The Semantic Web: A Defence of Floridi Against Berners-Lee et al.

Melissa Bruno

Faculty of Information, University of Toronto

This paper discusses the positions of Floridi (2009) and Berners-Lee et al. (2001) regarding the Semantic Web and suggests that the former's position is the stronger of the two. By providing a summary of their respective positions, it will be argued that Floridi's argument gains the upper hand by having a clearer understanding of the terms being used when arguing the viability of the Semantic Web, such as: meaning, understanding, and knowledge. As a result, Floridi's view of a "MetaSyntactic Web" bodes better for future developments concerning the Web by laying out a path that is realistic and viable within the parameters of current and probable computing capabilities and practices.

As Floridi (2009) notes, Berners-Lee was the first to introduce the idea of the Semantic Web in the nineties. According to Berners-Lee et al. (2001), the vast majority of online content is "designed for humans to read, not for computer programs to manipulate meaningfully. Computers can adeptly parse Web pages for layout and routine processing—here a header, there a link to another page—but in general, computers have no reliable way to process the semantics" (p. 36). As such, Berners-Lee et al. (2001) theorize that the Semantic Web will "bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users" (p. 36). To do so, Berners-Lee et al. (2001) suggest that augmenting Web pages with documents and data that are specifically targeted at computers will assist in translating the Web as we know it (i.e., a medium of documents created for the manipulation and use of *people*) into a Semantic Web, ultimately allowing/enabling computers to "process and 'understand' " (p.37) data that, today, is merely displayed. As an extension to the current Web, Berners-Lee et al. (2001) postulate that said Semantic Web computers will be able to "find the meaning of semantic data by following hyperlinks to definitions of key terms and rules for reasoning about them logically" (p. 36), thus

resulting in highly functional semantic agents. Moreover, ordinary users will assist in the creation of the Semantic Web by adding new definitions and rules using off-the-shelf software, helping to refine the semantic markup through encoded web pages (Berners-Lee et al., 2004).

A major tenant of Berners-Lee et al.'s (2001) Semantic Web is the prospect of universality of the World Wide Web in the sense that hypertext links can link anything to anything, thus creating an equal terrain between all content at all stages of production (i.e., rough drafts to finished final products), types (i.e., academic and commercial), formats, languages, and so forth. As traditional knowledge-representation systems have, for the most part, been centralized, "requiring everyone to share exactly the same definition of common concepts" (Berners-Lee et. al, 2001, p. 37), the challenge is being able to harness a decentralized model which will provide various resources for a language that expresses both data and rules for reasoning about the data, allowing for existing rules from current knowledge-representation systems to be integrated into the Web. In looking to add "logic to the Web," Berners-Lee et al. (2001) argue that two major technologies to facilitate the Semantic Web are already in place: eXtnesible Markup Language (XML), a tool that allows users to "add arbitrary structure to their documents but says nothing about what the structures mean" (p.38), and Resource Description Framework (RDF), a scheme for "expressing the meaning of terms and concepts in a form that computers can readily process" (p. 38). As Berners-Lee et al. (2001) note, RDF can use XML for its syntax and Uniform Resource Identifiers (URI) to specify entities, concepts, properties and relations; more explicitly, as "RDF uses URIs to encode this information in a document, the URIs ensure that concepts are not just words in a document but are tied to a unique definition that everyone can find on the Web" (p. 39). Continuously, as Berners-Lee et al. (2001) state, an ontology is a document or "file that formally defines the relations among terms" (p. 39), in which the most "typical kind of

ontology for the Web has a taxonomy and a set of inference rules" (p. 39), defining classes of objects and relations among them. As such, the "computer does not truly 'understand' any of this information, but it can now manipulate the terms much more effectively in ways that are useful and meaningful to the human user" (Berners-Lee et. al, 2001. p.39). Therefore, presently, sharing and combining information across databases is challenging as programs need to know how to link the unique language/terms used to convey the same thing; as such, Berners-Lee et al. (2001) state that, in order to discover common meanings between databases, ontologies (or other Web services), need to provide "equivalence relations" (p. 39), an act which Berners-Lee et al. (2001) suggest can improve the accuracy of web searches insofar as the search program would look for concepts as opposed to ambiguous keywords.

As such, Berners-Lee et al. (2001) postulate that Semantic Agents will take on the task of collecting, processing, and exchanging Web content from multiple sources. This provision of information via Semantic Agents (Berners-Lee et al., 2001) will prove more effective with the increase in the availability of machine-readable Web content and automated services/agents, thus affording software that was originally not deigned to be compatible the ability to share and transfer data vis-à-vis inference engines and semantics: the "Semantic Web's unifying language (the language that expresses logical inferences made using rules and information such as those specified by ontologies)" (p. 42). Berners-Lee et al.'s (2001) theorized service discovery is enabled by a "unifying language" (p. 42) combined with digital signatures (i.e., encrypted blocks of data), which can only happen when there is a common language to describe services in a way that lets other "agents 'understand' both the function offered and how to take advantage of it" (p. 42). There are obvious limitations as "standardization can only go so far, because [one] cannot anticipate all possible future needs" (Berners-Lee et. al, 2001, p. 42). Therefore, the consumer

and producer agents can reach a "shared understanding by exchanging ontologies, which provide the vocabulary needed for discussion" (Berners-Lee et. al, 2001, p. 42); moreover, agents can even " 'bootstrap' new reasoning capabilities when they discover new ontologies. And Semantics also makes it easier to take advantage of a service that only partially matches a request" (p. 42). As Berners-Lee et al. (2001) conclude, the Semantic Web, in naming every concept simply by a URI, lets anyone express new concepts that they invent with minimal effort: "its unifying logical language will enable these concepts to be progressively linked into a universal Web" (p. 43). The problem with realizing the Semantic Web is articulated by Floridi (2009) by making a clearer use of terms; furthermore, Floridi (2007) describes a viable path of development for present Web technologies which take into account the ambitions of Berners-Lee et al. without falling prey to unrealizable ideas. We turn now to an explanation and defence of Floridi's position.

As Floridi (2009) notes, it is crucial to take account of the discrepancy between Berners-Lee et al.'s (2001) articulation of their vision of the Semantic Web and the World Wide Web Consortiums (W3C). Floridi (2009) asks us to consider why there is a discrepancy between the former's vision, which states that:

> [t]he Semantic Web will bring structure to the *meaningful content* of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users. . . . all this *without needing artificial intelligence* on the scale of 2001's Hal or Star Wars's C-3PO…The Semantic Web will enable machines to *comprehend semantic documents* and data, not human speech and writings. (p. 26-27).

The latter's view states that: "the Semantic Web is a common framework that allows *data* to be shared and reused across application, enterprise, and community boundaries. . . . It is based on

the Resource Description Framework (RDF)" (W3C, 2008b, par. 6). It is precisely the poor use of language which ironically casts a shadow over the viability of a Semantic Web as espoused by Berners-Lee et al. Floridi (2009) outlines some problems with Berners-Lee et al.'s vision of the Semantic Web which will now be developed.

To begin, Floridi (2009) highlights that by poorly defining key terms such as "semantics", "meaning", "understanding", "comprehension", "information", "knowledge", and "intelligence" (p. 28), proponents of the Semantic Web lose sight of the fact that "languages, protocols, and ontologies for metadata and metasyntax already allow integration, aggregation, sharing, syndication, and querying of heterogeneous but well-circumscribed topic-oriented data across different databases"(p. 28), without having to ask of machines to create and/or decipher semantic meanings. The point being made here is that processors and advanced algorithms allow for the above mentioned activities already, and consequently, "no meaning or intelligence plays any role in this" (Floridi, 2009, p. 28). We will expand on this point latter in the paper with a discussion of IBM's super computer *Watson*, and Searle's (2011) critique of "thinking machines."

Extremely important to consider is the reality that "[s]emantic content in the Semantic Web is generated by humans, ontologized by humans, and ultimately consumed by humans…RDF, XML, URI, and all the other technical solutions are just the mid-stream syntax between a human upstream producer and a human downstream consumer" (Floridi, 2009, p.29). Within the technical solutions just outlined, the driving force is efficient and effective taxonomies (and more and more folksonomies which are developed by independent user communities) which should not be confused with understanding or semantics. At the very least, understanding requires self-referentiality, and computers/machines (at least at this stage in the game) are not conscious. The manipulation of symbols/taxonomies does not qualify as

knowledge, but is mere computation. We will return to this point in our discussion of *Watson* mentioned above. For now, it is well to observe that the degree of difficulty concerning the ontologies required to "furnish the semantics for the Semantic Web must be developed, managed, and endorsed by committed practice communities" (Shadbolt et al. 2006, p. 99). Also, ontologies "suffer from a limited degree of modularity: every bottom-up tag helps immediately, but systematic, top-down, exhaustive, and reliable descriptions of entities are useless without a large economy of scale" (Floridi, 2009, p. 30). Ultimately, Floridi (2009) asks us to be realistic about the possibilities of a Semantic Web which border on the fantasies of Artificial Intelligence, and suggests that we set our sights on more feasible projects which the W3C articulate as follows:

> The Semantic Web is a *web of data*. . . .The Semantic Web is about two things. It is about *common formats for integration and combination of data* drawn from diverse sources, where the original Web mainly concentrated on the interchange of documents. It is also about *language for recording how the data relates to real world objects*. That allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing. (W3C 2008a, par 3)

As Floridi (2009) notes, the vision of the Semantic Web thus articulated (or MetaSyntactic Web, as data is not confused for information, nor meaning with syntax) may not be as exciting as the vision of Berners-Lee et al., but it is a project which is already in gear and provides for "a genuinely useful development" (p.31).

Some other problems to consider in the realization of the Semantic Web can be gleaned through an analysis of Doctorow's (2001) article, *Metacrap: Putting the Torch to Seven Straw-Men of the Meta-Utopia*. Doctorow (2001) outlines seven key issues with realizing the Semantic

Web which are as follows: people lie; people are lazy; people are stupid; people are not their own best critics; schemas are not neutral; metrics influence results; and there is more than one way to describe something.

As Doctorow (2001) notes, because metadata "exists in a competitive world" (sec 2.1), and "people lie", it is naïve to expect the Semantic Web to be a vehicle of reliable metadata, for "when poisoning the well confers benefits to the poisoners, the meta-waters get awfully toxic in short order" (sec 2.1). Although it may be argued that collaboration amongst interests is often precisely the paradox encountered by self-serving agents seeking their advantage, it is clearly a huge challenge to get interests on board to create a "common language" when their interest is not immediately evident, and there is no guarantee that a collaborating party would not turn its back on such an enterprise once its interest is served in doing so.

Doctorow (2001) observes that "info-civilians are remarkably cavalier about their information" (sec 2.2), and as such, because of a lack of precision in tagging information, the Semantic Web will suffer from an ability to add metadata from average user files. The methods and language people use, or do not use, to describe content they create and/or disseminate online is volatile and highly subjective. For example, where some might painstakingly label all the content of their documents, others might both inconsistently label content, or fail to provide any metadata at all. As such, the varying level of metadata applications will disproportionality represent the content found online as novice as well as advanced metadata users are both equally capable of creating and providing content. Building on this last point, Doctorow (2001) highlights how "[e]ven when there's a positive benefit to creating good metadata, people steadfastly refuse to exercise care and diligence in their metadata creation" (sec 2.3). As the internet is ripe with spelling errors, as well as people intentionally using slang and other

abbreviations, these fine and gross points of [il]literacy (i.e., spelling, punctuation, grammar), will further thwart any attempt at achieving a consistent application of metadata to online content.

Moreover, as Doctorow (2001) notes, "people are lousy observers of their own behaviours" (sec 2.4); as such, unless "everyone engaged in the heady business of describing [content] carefully, weighs the [content] in the balance, and accurately divines the [content]'s properties, noting those results" (sec 2.4), the possibility of there being reliable metadata will remain unachievable. In other words, individual self-perception and value judgements placed on the significance and importance of one's own abilities of actions will, by extension, negatively translate itself into the metadata provided by said skewed understandings. That said, schemas are inherently subjective, not neutral, as there will always be debate on how content should be labeled beyond the immediacy of individual extrapolation and context.

As Doctorow (2001) delineates, any "hierarchy of ideas necessarily implies the importance of some axes over others" and assumes that there is a "'correct' way of categorizing ideas, that reasonable people, given enough time and incentive, can agree on the proper means for building a hierarchy" (sec 2.5). For the reasons discussed above (i.e., competing interests, individual contexts of use, and local understandings), the feasibility of there being a "correct" way of doing anything is next to impossible, and any attempt to create any level of detailed, agreed upon, *universally* accepted classification structure or metadata schema to satisfy all desired levels of granularity and use is an exercise in futility. Furthermore, as "metrics influence results" (Doctorow, 2001, sec. 2.6), agreeing to a common "yardstick for measuring the important [content] in any domain necessarily privileges the items that score high on that metric, regardless of those items' overall suitability" (sec 2.6). That said, every player in a metadata standards body wants to "emphasize their high-scoring axes and de-emphasize (or, if possible, ignore altogether)

their low-scoring axes" (Doctorow, 2001, sec 2.6). As such, like points raised earlier on, the possibility of a group of people looking to advance their agendas being universally pleased with any hierarchy of knowledge will not be a reality until they abandon said agendas and solely work towards the consistent application and representation of all content created and provided online.

Lastly, as everyone has their own unique understanding of the world, they, too, have their own interpretation and methods for describing said understanding. As Doctorow (2001) highlights, "[r]equiring everyone to use the same vocabulary to describe their material denudes the cognitive landscape, enforces homogeneity in ideas" (sec 2.7), ultimately precluding any exercise or instance of difference or newness that breaches the confines of pre-existent understandings afforded by said cognitive model. For example, think of the term Breakfast. Breakfast for one might mean a time of day, it might mean eggs and bacon, or pancakes, or a luxury when you have enough time in the morning. Alternately, think about Impressionist art—some might deem it inspiring, provocative, while others may consider it pedestrian, disordered, and sloppy. Until there is an agreed upon way to accurately describe and capture anything, reducing and relegating it to a singular understanding, or even a linked understanding [as suggested by Berners-Lee et al.(2001)], does not do justice to the individual interpretation, or unique conceptualization inherent in any articulation afforded by metadata. However, despite the negative points delineated above, metadata is still highly valuable as it is "often a good means of making rough assumptions about the information that floats though the Internet" (Doctorow, 2001, sec. 3). Without it, the sheer volume of content online would be unnavigable and inaccessible, reducing and rendering the information-retrieivability of the mass of the World Wide Web to ambiguous and broad keyword searches. Within the parameters of the above discussion, we can now expand on a problem that brings to light the deficiencies of the Semantic

Web as envisioned by Berners-Lee et al. (2001), namely, IBM's super computer *Watson* and its *Jeopardy!* adventure.

Watson is an AI computer which has the ability to answer questions posed in natural language. Developed with the specific intention of answering questions on the game show *Jeopardy!*, Watson competed with two former *Jeopardy!* winners in 2011 and won. As Florini (2009) comments, however, the victory is comparable to IBMs other super computer, *Deep Blue*, defeating chess master Garry Kasparov, "despite having the intelligence of a toaster" (Florini, 2009, p.26). But why is it the case that both computers lack intelligence despite their impressive victories? As philosopher John Searle (2011) notes, computers do not "think" but rather are devices "that manipulate formal symbols…An increase in computational power is simply a matter of increasing the speed of symbol manipulation" (par. 7). As such, an increase in the ability to compute does not amount to thinking, which can be explained via Searle's famous *Chinese Room Argument*, basically running as follows: suppose a person which lacks knowledge of Chinese were locked in a room with boxes containing Chinese symbols and an instructional booklet in English explaining how to manipulate the symbols. The boxes can be considered "the database", the instruction booklet "the program", and the person "the computer". Without the person knowing it, people outside the room pass in Chinese symbols that are questions, and by looking in the manual, the person is able to answer the questions using Chinese symbols. Over time the person in the room becomes so efficient at answering the questions that are passed inside the room that he gives the impression that he can speak Chinese, despite not actually knowing the language. As such, the point being made is that a computer functions in the same way, which cannot be considered understanding or knowledge, for the computers processing system is unable to extract meaning from the symbols it manipulates (i.e. it is incapable of

getting from syntax to semantics). Thus, the very notion of a Semantic Web breaks down at its base for confusing syntax with semantics, and why Floridi's (2009) suggestion of terming the Semantic Web as the MetaSyntactic Web is both more feasible and philosophically coherent as regards the use of words.

Floridi's (2009) position against Berners-Lee et al. (2001) regarding the Semantic Web is the stronger of the two. It is well to recall that Berners-Lee et al. (2001) suggest that a computer's manipulation of terms/taxonomies does not amount to "understanding" (p.39); as such, Floridi's (2009) assertion that their vision inflates rhetoric to the status of a strong possibility should be heeded. By employing a more rigorous language to articulate the possibilities of a viable evolution of the Web, Floridi undermines Berners-Lee et al.'s position and shows it to be an unpromising reality.

References

Berners-Lee, T, Hendler, J., & Lassila. O. (2001). The Semantic Web. *Scientific American 284* (5), 34-43.

Doctorow, C. (2001). *Metacrap: Putting the torch to seven straw-men of the meta-utopia.* Retrieved from http://www.well.com/~doctorow/metacrap.htm

Luciano, F. (2009). Web 2.0 vs. the Semantic Web: a philosophical assessment. *Episteme 6* (1), 25-37.

Searle, J. (2011). Watson doesn't know it won on 'Jeopardy!'. *The Wall Street Journal*. Retrieved from http://online.wsj.com/article/SB10001424052748703407304576154313126987674.html

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

W3C. (2008a). *W3C Semantic Web Activity*. Retrieved from http://www.w3.org/2001/sw/

----. (2008b). *W3C Semantic Web Frequently Asked Questions*. Retrieved from http://www.w3.org/RDF/FAQ