Welcome to this issue of IJSWIS on the theme of identity, reference, and the Web. Identity is at the heart of the semantic Web, yet there is no consensus on the question “What does a URI on the Web identify?” Tim Berners-Lee, the co-author of the original specifications, said that one great thing about the Web was that he never had to answer the question what a URI (uniform resource identifier) actually identifies (1998). Crucial to the Web’s evolving into the semantic Web is the development of a coherent way to determine what a URI identifies, since the vision of knowledge representation on the semantic Web requires multiple people to use the same URI when they need to identify the same thing. Despite seeming simple, this question has deep foundations and surprising ramifications.

THE IDENTITY CRISIS

A URI on the semantic Web is used as a globally disambiguating term, much like a foreign key in a global database. However, unlike classical databases, where there is one normative schema for the entire database that pins down the meaning of the data, on the semantic Web there may be multiple and even conflicting schemas created in a decentralized manner. So agreement must develop about what a URI identifies for someone to use it with confidence on the semantic Web. If URIs are reused the semantic Web will flourish; otherwise the semantic Web will become isolated islands of semantics, with each community having to reinvent the wheel by minting new URIs. It is precisely this latter option that currently dominates most work on the semantic Web, with the exception of the linked data initiative.

The problem of what precisely a URI identifies may seem like an arcane question at the intersection of Web architecture and philosophy, but it has important engineering ramifications for the semantic Web. In particular, what representation should one host at a URI that is used on the semantic Web? Should one just put an HTML Web page there, or a copy of the ontology in OWL or RDF, or redirect to another resource? This is one of the first questions anyone who is actually using the semantic Web will encounter. Until recently, most have not taken the question seriously but have put up one thing or another—or sometimes nothing at all—at a semantic Web URI. This has led to a wide variety of practices that make it difficult for both computer applications and humans to “follow their noses” in order to discover what URIs on the semantic Web mean. This problem has often been termed the “identity crisis” of the (semantic) Web.

The absence of a consistent means for specifying or determining what semantic Web URIs identify may very well be why the semantic Web has not experienced the exponential growth and widespread deployment that characterized the original Web. One of the reasons the original Web experienced such widespread growth was that a person could put up a Web page and tell
anyone—including those outside the inner circle of Web digerati—to visit the Web page, and those who visited could almost always easily determine what the page was about. There was an instant gratification that is missing from the semantic Web. The current situation discourages, instead of encourages, URI reuse. If a person is putting some data on the semantic Web, how do they know it should reuse, or even link to, other URIs on the semantic Web? Furthermore, while a few core semantic Web vocabularies such as FOAF (friend-of-a-friend) are heavily reused, the vast majority are left languishing in isolation. In particular, almost all the reuse of URIs seems to be of URIs that denote classes, or what description logicians would phrase the terminology, or T-Box. Almost no URIs for individuals or instance data, the assertions or A-Box in description logic, are ever reused (Bouquet, Stoermer, Mancioppo, Giacomuzzi, & Okka, 2006). This problem about URI reusage is the direct result of a stunning lack of clarity about how one can communicate what a URI identifies on the semantic Web.

The use of URI is what puts the “Web” into the semantic Web. Should a semantic Web URI wear its nature on its sleeve? If so, should we standardize what dereferencing a semantic Web URI results in? There may be important applications that could solve long-standing problems on the semantic Web if this were done, but only if done correctly.

Given the importance of the subject matter for the success of the semantic Web, a workshop at the 2006 World Wide Web conference was held in Edinburgh on “Identity, Reference, and the Web” (IRW) (Halpin, Thompson, & Hayes, 2006). The proceedings and recordings of this lively and well-attended workshop are available at http://www.ibiblio.org/hhalpin/irw2006/. The articles in this issue are refined versions of the presentations at the workshop.

OVERCOMING AND DEFENDING AMBIGUITY

Dan Connolly, co-author of the original HTML specification as well as a member of W3C Technical Architecture Group, presented a fully worked out example of how the identity crisis on the semantic Web presents practical difficulties for inference in this issue (2006). Interested readers should follow the example as given at http://www.w3.org/2006/04/irw65/urisym.html. The heart of the problem is that a person can make contradictory claims about whether a URI identifies a person or his or her Web page, and if these classes are defined by an ontology as disjoint, then a reasoner would find an inconsistency. To solve this problem, Web pages are defined as “information resources” but things like people are not. The official W3C-endorsed solution to the problem is to use 303 redirection to another URI for URIs that are not information resources if you want to “follow-your-nose” to discover more information. In a simplified nutshell, this states that URIs that refer to something that is not “an information resource” should redirect to another resource hosting a Web page.

This idea of an information resource is a term from the W3C TAG that seems to refer to things such as Web pages that are “on the Web,” although the details are controversial to say the least. Is a namespace document an information resource? What about Moby Dick? Regardless, the W3C TAG decision does mean most of the URIs on the semantic Web should use a particular redirection technique in order to avoid the problems pointed out by Dan Connolly (2006) or use a URI that ends in a fragment identifier, the hash, although this is an arguably an abuse of fragment identifiers as originally conceived. Furthermore, a number of forthcoming W3C Recommendations or Notes, such as “Cool URIs for the semantic Web” (http://www.w3.org/TR/cooluris/) and “Best Practice Recipes for Publishing RDF Vocabularies” (http://www.w3.org/TR/swbp-vocab-pub/), deploy this solution.

In response to the solution put forward by Connolly and TAG, in this issue Hayes and Halpin present a merger of their work that both outlines the philosophical issues and provides—rather surprisingly—a defense of ambiguity on the semantic Web (2008). In particular, most of the confusion and advantages of using URIs seem to come from the crucial ability of URIs in certain cases to access what they identify, not just to refer. Reference is always to an extent ambiguous, and so no URI can be fully disambiguating; but ambiguity can be a blessing as opposed to a curse, leading to a whole range of solutions to help make explicit what a URI both accesses and in “good practice” refers to Hayes et al. (2008).
DESCRIBING AND FORMALIZING IDENTIFICATION

Yet, is ambiguity always a blessing? There are certainly some cases where ambiguity may simply hide a true ontological distinction of importance. If every even minor ontological distinction led to minting a new semantic Web URI, the semantic Web would then devolve away from a space of linked data into islands of self-contained isolated data. In this issue, Ginsberg points this out via a brilliant example based on the various different meanings of marriage (2008). Interestingly enough, earlier at the workshop he had used the example of the status of Pluto as a planet, but shortly thereafter the astronomical community had decided that Pluto was indeed not a planet (Halpin et al., 2006). However, having the various concerned communities agree what the definition of a marriage constitutes appears to be a more long-standing problem. Ginsberg connects the issue of ambiguity to further issues in ontological determinacy and holism in philosophy, and then presents a solution in the form of a certain type of representation to be associated with a semantic Web URI, an “intended usage record” that can make explicit the disagreements while still maintaining the URI for reuse (2008).

The issues at hand may even cut to the heart of the formal meaning of URIs. If a URI may actually have different formal meanings depending on different interpretations of what it identifies, is monotonicity at risk? Is monotonicity non-negotiable? Patel-Schneider and Parsia have argued that the formal solution should be that meaning on the semantic Web should not be global, but instead be local, so as to allow divergence of meaning. To accomplish this, one would have to force ontologies from elsewhere to be explicitly imported via owl:imports to show acceptance of community norms and opinions explicitly (Parsia and Patel-Schneider, 2006).

Would it only be some special type of URIs, those the TAG thinks are do not identify “information resources,” that this would apply to? How can we define what an information resource is in a formal manner that is not vulnerable to disagreement? This is a symptom of deeper problem: there has been little work on formalizing the ontology of the Web itself. It’s not therefore surprising that modeling the distinction between the various types of resources that Web architecture depends on remains unclear. In the final article in this journal issue, Presutti and Gangemi present an ambitious formalization, the “identity of resources and entities on the Web” (2008). This formalization captures many of the distinctions currently being debated and—like any good formalism—produces a number of other distinctions not previously noticed, such as the use of proxies, which could be considered a generalization of foaf:topic. Furthermore, this work is built upon the widely respected DOLCE ontology and has connections to other work on identification such as tagging.

TAGGING AND SOCIAL MEANING

Looking beyond this journal and into the future, tagging may indeed be one way to establish identity and reference with a minimum of work. This brings up the important point of how identity and reference are socially constructed. One philosophical analysis of social meaning is given in terms of convention. A convention is used by a community to solve some coordination problem, such as determining how to list dates so as to schedule meetings, even if such a choice is arbitrary, as given by the American convention of listing the month before the day as opposed to the European method of listing the day before the month (Lewis, 1969). In its classic game-theoretic formulation by Lewis, convention is given as a non-normative solution to a recurring coordination problem where the solution itself is driven by precedence (1969).

The identity and meaning of data on the Web could be viewed as a coordination problem. A further general rule that falls out of a game-theoretic analysis is that people will in general use the minimum amount of convention to solve their coordination problem of identifying things on the Web. This rule of thumb might explain the slowness of the Web community to embrace model-theoretic semantics. After all, tagging is easier than using the semantic Web, requiring only that the user type in a few natural language words. An overlap of tags is taken as a sign of similarity of identity, and two pieces of data that share the majority of the same tags are thought to be about the same thing.

While it is unclear whether such a technique can be subsumed by more articulated, logic-based
semantic Web technology or are a low-cost alternative to the semantic Web, it seems that tagging is here to stay thanks to its large deployed user base. Developers found it trivial to implement tagging-based solutions to data sharing and identification problems, while many were probably unfortunately confused by the complex nature of semantic Web standards, even though the core conceptual basis of standards such as RDF are simple. Also such a loose and unconstrained mechanism as tagging, although useful for simple things, makes it difficult for humans to express complex relationships or for machines to use tagged data reliably. Luckily, recent work has shown that the tags users choose are not chaotic, but rather quickly converge to a common descriptive set of tags that is almost unchanging over time (Halpin et al., 2007). Perhaps once the tags have stabilized, coherent URI-based identification schemes could emerge.

What tagging does point to convincingly is the social aspect of naming. In a given natural language, many sorts of identifiers, such as common words, are socially centralized. Other sorts of identifiers, such as proper names, are socially decentralized, varying from local context to local context. Black has noticed a correspondence between this socially grounded identification process and the use of socially constructed Web sites (2006). If an identifier truly is social, then for an identifier to be adopted it should separate the process of creating the identifier and accepting the identifier, particularly for agents that did not create the identifier in the first place. Once an identifier, such as a URI on the semantic Web, is created, the identifier can then become common ground among a community of agents that accept the identifier (Black, 2006). If a standardized way of socially grounding identifiers emerges, then popular identifiers could be easily found and accepted, encouraging identifier reuse.

MEANING AS USE AND CONVENTION

There were many other important perspectives presented at the IRW workshop and under consideration by the W3C TAG that, due to the constraints of space in this journal issue, we were not able to include. Booth of Hewlett-Packard took the rather Wittgensteinian stance that URI identity is a myth, for the meaning of a URI should be completely defined by its use (2006). The main use of URIs is or should be that one can deference them to obtain authoritative descriptive information. Since all descriptive information is inherently partial by nature, what matters is that the partial information be sufficient to identify the resource for a given application (Booth, 2006). Yet on the Web-scale, one loses the network effect unless one can provide information sufficient for universal identification.

This notion of universality is important, for as Steve Pepper has pointed out, the fact that URIs constitute a universal, globally unique way of naming things is one of the most powerful aspects of the Web (2006). However, people are not using (and reusing) URIs as identifiers extensively enough to realize the potential that this offers. Pepper cited three main reasons: the confusion caused by the many flavors of URIs (URNs, TDBs, etc.), the lack of a discovery mechanism besides general-purpose engines like Swoogle, and the difficulty of knowing what a given identifier actually identifies (2006). Uniting around the use of http: URIs of course makes sense, as they are already the most popular and have a widely implemented resolution and access mechanism. Still, a common format for semantic Web URIs that uses some simple set of conventions to enable discovery through the use of search engines or repositories is still lacking. Pepper summarized the way the Topic Map conception of a semantic web addresses these issues by distinguishing subject identifiers from subject indicators (2006).

CONCLUSION

It is the precise nature of these conventions for identification—ranging from the 303 solution of the TAG to the resurrection of the “Link” entity response header—that is under debate. The solutions are varied, but all point to the need to have an explicit relation that connects URIs with descriptions of referents, be it in a HTTP entity response header, an RDF statement, or a special sort of representation actually returned by the format. Regardless, these conventions allow one, if necessary, to make the long-standing distinction between the sign and the signified, the symbol and its denotation, the identifier and the identified, on the Web.

We hope this journal issue of IJSWIS contributes to the ongoing discussion and eventual
resolution of these issues by presenting many of the initial arguments and positions in much greater depth than has been allowed in other workshops. There have been a number of follow-up workshops to the original “identity, reference, and the Web” conference on this topic at academic Web conferences, in particular the “identity, identifiers, identification” (I3) workshop at the 2007 World Wide Web conference in Banff, Canada (Bouquet, Stoermer, Tummarello, & Halpin, 2007). From these workshops has come a European Union funded project to ensure best usage of URIs on the Web, the OKKAM project (http://www.okkam.org/). We would like to extend special thanks to Amit Sheth for giving us the opportunity and the time to put this issue of IJSWIS together, as well as Paolo Bouquet for providing help.

At the identity, reference, and the Web workshop, Pat Hayes asserted that “the semantic Web languages would operate exactly unchanged if the identifiers in them were not URIs at all, and if the Web did not exist.” If we accept that proposition, then the semantic Web project is uninterestingly different from any number of contributions to the Knowledge Representation project of artificial intelligence over the last thirty-five years. If this is indeed true, that is dismal news, for the rate of progress of that effort has been slow. If the one way to distinguish the semantic Web from artificial intelligence is the use of URIs, then getting clear about what the value of using URIs as identifiers actually is of utmost importance. We offer this special issue as a concrete step toward doing just that.

REFERENCES


Harry Halpin is a postgraduate researcher at the University of Edinburgh in the School of Informatics at the University of Edinburgh in the Institute for Communicating and Collaborative systems. His background is in artificial intelligence, philosophy of the mind, machine-learning, and natural-language processing, and he has taught varied topics such as formal semantics and enterprise Web engineering. Recently he has focused work on understanding new Web phenomena such as collaborative tagging, and has shown how tagging can lead to consensus of description over large data-sets. He is the chair of the GRDDL (Gleaning Resource Descriptions from Dialects of Language) Working Group that has produced standards on how to extract semantics from XML documents and microformats, and serves of the W3C Semantic Web Co-ordination Group. Guiding all of his work is a long-standing interest in the philosophical background of Web architecture, he has published some of the first philosophical articles on the Web and was a research fellow in Information Science and Studies at Duke University from 2006-2007. He was the lead organizer of the first workshop on Identity, Reference, and the Web at WWW 2006.

Henry S. Thompson divides his time between the School of Informatics at the University of Edinburgh, where he is reader in artificial intelligence and cognitive science, based in the Language Technology Group of the Human Communication Research Centre, and the World Wide Web Consortium (W3C), where he works in the XML Activity. He received his PhD in linguistics from the University of California at Berkeley and was affiliated with the Natural Language Research Group at the Xerox Palo Alto Research Center, where he participated in the GUS and KRL projects. His current research is focussed on articulating and extending the architectures of XML. He was a member of the SGML Working Group of the World Wide Web Consortium which designed XML, a major contributor to the core concepts of XSLT and W3C XML Schema and is currently a member of the XML Core, XML Schema and XML Processing Model Working Groups of the W3C. He has been elected twice to the W3C TAG (Technical Architecture Group). He is interested in the philosophical foundations of computer science and AI, and helped organize the workshop on Identity, Reference, and the Web at WWW 2006.