An Ontology of Resources for Linked Data

Harry Halpin
Institute for Communicating and Collaborative Systems
University of Edinburgh
2 Buccleuch Place
Edinburgh, United Kingdom
H.Halpin@ed.ac.uk

Valentina Presutti
Semantic Technology Laboratory
ISTC-CNR
Via Nomentana 56, 00161
Rome, Italy
valentina.presutti@cnr.it

ABSTRACT

The primary goal of the Semantic Web is to use URIs as a universal space to name anything, expanding from using URIs for webpages to URIs for "real objects and imaginary concepts," as phrased by Berners-Lee. This distinction has often been tied to the distinction between information resources, like webpages and multimedia files, and non-information resources, which are everything from real people to abstract concepts like 'the integers.' Furthermore, the W3C has recommended not to use the same URI for information resources and non-information resources, and several communities like the Linked Data initiative are deploying this principle. The definition put forward by the W3C, that information resources are things whose "essential nature is information" is a difficult distinction at best. For example, would the text of Moby Dick be an information resource? While this problem could safely be ignored up until recently, with the rise of Linked Data and projects like OKKAM, it appears that this problem should be modelled formally. An ontology called IRW (Identity and Reference on the Web) of various types of resources and their relationships, both for the hypertext Web and Linked Data, is presented. It builds upon Information Object Lite (an extension of DOLCE Ultra Lite for describing information objects) and IRE (an earlier ontology of and aligns with other work in this area. This ontology can be used as a tool to make Linked Data more self-describing and to allow inference to be used to test for membership in various classes of resources.

Categories and Subject Descriptors

H.3.d [**Information Technology and Systems**]: Metadata

General Terms

Knowledge Representation

Keywords

Linked Data, ontology, resource, Web architecture

1. INTRODUCTION

The key feature of the Semantic Web is not its use of knowledge representation technologies like ontologies and inference per se, but the introduction of these technologies to operate over Web resources as defined by URIs. Early Semantic Web efforts forgot this, and treated URIs as just odd sorts of symbols. The Linked Data Tutorial provided a way for putting Semantic Web technologies in harmony with Web architecture, and now Linked Data is experiencing amazing growth. Yet, there is still debate within Web architecture circles as to what the definition of a 'information resource' is, a term crucial to Linked Data, and how terms like this relate to the pre-Semantic Web hypertext Web. We model the terms used in Linked Data and Web architecture using a lightweight formal ontology in OWL-DL, which we call *IRW*, for 'Identity of Resources on the Web.' The hope is this ontology will clarify these debates and allow further development of a provenance-aware and semantically verified Linked Data Web.

Before trying to figure out the difference between a 'non-information' and 'information' resource, what is a resource? The W3C TAG state in their *Architecture of the Web* that 'resource' is used in a general sense for whatever might be identified by a URI [**?**]. Previously, a resource was thought of as strictly to be for network-accessible objects such as webpages, since the term 'resource' is defined by Fielding in the first HTTP RFC as "a network data object or service, identified by a URI". However, Berners-Lee broadened the concept of resource in his RFC 2396, stating that "a resource can be anything that has identity. Familiar examples include an electronic document, an image, a service (e.g., 'today's weather report for Los Angeles'), and a collection of other resources. Not all resources are network 'retrievable'; e.g., human beings, corporations, and bound books in a library can also be considered resources" [**?**].

One distinction that has been upheld by Hayes and others is the distinction between reference and access [**?**]. Making an analogy between URIs and names, *access* means "that the name provides a causal pathway to the thing, perhaps mediated by the Web" while *reference* means that "the name is being used to mention the thing," which may or may not coincide with access [**?**]. Something is then 'Web-accessible' if it can accessed via the use of HTTP. This use of the term 'resource' for both referring to non-Web accessible things and for naming Web-accessible things is continued in URI RFC 3986, the current IETF RFC, which states that "this specification does not limit the scope of what might be a resource; rather, the term 'resource'... likewise, abstract concepts can be resources, such as the operators and operands of a mathe-

LDOW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

matical equation, the types of a relationship (e.g., 'parent' or 'employee'), or numeric values (e.g., zero, one, and infinity)" [?]. It is precisely this ability to name things with URIs that aren't Web-accessible that defines both the Semantic Web and Linked Data. However, unlike traditional Semantic Web applications, Linked Data allows Web-accessible *associated descriptions*, in both machine and human-readable forms, to be accessed from a URI for a non-information resource.

The most obvious distinction is between a resource that could *in principle* be Web-accessible, like a webpage, and a resource that is not in principle Web-accessible, like the Eiffel Tower itself. This distinction is given by the W3C TAG as the distinction between an information resource and something that may not be an information resource [?]. The W3C TAG then define an *information resource* as something "whose essential characteristics can be conveyed in a message," which is a controversial definition [?]. As noted by the Linked Data tutorial, this implies there is another kind of resource, *non-information resources*, for things that are not possibly Web-accessible, like a URI whose primary purpose is to refer to the Eiffel Tower [?]. Furthermore, one can distinguish 'Web resources' (a subset of information resources) that *are* usually Web-accessible, such as web-pages, from things that simply carry information, like the text of Moby Dick, regardless of whether it is on the Web or not. Again, let us emphasize that some find these distinctions very intuitive, while others do not. Lastly, in order to distinguish URIs for non-accessible things on the Semantic Web (the 'Cool URIs for the Semantic Web') from the normal use of URIs on the hypertext Web, we call the former *Semantic Web URIs* [?]. In Web architecture circles, what are typically called 'webpages' are just one kind of a 'representations' of a resource [?]. In order to distinguish the use of the word 'representation' in Web architecture circles from its normal usage, the word *Web Representation* is used in this paper to designate a more encompassing notion of representation of a resource, i.e. any set of bits that is 'coming down the wire' in response to the use of the Web.

2. LINKED DATA AND REDIRECTION

Linked data allows the access of associated descriptions from URIs for non-information resources by use of redirection. This was codified by the W3C TAG when it officially resolved *httpRange-14* by saying that the `303 See Other` HTTP header can serve to disambiguate between information resources and possible non-information resources. The official resolution by the TAG is given below as [?]:

- If an HTTP resource responds to a GET request with a `2xx` response, then the resource identified by that URI is an information resource;
- If an HTTP resource responds to a GET request with a `303 (See Other)` response, then the resource identified by that URI could be any
- If an HTTP resource responds to a GET request with a `4xx` response, then the nature of the resource is unknown.

One concrete example would be an agent is trying to access a URI that refers to the Eiffel Tower itself, `http://dbpedia.org/resource/Eiffel_Tower`. Upon attempting to access that resource with a HTTP GET request on a URI, since the Eiffel Tower itself is not an information resource, no Web representations are directly available. Instead, the agent gets a `303 See Other` that in turn redirects them to an information resource that hosts Web representations about the Eiffel Tower, such as `http://dbpedia.org/page/Eiffel_Tower`. When this URI returns the 200 status code in response to an HTTP GET request, the agent can infer that `http://dbpedia.org/page/Eiffel_Tower/` is actually an information resource. The Semantic Web URI used to refer to the Eiffel Tower itself, `http://dbpedia.org/resource/Eiffel_Tower`, could be any kind of resource and so *could* be a non-information resource [?]. This example is illustrated in Figure ??, using terms from the IRW ontology introduced in Section ??. An alternative to the `303` redirection is the *hash convention*, in which one uses the fragment identifier of a URI to get redirection 'for free' with smaller RDF vocabularies. If one wanted a Semantic Web URI that referred to the Eiffel Tower itself without the hassle of a 303 redirection, one would use the URI `http://www.tour-eiffel.fr/#it` to refer to the Eiffel Tower itself. Since browsers either dispose of or treat the fragment identifier as a fragment of a hypertext document or some other Web representation, if an agent tries to access via HTTP GET a Semantic Web URI that uses the hash convention, the server will not return a `404 Not Found` status code, but instead will resolve to the URI before the hash, `http://www.tour-eiffel.fr`, which can then be an Web resource capable of returning Web representations, which is called an 'associated description' in the Linked Data community [?]. In this way, Semantic Web inference engines can keep the Semantic Web URI that refers to the Eiffel Tower and an associated description about the Eiffel Tower separate by taking advantage of the predefined behaviour in web browsers. However, practically the `303` redirection of the W3C TAG and the hash convention leave the question of whether a resource is an information resource or non-information resource indeterminate, since there is nothing to prevent `303` redirection from being used to redirect from one information resource to another information resource, and the hash convention is dependent on media types, being more often used for named parts in the document in HTML instead of as a shortcut for distinguishing non-information resources and their associated descriptions.

3. RELATED WORK

There has been some related work in this area. Mogul has suggested that there are fundamental disagreements about what precisely the difference between an HTTP entity and a "representation of a resource" are, and that this leads to widespread problems with caching implementations in HTTP [?]. David Boorh has proposed an informal categorisation of what can be identified by a URI, noticing the confusion between 'naming' and 'identifying' and even 'describing' [?]. Hayes has long attempted to elucidate the fundamental difference between the use of resources to access webpages and the use of a URI to refer to some non-Web accessible thing [?]. Furthermore, the use of URIs to refer to physical entities and the subsequent clarification of the direct reference position has led to the OKKAM project, a project to build a catalogue of 'entity' URIs that is supposed to directly refer to physical entities [?]. This general line of thinking has led to a number of workshops at conferences

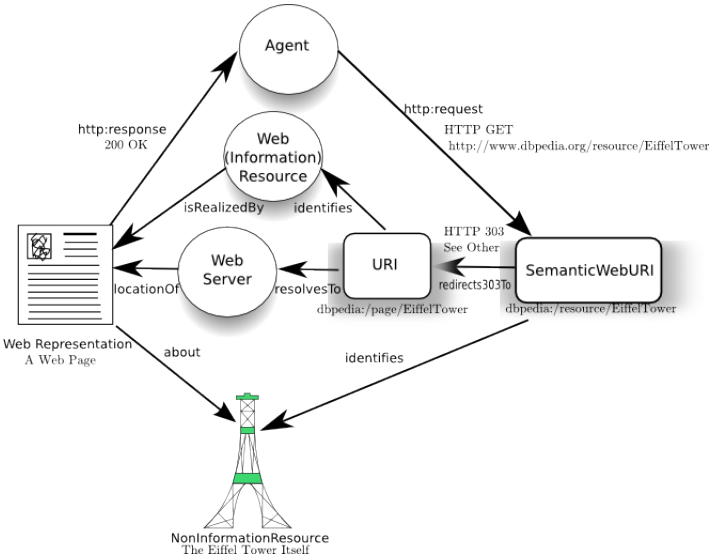

Figure 1: 303 Redirection for Semantic Web URIs

such as the World Wide Web Conference and the European Semantic Web Conference devoted to this topic [**?**, **?**].
Within the W3C, there is an informal activity of the W3C TAG called the 'Architecture of the Semantic Web' (AWWSW) that has for over a year attempted to decipher Web architecture, in part prompting by the need to model HTTP in RDF directly in order for HTTP transactions to be validated via EARL, the RDF-based *Evaluation and Report Language* used by the W3C to validate new W3C standards and describe test-cases [**?**, **?**]. Yet, HTTP in RDF currently does not model the notion of 'resource' except with a misuse of `rdf:Alt`, so it must be corrected by integrating an ontology of resources like IRW. While both EARL and the AWWSW are attempting a much more detailed and low-level description of HTTP transactions than we attempt, the lightweight IRW ontology described in this paper should allow specifications like HTTP in RDF to directly address the notion of a 'resource.'

4. THE USE OF A FORMAL ONTOLOGY

The primary use of a formal ontology in the context of Linked Data is to provide a foundation for the use of a common ontology to describe Linked Data and typical Linked Data transactions, currently being done by different ontologies in Section **??**. To this aim, IRW can be discussed, reviewed, and comment on the ontologydesignpatterns.org wiki[1]. To serve the aim of elucidating arguments, additional modules of IRW have been developed and are briefly introduces in Section **??**.
There have been previous attempts to model at least a subset of the notions outlined in a formal ontology, but all lack coverage of some crucial concepts. For example, while the ontology given by RDF Schema touches upon the vocabulary of resources via its term `rdfs:Resource`, it does not cover the distinction between information and non-information resources. The IRE (Identifiers, Resources, and Entities), based on Dolce Ultra Lite (DUL),[2] a light version of the widely-known DOLCE foundational ontology and its extension for describing information objects[3] (IOL, described in [**?**]), attempted to model some of these concepts earlier [**?**]. However, many aspects were not included in IRE, such as the distinctions between resources and their Web representations, or the concept of accessing a web-page via a web server, that are crucial to the efforts within the W3C and Web community, while many of the distinctions drawn by DUL+IOL were found to be too 'heavy-weight' for these communities [**?**]. In response to these concerns, the IRE ontology has been evolved into the IRW ontology.

5. THE IRW ONTOLOGY

The prefix `irw:` is for the namespace `http://purl.org/NET/irw/` of the IRW ontology. The stable version of the ontology can also be accessed via its PURL. The latest version of the IRW ontology may be accessed at: `http://ontologydesignpatterns.org/ont/web/irw.owl`. The prefix `rdfs:` is used for the RDF(S) namespace `http://www.w3.org/2000/01/rdf-schema#`. `ir:` is `http://www.ontologydesignpatterns.org /cp/owl/informationrealization.owl`. While the IRW ontology in full can not explicated due to lack of space, the primary classes and properties are given in Figure **??**. The IRW-related elements needed for the example of 303 redirection are given in Figure **??**. The IRW ontology starts with `irw:Resource`. While this class expresses the same intuition as `rdfs:Resource`, we have defined it because this version of IRW is within OWL-DL expressivity. In OWL Full, this class is equivalent to `rdfs:Resource`.

Identification and reference..

The notion of a URI is modeled as a class, `irw:URI` that has exactly one value for the datatype property `irw:hasURI` allowing to specify its value. Modelling URIs as a class allows us to talk about different kinds of URIs, such as IRIs (Internationalized Resource Identifiers) and Semantic Web

[1] `http://ontologydesignpatterns.org/wiki/Submissions:IRW`
[2] `http://www.loa-cnr.it/ontologies/DUL.owl`
[3] `http://www.loa-cnr.it/ontologies/IOLite.owl`

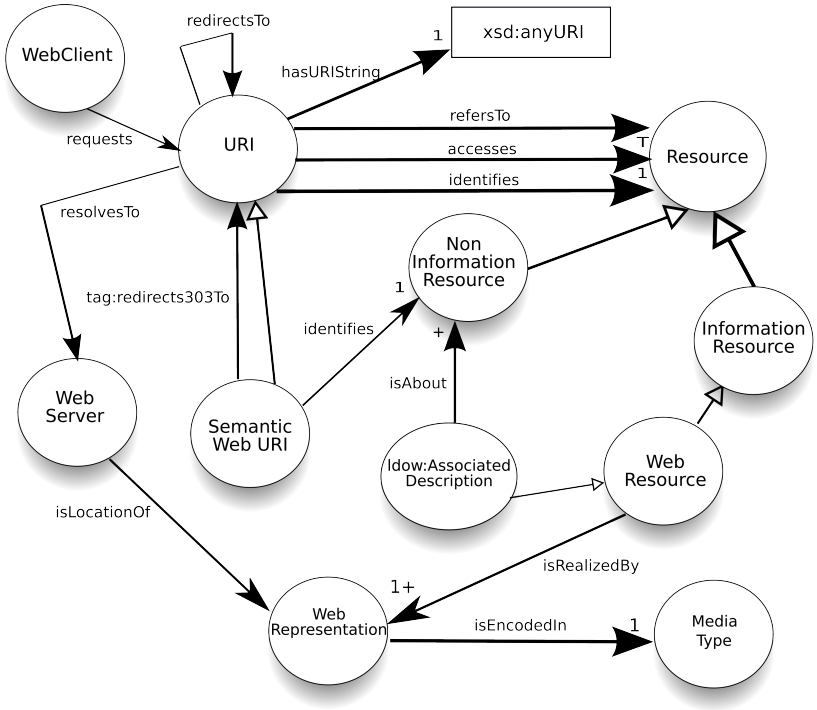

Figure 2: The IRW ontology illustrated as a graph. Rounded nodes are classes, while rectangular ones are datatypes. Arcs ending with an empty triangle are `rdfs:subClassOf` relationships. Arcs ending with a filled triangle are either object properties or datatype properties depending of the range node. Arcs' direction indicates the domain and range of the property. A '1' associated to a property means it is functional, a 'T' means it is transitive, '1+' means 'at least one'. Prefixes are indicated only if different from `irw:`.

URIs. According to some like Berners-Lee, URIs identify exactly one resource. This is modeled in IRW by the functional property `irw:identifies`, having range `irw:Resource` (and inverse property, `irw:isIdentifiedBy`). Of course, those that disagree with this viewpoint may not use `irw:identifies`, and so it is given sub-properties `irw:accesses` and `irw:refersTo`. The idea of reference as explicated by Hayes is modeled by the object property `irw:refersTo` (and inverse property, `irw:isReferencedBy`) [**?**]. One condition on this property is that the object of reference should be "immediately causally disconnected" from its subject [**?**]. This is important, as reference is the relationship to both URIs for non-information resources like the Eiffel Tower or integers, but also applies to the relationship of an information resource to some non-information resource, like the relationship of Tim Berner-Lee's homepage to Berners-Lee himself. So, the key point is that URIs can identify resources, and some of these URIs refer to non-information resources.

Access and redirection..

Distinct from reference is the `irw:accesses` relationship, which is a causal connection to the thing identified. This is modelled again as a relationship between URIs and resources, although it is transitive, unlike `irws:refersTo`. If one can access a and a accesses b then a accesses c (via b). Although a wide notion, access allows us to model the typical HTTP request-response Web transactions between a Web client and a server. A URI may also have a `irw:redirectsTo` property, a sub-property of `irw:accesses`, that we can use to model HTTP redirection. However, since redirection can be used between just information resources that have nothing to do with the Semantic Web, their domain and range says nothing about the type of resource. In order to model explicitly the redirection, two distinct sub-properties of this have been added in a TAG-specific module of IRW[4] that contains `tag:redirects303To` property and a `tag:redirectsHashTo` property. Obviously, `tag:redirects303To` models the TAG's 'solution' to *httpRange-14* while `tag:redirectsHashTo` represents the hash convention.

Types of resources..

Having defined reference and redirection, we can now categorize resources. There are two main disjoint sub-classes of `irw:Resource`. The first subclass is given as `irw:InformationResource`, which is an information object, such as a musical composition, a text, a word, a picture. An information object is an object defined at a level of abstraction, independently from how it is concretely realized. So an `irw:InformationResource` expresses the same intuition and is an equivalent class to the DUL+IOL information object [**?**]. This means an information resource has, via the `ir:realizes` property (with inverse `ir:isRealizedBy`), at least one `ir:InformationRealization`, a concrete *realization*. This term is again imported from DUL+IOL [**?**]. So an information resource's "essential characteristics can be conveyed in a single message" implies that everything from a bound book to an HTTP message can be a realiza-

[4] `http://www.ontologydesignpatterns.org/ont/web/tag2irw.owl` associated with prefix `tag:`.

tion for an information resource [?]. Furthermore, the property `irw:isAbout` (and inverse property, `irw:isTopicOf`) expresses the relationship of an information resource to a resource or resources the information is 'about.' Examples of this are descriptions of a resource using natural language or depictions of a resource using images. Information resources also can, but not necessarily, be identified (either accessed or referred to) with a URI. In this manner, the text of Moby Dick can be an information resource since it could be conveyed as a single message in English, and can be realized by both a particular book or a webpage containing that text.
Note `irw:NonInformationResource` complements `irw:InformationResource` from which it is disjoint with. Such class represents things that can not themselves – for whatever reason – be realized as a single digitally encoded message. A number of different kinds of things may be `irw:NonInformationResources`. Since this concept is the cause of much confusion and debate, it is detailed with three disjoint sub-classes. These kinds of IRW distinctions are not normative, as there are other possible plausible, more detailed modeling choices. Our aim here is of communicating the intuition behind the concepts of information and non-information resources without entering the philosophical debate about top-level ontologies. IRW contains three sub-classes of `irw:NonInformationResources`:[5]
`irw:PhysicalEntityResource`, is a resource that is 'touchable' like physical people, artifacts, places, bodies, chemical substances, biological entities;
`irw:ConceptualResource`, which refer to resources that are created in a social process that can not be completely realized digitally, such as legal entities, political entities, social relations, as well as the concept of horse and imaginary objects like unicorns; and finally `irw:AbstractResource`, which refers to abstract combinatorial spaces that cannot be located in space-time such as formal entities like functions or the integers as well as more mundane resources like the infinite set of names that constitute the resource identified by URIs themselves. A sub-class of `irw:InformationResource` is `irw:WebResource`, which is an information resource identified by at least one URI and realized by at least one `irw:WebRepresentation`, so that a Web resource is just an information resource that is realized by at least one accessible Web representation like a web-page. `irw:WebRepresentation` is a sub-class of `irw:InformationRealization` with constraints added to make the cardinality of `ir:isRealizedBy` and `irw:isIdentifiedBy` both at least 1. In this way IRW can distinguish between a resource for the text of 'Moby Dick' in general and a webpage about 'Moby Dick.'

Hypertext Web transactions..

The typical hypertext Web transaction can be modelled by IRW. We begin with `irw:WebClient`, which is some client in the context of the Web that can have a `irw:requests` relationship to a URI (note that `irw:requests` serves as an hook to the alignment of IRW with *HTTP in RDF* [?]), as exemplified by a typical `HTTP GET` request). The `irw:requests` property is a sub-property of `irw:access`. A `irw:WebClient` then `irw:requests` a `irw:URI`. We also introduce the class `irw:WebServer`, which has a `irw:isResolutionOf` property that relates a URI to a concrete Web server (inverse property `irw:resolvesTo`). This `irw:resolvesTo` property is currently implemented by mapping a URI to an IP address or addresses. So each `irw:WebServer` is the resolution of at least one `irw:URI`. Additionally, a `irw:WebServer` has a `irw:isLocationOf` property with at least one `irw:WebRepresentation` (inverse property, `locatedOn`), indicating the Web server concretely can respond to an HTTP request with a particular Web Representation.

Linked Data transactions..

The typical Linked Data transaction is also modeled. A new sub-class of `irw:URI`, `SemanticWebURI` is given, where the Semantic Web URI has a constraint that it must have at least one `irw:redirects` property. In the Linked Data Initiative, another important kind of resource is "associated descriptions," which is just an Web resource that can be accessed via redirection from a Semantic Web URI [?]. For example, in DBPedia[6] the resource `dbpedia:/resource/Eiffel_Tower` redirects to an associated description at `dbpedia:/data/Eiffel_Tower`, and to an HTML page at `dbpedia:/page/Eiffel_Tower` depending on the requested media type [?]. This scenario can be generalized:
a `irw:WebClient irw:requests` a `irw:SemanticWebURI` x and the request is redirected (e.g. via hash or 303 redirection) to another URI, where this second URI identifies an `ldow:AssociatedDescription`,[7] which has one `irw:isAbout` property to a non-information resource. We model `ldow:AssociatedDescription` as a subclass of `irw:WebResource`.

6. ALIGNING IRW TO OTHER ONTOLOGIES

In this section, we present a number of suggested alignments, as given in Table 2. The alignments are to the three primary other ontologies, the *RDF in HTTP* ontology [?], and the IRE ontology as well as an ontology for HTTP used by the Tabulator Browser [?, ?]. The namespaces for `ont` is `http://www.w3.org/2007/ont/http`. IRE, due to its modular construction and re-use of terms from DUL+IOL patterns, uses many namespaces, but they can be found at `http://www.ontologydesignpatterns.org/cpont/ire.owl`. The `http` namespace is `http://www.w3.org/2006/http#`.

7. APPLICATIONS

There are several applications of this ontology. The first is to solve the problem noted earlier that currently Linked Data resources are still not self-describing, such that there is no "definition, description, some other kind of indication of what the identifier is intended to identify" on the level of a resource [?]. If one gets a URI of Linked Data, how can one record that it for a non-information resource or an associated description, besides *actually* going to the URI and performing HTTP GET. Then, how should one record

[5]Note that the three classes does not constitute an exhaustive partition.

[6]Prefix `dbpedia:` is used for the namespace `http://dpedia.org`

[7]Typical Linked Data terminology is represented in a specific module of IRW represented here by the prefix `ldow:` referring to the namespace `http://ontologydesignpatterns.org/ont/web/ldow2irw.owl`

Class or Property	Alignments
irw:WebRepresentation	owl:equivalentClass http:Message owl:equivalentClass ont:ResponseMessage rdfs:subClassOf ire:InformationRealization rdfs:subClassOf ir:InformationRealization
http:Content	rdfs:subClassOf ir:InformationRealization
http:MessageHeader	rdfs:subClassOf ir:InformationRealization
irw:InformationResource	owl:equivalentClass ir:InformationObject
irw:SemanticWebURI	ire:SemanticWebURI
irw:identifies	ire:isExactProxyFor
irw:isAbout	ire:about

Table 1: Mapping of IRW to Other Ontologies

this provenance? The IRW ontology this in turn allows the *semantic validation*, to be able to describe and infer in detail the types of resources that can be interacted with via HTTP, which is useful for both tools like EARL that record validation of Web standards to be implemented in a reliable fashion, which is useful for error-reporting on the Web in general and HTTP in particular [**?**]. One facet of semantic validation is the description of Linked Data, where terms like non-information resource and associated description become important. This is useful for both semantic validation of Linked Data and Semantic Web Search engines [**?**].

7.1 Making Linked Data Self-Describing

There would be a number of advantages if webpages that have RDF content could distinguish themselves as such, in the same way that HTML 'valid' documents are currently validated by W3C Validators and often mark themselves by a computer graphic. This can be done by embedding a IRW statement in RDF/XML documents, RDF returned from SPARQL endpoints, and RDFa or GRDDL statement in XHTML or XML documents [**?**]. Ideally, this would be in conjunction with some sort of graphical logo to distinguish the page as 'Linked Data Enabled,' as detecting the RDF statement, even in RDFa, is difficult for humans. Second, for `irw:NonInformationResources` that are part of the Linked Data and thus have no Web Representation to embed such a statement in, or resources whose actual Web can not be changed or must be changed en mass, such a RDF triple can be embedded directly in HTTP via the use of the HTTP `Link` Header [**?**].

7.2 Semantic HTTP Validation

For EARL, we can then use the inference not only to detect the presence of Semantic Web URIs and Information Resources, but also to determine constraints and contradictions. For example, one constraint that EARL is interested in finding out is whether namespaces documents are employing either the hash convention or 303 redirection, since according to the W3C, namespace resources are not information resources but an abstract space of infinite names. According to IRW, a namespace resource would be an `irw:AbstractResource`. This is because a user can 'mint' a new namespace name without checking any namespace documents in any RDF and XML document and there is no ability of the namespace document to constrain names, but only to recommend them. One obvious use-case is to check every new namespace document and see if the namespace document can be reached through a `irw:redirectsTo` from a `irw:NonInformationResource`.

This same RDF records of what resources are Web resources or non-information resources, associated descriptions and their media-types (particularly RDF documents) is important information for any Semantic Web search engine. The proposed Semantic Web Site-maps allows authors to publish various characteristics of Semantic Web data, such as its update frequency and preferred method of access via an `HTTP response` [**?**]. However, it has to express *what kind* of data it is. This is important, as currently Semantic Web search engines often do specialise in different types of Semantic Web resources. For example, FALCON-S distinguishes between searching for what they call *objects* (`irw:PhysicalEntityResources`) and *concepts* (`irw:ConceptualResources`). As tools like Swoogle specialises in conceptual resources while the OKKAM project specialises in naming entities, by allowing publishers to describe what kinds of Semantic Web resources they have, a Semantic Web search engine can then specialise in searching and displaying for different kinds of resources [**?**, **?**]. Furthermore, the use of a Semantic Web search engine that searches *a*ll kinds of RDF like Sindice, along with some large-scale inference engine like SOAR that could run some kind of inference-based reasoning algorithm against a large data-set, would allow the different kinds of resources to be automatically annotated and categorised [**?**, **?**].

7.3 Linked Data Metadata

One use of IRW to systematise the process of Linked Data validation. Currently, the only Linked Data validator is *Vapour*, which is coded procedurally and whose results can not themselves be presented as RDF [**?**]. The IRW and the HTTP in RDF vocabulary can be used to record whether or not each Linked Data resource is properly redirected using 303 redirection, and the IRW vocabulary can be used to make sure that the `303` redirection can lead access *both* an associated description in HTML and in RDF [**?**]. Any errors over large linked data-sets are easily collected and tested via SPARQL. Furthermore, Linked Data publishers could add two RDF statements that let their associated description be self-describing, solving the identity crisis in the context of Linked Data, and possibly leading to less incorrect use of `owl:sameAs`. Just embedding `dbpedia:data/Eiffel_Tower irw:isAbout dbpedia:/resource/Eiffel/Eiffel_Tower` would work. The following statement: `dbpedia:data/Eiffel_Tower rdf:type ldow:AssociatedDescription` could be added, as

well as stating `dbpedia:resource/Eiffel_Tower` is of type `irw:NonInformationResource` or
even `irw:PhysicalEntityResource` for clarity. This class would be useful for determining whether or not the resource had a property such as latitude or longitude, since concrete physical entities will have them while concepts and abstract mathematical expressions will not.

8. CONCLUSION AND FUTURE WORK

Overall, the IRW ontology is a beginning, yet it should serve as foundational contribution of modelling Linked Data and so the "Dark Side of Semantic Web" that Hendler believes may give the Semantic Web a crucial advantage over previous efforts in knowledge representation [**?**]. IRW clarifies the interactions between the hypertext Web and Linked Data, allowing Linked Data spiders to keep track of important provenance regarding the identity of resources, and to characterise the resources correctly for semantic validation and error detection. Future work needs to be done to standardise IRW or a descendant thereof through the W3C, which will doubtless result in refinements to IRW, and to encourage its use within the Linked Data community in the context of various validators, debuggers, and search engines. By developing a consistent vocabulary for describing the identity of resources in IRW, the first step has been taken.

9. ACKNOWLEDGEMENTS

We would like to thank Aldo Gangemi for his insightful comments. Also, Harry Halpin was partially supported by a Microsoft 'Beyond Search' award. Valentina Presutti was supported by NeOn and IKS EU FP7 projects.

10. REFERENCES

[1] S. Abou-Zahra. Evaluation and Report Language (EARL) 1.0 Schema. W3C Working Draft, W3C, 2007. http://www.w3.org/TR/EARL10-Schema/.

[2] B. Adida, M. Birbeck, S. McCarron, and S. Pemberton. RDFa in XHTML: Syntax and Processing. W3C Recommendation, W3C, 2008. http://www.w3.org/TR/rdfa-syntax/.

[3] A. P. Aidan Hogan, Andreas Harth. SOAR: Authoritative reasoning for the web. aswc 2008: 76-90. In *Proceedings of the Asian Semantic Web Conference (ASWC2008)*, pages 76–90, Bangkok, Thailand, 2008.

[4] S. Auer, C. Bizer, J. Lehmann, G. Kobilarov, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *Proceedings of the International Semantic Conference and Asian Semantic Web Conference (ISWC/ASWC2007)*, pages 718–728, Busan, Korea, 2007.

[5] T. Berners-Lee, R. Fielding, and L. Masinter. IETF RFC 2396 Uniform Resource Identifier (URI): Generic Syntax, 1998. http://www.ietf.org/rfc/rfc2396.txt (Last accessed on Sept. 15th 2008).

[6] T. Berners-Lee, R. Fielding, and L. Masinter. IETF RFC 3986 Uniform Resource Identifier (URI): Generic Syntax, January 2005. http://www.ietf.org/rfc/rfc3986.txt(Last accessed on April 2th 2008).

[7] T. Berners-Lee, J. Hollenbach, K. Lu, J. Presbrey, E. Prud'hommeaux, and mc schraefel. Tabulator Redux: Browsing and Writing Linked Data. In *Proceedings of the WWW2007 Workshop on Linked Data on the Web*, 2008.

[8] C. Bizer, R. Cygniak, and T. Heath. How to publish Linked Data on the Web, 2007. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/ (Last accessed on May 28th 2008).

[9] D. Booth. URIs and the myth of resource identity. In *Proceedings of Identity, Reference, and the Web Workshop at the WWW Conference*, 2006. http://www.ibiblio.org/hhalpin/irw2006/dbooth.pdf.

[10] 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.

[11] P. Bouquet, H. Stoermer, G. Tummarello, and H. Halpin, editors. *Proceedings of the WWW2007 Workshop I^3: Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web, Banff, Canada, May 8, 2007*, CEUR Workshop Proceedings. CEUR-WS.org, 2007.

[12] P. Bouquet, H. Stoermer, G. Tummarello, and H. Halpin, editors. *Proceedings of the ESWC2008 Workshop on Identity, Reference, and the Web, Tenerife, Spain, June 1st, 2008*, CEUR Workshop Proceedings, 2008.

[13] 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 (Last accessed November 22nd 2008).

[14] R. Cyganiak, H. Stenzhorn, R. Delbru, S. Decker, and G. Tummarello. Semantic sitemaps: Efficient and flexible access to datasets on the semantic web. In *Proceedings of European Semantic Web Conference*, pages 690–704, 2008.

[15] S. F. Diego Berrueta and I. Frade. Cooking http content negotiation with vapour. In *Proceedings of Identity, Reference, and the Semantic Web Workshop at the European Semantic Web Conference*, 2008.

[16] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. C. Doshi, and J. Sachs. Swoogle: A Search and Metadata Engine for the Semantic Web. In *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*. ACM Press, November 2004.

[17] A. Gangemi. Norms and plans as unification criteria for social collectives. *Journal of Autonomous Agents and Multi-Agent Systems*, 16(3), 2008.

[18] A. Gangemi, N. Guarino, C. Masolo, R. Oltramari, and L. Schneider. Sweetening ontologies with DOLCE. In *Proceedings of International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, pages 166–181. Springer, 2002.

[19] P. Hayes and H. Halpin. In defense of ambiguity. *International Journal of Semantic Web and Information Systems*, 4(2):1–18, 2008.

[20] J. Hendler. The Dark Side of the Semantic Web. *IEEE Intelligent Systems*, 22(1):2–4, 2007.

[21] I. Jacobs and N. Walsh. Architecture of the World Wide Web. Technical report, W3C, 2004. http://www.w3.org/TR/webarch/ (Last accessed Oct 12th 2008).

[22] J. Koch, C. A. Velasco, and S. Abou-Zahra. HTTP Vocabulary in RDF. W3C Working Draft, W3C, 2008. http://www.w3.org/TR/EARL10-Schema/.

[23] J. Mogul. Clarifying the fundamentals of HTTP. In *Proceedings of the 11th International World Wide Web Conference*, pages 444–457, 2002.

[24] M. Nottingham. IETF Internet Draft HTTP Header Linking, 2008. http://www.mnot.net/drafts/draft-nottingham-http-link-header-01.txt.

[25] E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, and G. Tummarello. Sindice.com: a document-oriented lookup index for open linked data. *International Journal of Metadata, Semantics, and Ontologies 2008*, 3(1):37–52, 2008.

[26] 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.

[27] V. Presutti and A. Gangemi. Identity of resources and entities on the web. *International Journal of Semantic Web and Information Systems*, 4(2):49–72, 2008.

[28] L. Sauerman and R. Cygniak. HTTP Vocabulary in RDF. W3C Note, W3C, 2008. http://www.w3.org/TR/cooluris/.