expected and no way to assess the order in which to follow up on the results returned.

Provenance information can function similarly to certainties in providing direction for inferencing tools to distinguish between links. Knowing, for example, whether or not a triple was produced by an automated extraction tool or curated "by hand" (possibly even by a particular person's hand) could stand in for or supplement certainties.

Such approaches to providing metadata are particularly useful for what are known as "follow-your-nose" agents (Yu, 2011). These are Web crawlers that typically perform depth-first searches that may not ever return to the top-level pages of a site, due to an almost limitless bottom of links to follow. In these cases it can be particularly valuable to add a third form of metadata by typing links. The Semantic Web does this as a matter of syntax in every single triple, since the connection between nodes is explicitly declared by necessity, Oscar_Wilde *wrote* The_Portrait_of_Dorian_Grey is by its very nature a "wrote" link. Some triples are more useful than others though when it comes to collecting certain sorts of information, and these should be introduced so that they may be exploited as much as possible. Exactly which predicates are most valuable will often depend on context, but including connections via predicates like rdf:seeAlso, foaf:primaryTopicOf, and owl:sameAs are generally very useful for connecting to other resources with similar information and declaring that you are doing so. Some caution must be exercised with owl:sameAs though, since it treats the connection as a strict equivalence relation, meaning that all the logic in the linked-to repository that is related to the linked-to element is inherited by the linking repository. As has been documented elsewhere (Brown & Simpson, 2013), this can lead to mistaken claims that are not easy to diagnose without significant knowledge of the subject area.

Regardless, it should be remembered that "[c]oding a Follow-Your-Nose agent is sometimes more of an art than a technique: it does require creative heuristics to make the collected facts more complete. For a given resource, different Follow-Your-Nose agents can very possibly deliver different fact sets. A key factor, again, is how to find clues to discover more facts" (Yu, 2011, pp. 545–546).

## Conclusion

While an important component of improving inferencing on the Semantic Web is improving the capabilities and capacities of Semantic Web reasoners and inference engines, this is a technical task that can be difficult for humanities scholars to contribute to. The three steps outlined above – adding content, reciprocating links, and adding metadata – are ways that any repository, and in some cases even end users, can contribute. Even if they cannot be completed in full, it must be remembered that a little semantics already takes us a long way (Hendler, 2013). Carrying out these tasks will not immediately bring a fully inference-enabled Semantic Web into being, but it will lay the foundation for its arrival. When it does arrive, we will then be able to judge whether or not it is able to make good on the promise of allowing us to see what we would otherwise not be able to see and forever changing the face of scholarly research.

## Notes

[1] Charles and Diana were still married in 1994, and a Semantic Web search carried out on an appropriately marked-up set of information would reveal this. If the question was "Who was Charles married to in 2001?" then the response would be less clear. If the reasoner used to interpret the search was operating under the closed-world assumption, then it would act as if all and only the true things were contained within the system and return an answer equivalent to "No one." The closed-world assumption is a bit strong, and the alternative is a reasoner that makes the open-world assumption. Open-world reasoners act as if they do not know something that is not either declared explicitly in the data or that follows directly from the data using valid rules of inference, so in response to the question would return something equivalent to "I don't know."

[2] It is not entirely accurate to say that triples are bare of properties unless explicitly given them. More accurately, they have properties by default; it is just that these properties are negative in the sense that they are the absence of properties. Properties that every triple are assumed to have unless explicitly specified in an ontology include irreflexivity, asymmetry, and intransitivity.

## References

Akerkar, R., & Lingras, P. (2008). *Building an intelligent Web: Theory and practice.* Sudbury, MA: Jones & Bartlett Publishers.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001, May). The Semantic Web. *Scientific American, 284*(5), 28-37.

Brown, S., & Simpson, J. (2013). The curious identity of Michael Field and its implications for humanities research with the Semantic Web. *2013 IEEE International Conference on Big Data,* 77-85. Doi: 10.1109/BigData.2013.6691674

Cheng, X. Y., Yang, A.Q., & Cheng, X.Y. (2011). The study of ontology reasoning to Semantic Web. *Advanced Materials Research,* 204-210, 375-380. Doi: 10.4028/www.scientific.net/AMR.204-210.375 [November 24, 2014].

Geonames. (2014). URL: [Search on "Springfield"]. http://www.geonames.org/advanced-search.html?q=springfield&country=US&featureClass=P [November 24, 2014].

Golbreich, C. (2004). Combining rule and ontology reasoners for the Semantic Web. In G. Antoniou & H. Boley (Eds.), *Rules and rule markup languages for the semantic Web* (pp. 6–22). New York, NY: Springer. Doi: 10.1007/978-3-540-30504-0_2 [November 24, 2014].

Hendler, J. (2007). Where are all the intelligent agents? *IEEE Intelligent Systems, 22*(3), 2–3. Doi:10.1109/MIS.2007.62 [November 24, 2014].

Hendler, J. (2013, August 6). *A little semantics goes a long way* [Online article]. URL: http://www.cs.rpi.edu/~hendler/LittleSemanticsWeb.html [November 24, 2014].

Kopena, J.B., & Regli, W.C. (2003). DAMLJessKB: A tool for reasoning with the Semantic Web. In D. Fensel, K. Sycara, & J. Mylopoulos (Eds.), *The Semantic Web—ISWC 2003* (pp. 628-643). New York, NY: Springer. Doi: 10.1007/978-3-540-39718-2_40

Nelson, T.H. (1999). Xanalogical structure, needed now more than ever: Parallel documents, deep links to content, deep versioning, and deep re-use. *ACM Computing Surveys, 31*(4). Doi: 10.1145/345966.346033

Oldman, D. (2012, September 4). The British Museum, CIDOC CRM and the shaping of knowledge. [Blog post] URL: http://www.oldman.me.uk/blog/the-british-museum-cidoc-crm-and-the-shaping-of-knowledge [November 24, 2014].

*Semantic Web Challenge.* (2014). URL: http://challenge.semanticweb.org [November 24, 2014].

Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The semantic web revisited. *IEEE Intelligent Systems, 21*(3), 96-101.

Simpson, J., & Brown, S. (2013, September). *From XML to RDF in the Orlando Project.* Paper presented at the International Conference on Culture and Computing, Kyoto, Japan. Doi: 10.1109/CultureComputing.2013.61

Stoilos, G., Grau, B.C., & Horrocks, I. (2010). How incomplete is your Semantic Web reasoner? *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence.* URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/viewFile/1574/2226 [November 24, 2014].

Yu, L. (2011). *A developer's guide to the semantic Web.* New York, NY: Springer.