The Principle of Self-Description: Identity Through Linking

Harry Halpin[1]

Edinburgh University

Abstract. If one wants to have a scheme for identifying non-Web accessible entities, should it be centralized or decentralized? Given a URI, how can one tell if it refers to a web page or a non-Web accessible entity? We present an analysis of these questions, and come to the conclusion that identity can only be defined through relationships, which are given on the Web as *links*. This in turn leads us to a definition of self-description serves as a good practice for identifying any resource, including entities, on the Semantic Web. We then bring up a number of ways a resource can be *normatively* linked to other resources that help define its identity. This article is meant to bring recent work in the IETF and W3C Web standards community to the attention to Semantic Web researchers working on entities and identification.

1 Introduction: What Makes Identifiers Work?

In order for the Semantic Web to succeed, URIs will be extended from identifying documents to all sorts of real-world things, and so become a genuine universal identification system. As said by Tim Berners-Lee at the WWW conference in 1994, "to a computer, then, the web is a flat, boring world devoid of meaning...this is a pity, as in fact documents on the web describe real objects and imaginary concepts, and give particular relationships between them" so adding semantics to the web involves two things: allowing documents which have information in machine-readable forms, and allowing links to be created with relationship values."[1]

This paper looks at how previous attempts at providing universal identification schemes for non-Web accessible things, such as DOIs and URNs have failed. We claim that they have failed precisely because they failed to allow decentralized creation of identifiers and allow accessibility to descriptions of the identifiers. We then explain how URIs fulfill both of these principles. Furthermore, earlier identification schemes conceptualized identifiers as a "dictionary" that mapped an identifier directly to its referent. However, we explain how identification works not as a dictionary, but in a "web of meaning" that allows the interpretation of a identifier to mapped to a range of possible referents. We demonstrate how this can be implemented by using the Principle of Self-Description, which combines the notions of accessibility and linking to allow a representation to "contain"

[1] See *http://www.w3.org/Talks/WWW94Tim/*

normative links to a program or document capable of producing an interpretation of the language that the URI partakes in. We also show how recent work in the W3C and IETF around the “Link” header has allowed this mechanism to be present in underlying protocol of the Web itself, HTTP.

2 Decentralization

The issue at hand is that there is a key difference between documents on the old-fashioned hypertext Web and “real objects and imaginary concepts,” mainly that documents like web-pages can be directly accessed on the Web, while real objects and imaginary concepts can not. This distinction was codified originally in the distinction between Universal Resource Locations (URLs) and Universal Resource Names (URNs). URLs are meant to be locations for web accessible resources like web pages which may not be persistent [1]. In contrast, URNs are unique names for things that may not be accessible over the Web such that “the URN will be globally unique forever, and may well be used as a reference to a resource well beyond the lifetime of the resource it identifies or of any naming authority involved in the assignment of its name” [15]. Despite their cosmic pretensions, URN schemes have only really ever been used to map already centralized names such as “isbn” and “mpeg” and have never had the explosive growth traditionally associated with URLs, despite their advantage of being both globally unique and persistent. There are two precise reasons why URNs never succeeded, and the lessons remain pertinent to projects like OKKAM that seek to create a “Web of Entities” [4].

First, centralization prevents growth in open systems like the Web. URNs and other more subject-specific schemes like ISBNs can only exist within domains that are already highly centralized with an agreed upon subject matter, which is but a small fragment of the world that people want to communicate about. Furthermore, such centralization of subject matter is reflected in bureaucratic centralization, for while URIs allowed any organization to register a URN scheme through liaisoning with the IETF, very few organizations did due to costs in time and effort to do so, as well as the obscurity of the process. Second, once a URN scheme was registered, the body that registered the scheme had to guarantee any use of the URN would be persistent and unique. Since the registering body had complete and total control of the URN scheme, new use of the URN scheme had to be explicitly approved. This stands in stark contrast to URLs, where while there is a centralized managing authority in the form of the top-level domain name system, anyone can register a top-level domain by mere exchanging of a relatively small amount of money and giving some relatively limited contact information. Once someone has purchased a top-level domain, the amount of URLs that one can produce under that URL is infinite and not necessarily centrally governed by the owner of the top-level URL. While registering a new URL scheme does require going through the IETF, the ability to without any centralized control mint a new URL is key to the explosive growth of URLs. Another example would be the creation of new topics in Wikipedia.

In order for success to be achieved in any identification scheme, the creation of new identifiers must be decentralized for the average user and have little to no deployment costs.

3 Accessibility

The second common reason for the failure of an identification scheme was the inability for the average user to tell what was identified. This was true of the URN effort in particular. If one has a URN, there is absolutely nothing one can do with it besides use it as a "place-holder" for the identified entity. In closed centralized systems with well-agreed upon subject matters, this sort of behavior can be useful, since agreement about the subject matter can be presupposed. The ISBN "978-1-59593-820-6" unambiguously within the domain of publications refers to the "Proceedings of the 18th Conference on Hypertext and HyperMedia." Yet people just do not use ISBNs - or URNs - to talk about pets or poets. While these centralized schemes may be useful in particular domains, their proposition for use-value is null outside their domain. Yet much of what average users want to communicate about cannot be parceled into clearly demarcated domains. The only way to successfully communicate what a particular identification scheme identifies in a decentralized manner is to have some sort of description accessible from the identifier. While experimental URN to URL translation services were planned, they have never moved beyond the experimental stage, and so this can be one of the reasons for the lack of use of URNs [15].

One interesting alternative scheme to URIs is DOIs (Digital Object Identifiers), as proposed by Robert Kahn, one of the inventors of TCP/IP and foundational figure of the Internet [13]. A DOI identifier such as "10.1145" can identify the "Proceedings of the 18th conference on Hypertext and Hypermedia," much like the ISBN "978-1-59593-820-6." Unlike the ISBN and like URIs, a DOI can also identify sub-parts, such as "10.1145/1286240.1286288" that identify the article "Towards better understanding of folksonomic patterns" in the proceedings of the previously mentioned conference. Unlike URNs, DOIs have implemented a standardized mechanism to map DOIs to accessible data, the Handle system, available at *http://www.handle.net*. This allows the handle for "10.1145/1286240.1286288" to be mapped to the URI *http://portal.acm.org/citation.cfm?id=1286240.1286288* that in turn allows a hypertext web-page to be delivered to a user, with links to copies of the paper. The PURL initiative from OCLC serves a very similar purpose.[2] However, is a centralized system like DOI that in effect redirects one persistent identification to a non-persistent scheme much of an improvement? Or is it "forcing a mailman to dance a jig before delivering the mail" with no real effect [11]? Indeed, it would seem so.

[2] See *http://purl.org/* for the system.

4 The Principle of Universality

The first advantage of the Semantic Web over other identification schemes is its decentralized nature: anyone can mint a new URI. The ability to register a new URI requires interaction with the domain name system via a simple financial transaction with a domain name registrar. While there is centralization as domain names are served on a first-come, first-serve basis and that it requires some overhead to get a domain name, once a single domain name is purchased, a very large numbers of URIs may be minted underneath a top-level domain name. Furthermore, if one has access a server that already has a simple top level domain name, one can mint a new URI without having to ask for centralized permission from anyone. In this way, by allowing any URI to be used as an identifier, the Semantic Web can avoid the centralization inherit in the URN approach.

The second advantage of the Semantic Web over previous universal identification schemes is that by using the *http* URI scheme it lets an user actually retrieve something to help determine what the URI "identifies" [18]. In other words, despite being originally meant to deliver hypertext, HTTP is now a universal delivery protocol for representations of nearly any sort, and so is the best protocol to guarantee some form of accessibility on the Web. Since HTTP is now ubiquitous, deployment of this technology costs nothing and is usable by the vast majority of users, unlike URNs or DOIs. The main disadvantage to this approach is that *http* URIs are not persistent, so that the resource may disappear at anytime. If one wants exponential growth due to decentralization, one must accept both the conditions that the identifier should allow access to some description of what the identifier is intended to identify and also that that resource may not be persistent.

These two points, the first pertaining to the creation of identifiers for anything and the second to the ability for representations to be accessible from these identifiers, can then be combined into a single principle. The ***Principle of Universality*** can be stated that *any resource can have a URI, and a representation of that resource can be accessible using that URI.*

To formalize this principle, we will have to employ a technique to interpret the quantifiers as ranging over "possible worlds" since it is rather obvious that every resource does not have a URI and there are many URIs that do not have accessible representations. So, in fact the principle can be thought of that "any resource can *possibly* have a URI, and a representation of that resource can be *possibly* be accessible using that URI." However, we can avoid using the modal operators in any formalization by implicitly having any quantifier embody a variant of the Barcan Formulae, which allow us to say that all possible worlds are actually just part of a single actual world [14]. Normally stated in terms of the modal operator for necessity ($\Box$), it can be rephrased in terms of the modal operator for possibility ($\Diamond$). For universal quantification $\forall x.\Diamond Fx \rightarrow \Diamond\forall x.Fx$, which can be stated as "if everything is possibly F, then it is possible that everything is F." For existential quantification we can state that $\exists x.\Diamond Fx \rightarrow \Diamond\exists x.Fx$, so that "if something is possibly F, then it is possible that something is F." Since we hold these two formulae to be true, then we from here on out will use quantifiers normally and hold the possibility to be always implicit in our quantification.

A URI is u, and a representation is r. This is casting the net of URIs resources as wide as possible, to include both resources that Web-accessible and those that are not. A URI has an accessibility relationship a with a representation if a request using a URI returns a representation r. The accessibility function a is an abstraction over the range of possible protocol functions and headers of the request and follows the various status codes and conventions, and so is compatible with the W3C's use of the "hash convention" and 303 redirects as well as possible future status codes and conventions [5]. Accessibility (a) relationships are transitive. The Principle of Universality can be stated as follows (keeping possibilities implicit), that for any resource, there exists a URI, such that that URI "identifies" the resource. The predicate id will be given for "identifies."
$\forall x \exists u.id(u,x)$
The consequent of identifying x with u is that a representation may be accessible from the URI.
$\forall x \exists u \exists r.id(u,x) \rightarrow a(u,r)$
The URI then also denotes at least one resource. We signal this "denoting" or "referential" relationship with the Φ relationship, as defined and argued for in "In Defense of Ambiguity" [11]. A URI may then denote a non-Web accessible resource and may be accessible.
$\forall x \exists u.id(u,x) \rightarrow \Phi(u,x)$
A URI may then both denote both a resource and host a representation.
$\forall x \exists u \exists r.id(u,x) \rightarrow \Phi(u,x) \wedge a(u,r)$
One can imagine the situation where the URI denotes a resource and hosts a representation that is also "about" that resource.
$\forall x \exists u \exists r.id(u,x) \rightarrow \Phi(u,x) \wedge a(u,r) \wedge \Phi(r,x)$

5 The Dictionary-Theory of Meaning

How does one tell what thing or things a URI denotes? One approach to meaning would be a "dictionary theory of meaning" where the URIs map directly to objects in the world. Imagine a giant dictionary that matches URIs with real-world objects. In this dictionary one could look up the URI *http://dictionary.example.org/wordsworth* and get back "William Wordsworth" himself. This dictionary could list all the types of concepts and entities that its users normally agreed upon, and translate them into their "real-world" referents. Of course this dictionary is fictional, since there is no way to return anything over the Web that can not be reduced to bits. At best, such a dictionary could only return some sort of authoritative information, such as an authorized biography. Even then there would be many cases where rampant disagreement would make getting the "facts" straight about even a real-world entity difficult, such as the birthday of Jimmy Wales of Wikipedia fame [9]. So, the dictionary would not be a dictionary of identity, but a dictionary for translation, with URIs being translated into "enough" relevant information in both natural and perhaps a machine language like RDF to "pin down" the referent.

The dictionary would face a crucial difficulty in the fact that the language the dictionary was translating from is no natural language, but just the language of URIs. So our dictionary of URIs has a strange parallel to the problem a linguist would have in translating the language of a "hitherto untouched people" into a known language, a problem dubbed *the problem of radical translation* by Quine [19]. How does one know if the URI for Wordsworth is about William Wordsworth the poet, his corpus of poems, or Wordsworth Technology Limited? Obviously, this could be done by inspecting the accessible representation. While the accessible representation may allow us to determine whether or not a link is about "Wordsworth Technology Limited" and "the poetry of Wordsworth the Poet" with ease, it would be far harder to tell apart a more fine-grained distinction such as "the poetry of Wordsworth of the Poet Laureate era" (whose poetry is considered not of high quality) from "the poetry of Wordsworth of the Lake Poet era" (whose poetry helped launch the English Romantic movement). This distinction is not facetious, as one might want to add a "five star" review to "the poetry of Wordsworth of the Lake Poet era" but not to "the poetry of Wordsworth of the Poet Laureate era." So, if one found a "five-star" review, should one assume it was to "poetry of the Wordsworth of the Lake Poet era" or the poetry of Wordsworth regardless of the era? The user of the dictionary would have to assume the "principle of charity" that the URI means the simplest thing possible that explains its usage - likely Wordsworth regardless of the era. Even with a Web-accessible representations for every URI, ambiguity will always be lurking in the shadows, since the creator of the explanation can not possibly cover every case of its intended usage in a single representation. No agent can ever "really" determine unambiguously what the URI means. This is Quine's thesis of radical indeterminacy of translation, which defeats any attempt to think of meaning as a simple dictionary [19]. In the dictionary theory of identity, the identity of every term is a semantic island, isolated and unknowable. Yet this clearly isn't how natural languages, formal languages, or the Web work.

6 The Principle of Linking: The Relational Theory of Meaning

An alternative to the "dictionary theory of meaning" is possible, which is the *relational theory of meaning. Meaning in general, including any identity conditions, fundamentally is part of a web of relationships*, and this applies both on and off the World Wide Web. No identity is an island, all identity is relational. Identity is built upon meaning in general, so only once a term participates in a meaningful web of use with other terms can it then be said to identity anything. In Frege's classic presentation, what determines the identity of a number is not the number in itself, but its relationship to other numbers and mathematical expressions [8]. Similarly, this principle applies to natural language in its own substitution principle. If the meaning of a word can be judged by its context in a sentence, so a substituted word that preserves the meaning of the sentence can be judged to be equivalent, i.e. to have the same *identity* as the word that in replaced in the context of that sentence. If one then wishes to determine the meaning of a sentence across a language, this principle applies, so that the meaning of a substituted sentence that preserves the meaning in a discourse can be judged to be equivalent in the context of the discourse. Meaning derives not from the syntactic form of the word itself but its connections to other words, sentences, and non-linguistic usage. This semantic holism applies not only to words and numbers, but to the Web. In RDF, formally a URI by itself does not mean anything. It merely denotes a object in the world, and these objects are denoted by virtue of being able to satisfy the relationships (in RDF, predicates) specified for that URI. There are usually a large number of objects in the world that are satisfied by relationships, not a single unambiguous one [11].

To clarify our terminology, the *meaning* of a URI would be a mapping from the URI itself to a *world*. This mapping and the world are called the *interpretation*. This interpretation could be given formally, in terms of creating the world using set theory, but can also in be a mapping to "real world objects and imaginary concepts." This interpretation of a URI would then denote the objects in the world, and the set of these objects would be the *identity* of the URI. It is precisely the relationships that the URI partakes in that give constraints on the possible worlds that satisfy the interpretation, worlds where the interpretation is true. The identity conditions can then be thought of as being given by the meaning of the relationships, so that all identity follows from the relationships and whatever additional semantic constraints are specified by the language. Although this is how formal semantics work, it also seems to be the best bet we have as to how informal semantics work [11].

The relationships between resources are captured by the *linking* between resources on the Web. A link is a directed connection between resources. Links are what transforms lone resources into a web. Likewise, links can be what allows non-Web accessible entities to participate in a web of meaning. A RDF predicate can be considered a type of link since it connects two resources. Links themselves can also be resources if they are given a URI (as in RDF), but are not necessarily resources, for links in standard HTML documents do not obviously have URIs themselves.

The ***Principle of Linking*** can be stated as *for any two resources there may be a link between them.* Linking can happen at a number of levels, both for reference and access. For the relationship l, the first argument is the "source" of the link and the second is the "target."
$\forall x_1 \exists x_2 . l(x_1, x_2)$
The most common use of a link is accessing a representation. This can be derived from the above formulation by giving the resources URIs via the Principle of Universality.
$\forall x_1 \exists x_2 \exists u_1 \exists u_2 . l(x_1, x_2) \wedge id(x_1, u_1) \wedge id(x_2, u_2)$
Therefore, a representation may be accessible from the URI:
$\forall x_1 \exists x_2 \exists u_1 \exists u_2 \exists r . l(x_1, x_2) \wedge id(x_1, u_1) \wedge id(x_2, u_2) \wedge a(u_2, r)$
The point that the link is between two resources with URIs can be summarized

by stating that two URIs can be linked.
$\forall u_1 \exists u_2 . l(u_1, u_2)$
We may also want to note that the link can be between a representation and a URI, as in hypertext.
$\forall r \exists u . l(r, u)$
Since a link is itself "some thing" and so a resource, it can be given a URI (although this requires having l both as a relationship and an object, something possible on the Web). This explains the parallel between the abstract notion of "linking" and the concrete notion of RDF predicates as links.
$\forall u_1 \exists u_1 \exists u_2 \exists u_3 \exists l . l(u_1, u_2) \wedge id(u_3, l)$
Each of these URIs may then denote resources.
$\forall u_1 \exists x_1 \exists x_2 \exists x_3 \exists u_2 \exists u_3 \exists l . l(u_1, u_2) \wedge id(u_3, l) \rightarrow \Phi(u_3, l) \wedge \Phi(u_1, x_1) \wedge \Phi(u_2, x_2) \wedge \Phi(u_3, x_3)$
In this manner, linking has a "dual nature" just like URIs, since links describe possibly non-Web accessible relationships between resources and provide accessible representations of the resources.
$\forall u_1 \exists x_1 \exists x_2 \exists x_3 \exists u_2 \exists u_3 \exists r_1 \exists r_2 \exists r_3 \exists l . l(u_1, u_2) \wedge id(u_1, x_1) \wedge id(u_2, x_2) \wedge id(u_3, l) \rightarrow \Phi(u_3, l) \wedge \Phi u_3, x_3 \wedge \Phi(u_1, x_1) \wedge \Phi(u_2, x_2) \wedge a(u_1, r_1) \wedge a(u_2, r_2) \wedge a(u_3, r_3)$

7 The Principle of Self-Description

Ideally, every resource should be self-describing, in that it should provide links to other resources that determine its meaning. The practical question is then how many and what sort of links are necessary to adequately describe a resource? The solution put forward is that one of the goals of the Web is for resources to be "self-describing." This is a slippery concept, currently defined as "individual documents become self-describing, in the sense that only widely available information is necessary for understanding them" [16].

How many and what sort of links are necessary to adequately describe a resource? We need to re-inspect what it means to describe a resource. A resource is successfully described if an interpretation is a possible. An interpretation can be defined as broadly as one wishes, ranging from a logical interpretation that maps the use of the URI onto a mathematical model to an informal interpretation by a human that maps the URI to "real-world" referents - and ideally, both. Given the URI *http://www.example.org/wordsworth* in a series of a RDF statements, we can formally have an interpretation onto the model given by the RDF Formal Semantics, which in turn gives the valid (albeit mostly uninteresting) entailments, entailments that could be automated [10]. With higher level languages like OWL, the number of valid entailments increase. A human can also inspect whatever information is returned by the URI *http://www.example.org/wordsworth*, like when a web-page containing natural language and images is returned so a human could identify the URI with the poet William Wordsworth, although a machine would have difficulty at best interpreting images and natural language. This process of following whatever data is linked in order to determine the interpretation of a

URI is informally called “following your nose” in Web architecture.

The “Following Your Nose” idea basically states that if a user agent encounters a representation in a language that the user agent can not interpret, one has three alternatives:

- **Inspect the Media-Type:** The media type of a representation provides the foremost normative declaration of how to interpret a representation. Since the number of IETF media-types is finite and controlled by the IETF, a user agent should be able to interpret these media-types.
- **Follow any Namespace Declarations:** Many representations use media-types that may be customized to define certain languages, like XML. In this case, if the language has a some ability to declare namespace declarations for the vocabulary, then the user agent may follow these namespace declarations in order to get more information needed to interpret the representation.
- **Follow any normative links:** Although what precisely defines a normative link can vary from language to language, the idea of *normative linking* is that some form of link *should* be followed, rather than *optionally* followed as most links are. In RDF Schema these kinds of links could be given as *rdfs:isDefinedBy* links and in OWL by the *owl:imports* links, although their normative status is unclear.

An example should clarify this. The notion of URI opacity states that interpretations should not be guessed from the text string of the URI alone. In this case, if no representation is accessed from *http://www.example.org/wordsworth* it would not be warranted to assume that URI is about “Wordsworth.” Assuming RDF/XML is returned from a HTTP GET on the URI (perhaps via 303 or hash redirection), the user agent may be able to be connect the media type of the language (such as *application+rdf/xml*) with a processor for the language. In case the user agent cannot find a processor, it could retrieve the normative IETF specification for the media-type[3] in the baseline human-readable media-type of *text/plain.* If an XML format is returned and the document has no media type or the user agent thinks the media type may be wrong, the agent can also “follow its nose” by sniffing the topmost node of the document and seeing what namespaces are declared. So it could find a declaration of the RDF namespace,[4] and could follow it in order to find a specification or program capable of giving the language an interpretation. Namespace documents are in turn given media-types, as many W3C Recommendations are in HTML and many Semantic Web namespaces return RDF Schema or OWL. In this case, the user agent can begin recursively searching for some sort of specification. Eventually, even everything could bottom out in specifications given by the IETF in plain, human-readable text - in worse case giving the human instructions on how to build the programs

[3] See *http://www.iana.org/assignments/media-types/*, and a mapping from media types to URIs has been proposed at *http://www.w3.org/2001/tag/2002/01-uriMediaType-9*.

[4] *http://www.w3.org/1999/02/22-rdf-syntax-ns#*

necessary to interpret the resources. This is why plain and human-readable text is the natural bottoming point of self-description.

In order to get mileage from the namespace document story, one would want full-fledged programs that can interpret the language a representation of a resource is given in attached to the vocabulary's namespace document. On the Semantic Web, most namespaces like the RDF Syntax namespace just returns an RDF schema, which is useless to a machine incapable of understanding RDF or OWL in the first place. A much better option would be to use a namespace document that could in turn link to programs that can determine valid interpretations for the language. These should be for a variety of programming languages and platforms, ideally with some that can be installed either "on the fly" with permission (in the manner of browser plug-ins) or available as Web Services. While this does present security concerns, these could be addressed via authentication, trust, and local policy just as any program installation currently does. An informal standard for namespace documents that provides types of links to applications and normative references already exists, the RDDL (Resource Directory Description Language) standard [3]. RDDL gives standardized links available as both human and machine-usable XLink links in order to make accessible various programs and specifications in a way that any agent could follow. A version of RDDL in RDF with an associated GRDDL transform exists in order to make it even easier for Semantic Web agents to follow namespace documents to associated resources [12, 20]. Therefore, by combining the accessibility given by the Principle of Universality with the Principle of Linking we can now describe how the Principle of Self-Description can solve decentralized deployment of not only things, but the languages that given them their interpretation, and thus identity.

The ***Principle of Self-Description*** can then be stated that for *for any given representation, the representation should be able to provide access to an interpretation for that representation.* We can state that a language (such as RDF) is given by the variable n, and a representation has an *encoding* relationship (e) with a language. A program that gives an interpretation of a language is given by p, and it has a "filtering" relationship f with a language n, since the program p can be considered to "filter" out invalid interpretations. Then we can get a valid interpretation I with regards to (wrt) a language n for a representation r.
$\forall r \exists n \exists p \exists I . e(r,n) \wedge a(r,p) \wedge f(p,n) \wedge wrt(I,n) \rightarrow I \models r$
The above case covers the case where the representation itself contains the filter program to give it an interpretation. Combined with the Principle of Linking, the filter program itself could be accessible via a link.
$\forall r \exists n \exists p \exists l \exists u . e(r,n) \wedge l(r,u) \wedge a(u,p) \wedge f(p,n)$
More precisely, from the representations or other links available from a namespace document.
$\forall r_1 \exists r_2 \exists n \exists p \exists u . e(r_1,n) \wedge l(r_1,u) \wedge a(u,r_2) \wedge l(r_2,p) \wedge f(p,n)$
Since this can be repeated, it may take several following several links to get a filter for a language. Furthermore, in order to get a filter of a language one may have to have an interpretation of another language, so one can repeat this

process until a valid “filter” program that can produce an interpretation can be found for the language. So, by the definition of a filter program p for a language n, we can then get our original statement:
$\forall r \exists n \exists p \exists u \exists I . e(r,n) \wedge l(r,u) \wedge a(u,p) \wedge f(p,n) \wedge wrt(I,n) \rightarrow I \models r$

8 Linking on the Semantic Web

While this is an attractive theoretical picture for demonstrating the use of links to understand identity on the Web, there is a crucial problem: a “link” that defines the normative identity of a resource is a stronger sort of link than that normally used in hypertext, which merely denotes some other resource of interest. For example, the Wordsworth Trust might be happy to link to the Wikipedia page, but seeing as they do not control it, they may not trust Wikipedia. The Wordsworth Trust would want a different kind of link to an authorized biography written by someone at the Trust. On the Semantic Web, the RDF triples given by a particular representation also constrain the interpretation of triples and even the representation itself (as when the URI that the representation was returned from is the subject of a triple). It is unclear if all RDF predicates are considered to be normative links, but user agents generally assumes they are.

Ideally, normative linking could be accomplished in both the hypertext and the Semantic Web in a uniform manner, although this far from the case today. Normative linking should be considered just as normative as retrieving media-types, although the media-type should override any information given by a normative link if there is a conflict. The normative linking mechanism should be clarified and made equivalent in power to retrieving namespace documents in XML. If multiple normative links are given, the results of following those multiple normative links should be merged if possible.[5] Due to security concerns, the ability of a user agent to follow normative links and use those to interpret a resource should be a matter of local policy, so the user can override certain kinds of link following.

In HTML these normative links are given by the *link* relation and in XHTML also by the *profile* header, which are in turn used by GRDDL to transform HTML to RDF [12]. GRDDL also licenses the *transformation* link in generic XML. On the Semantic Web. However, no consistent cross-vocabulary linking mechanism has been decided upon. One can imagine it is safe to retrieve *owl:imports* and even *rdfs:isDefinedBy* resources, although if these are “special” compared to other RDF predicates is unclear. While a few programs like GRDDL rely on self-describing resources, it does not generalize the technique, since GRDDL looks only for GRDDL-specific authorized RDF links. Currently virtually no browsers and agents support such automated “following-of-your-nose” to discover more useful information.

The solution would be a type of link that tells agents that this link normatively associates a description to the resource and for this to be implemented

[5] This especially makes sense if the results of following the links are given as RDF.

uniformly across both the hypertext Web, XML, and the Semantic Web. The owner of a URI could easily then just deploy this type of link to show their endorsement of a particular description. We propose that all of these links, ranging from GRDDL's *transformation* link in XML to the use of the *link* element in HTML headers be given as sub-properties of *rdfs:isDefinedBy*, and that user agents be allowed to retrieve these links when fetching a representation. However, unlike Booth's URI Declarations, we do not say that "following your nose" to a URI forces one to hold as true whatever triples are accessed from that URI, as following these should be a matter of local policy. In contrast, *the Principle of Self-Description merely states that if an interpretation using that URI cannot be given by the user agent, then the URIs normatively linked to should be followed in an attempt to find an interpretation, and the owner of a URI should provide the links needed to interpret the URI.*

For example, a user agent may trust some normative links and not others. These normative links could be typed with the URI *rdfs:isDefinedBy.* It would be stronger than *rdfs:seeAlso*, so that an application can expect to find whatever data at the end of the link to be licensed by the owner of the URI as an accurate and trusted description of their resource, in a similar manner to Booth's "URI Declarations"[2]. Since this link should be followed by agents, it could be considered the reverse of the "nofollow" non-standard HTML attribute put on links. Lastly, for those that consider the distinction between information resources and non-Web accessible things to be important, we could deploy a subproperty of *rdfs:isDefinedBy* called *ex:thingDescribedBy* that denoted a relationship between a non-Web accessible thing and an information resource. It could have a converse, such as *ex:describesThing.* Lastly, in order to make the self-description story work, RDDL could have their links be sub-properties of *rdfs:isDefinedBy.*

We can now outline how to solve the Wordsworth Trust's problem about linking to normative descriptions to *http://www.example.org/wordsworth.* We could deploy the "303 redirection" convention to redirect to *http://www.example.org/wordsworth/data* and then serve a RDF document containing associated descriptions they endorse about Wordsworth, like *ex:wordsworth foaf:birthday "7-4-1770"*, and a link to a human readable biography that the Trust endorses at *http://www.example.org/wordsworth/bio.* An RDF statement could then authorize this as an associated descriptions via *http://www.example.org/wordsworth ex:thingDefinedBy ex:wordsworth/bio.* Regardless, then regular links or *rdfs:seeAlso* links can be given to the human-readable Wikipedia page and other resources the Wordsworth Trust does not officially endorse. Furthermore, since the user agent may not be able to interpret RDF, a series of normative RDDL links can be made to a number of reasoners and RDF-enabled plug-ins that would allow non-RDF enabled user agents to interpret the resource. What is be needed would be for a RDDL document linking to various RDF browsers and interpreters, ideally the sort that can be automatically installed as plug-ins, at the RDF namespace URI.

9 The Return of the Link Header

One argument against using normative links is that people may not have access to the document itself, and so can not place links directly in the document. There is no reason why something semantically identical with links cannot be using other levels of Web architecture. One could imagine the use of the "link" element in the header of a HTML document to specify a normative link. One could also use a revived *Link* HTTP header, as was included in an earlier version of HTTP given by RFC 2068 [7], but was left out of the latest RFC 2616 [6]. The main issue with the *Link* header is that it would lack the ability to type itself as a normative link or not, and so would be equivalent to a hypertext link.

The *Link* header could be given proper semantics if it could be used in combination with a link relation that allowed for the proper semantics to be specified. Each *Link* in HTTP can then be paired with two URIs, the target of the link and an optional URI specifying the type of the link. This second type of URI is called the *link relation.* In this manner, one could have a HTTP *Link* that specified its relation was *rdfs:isDefinedBy* or *ex:thingDescribedBy.* Link relations are superior than the use of a Profile header as given in HTML, since in HTML (or even a HTTP header for *Profile*) can specify both multiple links and multiple profiles, leading to ambiguity about which profile URI is matched with a particular *Link.*

Imagine the Wordsworth Trust does not want to disturb whatever delicate HTML is currently served by *http://www.example.org/wordsworth/bio* but wants to add a statement that this page describes Wordsworth the person as given by *http://www.example.org/wordsworth.* Assuming a world where the *Link* header with URI link relations are valid parts of HTTP, we can fix this problem without altering any actual content hosted at *http://www.example.org/wordsworth/bio.* All it then needs to do is to create a *Link* with a link relation of *ex:describesThing* to *http://www.example.org/wordsworth/bio.* Various information, like Creative Commons licensing and GRDDL transforms of HTML to RDF, can also be linked using *Link* headers with the appropriate link relations in order to make the URI self-describing, all without any change to any HTML or RDF documents.

There is no reason to restrict the fundamental predicates for associating descriptions to RDF or even the *Link* header. They should be able to be inserted in HTML headers for those unfamiliar with RDF and even inside HTML bodies in a style similar to microformats, in order to allow normative links to be used by those without access to any headers or knowledge of RDF. So all the following should be equivalent, with the document given implicitly having the URI *http://www.example.com/document* :

1. As a typed link header, as in *Link: http://www.example.org/thing rel="http://www.example.org/describesThing"*

2. As a normal RDF statement, as in *http://www.example.com/document http://www.example.org/describesThing http://www.example.org/thing.*

3. In HTML (and arbitrary XML):

```
<HTML><HEAD>
<LINK rel="http://www.example.org/describesThing"
href="http://www.example.org/thing">
</HEAD>
....
```

4. In HTML body RDFa style:

```
<BODY>
<DIV rel="ex:describesThing" class="ex:thing"
xmlns:ex="http://www.example.org/">
Some Text
</DIV>
</BODY>
```

This use of the Link header with URIs for link relations is currently supported in Mark Nottingham, Chair of the IETF HTTP Working Group's, current IETF Draft, and so is likely to become part of HTTP [17]. The RDFa extension is already a valid part of RDFa, as it the use of URIs in the *rel* attribute in HTML, although it needs to be altered in HTML 5 to use URIs.

10 Conclusion: Working in the Wild

Regardless of the details, the use of any technology in Web architecture for identification does nothing more than allow the owner of a URI to explain what they intend a URI to identify. Ultimately, there is nothing that Web architecture can do to ensure that a URI identifies one thing. However, by giving an machine agent normative links to the necessary programs that can give an interpretation allows an agent to at least give a best effort at aligning their interpretation with that of the owner of the URI. This would vastly improve the situation from where it is today, where implementing some level of self-description to even communicate what the owner thinks is difficult in many circumstances without making huge assumptions about the user agent. The proposed return of the *Link* Header in HTTP is just one symptom of a wider return to linking, as opposed to centralized dictionaries, to solve the identity crisis on the Web. Instead of relying on a single centralized dictionary or placing hope in using arcane redirection techniques by themselves, the creators of URIs for entities should easily deploy links to accessible resources on a variety of levels, a tried and tested method of creating self-describing identity on the Web.

References

1. T. Berners-Lee, R. Fielding, and M. McCahill. IETF RFC 1738 Uniform Resource Locators (URL), 1994. http://www.ietf.org/rfc/rfc1738.txt.

2. D. Booth. URIs declaration versus use. In *Proceedings of Identity, Reference, and the Semantic Web Workshop at the ESWW Conference*, 2008.
3. J. Borden and T. Bray. Resource Directory Description Language (RDDL), 2002. http://www.rddl.org/.
4. P. Bouquet, H. Stoermer, and D. Giacomuzzi. OKKAM: Enabling a Web of Entities. In *i3: Identity, Identifiers, Identification. Proceedings of the WWW2007 Workshop on Entity-Centric Approaches to Information and Knowledge Management on the Web, Banff, Canada, May 8, 2007.*, CEUR Workshop Proceedings, ISSN 1613-0073, May 2007. online http://CEUR-WS.org/Vol-249/submission_150.pdf.
5. D. Connolly. A pragmatic theory of reference for the web. In *Proceedings of Identity, Reference, and the Web Workshop at the WWW Conference*, 2006. http://www.ibiblio.org/hhalpin/irw2006/dconnolly2006.pdf.
6. R. Fielding, J. Gettys, J. Mogul, H. Frystyk, and T. Berners-Lee. IETF RFC 2068 hypertext transfer protocol - HTTP 1.1, 1999. http://www.ietf.org/rfc/rfc2616.txt.
7. R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. IETF RFC 2068 hypertext transfer protocol - HTTP 1.1, 1997. http://www.ietf.org/rfc/rfc2068.txt.
8. G. Frege. Uber sinn und bedeutung. *Zeitshrift fur Philosophie and philosophie Kritic*, (100):25–50, 1892.
9. A. Ginsberg. The big schema of things. In *Proceedings of Identity, Reference, and the Web Workshop at the WWW Conference*, 2006. http://www.ibiblio.org/hhalpin/irw2006/aginsberg2006.pdf.
10. P. Hayes. RDF Semantics, 2004. W3C Recommendation. http://www.w3.org/TR/2004/REC-rdf-mt-20040210/.
11. P. Hayes and H. Halpin. In defense of ambiguity. *International Journal of Semantic Web and Information Systems*, 4(3), 2008.
12. D. Hazael-Massieux and D. Connolly. Gleaning resource descriptions from dialects of language. In *Proceedings of XTech*, Amsterdam, Netherlands, 2005.
13. R. Kahn and R. Wilensky. A framework for distributed digital object services. *International Journal on Digital Libraries*, 6(2), 2006.
14. R. Marcus. A functional calculus of first order based on strict implication. *Journal of Symbolic Logic*, pages 1–16, 2946.
15. M. Mealling and R. Daniel. IETFRFC (Experimental) 2483 URI resolution services necessary for URN resolution, 1999. http://www.ietf.org/rfc/rfc2483.txt.
16. N. Mendelsohn. The self-describing web. Draft tag finding, W3C, 2006. Last accessed on Monday November 26th.
17. M. Nottingham. IETF Internet Draft HTTP Header Linking, 2008. http://www.mnot.net/drafts/draft-nottingham-http-link-header-01.txt.
18. S. Pepper. The case for published subjects. In *Proceedigs Identity, Reference, and the Web Workshop at the WWW Conference*, 2006. http://www.ibiblio.org/hhalpin/irw2006/spepper2.pdf.
19. W. Quine. *Word and Object.* MIT Press, Boston, Massachusetts, 1960.
20. N. Walsh and H. Thomsppn. Associating resources with namespaces, 2007. Draft Tag Finding. http://www.w3.org/2001/tag/doc/nsDocuments/.

This article was processed using the LaTeX macro package with LLNCS style