Functional Requirements for Information Resource Provenance on the Web

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

<Timothy Lebo> prov:actedOnBehalfOf <James P. McCusker> .

What do we use a URL to identify?

(Answer: too many things)

(Problem: we’re getting confused!)

Two provenance use cases

How do we use URLs to track the provenance of the information we get from them?

Image analysis: who has reviewed this image?

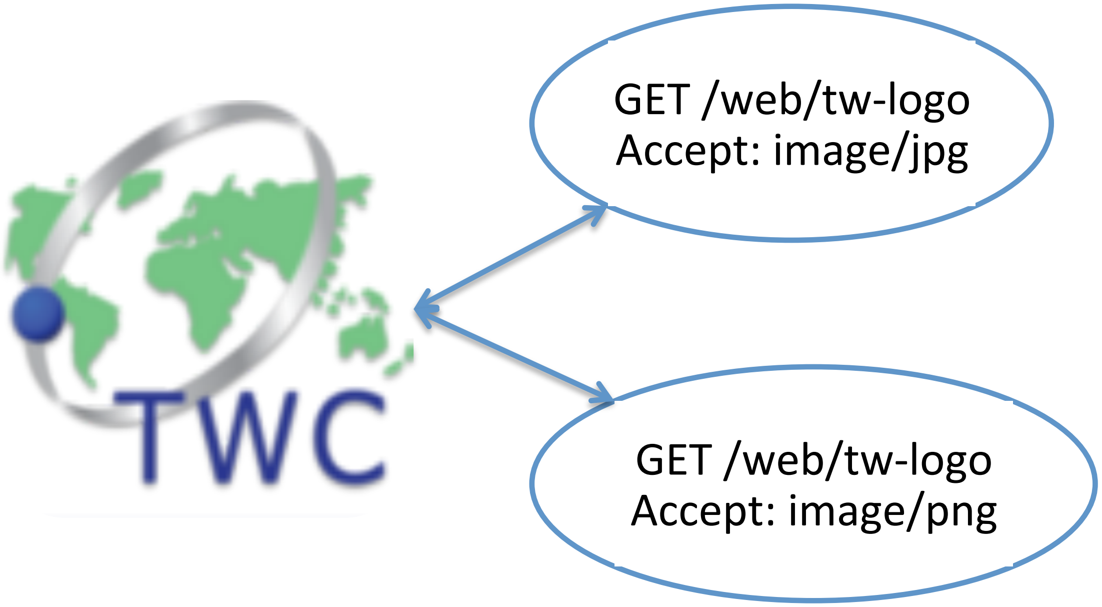
- Different tools identify and use different formats and resolutions.
- How do we know when Dr. Smith reviewed the image we're seeing?

Weather: weather.gov offers ongoing forecasts at an unchanging URL.

Use case 1: Image Analysis

What is a URL?

Is it the data?



GET /web/tw-logo
Accept: image/jpg

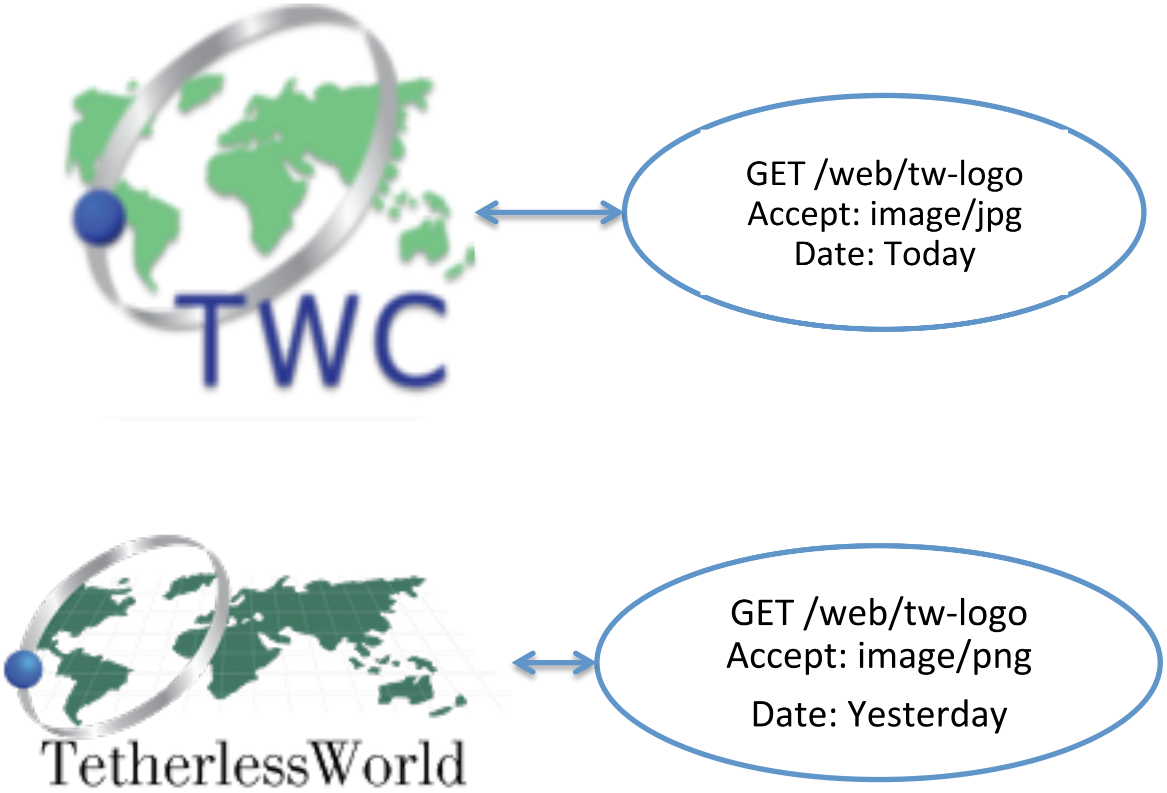
GET /web/tw-logo
Accept: image/png

Use case 1: Image Analysis

What is a URL?

~~Is it the data?~~

Is it the content?



GET /web/tw-logo
Accept: image/jpg
Date: Today

GET /web/tw-logo
Accept: image/png
Date: Yesterday

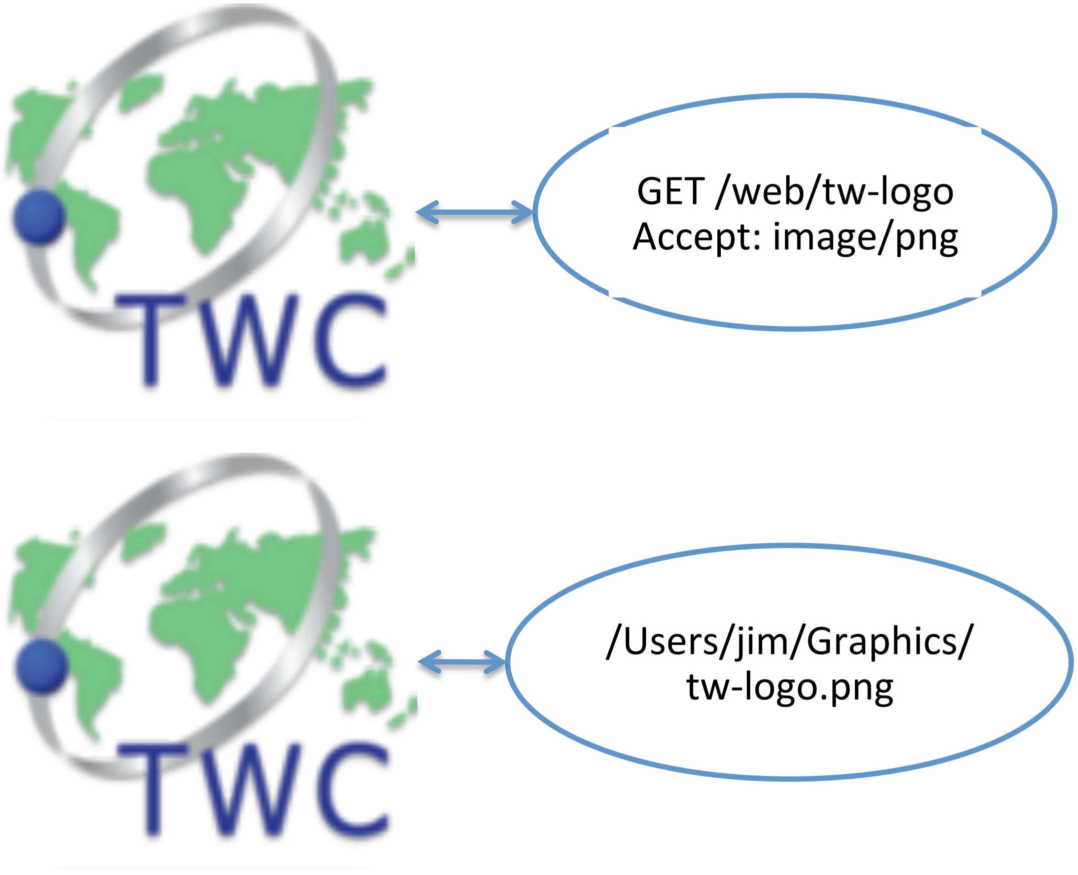
Use case 1: Image Analysis

What is a URL?

~~Is it the data?~~

~~Is it the content?~~

What about multiple copies of the same file?



GET /web/tw-logo
Accept: image/png

/Users/jim/Graphics/
tw-logo.png

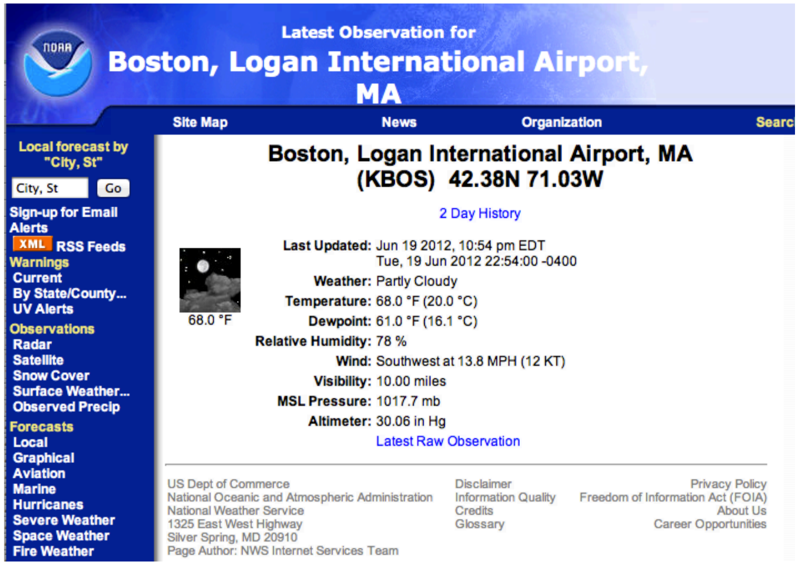
Use case 2: Weather

A Little More on Weather

It changes over time

It has multiple representations

Copies might be saved for historical analysis



Latest Observation for
Boston, Logan International Airport, MA

Boston, Logan International Airport, MA
(KBOS) 42.38N 71.03W

2 Day History

68.0 °F

Last Updated: Jun 19 2012, 10:54 pm EDT
Tue, 19 Jun 2012 22:54:00 -0400

Weather: Partly Cloudy

Temperature: 68.0 °F (20.0 °C)

Dewpoint: 61.0 °F (16.1 °C)

Relative Humidity: 78 %

Wind: Southwest at 13.8 MPH (12 KT)

Visibility: 10.00 miles

MSL Pressure: 1017.7 mb

Altimeter: 30.06 in Hg

Latest Raw Observation

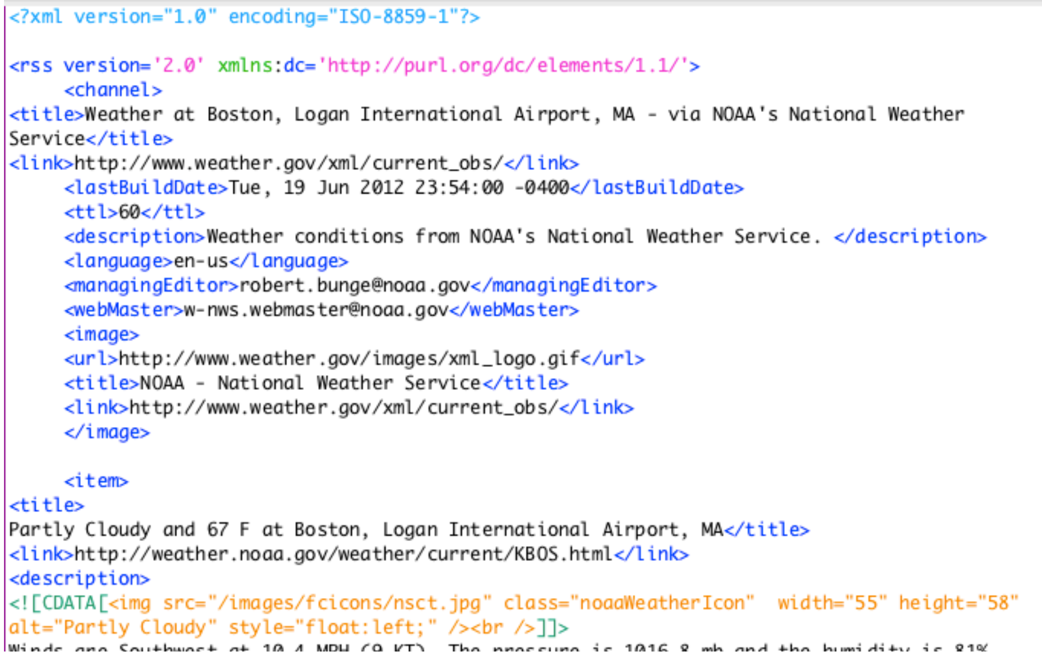
US Dept of Commerce
National Oceanic and Atmospheric Administration
National Weather Service
1325 East West Highway
Silver Spring, MD 20910
Page Author: NWS Internet Services Team

Disclaimer
Information Quality
Credits
Glossary

Privacy Policy
Freedom of Information Act (FOIA)
About Us
Career Opportunities

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

But it’s all the same URL!



```
<?xml version="1.0" encoding="ISO-8859-1"?>

<rss version='2.0' xmlns:dc='http://purl.org/dc/elements/1.1/'>
     <channel>
<title>Weather at Boston, Logan International Airport, MA - via NOAA's National Weather
Service</title>
<link>http://www.weather.gov/xml/current_obs/</link>
     <lastBuildDate>Tue, 19 Jun 2012 23:54:00 -0400</lastBuildDate>
     <ttl>60</ttl>
     <description>Weather conditions from NOAA's National Weather Service. </description>
     <language>en-us</language>
     <managingEditor>robert.bunge@noaa.gov</managingEditor>
     <webMaster>w-nws.webmaster@noaa.gov</webMaster>
     <image>
     <url>http://www.weather.gov/images/xml_logo.gif</url>
     <title>NOAA - National Weather Service</title>
     <link>http://www.weather.gov/xml/current_obs/</link>
     </image>

     <item>
<title>
Partly Cloudy and 67 F at Boston, Logan International Airport, MA</title>
<link>http://weather.noaa.gov/weather/current/KBOS.html</link>
<description>
<![CDATA[<img src="/images/fcicons/nsct.jpg" class="noaaWeatherIcon"  width="55" height="58"
alt="Partly Cloudy" style="float:left;" /><br />]]>
Winds are Southwest at 10.4 MPH (9 KT). The pressure is 1016.8 mb and the humidity is 81%
```

We need to solve these confusions to get good provenance for information on the web.

Okay, fine. Content changes.

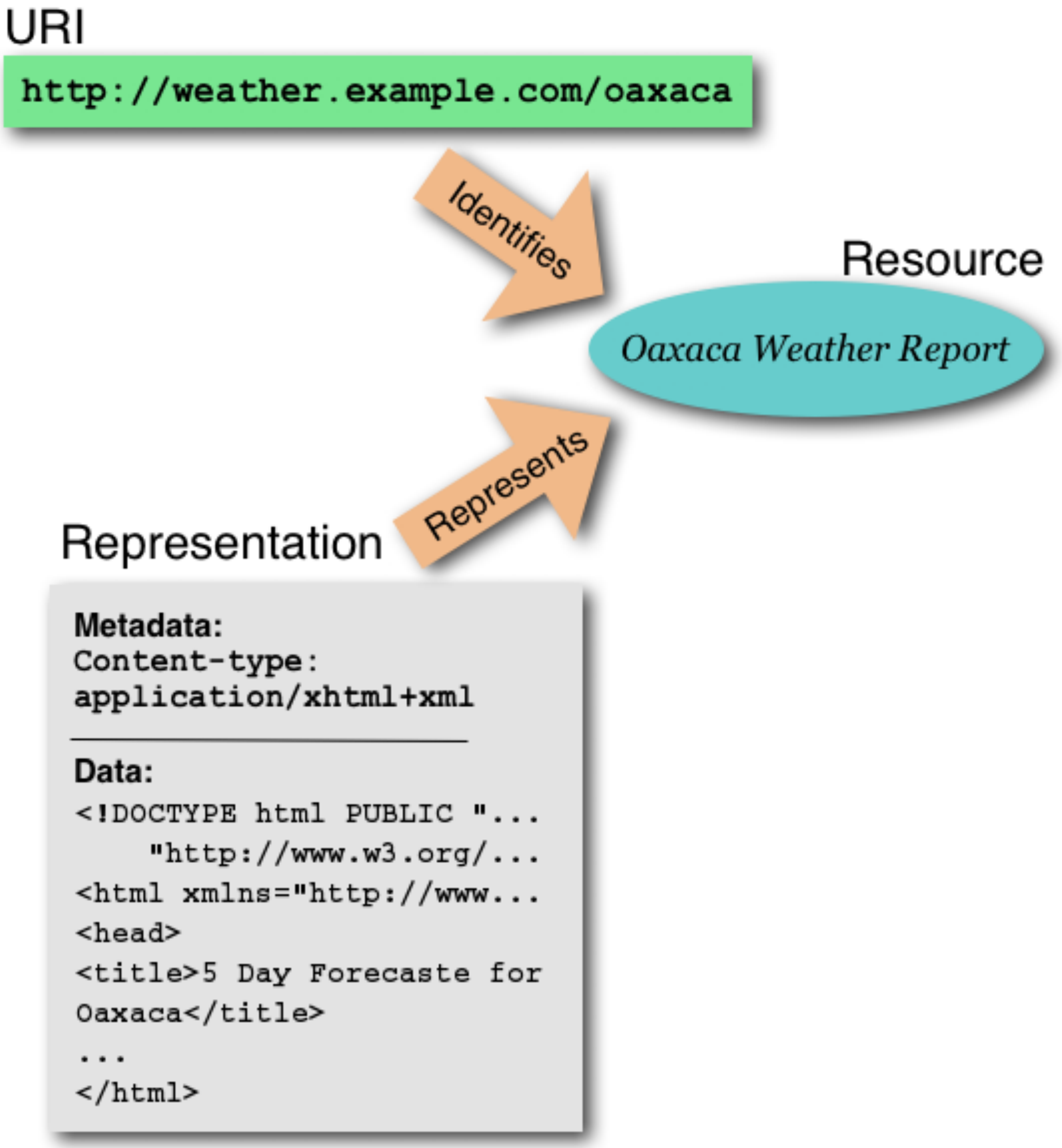
Let's be extremely conservative: unique identifiers to identify the content?

There are 4 steps to data sanity.

4 Steps to Data Sanity

Step 1: Distinguish resources from their representations

The Architecture of the World Wide Web [1] provides a framework for this.



```
Metadata:
Content-type:
application/xhtml+xml

Data:
<!DOCTYPE html PUBLIC "...
    "http://www.w3.org/...
<html xmlns="http://www...
<head>
<title>5 Day Forecaste for
Oaxaca</title>
...
</html>
```

[1] http://www.w3.org/TR/webarch/

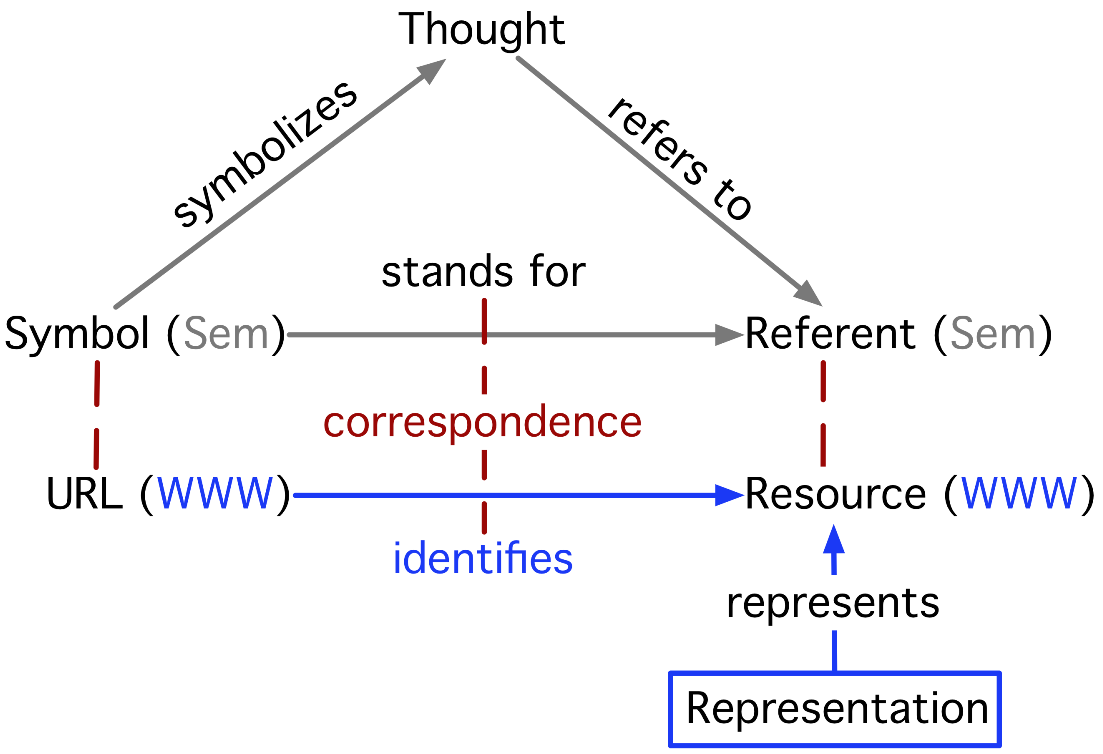
4 Steps to Data Sanity

Step 1: Distinguish resources from their representations

Actually, this is very close to the semiotic triangle.

We can now separate out symbols, referents, and representations.

This leads to fragmentation as identifiers proliferate.



Wait, what is FRBR?

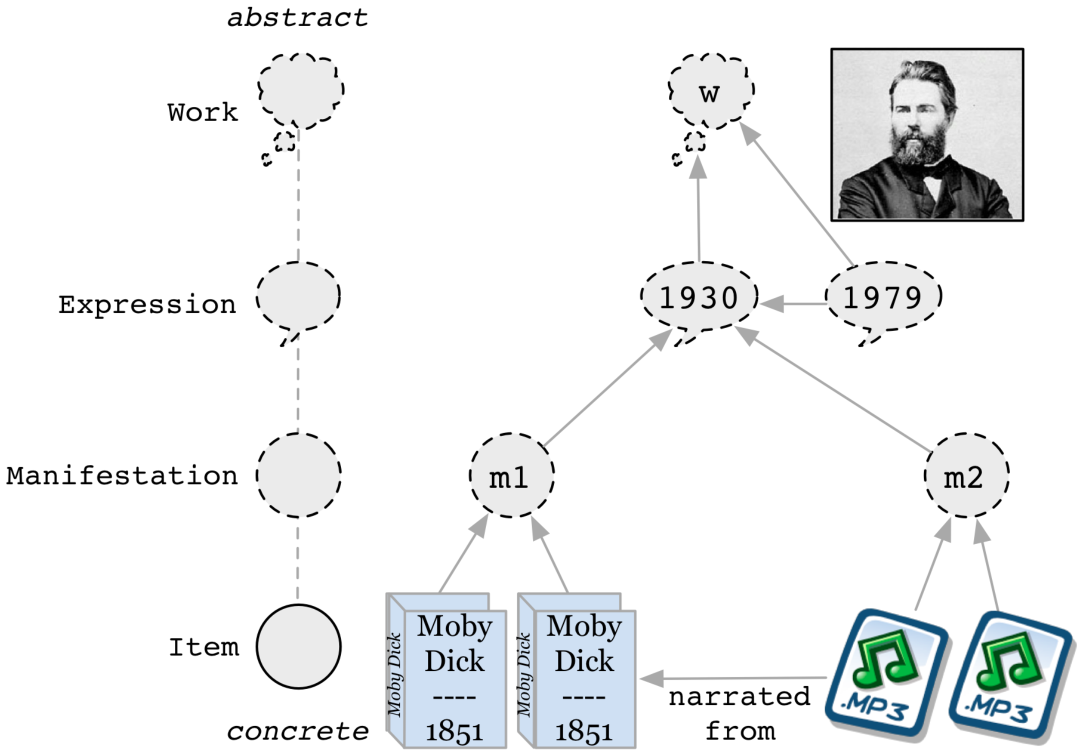
Functional Requirements for Bibliographic Records

Work: a "distinct intellectual or artistic creation."[1]

Expression: "the specific intellectual or artistic form that a work takes each time it is 'realized.'"[1]

Manifestation: "the physical embodiment of an expression of a work. As an entity, manifestation represents all the physical objects that bear the same characteristics, in respect to both intellectual content and physical form."[1]

Item: "a single exemplar of a manifestation. The entity defined as item is a concrete entity."[1]

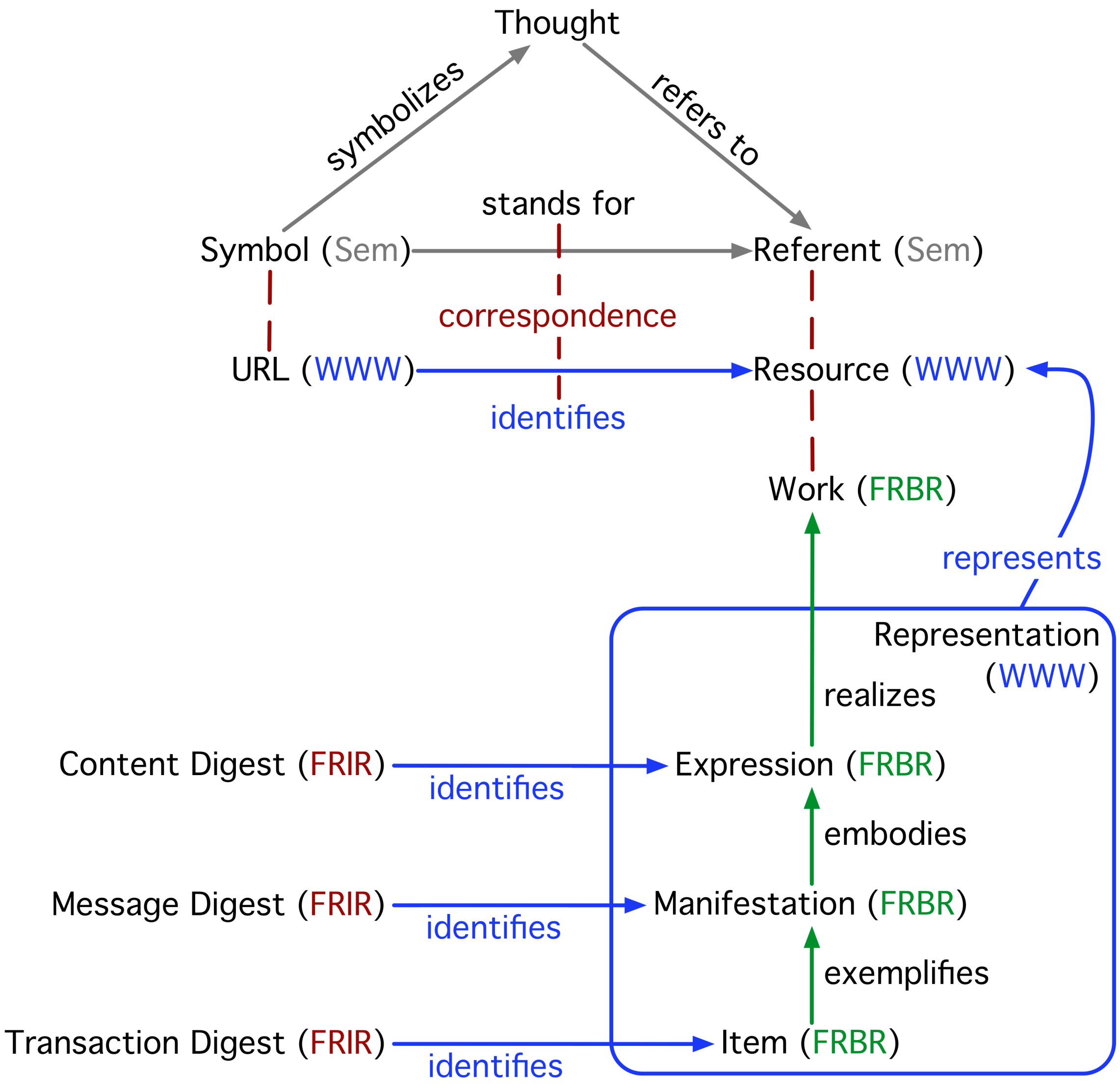


1. Functional Requirements for Bibliographic Records: http://archive.ifla.org/VII/s13/frbr/frbr1.htm#3.2

4 Steps to Data Sanity

Step 1: Distinguish resources from their representations.

Step 2: Avoid fragmentation by using FRBR and message/content digests.



4 Steps to Data Sanity

Step 1: Distinguish resources from their representations.

Step 2: Avoid fragmentation by using FRBR and message/content digests.

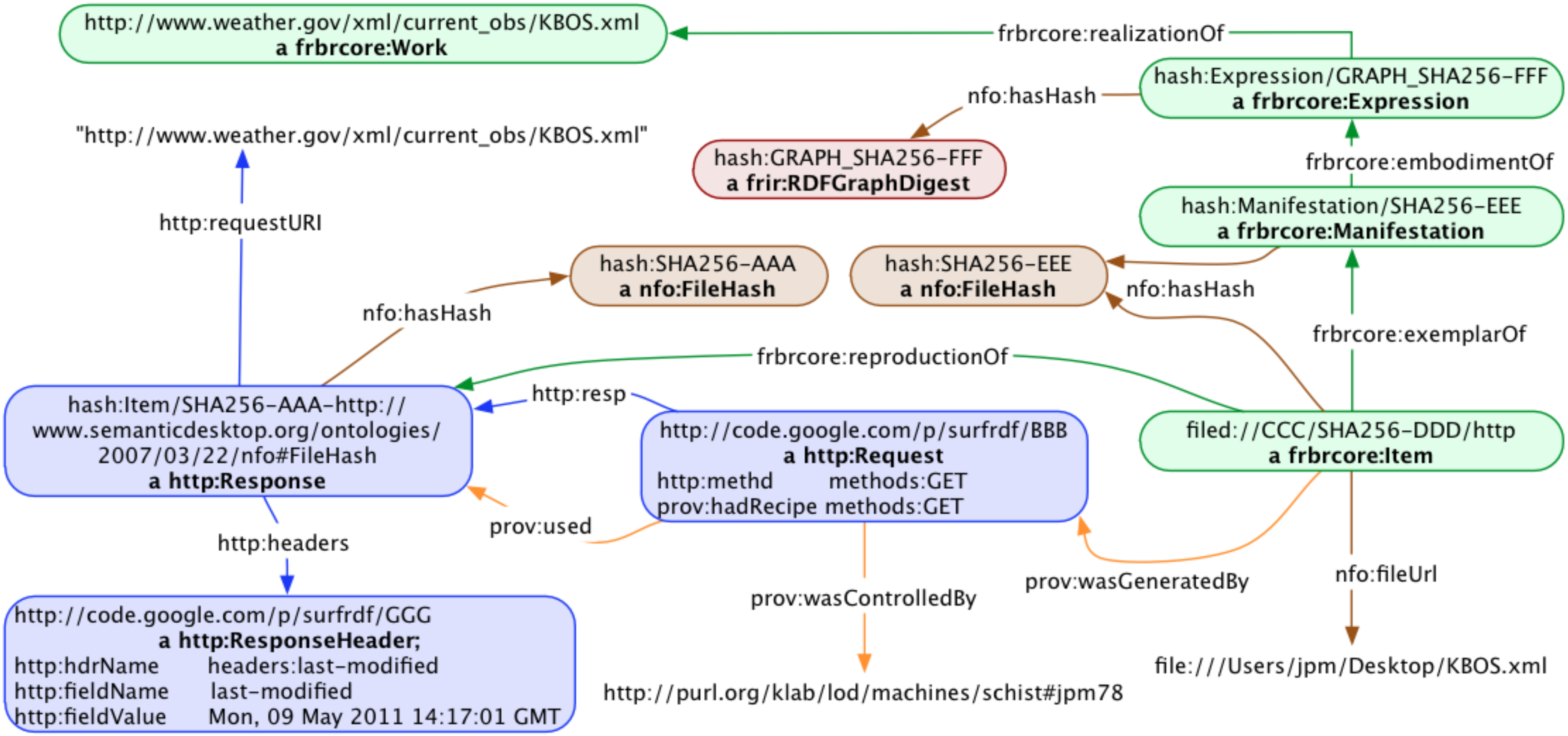
Step 3: Applying PROV

This is **Functional Requirements for Information Resources** (FRIR).

We found that 14 of 18 frbr:relatedEndeavour subproperties mapped to one or more PROV properties.

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

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



http://www.weather.gov/xml/current_obs/KBOS.xml
a frbrcore:Work

frbrcore:realizationOf

hash:Expression/GRAPH_SHA256-FFF
a frbrcore:Expression

nfo:hasHash

hash:GRAPH_SHA256-FFF
a frir:RDFGraphDigest

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

frbrcore:embodimentOf

hash:Manifestation/SHA256-EEE
a frbrcore:Manifestation

http:requestURI

hash:SHA256-AAA
a nfo:FileHash

hash:SHA256-EEE
a nfo:FileHash

nfo:hasHash

nfo:hasHash

frbrcore:exemplarOf

frbrcore:reproductionOf

http:resp

hash:Item/SHA256-AAA-http://
www.semanticdesktop.org/ontologies/
2007/03/22/nfo#FileHash
a http:Response

http://code.google.com/p/surfrdf/BBB
a http:Request
http:methd methods:GET
prov:hadRecipe methods:GET

filed://CCC/SHA256-DDD/http
a frbrcore:Item

http:headers

prov:used

prov:wasGeneratedBy

nfo:fileUrl

prov:wasControlledBy

http://code.google.com/p/surfrdf/GGG
a http:ResponseHeader;
http:hdrName headers:last-modified
http:fieldName last-modified
http:fieldValue Mon, 09 May 2011 14:17:01 GMT

file:///Users/jpm/Desktop/KBOS.xml

http://purl.org/klab/lod/machines/schist#jpm78

Explaining a HTTP transaction for the Weather.

Green: Information aspects of information accessed from the URL

Blue: The HTTP request and response

Brown: Content and message digest hash entities that identify the aspects.

4 Steps to Data Sanity

Step 1: Distinguish resources from their representations.

Step 2: Avoid fragmentation by using FRBR and message/content digests.

Step 3: Use PROV

Step 4: Use our tools.

pcurl.py: provenance-enabled curl.

fstack.py: build a FRBR stack for any file.

pcurl.py and fstack.py:

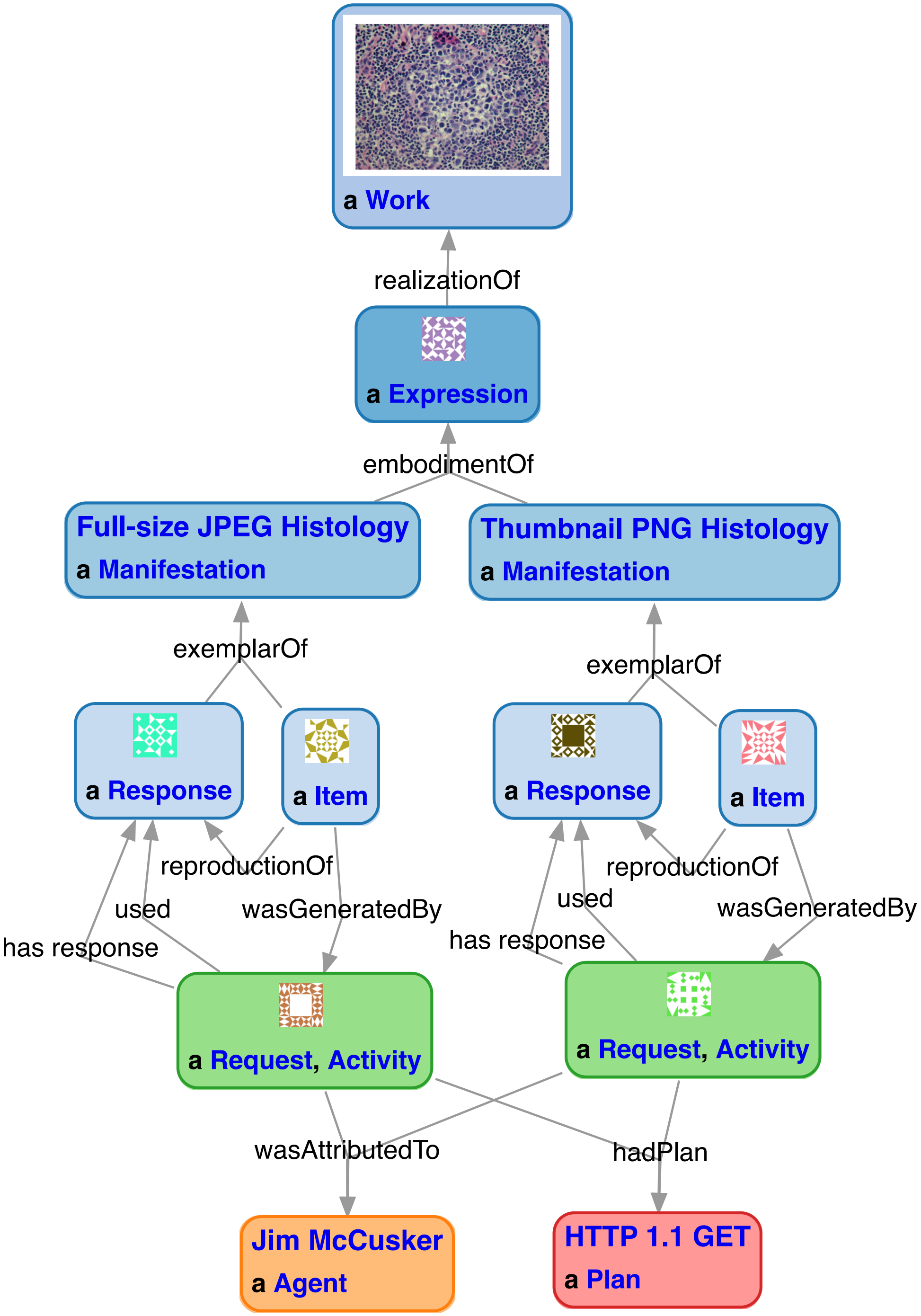
- Use cryptographic identities based on message digests and content digests.
- Have Java and Python implementations.
- Are open source.

https://github.com/timrdf/csv2rdf4lod-automation/wiki

Example: Files as different formats

Two files with the same image at different resolutions in different formats still have the same Expression and Work.

We can now link the high-resolution file used by the pathologist to the low resolution summary image the patient sees.



It also works for HTTP POST

(See the paper for details)

4 Steps to Data Sanity

Step 1: Distinguish resources from their representations.

Step 2: Avoid fragmentation by using FRBR and message/content digests.

Step 3: Use PROV

Step 4: Use our tools.

Conclusions

- Know what is being identified.
- Separate out content, data, and work.
- Relate them to each other using FRBR/FRIR
- Walk that abstraction hierarchy to talk about things in the detail you need.

Avoid confusion and Avoid fragmentation!

Questions?

(Thank you)