

Functional Requirements for Information Resource Provenance on the Web

James P. McCusker, Timothy Lebo, Alvaro Graves, Dominic Difranzo, Paulo
Pinheiro, and Deborah L. McGuinness

`{mccusj,lebot,gravesa3,difrad}@rpi.edu,paulo.pinheirodasilva@pnnl.gov,`
`d1m@cs.rpi.edu`

Tetherless World Constellation
Department of Computer Science
Rensselaer Polytechnic Institute
110 8th Street Troy, NY 12180, USA
<http://tw.rpi.edu>
Pacific Northwest National Labs
Richland, WA, USA
<http://www.pnnl.gov>

Abstract. HTTP transactions have semantics that can be interpreted in many ways. At a low level, a physical stream of bits is transmitted from server to client. Higher up, those bits resolve into a message with a specific bit pattern. More abstractly, information, regardless of the physical representation, has been transferred. While the mechanisms associated with these abstractions, such as content negotiation, are well established, the semantics behind these abstractions are not. We extend the library science resource model Functional Requirements for Bibliographic Resources (FRBR) with cryptographic message and content digests to create a Functional Requirements for Information Resources (FRIR) ontology that is integrated with the W3C Provenance Ontology (PROV-O) to model HTTP transactions in a way that clarifies the many relationships between a given URL and all representations received from its request. Use of this model provides fine-grained provenance explanations that are complementary to existing explanations of web resources. Furthermore, we provide a formal explanation of the relationship between HTTP URLs and their representations that conforms with the existing World Wide Web architecture. This establishes the semiotic relationships between different information abstractions, their symbols, and the things they represent.

Keywords: World Wide Web, Information Resources, Data Management, multi-level granularity provenance

1 Introduction

The architecture of the World Wide Web [1] defines the relations between URLs, Resources, and Representations, which is illustrated in Figure 1¹. However, these relationships are incomplete, since the content of representations can change over time and content negotiation can result in different data being transferred. For example, the temperature reading in a weather report will change regularly, while different requests for the same weather report can return a variety of formats such as HTML, XML, RDF, and JSON. The ability to explain what an HTTP client sees as a result of a transaction and how, exactly, it relates to the URL that it requested is critical to the understanding of both how information resources² work on the web and how the provenance of web information resource access should be represented. We look to library science and provenance models to help provide these explanations, along with some help from the field of semiotics.

There are many reasons to clarify these semantics. For instance, the content of an image is more important than its format. Validating that a pathologist reviewed a particular image relies on the fact that the pathologist saw a particular image, not what file format it was saved in. In fact, transcoding of that image from a database to the client may happen as part of a web application. If it were possible to identify content regardless of format, our doctor would be able to make verifiable claims that she not just read data from a particular file, but that she saw a particular image. Similarly, web site mirroring mechanisms allow the same content to be available from multiple locations. Content-based identity of information would allow users to discover alternative locations for data, and validate that the information is actually the same regardless of source or format.

1.1 A Weather Example

To illustrate some of the issues regarding the relationship between a URL and the variety of representations that its request may return, we use a weather report provided by the National Oceanic and Atmospheric Administration's (NOAA) National Weather Service. Current weather conditions are provided for locations across the United States and include fundamental measures such as time, temperature, wind direction, and visibility distance. The latest hourly reports for Boston are provided in both RSS³ and XML⁴ formats. Although the service reports that it updates every hour on the hour, updates occur at unpredictable

¹ Copyright ©2004 World Wide Web Consortium, (Massachusetts Institute of Technology, European Research Consortium for Informatics and Mathematics, Keio University). All Rights Reserved. <http://www.w3.org/Consortium/Legal/2002/copyright-documents-20021231>

² Because we consider URLs returning status codes other than 200 to be non-information resources, they are out of scope in this paper.

³ http://www.weather.gov/xml/current_obs/KBOS.rss

⁴ http://www.weather.gov/xml/current_obs/KBOS.xml

intervals. In this particular example, the service updated at 3:00 and 4:00, handled RSS requests at 3:05 and 4:05, and handled XML requests at 3:10 and 4:10.

Given the current Web Architecture, what can we say about these two URLs and the four representations retrieved by their request? According to the AWWW, [1] the two RSS files *represent* the referent *identified* by the URL, while, because the URL is different, the XML files represent another referent. That these are alternative representations for the same referent means that we need a more sophisticated understanding of how the four files relate to one another and whether each relates to its URL *differently*. How can this be accomplished? We could compare files, but different formats would make it impractical to see their similarities. We could look to the files' creation date to learn when each file was received, but we cannot know how content has changed over time or if two transactions returned the same content in different representations. If different clients received the different representations, how can they begin to rationally discuss, compare, and share their individual representations?

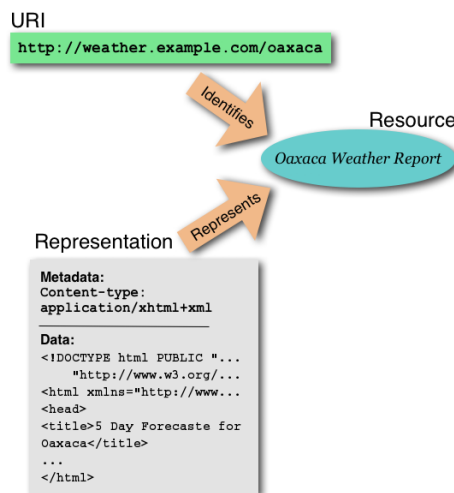


Fig. 1. The relationships between identifier, resource, and representation from Architecture of the World Wide Web.

2 Background: Existing W3C Recommendations

This leads us to wonder if there are any other existing semantics defined in W3C recommendations relating to how URIs, XML entities and RDF resources are related. This may appear to be a surprising question after years of success of W3C recommendations. However, the latest recommendations for XML [2] and RDF/XML [3] do not illuminate the issue. The XML recommendation [2] comes no closer to the issue than to state the following:

“Attempts to retrieve the resource identified by a URI may be redirected at the parser level (for example, in an entity resolver) or below (at the protocol level, for example, via an HTTP Location: header). In the absence of additional information outside the scope of this specification within the resource, the base URI of a resource is always the URI of the actual resource returned. In other words, it is the URI of the resource retrieved after all redirection has occurred.”

From this definition, one can infer that more than one resource may be returned for a URL and that the exact nature of this resource can be unpredictable. This is because an HTTP-based entity resolver implies the ability to return multiple representations of the same content. Similarly, the RDF/XML recommendation [3] states that:

“nodes are RDF URI references, RDF literals or are blank nodes. Blank nodes may be given a document-local, non-RDF URI references identifier called a blank node identifier. Predicates are RDF URI references and can be interpreted as either a relationship between the two nodes or as defining an attribute value (object node) for some subject node.”

but goes no further. An “RDF URI reference” is syntactically described, and the recommendation further discloses that “RDF URI references are compatible with the anyURI datatype as defined by XML schema datatypes [4], constrained to be an absolute rather than a relative URI reference.” Again, this leaves the recommendation reader without an explanation of what is the meaning of a URI in an RDF graph.

3 The semiotics of HTTP URLs

The dereferencing of a URL can be mapped to a semiotic interpretation. For example, it is possible to use Ogden and Richards’ Semiotic Triangle [5], a model of how real world objects are related to symbols and how people think about those objects from a linguistic perspective. In order to consider HTTP operations in these terms, it is important to remember that a URL is not only a symbol but also an address for information about that symbol. For example, http://www.weather.gov/xml/current_obs/KBOS.xml indicates that a web page can be accessed using the HTTP protocol against the server denoted by the name www.weather.gov and requesting the document `/xml/current_obs/KBOS.xml`. The document obtained is a representation (an XML document) of the thing identified by this URL. Figure 2 illustrates the partial correspondence between the semiotic triangle and the web architecture.

While a URL is a Symbol that stands for and identifies a Referent Resource, the correspondence to thoughts (from the Semiotic Triangle) or representations (from the AWWW) isn’t immediately clear. The major issue is that the document retrieved cannot be defined only as a representation of a resource: The document can be described in terms of either its content or the set of bytes used to represent it – or both. So, the document needs to be described further. A potential solution is to refine the representation into its constituent identities that are based on different levels of abstraction. In the next section we will introduce a model that, when paired with a provenance model, can provide the necessary distinctions to fully satisfy both the semiotic relationships inherent within HTTP and the means to provide provenance traces for HTTP transactions at the levels of abstraction that are inherent within the protocol.

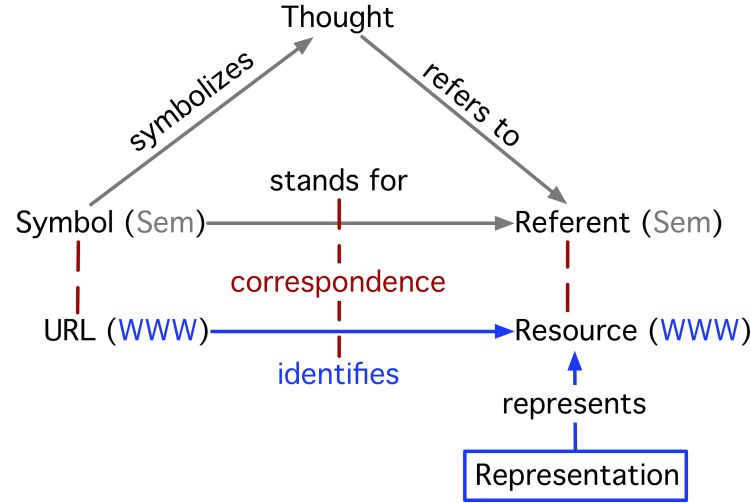


Fig. 2. AWWW’s URL and Resource correspond to the semiotic triangle’s Symbol and Referent, respectively. A representation is itself another referent that is not identified here, but will be elaborated on in Section 4.

4 FRBR and FRIR

Functional Requirements for Bibliographic Resources (FRBR) [6] is a mature model from the library science community that distinguishes four aspects of an author’s literary work, ranging from purely concrete to completely abstract. For instance, FRBR can describe how different copies of the same book, or different editions of the book, relate to each other. The most concrete aspect is the Item – the physical book that exists in the world. Items are singular entities; making a copy of an Item results in a new Item. Items are exemplars of Manifestations, which represent similar physical structure. For instance, an exact copy of an Item preserves the original Manifestation. If the copy is inexact, or if the book is turned into an audio book, then the Manifestation changes. However, the Expression of the paperback and audio book remains the same, because the Expression reflects particular content regardless of physical configuration. An Expression in turn realizes a Work, which is “a distinct intellectual or artistic creation.” [7] A Work remains the same through different realized Expressions that result from translation, revision, or any other change. To facilitate discussion, we use the term *FRBR stack* to refer to a tuple (*frbr:Work*, *frbr:Expression*, *frbr:Manifestation*, *frbr:Item*) that represents these four distinct aspects of a resource.

Functional Requirements for Information Resources⁵ (FRIR) extends the use of *frbr:Work*, *frbr:Expression*, *frbr:Manifestation*, and *frbr:Item* to electronic resources, and therefore any information resource. Within electronic resources,

⁵ <http://purl.org/twc/pub/mccusker2012parallel>

a *frbr:Work* remains a distinct intellectual or artistic creation. A *frbr:Work* corresponds to the Resource or Referent in the semiotic framework discussed above, and is identified by a URL, as was shown in Figure 2. Taken together, *frbr:Expression*, *frbr:Manifestation*, and *frbr:Item* are all aspects of the Representation, and are each Referents in their own rights. Inasmuch as they can be identified or symbolized, they have symbols that identify them. *frbr:Expression* corresponds to a specific set of content regardless of its serialization. For instance, two files would have the same *frbr:Expression* if they are the same picture stored in two different formats (e.g., JPG and PNG). Similarly, a spreadsheet stored in both CSV and Excel would still have the same *frbr:Expression*. *frbr:Manifestations* correspond to a specific bit pattern. If a file is an exact copy of another file, they have the same *frbr:Manifestation*. An *frbr:Item* is a specific copy of information stored somewhere or transmitted through a communication link. If a copy of the *frbr:Item* is made, it results in a new *frbr:Item*.

FRIR also integrates FRBR with the W3C Provenance Ontology (PROV-O) by declaring *frbr:Endeavour* to be a subclass of *prov:Entity* and mapping 14 of 18 *frbr:relatedEndeavour* subproperties as subproperties of one or more of *prov:wasDerivedFrom*, *prov:alternateOf*, and *prov:specializationOf*, as shown in Figure 3.

As part of FRIR we have identified two levels of cryptographically computable identity: content and message. Conventional message digests such as MD5 or SHA-1 produce identifiers where the probability of creating the same identifier using different data is vanishingly small. This corresponds very closely to our definition of *frbr:Manifestation* for electronic resources, so we make it possible to identify *frbr:Manifestations* using message digests. Similarly, a number of *content digests* have been developed for RDF graphs, spreadsheets, images, and XML documents that provide the same digest hash regardless of any particular serialization. We use this to computationally identify *frbr:Expressions*. Further work on creating content digests will allow us to incrementally improve the ability to identify common *frbr:Expressions*. These identifiers fill out the means by which to identify the representation referents, as shown in Figure 4.

5 Explaining HTTP with FRBR, FRIR, and PROV-O

When explaining what is retrieved from a URL, the URL denotes a single *frbr:Work*. We implement these explanations in RDF, which follows the non-unique naming assumption. That is, unless otherwise specified, two identifiers can potentially denote the same thing. URLs are perfect examples of this. If a web site is mirrored, a page on the mirror corresponds to a page on the original. Those two pages can be thought of as the same *frbr:Work* within the FRBR/FRIR perspective. Content retrieved from URLs can change over time, but are expected to have a similar sort of coherence as defined by *frbr:Work* as “a distinct intellectual or artistic creation.” [7]

HTTP 1.1 [8] introduced content negotiation, which makes it possible to abstract a URL away from any one particular file format. When a client asks

Subclass	Superclass
frbr:Event	prov:Activity
frbr:ResponsibleEntity	prov:Agent
frbr:Endeavour	prov:Entity
nie:DataObject	prov:Entity

(a)

Subproperty	wasDerivedFrom	alternateOf	specializationOf
frbr:adaptionOf	X		
frbr:imitationOf	X		
frbr:reconfigurationOf	X		
frbr:transformationOf	X		
frbr:abridgementOf	X	X	
frbr:arrangementOf	X	X	
frbr:reproductionOf	X	X	
frbr:summarizationOf	X	X	
frbr:translationOf	X	X	
frbr:alternateOf		X	
frbr:revisionOf		X	
frir:redirectsToTransitive		X	
frbr:embodimentOf			X
frbr:exemplarOf			X
frbr:realizationOf			X

(b)

Prefix	URI
frbr:	http://purl.org/vocab/frbr/core#
frir:	http://purl.org/twc/ontology/frir.owl#
prov:	http://www.w3.org/ns/prov#
nie:	http://www.semanticdesktop.org/ontologies/2007/01/19/nie#

(c)

Fig. 3. (a) Class mappings between FRBR and PROV-O. (b) Property mappings between FRBR, FRIR, and PROV. PROV super properties are columns and FRBR and FRIR subproperties are rows. (c) Prefix mappings for (a) and (b).

an HTTP server for a mime type at a URL, the server can respond with many different possible files depending on how the content is negotiated. If the client asks for plain text, the server will try to find the best way of representing the content of the URL in plain text. This idea of “same content regardless of format” is built into *frbr:Expression*. As previously discussed, the bit sequence of a file aligns very closely with *frbr:Manifestation*, so we use message digests to express this. *frbr:Items* can be files on disk, but they can also be data as streamed over a network connection. We uniquely identify the data streamed over a particular HTTP transaction using the combined message digest of the HTTP header and content. Since the header includes the exact time that the transaction occurred, the likelihood of a *frbr:Item* collision is very low. This enables provenance trace

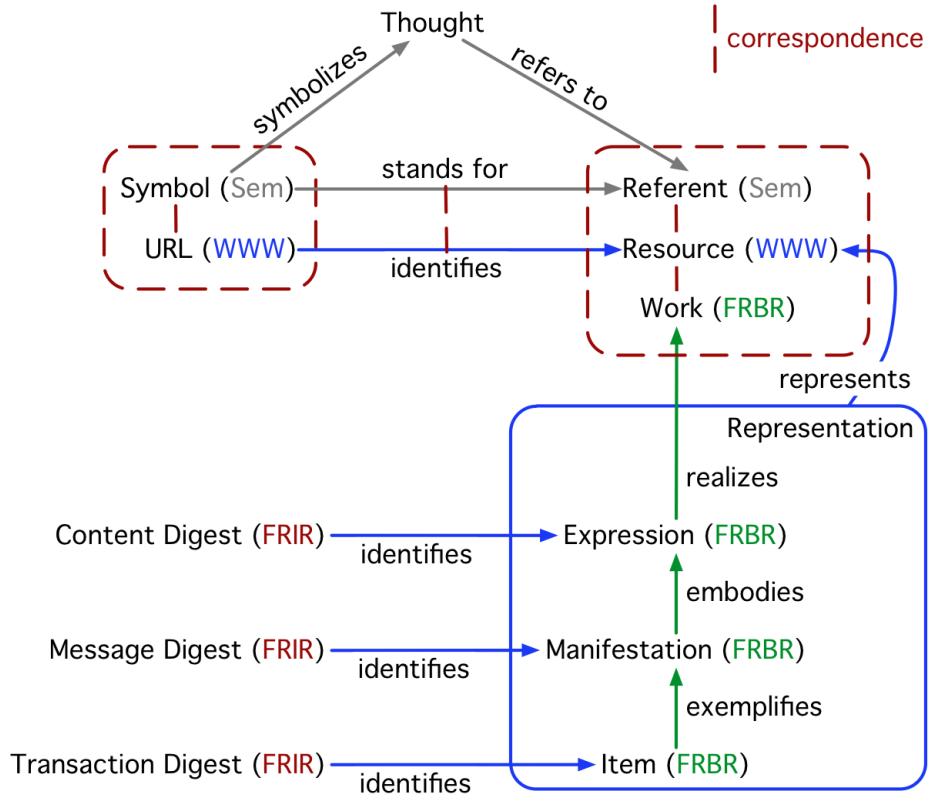


Fig. 4. Relating URIs, Resources, and Representations using FRIR, FRBR, and the semiotic triangle. URIs are symbols that identify resources, which in semiotics are referents and considered *frbr:Works* in FRBR. The representation of that resource is the content that comes from dereferencing the URL, and is composed of an *frbr:Expression*, *frbr:Manifestation*, and *frbr:Item*. The proposed content identities create implicit symbols (URIs) for each level of representation. Users can then use the level of abstraction that suits their task.

assertions to be applied to individual HTTP transactions without having to store the entire transaction.

An HTTP GET can be a very simple transaction. A client makes a request to a server for a particular URL, the server looks up which file corresponds to that URL, and copies it to the network channel in response. The client then copies the data it sees from the network channel and either saves it to disk or displays it on screen. Things can become much more complicated on both ends, but these complications can be explained using current provenance representations, including the emerging W3C Prov standard [9]. This simple case, however, belies the subtleties that we discuss above. The following is a formalization of an HTTP GET request and response composed of common provenance constructs (events, generated by, used, etc.) that are under development in W3C's Prov standard:

HTTP GET: The server and client both share an event, the HTTP connection, which is composed of a request and response. The request is generated by the client and transmitted to the server. It is itself an Item with a singular FRBR stack. The request is for a specific *frbr:Work*, and if there are Accept headers sent, then the request is for a *frbr:Manifestation* with specific properties (the file format). The server then uses the request to generate a response, which is an Item of the URL's *frbr:Work*. This *frbr:Item* only exists on the network channel, and if the client saves the *frbr:Manifestation* to disk, it produces another *frbr:Item*. The response Item is derived from the server's file *frbr:Item*, and the client's file *frbr:Item* is derived from the response Item. All three items share the same Manifestation, Expression and Work (the URL).

HTTP POST: A similar explanation can be made for HTTP POST requests, which send a document as input content. In this case, both request and response content can be represented as FRBR stacks with no explicitly identified *frbr:Work*. Because web servers that handle POST requests derive their responses from the request, their handling can be formalized as a derivation edge in a provenance graph using the POST URL as an agent controlling the transaction process. Two HTTP request methods, PUT and DELETE, are used specifically to change the value of the *frbr:Work* by creating a new *frbr:Expression* (PUT) or invalidating existing *frbr:Expressions* (DELETE).

HTTP also provides other request methods to ask for services and information about a particular resource. These metadata request methods, HEAD and OPTIONS, do not provide information resources as discussed, and so are not in the scope of the paper. Similarly, the HTTP request methods TRACE and CONNECT are more functional in nature and deal more with the actual server than its content and are also outside the scope of this paper.

6 Implementation

We provide an implementation of curl called `pcurl.py`⁶ that will record the provenance of an HTTP GET transaction using FRBR⁷, FRIR⁸, Nepomuk File Ontology (NFO)⁹, PROV-O¹⁰, and HTTP-in-RDF¹¹. We show a retrieval of the HTTP-in-RDF core classes as an example in Figure 5. We use message and content hashes to generate URIs for *frbr:Expressions*, *frbr:Manifestations*, and *frbr:Items* to allow for automatic aggregation of endeavors that share the same hash. Future use of OWL keys and multiple digest algorithms is enabled through creation of *frir:ContentDigest* and *nfo:FileHash* instances. In Figures 6 and 7 we also show how transcoding and mirroring are represented in the FRIR model.

7 Discussion

Using a four-part FRIR stack to identify web resources makes it possible to do a number of useful things. For instance, in cases such as the weather report, RSS feeds and XML files, the same information is conveyed in multiple formats at different URLs. FRIR naturally expresses this by asserting that the *frbr:Works* (the page URL) are *owl:sameAs* each other. More concrete levels of the FRBR stack, such as *frbr:Manifestation* and *frbr:Item*, however, will be distinct because of the differing formats and different physical locations of the representations. The identity between the *frbr:Works* and *frbr:Expressions* of these two URLs can be expressed in semantic site maps so that tools that prefer one format of data over another can discover which URLs can be accessed without concern for missing content from varying formats. In fact, we have previously argued that *owl:sameAs* has been overextended to assertions in Linked Data [11]. By linking *frbr:Works* and *frbr:Expressions* manually while keeping *frbr:Manifestations* distinct we provide a means to show exactly how two information resources on the web relate to each other. Additionally, we have shown how FRIR can be used to provide clarity to management of Open Government Data (OGD) [12] and have argued that FRBR constructs can be used to provide a clear description of sources of information on the web [13].

By formally modeling the description of web retrieval, we can compare the content received at different levels. In our weather report example, as the weather changes, so does the data. Two clients may see different data if they access it at different times. *frbr:Works* and *frbr:Expressions* can be used to show that content has changed, even when accounting for potentially different data formats, as is the case with the weather (XML vs RSS). Similarly, using content digests

⁶ <http://purl.org/twc/software/pcurl.py>

⁷ <http://purl.org/vocab/frbr/core>

⁸ <http://purl.org/twc/ontology/frir.owl>

⁹ <http://www.semanticdesktop.org/ontologies/2007/03/22/nfo>

¹⁰ <http://purl.org/twc/page/prov-o>

¹¹ <http://www.w3.org/TR/HTTP-in-RDF/>

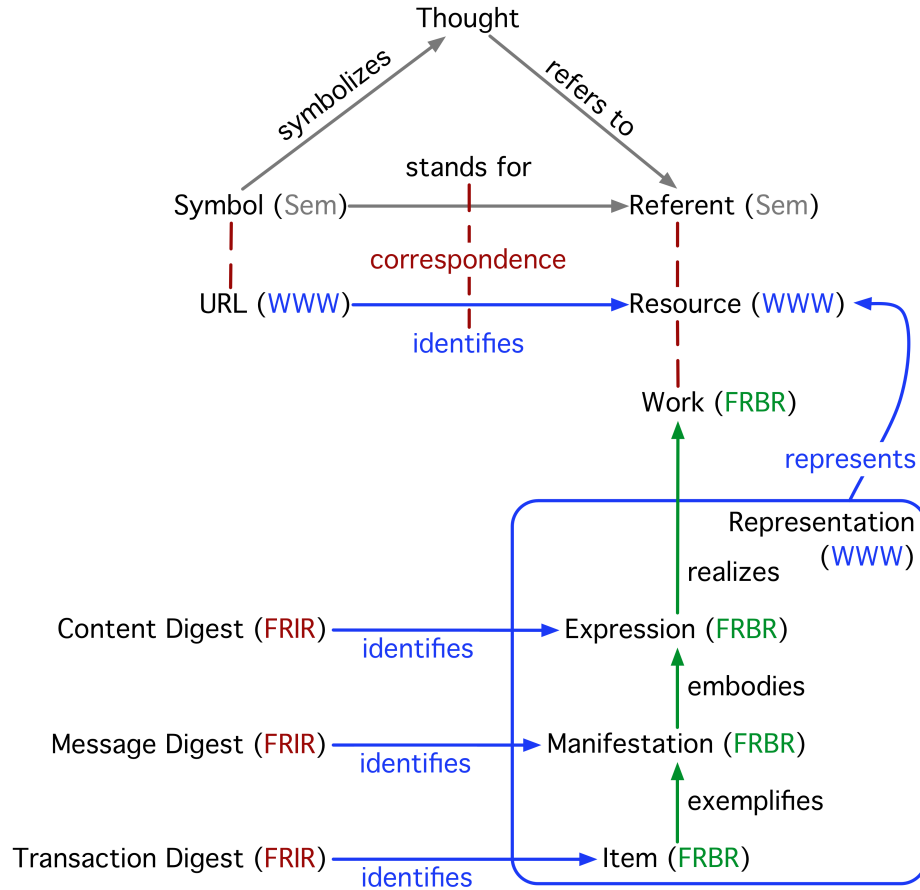


Fig. 5. Results of applying `pcurl.py` to retrieve the weather result example. The HTTP-in-RDF, FBIR, FRIR, PROV-O, and NFO vocabularies are used to create RDF descriptions of the representation received when the URL is requested. Entities are named using message and content digests, the HTTP transaction Item is associated to the file Item, which in turn has a FRBR stack representing all four aspects from the concrete file to the abstract URL/Referent/Work.

and cryptographic signatures, clients can assert that they have seen specific content, regardless of format, on a particular web page. This makes it possible for clients to refer to specific content regardless of format. This is critical for access to scientific databases. Available information is changing daily and released in different representations, which is convenient but can hinder experimental repeatability. Being able to assert which data was used in an experiment improves the transparency and veracity of the sciences that take advantage of that data. Finally, this theory of web information and access provides a way to create consistent provenance assertions about access of web information resources, which can improve interoperability of provenance statements about access of information resources.

8 Conclusion

We have shown how the use of FRBR and FRIR can help to describe the relation between a URL and the representation obtained using HTTP. We have also shown how this new representation describes a richer set of entities that can be identified by different elements from FRIR. Thus it is possible to use Content, Message and Transaction digests to identify the Expression, Manifestation and Item aspects of the representation. This can lead to a semantically richer description of an HTTP GET operation that includes provenance about the information published and transmitted on the web at each level of abstraction.

As future work, there are several paths we may take: In this paper, our focus has been on URLs that identify information resources. There is also the question of what (and how) non-information resources can be described in terms of FRBR and FRIR. This is particularly interesting when considering the relationships between URL *frbr:Works* that are associated by HTTP 303 redirections. Additionally, the solutions we present here are in principle compatible with the proposed changes to what has been called the *HttpRange-14* issue.¹² Applying FRBR and FRIR to address the relationships between a resource, its representations, and its identifiers in a clear manner can serve as a standard pattern for the provenance of web resource access, comparison, and integration.

9 Acknowledgments

The Tetherless World Constellation is partially funded by grants and/or gifts from DARPA, IARPA, U.S. Department of Energy, Fujitsu, LGS, Lockheed Martin, Microsoft Research, NASA, the National Science Foundation, and Qualcomm. This research was partially funded by the National Science Foundation under CREST Grant No. HRD-0734825.

¹² <http://lists.w3.org/Archives/Public/public-lod/2012Mar/0115.html>

References

1. Jacobs, I., Walsh, N.: Architecture of the World Wide Web, Volume One (December 2011) <http://www.w3.org/TR/2004/REC-webarch-20041215/>.
2. Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., Francois Yergeau: Extensible Markup Language (XML) 1.0 (Fifth Edition) (November 2008) <http://www.w3.org/TR/2008/REC-xml-20081126/>.
3. Klyne, G., Carroll, J.J., McBride, B.: Resource Description Framework (RDF): Concepts and Abstract Syntax (February 2004) <http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/#section-Graph-URIref>.
4. Biron, P.V., Malhotra, A.: Xml schema part 2: Datatypes (May 2001) <http://www.w3.org/TR/2001/REC-xmlschema-2-20010502/>.
5. Ogden, C.K., Richards, I.: The meaning of meaning. Trubner & Co, London (1923)
6. Madison, O., John Byrum, J., Jouguelet, S., McGarry, D., Williamson, N., Witt, M.: Functional requirements for bibliographic records final report. Technical report, International Federation of Library Associations and Institutions (February 2009) <http://www.ifla.org/VII/s13/frbr/>.
7. Madison, O., John Byrum, J., Jouguelet, S., McGarry, D., Williamson, N., Witt, M.: Functional Requirements for Bibliographic Records (February 2009) <http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records>.
8. Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T.: Hypertext transfer protocol-HTTP/1.1. Technical report, RFC 2616, June (1999)
9. Lebo, T., Sahoo, S., McGuinness, D.L., Mike Lang, J., Belhajjame, K., Cheney, J., Garijo, D., Soiland-Reyes, S., Zednik, S.: The PROV Ontology: Model and Formal Semantics (December 2011) <http://www.w3.org/TR/prov-o>.
10. Xu, S., McCusker, J.P., Krauthammer, M.: Yale Image Finder (YIF): a new search engine for retrieving biomedical images. *Bioinformatics* **24**(17) (2008) 1968–1970
11. McCusker, J.P., McGuinness, D.L.: Towards identity in linked data. Proceedings of OWL: Experience and Directions, San Francisco, USA (June 21-22, 2010) (2010) http://www.webont.org/owled/2010/papers/owled2010_submission_12.pdf.
12. McCusker, J., Lebo, T., Chang, C., Pinheiro da Silva, P., McGuinness, D.: Parallel Identities for Managing Open Government Data. *IEEE Intelligent Systems Open Government Data Special Issue* (2012)
13. McCusker, J., Lebo, T., Ding, L., Chang, C., Pinheiro da Silva, P., McGuinness, D.: Where did you hear that? Information and the Sources They Come From. *Proceedings of Linked Science 2011* (2011)

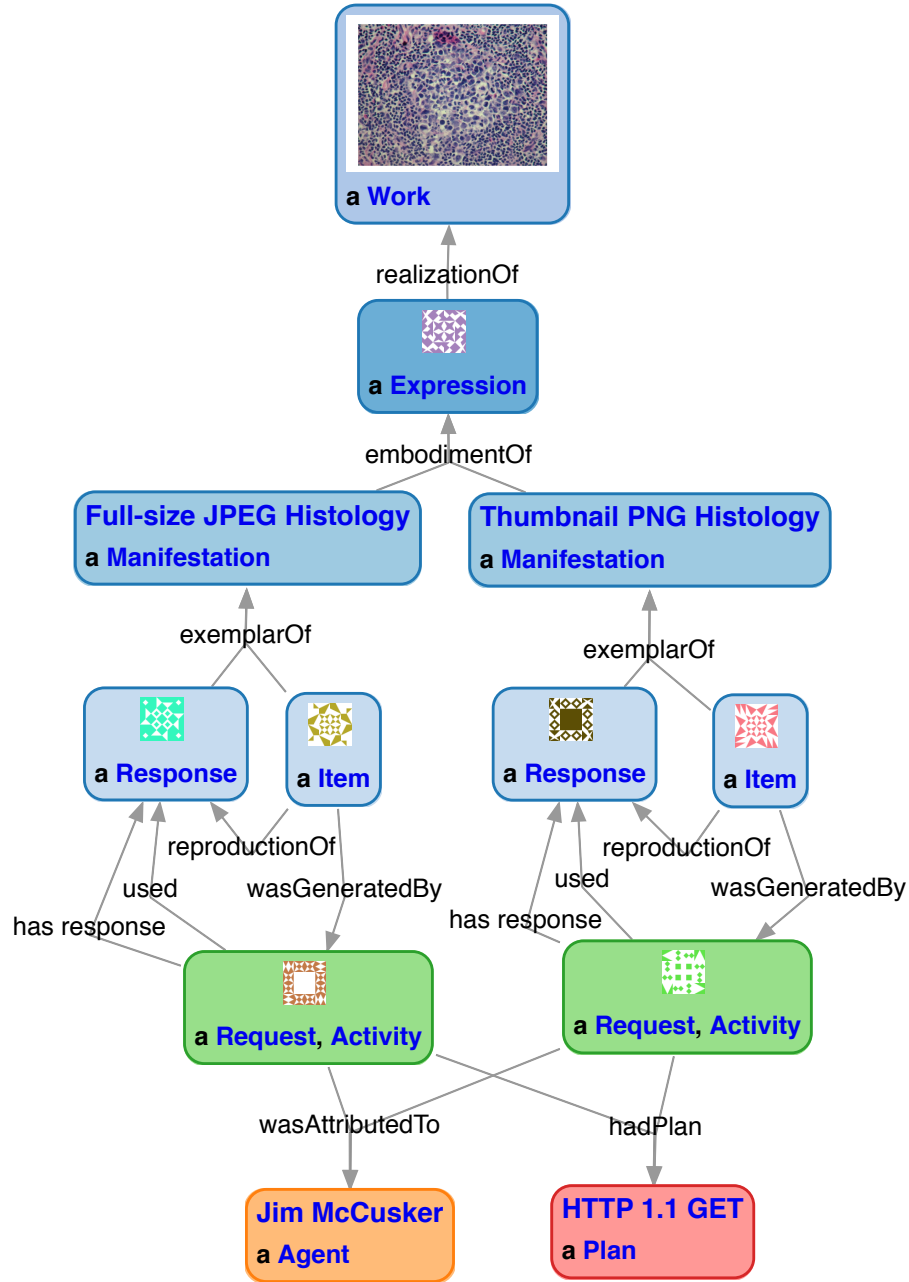


Fig. 6. An example of transcoding a histogram image from a large JPEG to a small thumbnail PNG. The *frbr:Expression* and *frbr:Work* are the same across the transcoding, but the *frbr:Manifestations* and *frbr:Items* are all distinct. This allows, for instance, a patient to verify that the low resolution image shown to them is the same content as the higher resolution image used to actually perform the analysis, even though the format and sizes are different. This graph was produced using pcurl.py

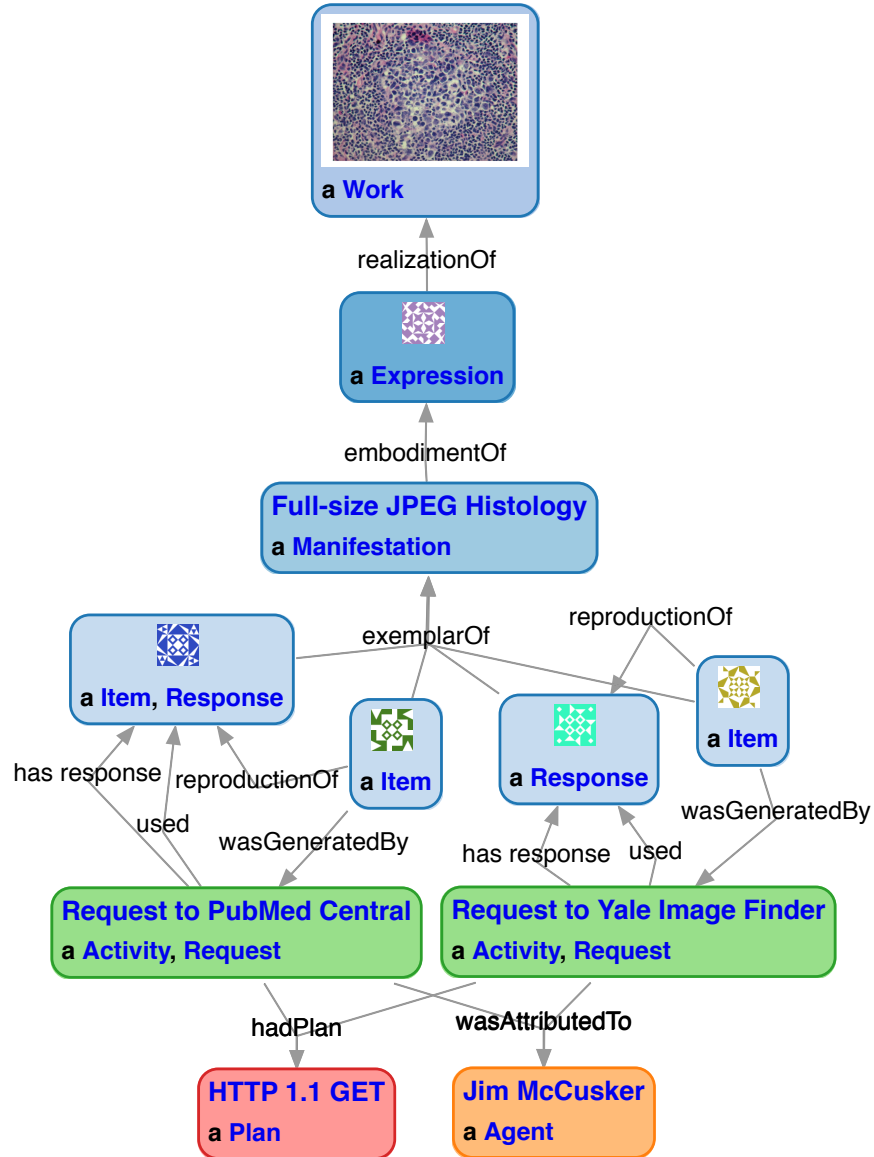


Fig. 7. An example of mirroring content between web sites. Here the Yale Image Finder [10] provides a mirror of an image published at PubMed Central. Since the file is an exact copy, the *frbr:Work*, *frbr:Expression* and *frbr:Manifestation* align, while the individual copies are different. This graph was produced using pcurl.py