A Pragmatic Theory of Reference for the Web

Dan Connolly
MIT Computer Science and Artificial Intelligence Laboratory (CSAIL)

The World Wide Web is a "global information universe"[4]. Links are used to represent all sorts of subtle, rich, and and even ambiguous references. One one web page, we might find...

```
I saw a great <a
href="http://www.imdb.com/title/tt0091605/"
>movie starring Sean Connery</a>
```

... while another says...

```
The <a
href="http://www.imdb.com/title/tt0091605/"
>IMDB page on "The Name of the Rose"</a>
is a great source of information.
```

Meanwhile, the Web Architecture document[15] says, "By design a URI identifies one resource." Which does `http://www.imdb.com/title/tt0091605/` identify, then, the movie, or the page about the movie? The pragmatic answer is: it doesn't matter that much, provided the community of people making the links and serving the information "agree (to a reasonable extent) on a set of terms and their meanings."

Human readers are quite robust when it comes to understanding puns and ambiguously indirect references, but computers are not; in a C program, the difference between `*p` and `**p` is the difference between a useful computational result and a crash. Even for human readers, there are limits. If visiting `http://www.imdb.com/name/nm0000225/` with a web browser showed Christian Slater's photo and filmography, and a writer used that address to refer, indirectly, to Sean Connery, readers would likely feel that Grice's Maxim of Manner[13], "Avoid ambiguity", had been violated.

The design of the Web of documents we have today is the result of taking the simplest features of hypertext designs from 15 to 20 years ago, adding globally scoped Uniform Resource Identifiers (URIs), and relaxing link consistency constraints. The social dynamics of the Web include lots of people agreeing to just a few design constraints in order to get a significant return on their investment, whether from reading or writing or both.

By analogy, the Semantic Web involves starting with simple database and logic designs and using URIs for column names and symbol terms. Which constraints need relaxing and which social norms will result in exponential growth are still open questions.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

The W3C Technical Architecture Group (TAG) is chartered to "document and build consensus around principles of Web architecture". This paper gives a pragmatic theory of reference in the form of some principles established by the W3C TAG and some personal conjectures about issues that are still open.

1. AN ANALYSIS OF HTTPRANGE-14

One of the most intensely debated TAG issues is httpRange-14: What is the range of the HTTP dereference function?. The dicussion of the issue is almost all publicly recorded, but following it is challenging, not only because of the quantity, but because of the diverse backgrounds of the participants, leading to much miscommunication.

Perhaps a formal analysis is an effective way to summarize. For example, the simple logical statement that `http://purl.org/dc/elements/1.1/title` is an `rdf:Property` was a matter of some dispute.

The Dublin Core Metadata Initiative (DCMI) publishes a schema[1] that says:

```
<rdf:Property
  rdf:about="http://purl.org/dc/elements/1.1/title">
  <rdfs:label xml:lang="en-US"
  >Title</rdfs:label>
  <!-- ... details elided ... -->
</rdf:Property>
```

Or, equivalently, using turtle[2] notation:

```
@prefix rdf:
  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs:
  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dc:
  <http://purl.org/dc/elements/1.1/>.

dc:title rdf:type rdf:Property.
```

Meanwhile, Tim Berners-Lee argued that "HTTP URIs (without `#`) should be understood as referring to documents, not cars". To formalize this position, we'll use some string manipulation properties from N3[5] and set aside the prohibition against literal subjects, going just beyond turtle into N3. Also, we use the `log:uri` property, which is similar to the `name` quoting function in KIF[12].

[1] `http://purl.org/dc/elements/1.1/`

```
@prefix dc:
  <http://purl.org/dc/elements/1.1/>.
@prefix log:
  <http://www.w3.org/2000/10/swap/log#> .
@prefix str:
  <http://www.w3.org/2000/10/swap/string#> .

dc:title
  log:uri "http://purl.org/dc/elements/1.1/title".
"http://purl.org/dc/elements/1.1/title"
  str:startsWith "http:".
"http://purl.org/dc/elements/1.1/title"
  str:notMatches "#".
```

We can use OWL[9] to formalize that something is a document and not a car, once we choose, for the purpose of this discussion, URIs for the concepts of documents and cars:

```
@prefix owl:
  <http://www.w3.org/2002/07/owl#>.
@prefix tbl:
  <http://example/tbl-terms#>.

tbl:Document owl:disjointFrom tbl:Car.
```

Berners-Lee also argued that properties are not documents:

```
tbl:Document
  owl:disjointFrom tbl:Car, rdf:Property.
```

To state his position generally, we add an N3 rule:

```
tbl:Document
  owl:disjointFrom tbl:Car, rdf:Property.

{
 ?X log:uri [
   str:startsWith "http:";
   str:notMatches "#"
 ] }
 => { ?X a tbl:Document }.
```

This position is quite clearly inconsistent with the DCMI schema. Berners-Lee was unable to persuade a critical mass of the TAG to accept this position. The utility of the dublin core vocabulary was apparent and the argument against its use of hashless HTTP URIs included few practical consequences. Plus, the constraint seems to encroach on the very important principle of opacity of URIs. Any general-purpose algorithm for finding out the nature of a resource starting from only its URI is a constraint on how URIs are minted.

Further, Mark Baker asserted the right to assign any meaning at all to URIs that he owns. In particular, he said that `http://markbaker.ca/` denotes his very self. To capture this position, we will borrow from the FOAF[10] vocabulary:

```
<http://markbaker.ca/>
  a foaf:Person;
  foaf:name "Mark Baker".
```

This is also inconsistent with Berners-Lee's position, since Berners-Lee held that `foaf:Person` is disjoint with `tbl:Document`.

Meanwhile, the DCMI showed some willingness to cooperate with those who hold that RDF Properties and web pages are disjoint; they arranged for HTTP redirections [11], rather than 200 OK repsonses, in reply to GET requests to `http://purl.org/dc/elements/1.1/title`.

Let's introduce some terms for discussing the HTTP protocol. To dereference `http://site.example/path`, it's typical to make a TCP connection to port 80 of `site.example` and send a `GET /path` request. If the reply is...

```
200 OK
content-type: text/plain

hello world.
```

Then we'll say that:

```
@prefix http:
  <http://example/http-terms#>.
@prefix mime:
  <http://example/mime-terms#>.

_:reply1 a http:OKResponse;
  http:about <http://site.example/path>;
  mime:body "hello world.";
  mime:content-type "text/plain".
```

It is stipulated by all parties in the httpRange-14 discussion that in this case, the `hello world.` body of type `text/plain` is a *representation* of `http://site.example/path`. We can state the general rule as:

```
@prefix w:
  <http://example/webarch-terms#>.

{
_:m a http:OKResponse;
  http:about ?R;
  mime:body ?BYTES;
  mime:content-type ?TYPE.
} => {
?R w:representation [
  mime:content-type ?TYPE;
  mime:body ?BYTES ].
}
```

When asked if `tbl:Document` included the targets of http POST messages, Berners-Lee said yes, and agreed, to some extent that the that the term *document* is misleading. The TAG coined the term *Information Resource*. The term is not completely defined, but the 15 Jun 2005 decision of the TAG to address httpRange-14 says:

> The TAG provides advice to the community that they may mint `http` URIs for any resource provided that they follow this simple rule for the sake of removing ambiguity:
>
> - 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 resource;

- If an `http` resource responds to a GET request with a 4xx (error) response, then the nature of the resource is unknown.

We can state this formally as:

```
{
?M a http:OKResponse;
  http:about ?R.
} => { ?R a w:InformationResource }.
```

We consider it implicit in this decision that it applies to the case of a dereferencing an `ftp:` URI as well, so the following single triple states the position of the TAG quite concisely with respect to the terminology developed so far:

```
w:representation
  rdfs:domain w:InformationResource.
```

Since DCMI does *not* claim that `dc:title` is a `w:InformationResource`, Berners-Lee is able to endorse their claim that `dc:title` is an `rdf:Property` (for example, in his tabulator implementation, announced in [3]) while maintaining his position that `rdf:Property` is disjoint from `w:InformationResource`.

Note that the TAG has not taken a position on whether `w:InformationResource` intersects with `rdf:Property`.They do say, "Other things, such as cars and dogs [...] are resources too. They are not information resources, however [...]" which strongly suggests that `w:InformationResource` is disjoint with `foaf:Person`. Since Mark Baker's server responds with ordinary 200 OK replies when asked about `http://markbaker.ca/`, we have:

```
<http://markbaker.ca/>
  a w:InformationResource.

foaf:Person owl:disjointWith w:InformationResource.
```

So we are have an inconsisitency between the definition of `w:InformationResource` and Mark Baker's claim that `http://markbaker.ca/` is a `foaf:Person`. While "anyone can say anything about anything"[6], there are consequences to making disagreeable claims.

2. THE VALUE OF AGREEMENT

Consider a formalized movie database, where a request to GET `http://fmdb.example/title/tt0091605/` gives:

```
<#title0091605>
  dc:title "The Name of the Rose";
  film:rating 7.7;
  film:star <#nm0000225>.

<#nm0000225>
  dca:agentName "Christian Slater".
```

And suppose John Doe wants a widely-understood identifier for Christian Slater so that he can use it in a description of a photo that Jim took of Christian. Jim writes, in `http://photohost.example/jim/photo20.ttl`:

```
<photo20>
  dc:title "Smile, Christian!";
  foaf:depicts
   <http://fmdb.example/title/tt0091605/#nm0000225>.
```

Clearly, then, a SPARQL query[20] over the merge of those documents of the form...

```
SELECT ?photo, ?name
WHERE  {
 ?photo foaf:depicts [
   dca:agentName ?name
 ]
}
```

...will return:

photo	<http://photohost.example/jim/photo20>
name	"Christian Slater"

But suppose, meanwhile, that Jim does not think The Name of the Rose merits 7.7 out of 10 stars. A SPARQL query over the same merge of those documents of the form...

```
SELECT ?photo ?title ?rating
WHERE {
 ?photo foaf:depicts _:who.
 [] dc:title ?title;
    film:star _:who;
    film:rating ?rating
}
```

...will return:

photo	<http://photohost.example/jim/photo20>
title	"The Name of the Rose"
rating	7.7

How can Jim reuse the information such as the name of the person depicted in his photo without implyling that the movie merits 7.7 out of 10 stars? One approach is to not merge the sources, but rather treat the fmdb data as a separate graph in the SPARQL dataset. The photo subject's name can be computed using a more explicit query:

```
SELECT
FROM <http://photohost.example/jim/photo20.ttl>
FROM NAMED <http://fmdb.example/title/tt0091605/>
 ?photo, ?name WHERE
  { ?photo foaf:depicts ?who.
    GRAPH <http://fmdb.example/title/tt0091605/> {
      ?who dca:agentName ?name }.
  }
```

This approach is akin to lifting between contexts[14]. It is a workable apprach to integrating data from separate contexts, but it is clearly not as straightforward as merging. The cost of keeping contexts separate demonstrates that *agreement is valuable*. The providers of `fmdb.example` can lower Jim's cost to use their information if they publish film ratings that Jim agrees with.

3. DELEGATION, CONSENT AND CAUSAL CHAINS

Jim could, of course, make up his own URI for the subject of his photo. But then Jim would have to maintain the information about the actor's name, the movies he starred in, and their titles, at his own cost. And if others in Jim's position did likewise, consumers of all this data would be able to correlate photo subjects only at a significant cost of dealing with *URI aliases*. We conjecture that overall utility for the community is maximized if we adopt a *causal theory of reference*[16]. In particular:

1. To mint a term in the community, choose a URI of the form `doc#id` and publish at doc some information that motivates others to use the term in a manner that is consistent with your intended meaning(s).

2. Use of a URI of the form. `doc#id` implies agreement to information published at `doc`.

Justification of this conjecture is in progress[8], using terms such as *intent* and *impact* from the TAG discussion of extensibility and versioning in Edinburgh in September 2005.

4. ADVICE: USE HASH URIS FOR PROPERTIES AND CLASSES

If you want to write RDF schemas that are consistent with the TAG's position on httpRange-14, you have three options:

1. Use the `doc#id` pattern as above.

2. Set up HTTP redirects a la `dc:title`.

3. Populate the intersection of `w:InformationResource` with `rdf:Property`.

The third option is like publishing movie reviews that people disagree with. The second option is more trouble than the first, unless the vocabulary you're describing is very large. So I advise the first option.

4.1 Fragments as sections vs. people

Some argue that "Using `#` [in this way] makes it impossible to make assertions about parts of documents (e.g. Person A authored Section #3)."[1]. Indeed, this is a concern. Let's consider it formally, using FRBR[18], [19]. Suppose `http://fansite.example/baseball` is a little database of great baseball players, with some statistics on `http://fansite.example/baseball\#FredPatek`, among others:

```
@prefix foaf:
  <http://xmlns.com/foaf/0.1/>.
@prefix baseball:
  <http://fansite.example/baseball#>.

baseball:average
  rdfs:domain foaf:Person.

baseball:FredPatek
  baseball:average 0.250 .
```

These data might be published in RDF/XML:

```
<foaf:Person rdf:ID="FredPatek">
    <baseball:average
      rdf:datatype=
"http://www.w3.org/2001/XMLSchema#decimal"
      >0.25</baseball:average>
</foaf:Person>
```

And then someone might use the ID to say that the Fred Patek section was written by somebody named John Doe:

```
@prefix frbr:
   <http://purl.org/vocab/frbr/core#>.

<http://fansite.example/baseball#FredPatek>
 frbr:creator [ foaf:name "John Doe" ].
```

These two are inconsistent, since the domain of `frbr:creator` is `frbr:Work`, which is disjoint with `foaf:Person`, the domain of `baseball:average`.

So indeed, if we choose URIs of the form `doc#id` for people, it leads to inconsistencies with quite reasonable ontologies if we also use them as document section identifiers. I advise authors to choose one or the other for each fragment identifier they publish and be consistent.

In order for this to work with documents published both in RDF/XML and XHTML, the XHTML media type specifications may need to be ammended so that authors can "opt out" of the section-of-the-document meaning of fragment identifiers that they publish. For example, the `profile` attribute from section 7.4.4.3 Meta data profiles of the HTML 4 specification[17] seems like a reasonable opt-out signal.

Populating the intersection of `w:InformationResource` with `foaf:Person`, the way Mark Baker seems to, seems likely to conflict with useful and reasonable ontologies. I suggest adopting `w:InformationResource rdfs:subClassOf frbr:Work` as a practical constraint. The `foaf:primaryTopic` relationship seems particularly useful for relating web pages to things. Rather than..

```
<http://markbaker.ca/> a foaf:Person;
foaf:name "Mark Baker"
```

... I suggest:

```
<http://markbaker.ca/> foaf:primaryTopic
  [ foaf:Person; foaf:name "Mark Baker"].
```

5. CONCLUSIONS AND FUTURE WORK

While the theory presented here gives little by way of rigorously justified theorems, we hope it gives a coherent and pragmatic approach to nagivating many issues of identity and reference in the Web.

We hope the formal analysis of the httpRange-14 discussion and decision, down to a single RDF triple, makes the issue clear without distorting the positions that it summarizes. We also hope it demonstrates the utility of RDF, turtle, and N3 as analytic tools.

The justification of the causal theory of reference is ongoing work. See, for example, Berners-Lee's Total Cost of Ontologies (TCO) argument in his ISWC 2005 presentation.

When considering which constraints need relaxing and which social norms will result in exponential growth of the Semantic Web, mechanisms for expressing trust seem to be critical. We are exploring approaches using quoting a la "if `http://weather.example/ny` contains a formula of the form `<ny#weather> nws:temp ?X` then lift that claim, `<ny#weather> nws:temp ?X` into the knowledge base as a fact." We are exploring these quoting techniques along with digital signature and proof exchange as a general-purpose trust infrastructure. These explorations suggest that the law of the excluded middle, i.e that every formula is either true or false, should be relaxed in order to allow indirect self-reference. Constructive proofs seem more promising than those that use classical fist order reasoning[7].

6. REFERENCES

[1] Anonymous. HashURI. `http://esw.w3.org/topic/HashURI`, 2003.

[2] D. Beckett. Turtle - Terse RDF Triple Language. http://www.dajobe.org/2004/01/turtle/, 2003-2006.

[3] T. Berners-Lee. Links on the Semantic Web. http://dig.csail.mit.edu/breadcrumbs/node/62, 2005.

[4] T. Berners-Lee, R. Cailliau, J.-F. Groff, and B. Pollermann. World-Wide Web: The Information Universe. *Electronic Networking: Research, Applications and Policy*, 1(2):74–82, 1992.

[5] T. Berners-Lee, D. Connolly, and S. Hawke. Semantic Web Tutorial Using N3. http://www.w3.org/2000/10/swap/doc/, 2003.

[6] J. J. Carroll and G. Klyne. Resource Description Framework (RDF): Concepts and Abstract Syntax . Technical Report http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/, W3C, February 2004.

[7] D. Connolly. Investigating logical reflection, constructive proof, and explicit provability. http://dig.csail.mit.edu/breadcrumbs/node/89, Feb 2006.

[8] D. Connolly. Using RDF and OWL to model language evolution. http://dig.csail.mit.edu/breadcrumbs/node/87, Feb 2006.

[9] M. Dean and G. Schreiber. OWL Web Ontology Language Reference. Technical Report http://www.w3.org/TR/2004/REC-owl-ref-20040210/, W3C, 2004.

[10] E. Dumbill. Finding friends with XML and RDF. http://www-128.ibm.com/developerworks/xml/library/x-foaf.html, 2002.

[11] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. Hypertext Transfer Protocol – HTTP/1.1 . Technical Report RFC2616, IETF, June 1999.

[12] M. R. Genesereth and R. E. Fikes. Knowledge Interchange Format (KIF) Version 3.0 Reference Manual. Technical Report Logic-92-1, Computer Science Department, Stanford University, Stanford, California, 1992.

[13] P. Grice. *Studies in the Way of Words*. Harvard University Press, 1989.

[14] R. V. Guha. Contexts: A Formalization and Some Applications. Technical Report STAN-CS-91-1399, Stanford Computer Science Department, Stanford, California, 1991.

[15] I. Jacobs and N. Walsh. Architecture of the World Wide Web, Volume One . Technical Report http://www.w3.org/TR/2004/REC-webarch-20041215/, W3C, 2004.

[16] S. Kripke. *Naming and Necessity*. Harvard University Press, Cambridge, Mass, 1980.

[17] A. Le Hors, D. Raggett, and I. Jacobs. HTML 4.01 Specification . Technical Report http://www.w3.org/TR/1999/REC-html401-19991224, W3C, December 1999.

[18] R. Newman and I. Davis. Expression of Core FRBR Concepts in RDF. http://vocab.org/frbr/core, 2005.

[19] K. Saur. IFLA Study Group on the Functional Requirements for Bibliographic Records. Functional Requirements for Bibliographic Records: Final Report. *UBCIM Publications-New Series*, 19, 1998.

[20] A. Seaborne and E. Prud'hommeaux. SPARQL Query Language for RDF . Technical Report http://www.w3.org/TR/2006/CR-rdf-sparql-query-20060406/, W3C, April 2006.