

Web Proper Names: Naming Referents on the Web

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

Henry S. Thompson
Human Communications Research Centre
School of Informatics
University of Edinburgh
and
World Wide Consortium
2 Buccleuch Place
Edinburgh, United Kingdom
ht@inf.ed.ac.uk

ABSTRACT

When *http:* URIs are used to refer to things not on the Web as opposed to web pages, ambiguity arises on the Web. Both humans and machines need to solve the question of what someone is actually referring to when they use a *http:* URI, especially if that URI can be used both to refer to a web page and a thing that is not on the Web. We propose a multi-tiered solution we called “Web Proper Names” to this problem based on insights from the philosophy of language and the ability of humans to verify the results of a search engine request. Our solution provides a distributed and interoperable approach to creating and sharing Web names for things.

1. INTRODUCTION

1.1 The Web is about things

The value of the World Wide Web stems in large part from the fact that the varied constituents of the Web are *about* things—they describe things or picture things or discuss things. Often, although not always, these things are not themselves on the Web, rather they exist in the physical world outside the Web. The ability to understand some thing as being about some other thing, as being oriented towards something else without any direct connection to it, is crucial to human intelligence. Any effort to make the Web more intelligent, for example by automating the exploitation of resources on the Web, will have to somehow reproduce the human ability to understand what things are about.

This is an issue of immense practical importance: when someone searches the Web, they are looking for information *about* something. At present no widespread automatic processes exist to index, organize, share, or even decide what web pages are about—all searches have to work with is text. The effort to provide machine-readable metadata through standards such as RDF and description logics as embodied in OWL are efforts to improve this situation. Although such efforts do allow a human to express what they believe a web-page is about in a standard way, they still beg the question of how to *interoperably* identify real-world things in such metadata.

Copyright is held by the author/owner(s).
WWW2005, May 10–14, 2005, Chiba, Japan.

Unfortunately, no-one from expert logicians to philosophers of consciousness have a solid idea about how we determine whether or not a thing is actually about something else. On the surface this *aboutness* seems physically spooky: I can think about the Eiffel Tower in Paris without being in Paris, or even having ever set foot in France. I can imagine what the Eiffel Tower would look like if it was painted blue. I can even think of a situation where the Eiffel Tower wasn’t called the Eiffel Tower. Most importantly for our purposes, I can view a web page, either by typing a URL such as *http://www.tour-eiffel.fr* into a browser or by typing **Eiffel** into a search engine and following one of the links it provides. Having done this, I know at a glance if the page is actually about the Eiffel Tower, or a hotel near the Eiffel Tower, as opposed to the object-oriented programming language Eiffel, or the film **The Lavender Hill Mob**, and so on. Yet this knowledge depends on fundamental aspects of human intelligence such as language understanding, scene recognition and so forth, which have proved distressingly resistant to automation.

1.2 Names for things

As presently constituted, the effort to automatically exploit the content of the Web is a broad movement, ranging from information retrieval performed by term-based search engines to the Semantic Web and Topic Map standardization efforts. Some of these approaches use URIs as the primary terms in the languages they use to express *metadata*, that is, information intended for machine processing. Metadata is composed of logical *sentences* which in turn use URIs to stand for things, for example:

```
http://www.tour-eiffel.fr/  
dc:creator  
http://www.vitruvio.ch/arc/masters/eiffel.htm
```

All the metadata sentences we use for examples in this paper have this form, that is, three URIs, to be understood as subject, predicate, object. The predicate *dc:creator* resolves to *http://www.purl.org/dc/elements/1.1/*, which defines its meaning as “an entity primarily responsible for making the content of the resource.” Since a resource can be anything, it is unclear if the content of the resource is the Eiffel Tower or the web page about the Eiffel Tower, or what “making”

means. On one reading, the first URI of this triple stands for the Eiffel Tower, the third URI stands for Gustave Eiffel, its architect. One could imagine an alternate reading in which one could think this statement said Gustave Eiffel created a web-page called *http://www.tour-eiffel.fr*.

This ambiguity about whether a URI refers to a web page or a thing is everywhere. While this may be easy enough for a human to figure out using context (such as their historical knowledge of the Eiffel Tower), there is no obvious way for a machine to detect if a URI stands for a web page or for a thing. One way is to partition URIs into those URIs that only about things and those that are only about web pages, yet there is no mechanism for distinguishing between the two, such as a new URI scheme to tell them apart. This problem has been brought up before by the Semantic Web and Topic Map communities before, as the RDF predicates like *foaf:page* in the Friend-of-a-Friend vocabulary (*http://www.foaf-project.org*) and the *subject indicator* parameter of Topic Maps demonstrate. However, these solutions only solve the problem in small domains through human-readable documentation. Providing one well-specified RDF vocabulary for things as opposed to web pages does not solve the problem in general. RDF vocabularies that could be created for this problem in general (such as a theoretical *DocumentIsAboutThisThing*) only cope with the issue by using an inverse functional property of the thing (which are quite rare for most things), or more likely another potentially ambiguous *http:* URI or a literal. In the terms of philosophy, all these approaches fall victim to the regress of the Symbol Grounding Problem[6].

If the same thing is given multiple URIs by differing authors, in the absence of any agreed central authority which decides what URIs should be used to stand for what things, there is a real risk that the Semantic Web will consist of a vast number of self-consistent but mutually incommensurable collections of metadata. Speaking informally, the URIs in the above example just address web pages. When a software agent fetches a web page from a URI, it's the web page addressed by the URI, as rendered by the agent on the basis of the encoding (such as HTML) returned by a server, that is actually about the Eiffel Tower or Gustave Eiffel.

The first challenge for the project of automating the exploitation of the Web is thus not to know what web pages are about—that's too hard for the time being. Just separating URIs for things from URIs for web pages and knowing when two URIs are about the *same* thing would be a huge step forward. Other researchers[5] are also trying to address this problem in the context of the Semantic Web.

We believe the Web needs a solution to this problem which

1. Provides a distributed approach to creating and sharing Web names for things;
2. Allows *use* of Web names for things to be easily distinguished from the *use* of URIs to address web pages;
3. Allows for efficient and reliable determination of whether Web names describe the same thing;
4. Does not require a single canonical name, while still achieving interoperability of names.

Our solution to this problem exploits the pervasive availability of search engines with substantial coverage by using them to find sets of pages that human users judge to describe

certain things. In an loose *analogy* with natural language, phrases such as **the Eiffel Tower** are called proper names, so we call our approach **Web Proper Names**, and use *wpn:* in our examples as a candidate URI scheme for Web Proper Names. Although the concept can be refined further, a **Web Proper Name (WPN)** for something is usually composed of a set of search terms known to return primarily URIs of web pages which describe that thing. It's at least initially plausible that such an approach to naming things for the Web should satisfy the requirements listed above—the rest of this paper is devoted to spelling out the details and demonstrating that in fact this is the case.

Note that we do *not* require that it be possible given a web page to automatically determine the Web Proper Name of whatever it describes. This would set the bar too high—even names in the real world don't have this property. We do not believe our solution is the only possible one or even best-suited for all tasks. In specialized domains this problem may not even occur. Yet for the many domains in which it does occur we have a general solution that can suit many needs.

2. PHILOSOPHICAL UNDERPINNINGS

2.1 Terminology

Although terminology in this field often is confused, the underlying phenomena are reasonably clear: Generally, when something is understood to be about something else we talk about *reference*, and the thing referred to is called the *referent*. Whenever the word *thing* is used we mean a thing that is not by nature on the Web, such as Henry S. Thompson as opposed to his web page. The reference relation is considered *semantic*, a relation of *meaning*. One kind of reference is that which starts from *names*—a class of linguistic expressions that are about something else. *Proper names* are names that refer uniquely to one referent, at least in an ideal situation. Figure 1 illustrates an example of this relationship.

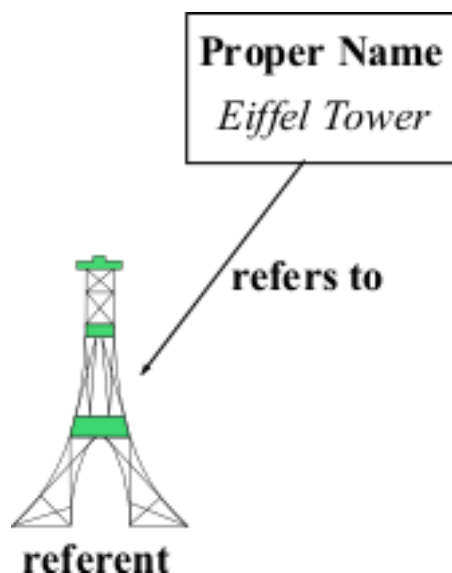


Figure 1: The reference relation

On the Web this relationship becomes more complex. The

W3C's Technical Architecture Group[7] says that a URI *identifies a resource*, and that browsers can *retrieve representations* of that resource.

As it stands a *resource* can be *anything*, including both things and web pages. The URIs by which resources are identified do not seem to be connected to them in the way that names are to their referents (see §2.4 below). **Henry S. Thompson** in the correct context refers to a real person, while the URI of his web-page may refer to him as a rigid designator or just his web page, even in the context of the Semantic Web.

Our take on the ordinary understanding of URIs is that a URI *addresses* a Web-based *encoding* of a *description* or *depiction* of a *denotation*. An *encoding* is the character sequence that is actually retrieved, along with a specification of its media type, e.g. HTML or SVG. Informally it is the source for a web page, although the term is intended to be broad enough to cover non-web standards that encode their data more directly, such as JPG for images or MP3 for sound.

A linguistic *description* or pictorial *depiction* is the rendered output of a program given an encoding. Henceforth we will use *expression* as a cover term for the whole range of humanly-perceivable forms whose standardized encoding is addressed and retrievable by URIs—in other words *expression* is a cover term for HTML pages, SVG and JPG images, MP3 audio streams, and so on as presented to humans by software. Also *web pages* will be used informally to cover both encodings and expressions in one term, and so will both cover the everyday language use of the term (as for HTML pages) but also refer to a wider set of phenomena (such as a URI addressing an audio stream). Following Goodman[4], we use *denotation* for that which is depicted or described by an expression, where the philosophical treatment of reference would use *referent*.

Subject to connectivity, the encoding addressed by a URI can be fetched, rendered as an expression by a software agent and seen or heard by a human, who can then determine what if anything the expression denotes. Figure 2 illustrates this. To summarize: the URI <http://www.tour-eiffel.fr> addresses an encoding in HTML, which can be retrieved by a web-browser, which renders the encoding as an expression composed of text and pictures, and these text and pictures will be recognized by a human being as denoting the actual Eiffel Tower in Paris.

We can now be more precise about what's going on with respect to Web searches. When searching, a user typically wants to fetch *expressions* constituting *descriptions* (such as HTML or XML pages) or *depictions* (such as JPG or SVG images) that actually describe or depict some *denotation* they are interested in. When searching the Web, many expressions can be found which are not about the item of interest, and distinguishing those that denote the item of interest from those that do not is not straightforward. The human ability to do this, as remarked above, is evidently based on a wide range of linguistic and cognitive abilities, which machines have so far proved unable to reproduce.

2.2 The Use-Mention Distinction

A note of caution, in the guise of introducing one more bit of terminology. The connection from URI to expression and from expression to denotation encourages a confusion which is analogous to the *use-mention* confusion familiar

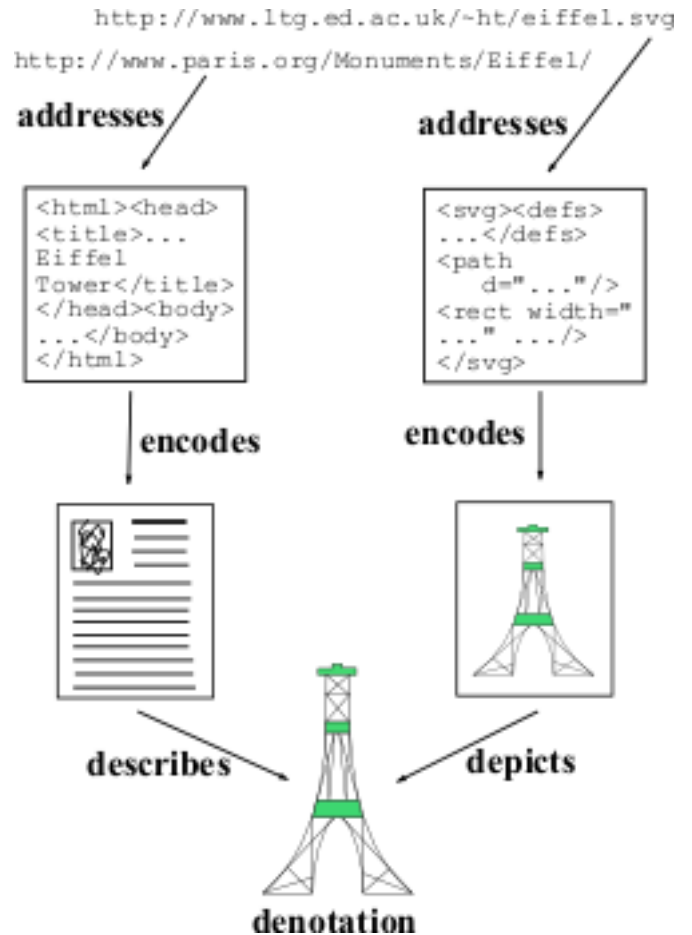


Figure 2: Description and denotation on the Web

to philosophers of language. Consider the difference between “Rice is tasty” and “**rice** is a one-syllable word.” The first sentence *uses* the word **rice** to refer to a foodstuff in the world. The second sentence *mentions* the word **rice** in order to discuss a property *of the word itself*. There is an analogous problem on the Web when a URI occurs in metadata. In practice we observe that such metadata may *either* be understood as saying something about the expression whose encoding is addressed by the URI (a *mention*), *or* as saying something about the denotation of said expression (a *use*). For example, in order to understand the following as saying that Henry Thompson’s W3C home page was created by Henry Thompson, we have to interpret the first URI as a *mention* but the second as an *use*:

```
http://www.w3.org/People/thompson/
http://purl.org/dc/elements/1.1/creator
http://www.ltg.ed.ac.uk/~ht/
```

For a range of reasons, which we will return to below, we think it’s a mistake in metadata to *use* URIs in general for things—most URIs should be understood as being *mentioned* in metadata, that is, as being used to refer to the web pages they address. To refer to things in the world in metadata, we offer Web Proper Names as a particular kind of URI intended for this purpose.

2.3 Search engines and descriptions

Although the philosophical story and the Web story (see Figure 1 and Figure 2 above respectively) appear to be different, in that in the one case reference is unmediated, but in the other mediated by a web page, in fact the parallel is much stronger.

The classic approach of Frege posits *three* elements to any reference: the *name*, the *sense*, and the *referent*[3]. The actual thing in the world is still the referent, and a name is a symbol that has a referent. The *sense* is the mode of presentation, a type of public, perhaps objective knowledge about the item. Frege himself would likely judge this to be Platonic in nature. Russell and others analyzed proper names as “abbreviated” descriptions. Their *descriptivist* theory of names analyzes a name as identifying a set of definite descriptive terms [12]. These descriptive terms could be logical or linguistic in form. On the descriptivist account a name maps in the head of its user to a private concept of what the referent is. *Sense* is the public projection of that private concept among a shared community. The third party of *sense* mediates the reference relationship. In Frege’s classical example, **Hesperus** has a sense (“the morning star”) different from that of **Phosphorus** (“the evening star”), yet both have the same referent, the planet **Venus**.

The descriptivist notion of sense is parallel to the place of search terms in the Web story. A web page addressed by an URI can thereby be fetched and shared among the community of Web users. The notion of a sense as composed of definite descriptive terms also has an intriguing connection to the contemporary use of search engines. Typing descriptive terms such as **Eiffel**, **Tower** and **Paris** into a search engine returns URIs that address descriptions of the actual Eiffel Tower. In the context of the Web, there is usually a non-arbitrary, although not strictly necessary, relationship between the descriptive terms and whatever the recovered web pages denote. Insofar as we’ve hinted that a Web Proper Name is a collection of search terms, this analogy is encouraging, particularly because the first step, from search terms to URIs, is automated and distributed.

It is important to note, however, that there are problems treating *sense* as a set of descriptive terms. It is in practice very difficult to come up with a set of descriptions that identifies exactly one referent. The Eiffel Tower is “a large metal monument.” To distinguish it from the multitude of other large metal monuments in the world, the Eiffel Tower is “a large metal monument in Paris.” There are other large metal monuments in Paris, and the Eiffel Tower would still be the Eiffel Tower if it were moved to Lake Havasu City. Searle addresses this issue in his *cluster theory* of names[13], in which he suggests that only some or most of the terms intended to identify a referent need do so. Furthermore, many of the descriptive terms, or indeed all of them, may also describe things which are *not* the intended referent.

Analogously, when using Google, typing in search terms for the Eiffel Tower such as *Eiffel Tower Paris* results in *some* web pages about the actual Eiffel Tower in Paris, but not all of them, and also web pages of things only marginally connected to the Eiffel Tower, such as hotels with views of the Eiffel Tower, or worse, something as inappropriate as an Eiffel programming language conference in Paris. The size of the retrieved set will also be *quite* large (“about 379,000” according to Google on the day of writing).

This suggests a refinement not usually found in philosoph-

ical accounts: the use of negative search terms. For example, the fact that the Eiffel Tower is not a hotel can be reflected by using *Eiffel Tower Paris -hotel* as the set of search terms. This has a dramatic effect—at the time of writing the size of the set Google returns for these terms is “about 166,000”.

The analogy we are developing looks like this—a Web Proper Name should function like a natural language *name*, identifying a referent. It consists of a set of search terms, including negative ones. Courtesy of a search engine, it determines a set of URIs that address web pages. At least a subset of those in turn denote the *referent* of interest.

When someone uses a search engine, if the majority of the descriptions retrieved for a given set of search terms, particular the high-ranking ones, do in fact describe the desired referent, then the search is generally considered successful. Analogously, a set of search terms is a good candidate for a Web Proper Name if the majority of the URIs retrieved for those terms, particular the high-ranking ones, do in fact address web pages with the same denotation.

2.4 Names and descriptions

It’s important not to confuse a *name* with *descriptions* of its referent. In the real world, we use the *name* **Eiffel Tower** to uniquely determine the Eiffel Tower referent. We use names, not descriptions, to identify people. For example, the *name* **Tim Berners-Lee** identifies a certain man in Boston who is the Director of the W3C and wrote the book called *Weaving the Web* about his part in the creation of the World Wide Web. Moreover, when we want to refer to Tim Berners-Lee, we don’t have to redescribe him using his title or the book he’s written. A name *alone* determines its referent, at least where all parties involved attach the name to the same referent. Furthermore, this is achieved without appeal to descriptions.

In *Naming and Necessity*, Kripke says that names function to *fix a referent* without being a shorthand for sets of descriptive terms[9]. This is in tension with both the descriptivist and cluster theories of names discussed above. Descriptions aren’t entirely out of the picture on Kripke’s account—they are necessary for disambiguation when the context of use allows more than one interpretation of a name, and they may figure in the process by which things actually get their names.

In Kripke’s account an agent or agents fix a name to a referent by a process called *baptism*, in which a thing and a word are directly associated. Afterwards one can use a name by virtue of being in a causal chain with the baptism. If the agent, the thing being named, and the listener are all co-present, the thing being named can be directly identified, otherwise careful use of descriptive terms will be required in order to adequately identify the object.

Sometimes proper names include ordinary words which themselves contribute to our understanding, for example **Prime Minister**, **Crystal Palace**, **Big Island**. The use of search terms in Web Proper Names parallels this to some extent. Using these terms a search engine can select from the vast number of web pages available on the Web a set which may describe the referent one is interested in.

The lesson here for naming on the Web are that names and search terms are not the same, but that search terms can be used to create names for the Web, via web pages, in a productive and interoperable way. Baptism on the Web can be achieved by an appeal to a set of search terms and

a search engine which can recover appropriate expressions, which in turn denote the intended referent. The baptizing agent of a Web Proper Name is the owner of the Web Proper Name. The referent is whatever thing the owner is interested in. A Web Proper Name is composed of search terms that given to a search engine will recover a set of URIs which address expressions which can in turn be verified by the baptizing agent as denoting the referent. We can effectively merge our two earlier pictures, as shown in Figure 3.

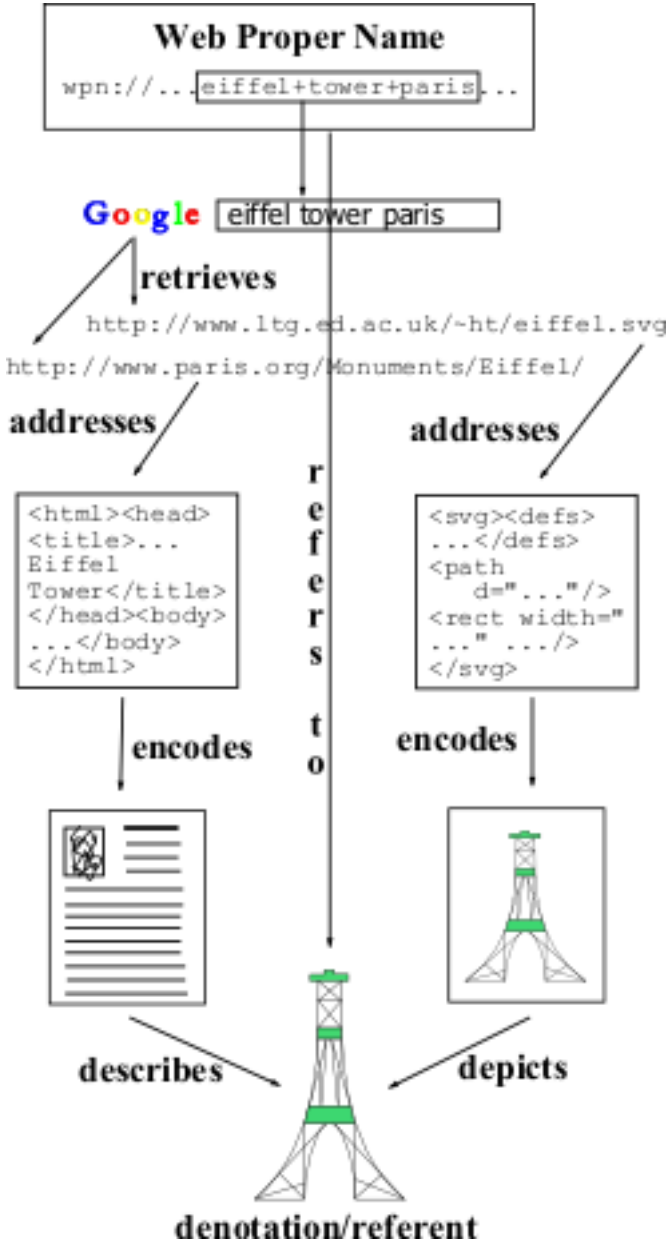


Figure 3: Web Proper Names

It would be difficult if not impossible to select a set of search terms that uniquely determine a referent, that is, terms which recover a set of URIs such that *all* the web pages addressed thereby denote the intended referent. That's why the role of the baptizing agent is crucial: It's their job to determine whether the denotation of each web page is *really*

the intended referent. Bar the creation of genuine artificial intelligence, currently only human inspection can check whether or not a given web page denotes a particular referent. A human agent with a referent to baptize must refine a set of search terms until an appropriate subset of the expressions addressed by the URIs recovered by a search engine from those terms denote that referent.

3. WEB PROPER NAMES SPECIFICATION

A *Web Proper Name* (WPN) is a Web-usable name for a referent, based on a set of search terms which recover a set of web pages that denote that referent. A Web Proper Name not only may determine many web pages, but a single web page may participate in many Web Proper Names. A Web Proper Name should not be confused either with the set of search terms, the referent itself, the set of descriptions, or the additional information needed to situate the context of its baptism. A Web Proper Name is unique by virtue of the conjunction of all these parameters. WPNs are not limited to naming things that already have ordinary proper names, such as the **Eiffel Tower** or **Tim Berners-Lee**, but can be constructed to name virtually anything, such as **my eldest sister-in-law** and **lambda calculus**, as well as fictional referents such as **unicorns**. Note since there are no restrictions on the referent we allow what someone might consider to be multiple referents to be taken as the singular referent of Web Proper Name.

The creation of a Web Proper Name via a search engine can be described concisely in terms of information retrieval. Assume a collection of web pages, T , which is the total number of web pages on the Web. There is an agent, assumed to be human in this paper, that has an information need about referent n that we are unable to directly access, however, they express this as a request to a search engine in terms of a linguistic object, the **search terms** that the search engine transforms into a query (the engine may stem input words to obtain terms, add system-derived weights to terms, and so on) over its document index descriptions. A search retrieves some subset R of T , which we call the **result sequence**. We call it a "sequence" as opposed to a "set" since it is ordered. The agent inspects and assess some set of web pages C , the **checked sequence**, that is a subset of R . Each web page in C is determined to either be or not be about referent n by the agent, partitioning C into two sequences, the **correct checked sequence** C_c whose members are about referent n and the **incorrect checked sequence** C_i whose members are not about n , which are equivalent to relevant and irrelevant web-pages respectively in the terminology of information retrieval.

Since what a web-page is *about* is on some level inaccessible and may vary over agents, only inspection of the content will suffice to characterize the web-page as relevant or not. We allow multiple copies of the content of a web-page to be in the **correct checked sequence** (C_c), although we do require unique URIs. Notice that there can be relevant web pages in the whole Web (T) that are not in the **result sequence** (R) and relevant web pages in that sequence R that are not in the **checked sequence** (C), making it impossible to establish a recall value for the **checked sequence** (C) in terms of the whole Web (T) or the **result sequence** (R). Also note that we are **not** measuring the performance of a search engine. While there might be other useful pages about n , we want only a sufficient, not necessary, set of web

pages to characterize referent n for the agent. Necessary is far too strong for many referents, for referents usually are not fully characterized by anything on the web. However, a collection of web pages can be *good enough* to characterize them to the degree deemed needed by the agent. It would be trivial to form a query that just retrieved the corrected checked sequence (C_c); all it would perhaps require is the pathological concatenation of the text of each member of that sequence. This does not reflect the actual search process used to find the **correct checked sequence** (C_c). To capture the actual search used to find C_c , we allow the user to “freeze” the relevance feedback of the search engine at any time. After iterating and improving the initial search to get a R of reasonable size, the agent pursues the inspection process to create C_c , saving relevant information such as the search terms. These search terms may not be optimal, but just good enough to begin inspection for an agent.

We define a Web Proper Name as a nine-tuple, as follows, with abbreviations for the components in parentheses:

Owner: Identification of the baptizing authority, in a form usable as a *http:* URI.

Short name: A short mnemonic for the WPN, in a form allowing it to be combined with the Owner to give a valid *http:* URI.

Engine (se): The domain name of the search engine used.

Date (dt): The date the search was done, in YYYY-MM-DD form.

Terms: The positive and negative search terms used, combined with plus signs, phrases surrounded by double-quotes, terms separated by a plus sign, negative terms marked with minus signs.

Language (ln): The natural language of the terms—for inclusion in the query if the search engine supports language-filtering.

Result Sequence Size (rs): The binary order of magnitude of the cardinality of the sequence of URIs retrieved by the search engine.

Checked Sequence Size (cs): The binary order of magnitude of the cardinality of the subsequence of the Result Sequence that have been checked to determine whether they describe the referent.

Percent Correct (pc): The percentage of the Checked Sequence found to actually describe the referent.

For use in metadata, a Web Proper Name must be recognizable as such. Accordingly we package the constituents defined above into a URI using the hypothetical *wpn:* scheme as follows:

```
wpn://owner/shortName?terms=terms&
se=engine&dt=date&rs=resultSequenceSize&
cs=checkedSequenceSize&pc=percentCorrect
```

Since the size of some of the sequences, particularly the Result Sequence, might be quite large, the size of the sequence is expressed as a binary order of magnitude in integer form. A Web Proper Name with the following composition:

Owner: www.ltg.ed.ac.uk/~ht/WPN

Short name: EiffelTower

Engine: www.google.com

Date: 2004-04-29

Terms: eiffel+tower+paris+-hotel+-webcam

Language: en

Result Sequence Size: 17

Checked Sequence Size: 5

Percent Correct: 84

is expressed in a *wpn:* URI like this:
*wpn://www.ltg.ed.ac.uk/~ht/WPN/EiffelTower&
terms=eiffel+tower+paris+-hotel+-webcam&
ln=en&se=www.google.com&dt=2004-05-21&
rs=17&cs=5&pc=84*

WPNs do not *require* search engines, and so all parameters except **owner** and **date** are optional. URIs may be gathered from many places; they can be e-mailed directly, seen on the sides of cars, written in ads in magazines, and found by casually poking around some web-pages. The concept of a Web Proper Name and the information embodied in its parameters are independent of any particular format although certain formats are ideal for certain uses as detailed in §5.

With regards to exactly what a referent is, WPNs and this proposal are agnostic. A WPN should not be confused with its denotation. It is not the task of WPN to define what a referent is, it is the task of the human who created the WPN. WPNs do not claim to create a universal and centralized ontology as Cyc does[10], but rather aims to enable a distributed and cooperating ontology fragments. The class of referents is as diverse as the possible interests of humans and world itself[14].

3.1 Requirements and design choices

How do we stand then with respect to the four goals stated in §1.2?

3.1.1 Provides a distributed approach to creating and sharing Web names for things

Anyone can create a Web Proper Name, and the components described above can be either published using the *wpn:* scheme or in an expanded form described below in §4. The fact that anyone can create a Web Proper Name does not distinguish it from URIs in general. What makes Web Proper Names as defined here independently creatable and sharable for the purpose of naming things on the Web in a way that arbitrary URIs are not is that it is easy for independently created Web Proper Names to be compared.

3.1.2 Allows the mention of URIs to be easily distinguished from the use of URIs

By *mention* we mean the use of URIs as names for things as opposed to their *use* for retrieving web pages. Web Proper Names evidently satisfy this by definition—the use of the *wpn:* URI scheme ensures this, and this use is the primary justification for the creation of the *wpn:* scheme. An expanded form detailed in §4 also allows a WPN web page to be made for *http:* URIs, but this can not be distinguished from non-WPNs at purely the URI level. By definition a WPN always denotes its referent.

3.1.3 *Allows for efficient and reliable determination of whether two Web names are about the same thing*

The design given here for Web Proper Names satisfies this goal at three levels:

1. by including the **Short Name** constituent, which can be used to signal the baptizer's intent;
2. by including the **Terms** constituent, which specifies the baptizer's intent much more explicitly;
3. by allowing for much more detailed information about the **Checked Sequence**, including a partition of its member URIs into correct and incorrect URIs, to be fetched if using the Expanded WPN (see §4 below).

Like reference in real life, there is no absolute guarantee that two WPNs are about the same thing. However, heuristic solutions will be able to within reason find out if WPNs are referring to the same or similar things. Significant overlap between the membership of the correct checked sequence of two WPNs gives a strong presumption of identity of intended referent.

3.1.4 *Does not require a single canonical name, while still achieving interoperability of names*

The implicit contrast here is with an approach to naming on the Web that requires or assumes some form of centralization, either of names themselves, or of assertions of equivalence of names. Web Proper Names are interoperable without such centralization, because two Web Proper Names can be compared on the basis of their constituents. This is achieved by appeal to the web-accessible form of the WPN itself, not to a universal central authority.

4. EXPANDED WEB PROPER NAMES

While Web Proper Names are not *universal* in the sense that a WPN uniquely identifies its referent over the Web for everyone, its format should be *uniform*, so WPNs may be exchanged and processed in a uniform manner by everyone. We use the name *Expanded Web Proper Name (EWP)* for this packaging and expansion of WPN information. EWPNs are especially made to be packaged and used over the *http:* URI scheme, and as such may be deployed currently without any change to current software.

For many purposes, such as re-checking a WPN or comparing WPNs, the exact URIs recovered from its search terms are crucial. If two EWPNs have a majority of recovered URIs in common, then there is a strong presumption that they are about the same thing or closely related things. However, this can not be determined unless the actual URIs or their content are available. The original specification of WPNs is accordingly modified with the additional information detailed above to make the Expanded Web Proper Name specification.

In an EWPN, the original WPN nine-tuple has the following two parameters of **Result Sequence Size** and **Checked Sequence Size** changed to be exact cardinality.

These are new parameters added to an EWPN:

1. **Correct Checked Sequence Size:** Total number of URIs in the Checked Sequence that has been checked and verified by an agent, such as the owner, to actually be about the referent. This means that they have

been verified by some investigation of the web page addressed by the URI, or in the case of URIs without web pages, the URI itself or its use in other contexts.

2. **Correct Checked Sequence:** A list of URIs in the Result Sequence that have been checked and *are* about the referent. The number of URIs in this list will be equal to the Correct Checked Sequence Size.
3. **Incorrect Checked Sequence:** A list of URIs in the Result Sequence that have been checked and are *not* about the referent. The number of URIs in this list will be equal to the Checked Sequence Size minus the Correct Checked Sequence Size.
4. **Further Information:** Any further potentially useful information. Metadata could give version history, such as how often the WPN is updated. More metadata would be crucial if one were merging WPNs, such as one would want to do when building multilingual WPNs.

The entries in the two lists of URIs may also include optional **relevance**, **comment**, and **number** parameters. The **relevance** parameter allows the inspecting agent to rate a URI on an ordinal scale as to how relevant to the WPN they are, as well as an optional **comment** for any additional potentially relevant information on the URI. Search engines return the URIs in a sequence, and so it is recommended that the order of the URI lists be in the same order that the search engine returned. A **number** parameter is provided for each URI to preserve the order returned by the search engine. A number of zero indicates the URI has been added by manual augmentation. The URIs or the content of the URIs may be stored by the user.

The information content of an Expanded WPN can be encoded as an XHTML RDDDL file[1], allowing things to be distinguished from web pages at the level of retrieved representation, not just at the URI scheme level. This RDDDL format provides *a web page that is guaranteed to be about one thing*. A few of the more obvious encodings (XML, RDF, OWL) are explored as examples of WPN use in §5. A WPN RDDDL is available at <http://www.webpropernames.org>, along with schemas and examples for other encodings.

5. USES OF WEB PROPER NAMES

5.1 As Authoritative Web Pages for Things

The information in an Expanded WPN can be stored as a RDDDL document that can usefully be displayed in a web browser for human perusal. Currently, while *http:* URIs are allowed to denote things not on the Web, it is unclear what type of representation should be retrieved for such URIs. The RDDDL form of a Web Proper Name can be used as a standard representation for URIs that denote things, eliminating ambiguity and providing useful information for the user. The creation, storing, and collection of EWPNs can be integrated into web browsers in the same fashion as bookmarks. There are already several established XML-based bookmark schemes like XBEL (XML Bookmark Exchange Language), yet an EWPN can do more than a conventional bookmark scheme[2] by offering an optional set of search terms that can be used to find out more information about a referent than any single page. An example screenshot of an EWPN RDDDL for the Eiffel Tower is included in Figure 4.

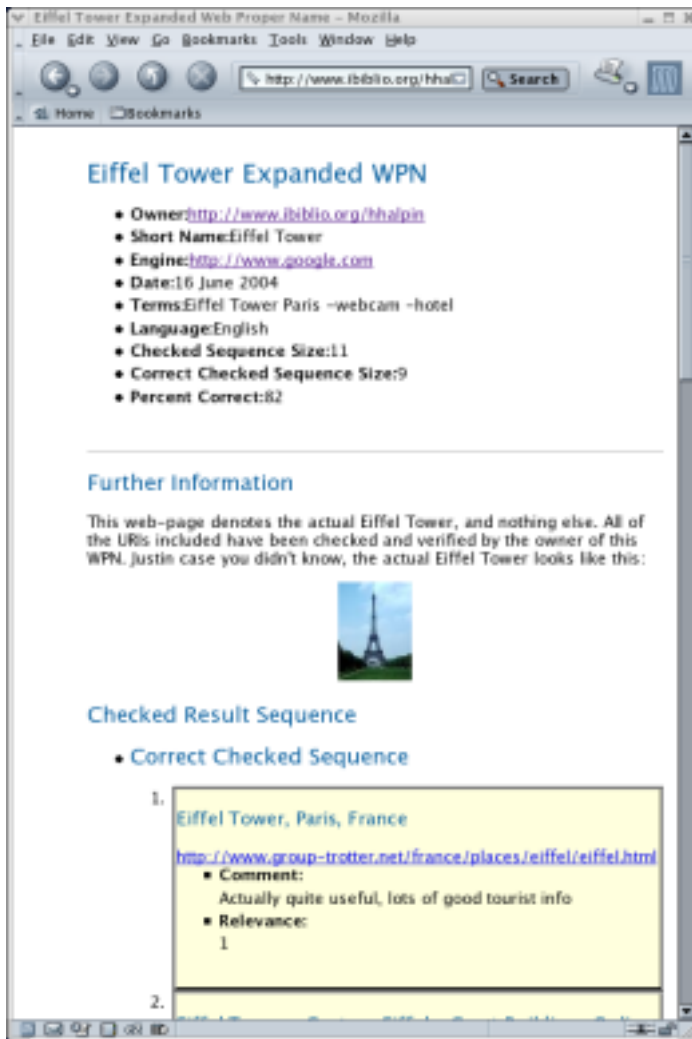


Figure 4: Screen Shot of an Expanded WPN RDDL

5.2 The Semantic Web from the Bottom-Up

The WPN proposal is crucial in the framework of the Semantic Web, if only to distinguish things from web-pages. Already, there is movement to store bookmarks as RDF as exemplified by Annotea's bookmark scheme[8]. A bookmark can be stored as a metadata about in a particular web page, and in a similar manner an EWP can be stored as RDF; all EWP formats can be easily transformed into RDF, since the base component of a WPN are URIs. This would allow the expressive power of OWL to be used in the management of EWPNs. OWL *unionOf* and *intersectionOf* can then be used automatically merge EWPNs and find difference sets of EWPNs. WPNs provide a natural way for everyday users of the Web to build ontologies in an analogous way that they currently build hierarchies of Web bookmarks. This use of WPNs provides an alternative and complementary methodology for the development of the Semantic Web other than the top-down methodology that hopes large organizations will come to agreement on standard ontologies for various domains. In contrast, the bottom-up methodology notes users are already creating rough and ready ontologies at home through their web searches.

6. CONCLUSION

There is much work to be done. Since WPNs have yet to be tested on a large scale, the exact form of the *wpn:* URI scheme, as well as the inventory of information included therein, cannot be confidently said to be optimal. Likewise the shape and contents of EWPNs will probably be in need of extensions and revisions. The problem of reference is fundamental to the Web, involving the crucial aspects of co-reference and identity. The Web Proper Name proposal, by making a clear distinction between a referent and web pages about that referent, adds to the conceptual apparatus needed to tackle this problem. By offering a series of concrete formats, applications that exploit this distinction can be built. It is in all our best interest, from the everyday user to professional ontologists, to put reference, in all of its mystery and power, back into the Web.

7. ACKNOWLEDGMENTS

Thanks to Karen Spärck Jones and Norm Walsh for comments, as well as the feedback by participants in *www-rdf-interest@w3.org*, *www-tag@w3.org*, and many others.

8. REFERENCES

- [1] J. Borden and T. Bray. Resource Directory Description Language. Technical report, RDDL Group, February 18 2002. <http://www.rddl.org/>.
- [2] F. Drake. The XML Bookmark Exchange Language. October 28 1998. <http://pyxml.sourceforge.net/topics/xbel//>.
- [3] G. Frege. Uber sinn und bedeutung. *Zeitschrift fur Philosophie and philosophie Kritik*, 100:25–50, 1892.
- [4] N. Goodman. *Languages of Art: An approach to a theory of symbols*. Hackett Publishing, 1976.
- [5] R. Guha. Semantic Negotiation: Co-identifying Objects across Data Sources. In *In Proceedings of Semantic Web Services Symposium*, Palo Alto, CA, 2004.
- [6] S. Harnad. The Symbol Grounding Problem. *Physica*, 1990.
- [7] I. Jacobs. Architecture of the World Wide Web. W3C, July 5 2004. <http://www.w3.org/TR/webarch/>.
- [8] M.-R. Koivunen, R. Swick, J. Kahan, and E. Prudhommeaux. An Annotea Bookmark Schema. 2003. <http://www.w3.org/2003/07/Annotea/BookmarkSchema-20030707>.
- [9] S. Kripke. *Naming and Necessity*. Harvard University Press, 1972.
- [10] D. Lenat and E. Feigenbaum. On the thresholds of knowledge. In *In Proceedings of International Joint Conference on Artificial Intelligence*. 1987.
- [11] L. Masinter. Dated URI and Thing Denoted By URI scheme. IETF Draft, 2004. <http://www.larry.masinter.net/duri.html>.
- [12] B. Russell. On Denoting. *Mind*, 14:479–493, 1905.
- [13] J. R. Searle. Proper Names. *Mind*, 67:166–173, 1958.
- [14] B. C. Smith. The Owl and the Electric Encyclopedia. *Artificial Intelligence*, 47:251–288, 1991.