Naming on the Web: What scholars should want, and what they can have

Henry S. Thompson
School of Informatics
University of Edinburgh

OAI 8, Geneva
19 Jun 2013

Table of Contents

1. Acknowledgements

Almost everything I have to say today has been developed over a number of years of collaboration with Jonathan Rees

- In particular, the distinction between binding and resolution is his
- But today's presentation, and in particular any faults and flaws, are my responsibility

I've been fortunate to serve on the W3C's Technical Architecture Group with Tim Berners-Lee, Dan Connolly and Larry Masinter, among others

- None of them will agree with everything I have to say today about Web Architecture
- But I owe a lot of it to them none-the-less

2. Introduction

The obligatory joke:

- It's alleged that when Zhou Enlai was asked what he thought of the French Revolution he replied
    - "It's too early to tell"
- The scholarly community could be forgiven for wishing they could say the same about the Web
- For better or worse, we don't have that luxury

The incentives for moving scholarship, *all* scholarship, onto the Web are enormous

Perhaps more to the point

- The penalties for failing to do so are rapidly increasing
- For the current generation of students, it is increasingly true that
    - "If it's not on the Web, it doesn't exist"

3. The reliable reference problem

Scholarship depends on reliable references

When I read (in [Horn 2007])

> [T]his theoretical construct has become controversial of late (Bach 1999 consigns it to a chimerical status, while Potts 2005 attempts a partial rehabilitation, as we shall see below)

I can depend on scholarly convention and real libraries working together if I want to assess the accuracy of Horn's summary

- My university library may not have a copy of [Potts 2005] (*The Logic of Conventional Implicatures*, OUP)
- But they will surely have Bach 1999 ("The myth of conventional implicature". *Linguistics and Philosophy* 22: 327-66

And there will be no doubt whatsover when I get my hands on them that they are what I was looking for.

Can the Web match this?

- Will libraries be able to match it for much longer?

4. How the Web fails us

If we do get a copy of [Potts 2005], we find the following (on p. 6):

> **1** Judith Thurman, 'Doing it in the road'. *New Yorker*, June 10 2002 (p. 86)
>
> **2** http://www.hamline.edu/apakabar/basisdata/1997/03/21/0066.html
>
> **3** http://jjdavis.net/blog/arc20010325.html

Neither of those links works today

- The first is at least still handled by Hamline University
  - but they don't have that page anymore
- The second is also, after several redirects, a 404, *but*
  - Further investigation shows that the `jjdavis.net` domain is now owned by someone in Japan
  - Back in 2011 JJ Davis started redirecting `jjdavis.net` to a new `.com` domain
  - And then in 2012 he didn't renew his lease (!) on the `jjdavis.net` domain
  - And it was taken up by someone else

The Web has failed Potts

- And it continues to fail scholars every day

5. Old news

None of this is surprising to any of you

- The vulnerabiity of the `http:` URI resolution process
  - To website abandonment or reorganisation
  - To more or less voluntary loss of domain control
- Has been recognised and criticised for many years now

Around the beginning of this century this led to calls for 'persistent' identifiers on the Web

- And to a number of proposals which aimed to supply them
  - Some implemented
  - Some not so much

But don't worry, I'm not going to drag you through yet another survey :-)

6. Preservation of content

Something else I'm not going to talk about!

- The Information Science/Digital Library folks have got this covered

That's not to say there aren't issues here

- Just that they are of a different order from the naming issues I'm focussing on today

7. The good news

So much for the past

- Actually, we'll come back to it again in a little while

The good news is that there is progress to report

I want to try to open up some new perspectives

- On how we think about naming
- On how the Web changes naming
    - And how it doesn't
- On how the received wisdom wrt Web Architecture may need to change

8. What do we know about names?

There is a lot of very good historical/linguistic/philosophical literature on names

- Perhaps too *much* literature

However, as far as I've been able to see

- Almost all of this work is about what I'll call **unmanaged** names
    - Names in ordinary language
    - Names governed in no explicit way

But there are huge numbers of **managed** names as well

- That is, names which *are* governed in some way

It seems unlikely that something so important is so little studied

- But I've found very little discussion of what I'm now calling **managed** naming systems

9. Managed naming systems

Once you start looking, you realise these are everywhere:

- Some are privately managed, typically within narrow scopes:
  - Product names (cars, operating systems, perfumes, . . .)
  - Sports franchises (football teams (some leagues), baseball teams, ...)
  - Variable names in programs
- Some are more-or-less publicly managed:
  - Asteroids and comets (IAU Committee for Small-Body Nomenclature)
  - Airport codes (FAA?)
  - Köchel numbers (?)
  - Generic drug names (?)
  - Internet domain names (ICANN/IANA)
  - Laws and regulations (governments etc.)
  - RFCs (IETF)
  - Binomial Nomenclature (ICN, ICZN)
- And there are hybrids:
  - Street names (local gov't plus local populace)
  - Country codes (ISO-3166 plus local language)

10. What is managed?

We can identify at least three aspects of a naming system which might be managed:

syntax
: Constraints on the *form* of names:
  - Country and language codes
  - Variable names
  - Domain names

binding
: How a name gets (and maybe loses/changes) its meaning
  - Legislative process
  - More-or-less explicit constitution of governing abodies
  - Individual fiat

resolution
: Given a name, how do you find out what it means?
  - Official publications, a.k.a registries
  - More-or-less well-defined search procedures
  - Online lookup (`bind`, search engines)

11. What *isn't* managed?

Once you start thinking in these terms

- Not only do you start seeing the commonalities
- But also the divergences and gaps

So, what about the Web?

Domain names are pretty straightforward:

- Syntax is specified by the IETF's RFCs 1034 and 1035
- Binding (and unbinding) is managed by IANA
  - under guidelines from ICANN
  - operated by registrars under contract
- Resolution is governed by IETF RFCs and implemented by `bind` and friends

When we turn to URIs we're in for a bit of a shock

12. URIs as managed names

The syntax of URIs is managed in a federated fashion:

- The overall syntax is governed by IETF RFCs (3986 and 3987)
- These devolve additional authority to scheme-specific RFCs
  - 2616 for `http:`
  - 3261 for `sip:`
  - 6068 for `mailto:`
  - 2141 for `urn:`
- The individual scheme RFCs may impose additional syntactic constraints

But the RFCs actually say surprisingly little about binding or resolution

- Although you may have thought they must have

13. URIs and binding

After all, URI stands for Uniform Resource *Identifier*

- So you might expect some discussion of identification in the goverrning standards
- And indeed there is some, in 3986:
  - > An identifier embodies the information required to distinguish what is being identified from all other things within its scope of identification. Our use of the terms "identify" and "identifying" refer to this purpose of distinguishing one resource from all other resources, regardless of how that purpose is accomplished (e.g.,

> by name, address, or context). These terms should not be mistaken as an assumption that an identifier defines or embodies the identity of what is referenced, though that may be the case for some identifiers.

- So All URIs do, officially, is identify in a *very* narrow sense
  - That is, guarantee (nearly) definitive (positive) answers to the question "Do these two URIs identify the same thing?

14. The official story for `http:` URIs

In the `http:` RFC (revised draft version), we don't find much more:

> The HTTP origin server is identified by the [domain name]

> The remainder of the URI . . . serves as an identifier for a potential resource within that origin server's name space

> HTTP does not limit the nature of a resource; it merely defines an interface that might be used to interact with resources. Each resource is identified by a Uniform Resource Identifier (URI)

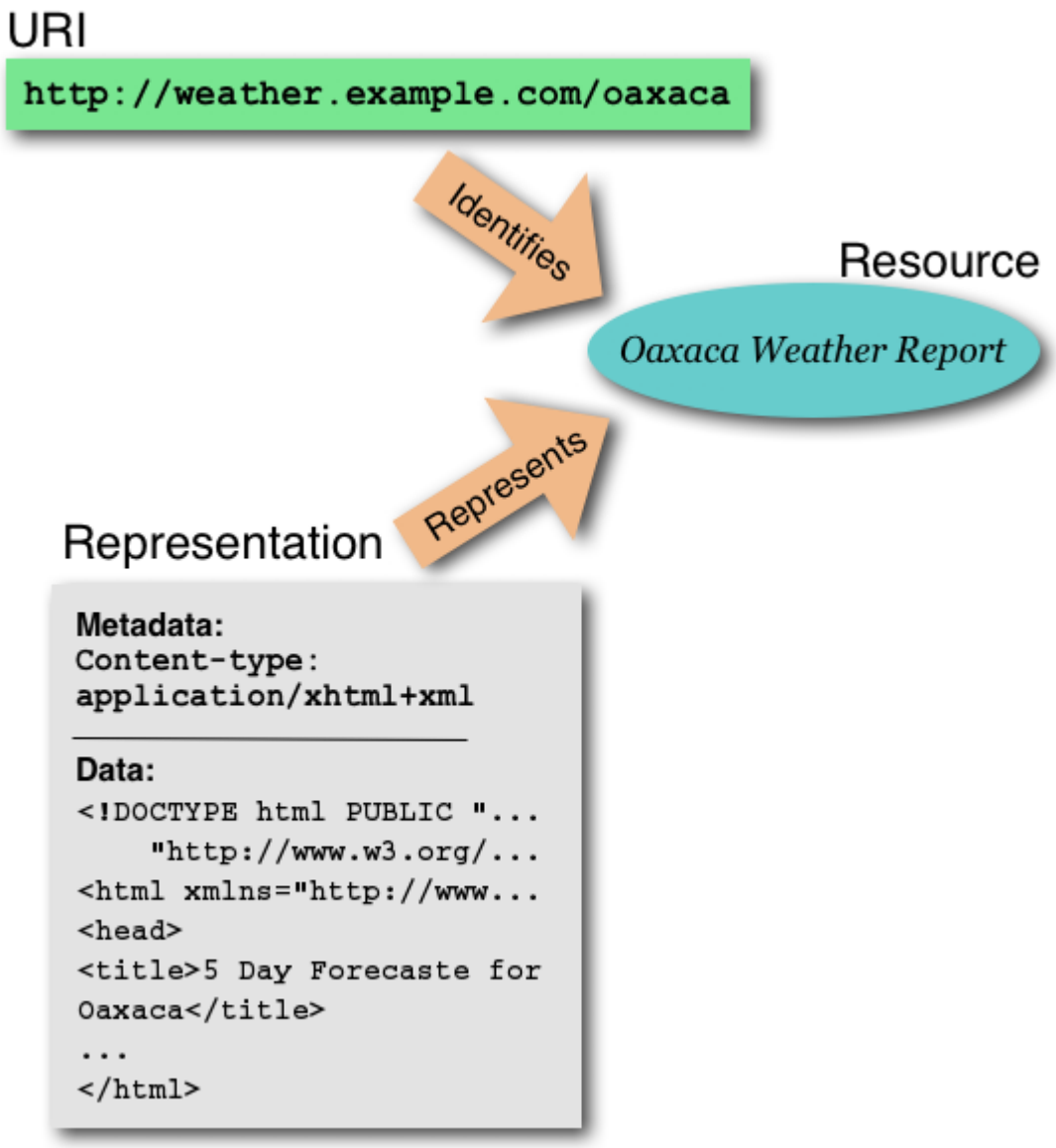
> HTTP provides a uniform interface for interacting with a resource . . . via the manipulation and transfer of representations

> [A] 'representation' is information that is intended to reflect a past, current, or desired state of a given resource

At best this gives us some *indirect* sense of what the intended story is with respect to (`http:`) URIs and their meaning

15. A picture may help

This illustration of the above is taken from the W3C TAG's *Architecture of the World Wide Web*



16. URIs and resolution

Resolution, remember, is the process whereby given a name you can find out what it means

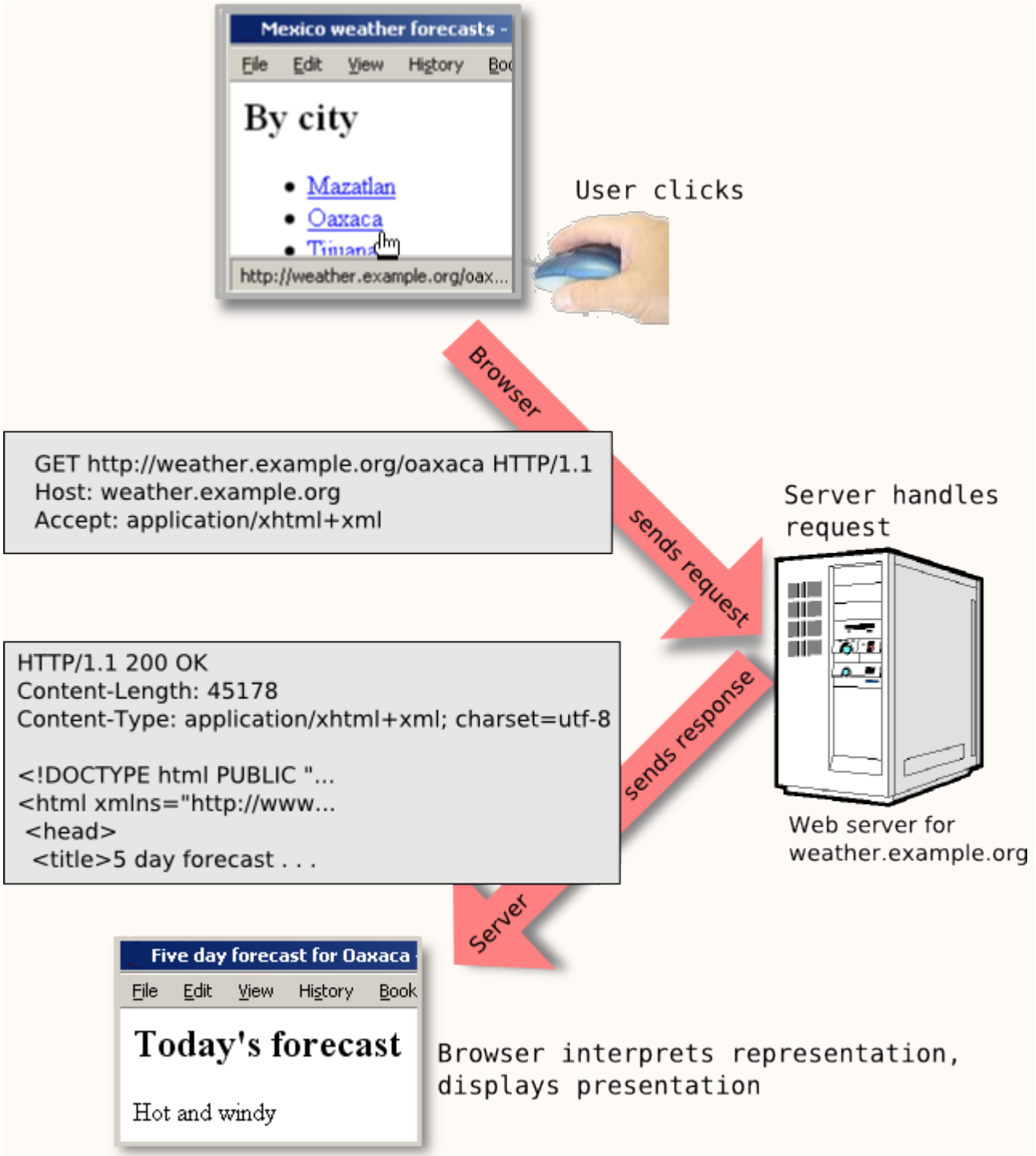
Here we're on safe ground, at least for `http:` URIs

- Right?
- Well. . .
- Again, perhaps surprisingly, no

What we saw above says

1. URIs identify resources
2. HTTP retrieves representations

Here's another picture



17. The resource/representation gap

Where in the specs is it required that what is retrieved bears any relationship to what is identified?

- I quoted that above, right?
  - > [A] 'representation' is information that is intended to reflect a past, current, or desired state of a given resource

But wait

- Intended by whom?
- Reflect for whom?
- How could we tell if this statement was or was not true in any given case?

18. The Social Contract

The problems we've encountered are not surprising

- You can't expect technical specifications to address issues of meaning

Berners-Lee describes the Web as a "shared information space"

Shared spaces need social norms to work well (or at all)

In the case of the Web, those social norms have not been well articulated

Here's a candidate norm, which scholars will recognise

> Owners of URIs described in their bibliographies as identifying their own published papers should arrange that, in response to retrieval requests for those URI, representations are served which when rendered correspond as closely as possible to the corresponding paper as published

19. Digression: A system which *does* work

Way back near the beginning I mentioned binomial nomenclature

- The Linnaean system for naming organisms

It carefully distinguishes binding from resolution

- With a clear story about each

Originally governed by an informal social process

- Institutionalised over subsequent *centuries*

20. A lesson about resolution from Linnaeus

The success of the binary nomenclature system for resolution depends on a very simple strategy

- It even has a contemporary acronym
- Most people don't realise Linnaeus invented it
- LOCKSS

That appears to be a problem for Web Architecture

- Which asserts that there *should* only be one name (URI) per resource

At worst, in a future world where we have an effective universal LOCKSS-based fallback resolution system for scholarly references

- This means overriding the DNS system in cases such as the `jjdavis.net` one we started out with

21. A lesson about binding from ...

Many of the most successful managed naming systems share an approach to binding

- IETF RFCs
- IAU "small bodies"

Binding is achieved by a publicly-visible multi-step process

- Governed by a constitution created by (representatives of) the consumers of the names
- Operated by representataives of that same constituency

22. Conclusions

Persistence doesn't have *technical* solutions

- `http:` URIs *can* be the *technical* basis for Web naming system adequate to scholarly needs

Good managed naming systems are backed up by robust social contracts

Distinguishing binding from resolution is a key step in designing a naming system which maximises persistence

- Quoting Jonathan Rees again
    - > Binding and ownership are different. Ownership is in principle the right to change binding, but in a persistent system, you won't be changing bindings.

There's a *lot* of work still to be done:

- To articulate the Social Contract
    - So far I've identified at least *six* actors who have to sign up in various ways, and I'm sure there are more
- To agree a socially-moderated non-repudiatable binding mechanism (SMNRBM) for scholarly reference URIs
    - Starting with a SMNRBM for robust domain names
- To provide a companion fallback resolution process