Building Metadata Aggregation Services for Resource Discovery

Paul Walk

p.walk@ukoln.ac.uk

UKOLN is supported by:

JISC

www.ukoln.ac.uk

A centre of expertise in digital information management

UNIVERSITY OF BATH

aggregating

metadata

why aggregate metadata?

- to address systems/network latency - a **cache**
  - supporting resource-discovery
- for 'Web Scale *concentration*'
  - 'gaming' Google - raising 'visibility' of content
    - network effects **if** user facing services also developed
- to showcase resources
- to create middleman business opportunities
- as *infrastructure* to support 3rd-party services
- as an approach to preservation

patterns

- harvest from network, aggregate and re-expose
  - discovery.ac.uk, Europeana, RepUK
- collect from offline sources and make available in aggregate on the network
  - Collections Trust (UK)
- harvest without re-exposing, build services on top of aggregation
  - Google et.al.
- expose as a ‘data dump’, or expose through an API

the big question facing data providers:

do you want to provide a data service, or just data?

current work in the UK

discovery A metadata ecology for UK education & research

Home | Vision & Approach | The Business Case | Developer Zone |

Welcome to Discovery

Towards a thriving metadata ecosystem

In 2010, the JISC and RLUK Resource Discovery Taskforce (RDTF) worked with stakeholders from the libraries, archives and museums to set out a Vision for making the most of our resources by effectively positioning their metadata for discovery and reuse within the global information ecosystem.

Find out about Discovery

More on Discovery, the Resource Discovery Taskforce Vision, and the approach we're taking.

Vision & Approach

- a metadata 'ecosystem'
  - aggregation is a major component
  - preparing resources for aggregation
  - **http://www.discovery.ac.uk**

- support innovation
- develop some ‘business intelligence’
- develop infrastructure component for services

issues with aggregation

distribution

- state management is a challenge! (deletions, changes)
- aggregation of aggregations is consequently non-trivial
  - e.g. federated models
- linking?
  - should records in an aggregation ever be the target of a link? Or, should such links point to the source?
  - can/should we make aggregations into Google-friendly targets?
  - if we succeed with SEO, are we undermining source repositories?
- 'attribution stacking' (http://sciencecommons.org/projects/publishing/open-access-data-protocol/)

openness and usability

- ‘open’ in danger of becoming synonymous with ‘permissively licensed’
- can be both ‘open’ but very difficult to use
  - needs periodic review - right now SPARQL is barrier to wide adoption
  - remember all those SOAP interfaces....
  - a well supported API **might** be more open than a completely freely available dump of gigabytes (or more) of data in the sense that it might allow open engagement from more people
- we need a richer understanding of openness

in other words...

be **open, usefully**

character encodings....

- huge number of XML records from UK IRs are invalid due to character encoding issues....

- there is a special place in hell for developers who ignore character encodings...

http://www.flickr.com/photos/10661825@N07/

a distributed system is one in which the failure of a computer you didn't even know existed can render your own computer unusable

Leslie Lamport

are we creating a new version of this with data....?

shifting landscape

- Google was previously seen as in opposition to a rich metadata approach...
  - *recall* versus *precision*
  - Google's abandonment of OAI-PMH
- but now...
  - Google, Microsoft & Yahoo committed to improving precision through harvesting of Microdata
  - schema.org and others bridging this divide
- so, is there still a need for other 'concentrations' or can we rely on the global search engines?

good practice

licensing!

- use explicit licenses
- this means *requiring* explicit licenses from sources
- if at all possible work with extremely open licenses such as CC0
- in data aggregation, especially when using a Linked Data approach, 'share alike' might be *easier* than 'attribution'

“build for normal users,
developers and
machines”

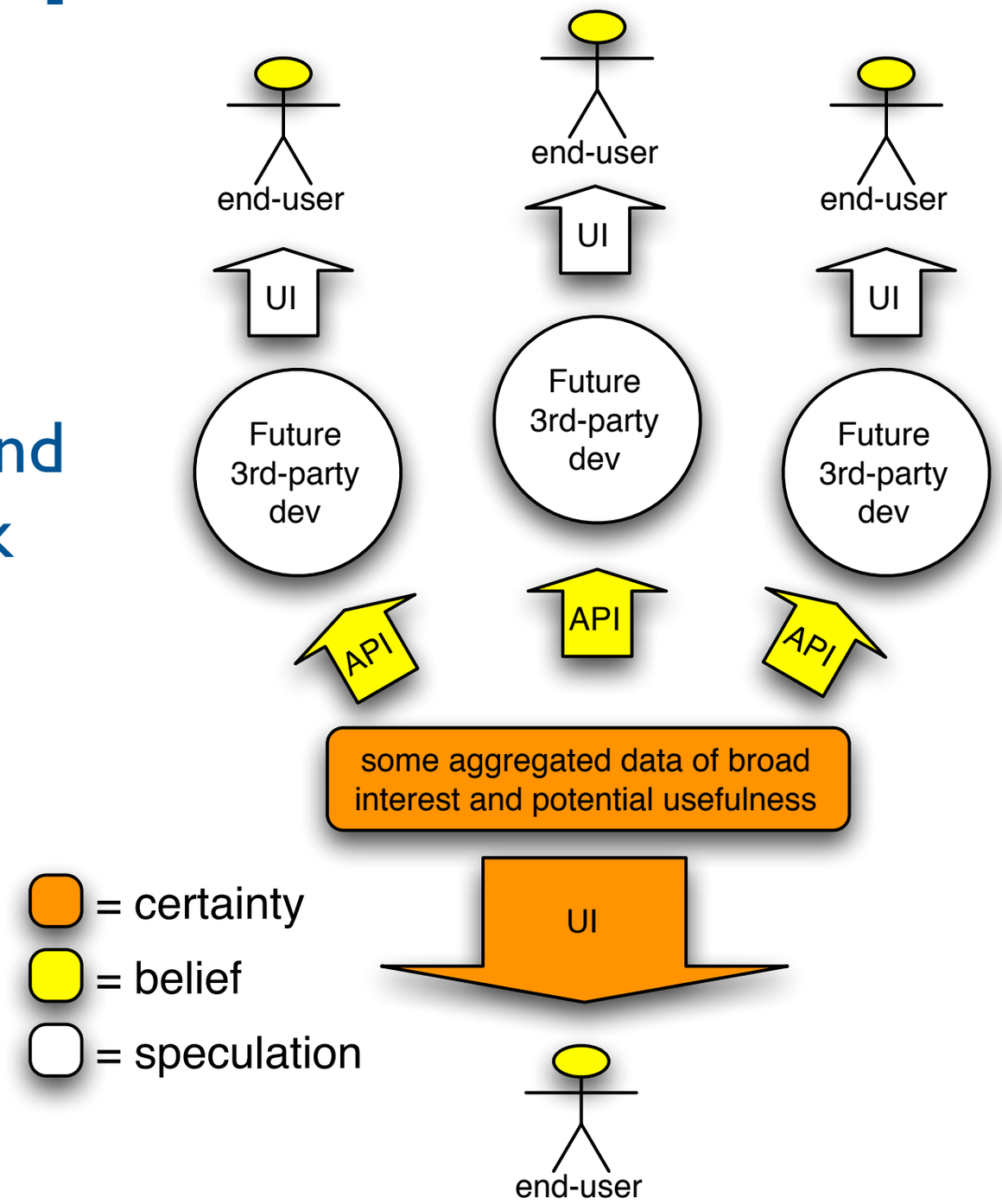
Tom Coates
http://www.plasticbag.org/archives/2006/02/my_future_of_web_apps_slides/

developer-friendly formats

- XML has a lot going for it:
  - very well supported with **tools**, libraries etc.
  - well understood & often fits the info models we're used to
- but it has some issues:
  - validation is a pain and is very often ignored
  - it's verbose - it takes up a lot of bandwidth
- JSON has gained rapid adoption
  - less verbose - good for simple client-side manipulation
    - curl -D - -L -H "Accept: application/rdf+xml" "http://dx.doi.org/10.1126/science.1157784"
    - curl -D - -L -H "Accept: application/**json**" "http://dx.doi.org/10.1126/science.1157784"

service (anti)patterns

- design your API to be developer-friendly
- be aware of what works, and of what **appears** to work but actually might not...
- share this understanding



Paul Walk, An infrastructure service anti-pattern
http://blog.paulwalk.net/2009/12/07/an-infrastructure-service-anti-pattern/

expect & enable users to **filter** - give them *feeds* (RSS/ Atom)

http://www.flickr.com/photos/httpwwwflickrcompeoplenadar/3349883/ (CC BY-NC-ND 2.0)

workshop tomorrow!

tomorrow at 16:15

- (Thursday, 23rd June, 16:15-18:15)
- short presentations from UKOLN on LOCAH and RepUK, and from Edina on aggregating services
- open discussion on the way forward for metadata aggregation, addressing questions such as:
  - is Linked Data the future for metadata aggregation services?
  - do initiatives like Microdata & schema.org reduce the need for our investment in metadata aggregation services?
  - does usability matter as much as 'openness'?
- please join us, and feel free to bring your own questions & issues to discuss

summing up in a sentence....

we should use aggregation
*[applying a **tool**]*

to balance the creation of opportunity
*[building **infrastructure**]*

with the solving of problems
*[developing & providing **services**]*

thank you