Resilient Linked Data

Dave Reynolds, Epimorphics Ltd
@der42

Outline

- What is Linked Data?
- Dependency problem
- Approaches:
  - coalesce the graph
  - link sets and partitioning
  - URI architecture
  - governance and registries

Linked data ...

publishing data on the web ...

... to enable integration, linking and reuse across silos

Linked data

Apply the principles of the web to publication of data

The web:

- is a global network of pages
- each identified by a URL
- fetching a URL gives a document
- pages connected by links
- open, anyone can say anything about anything else

Linked data

Apply the principles to the web to publication of data

The linked data web:

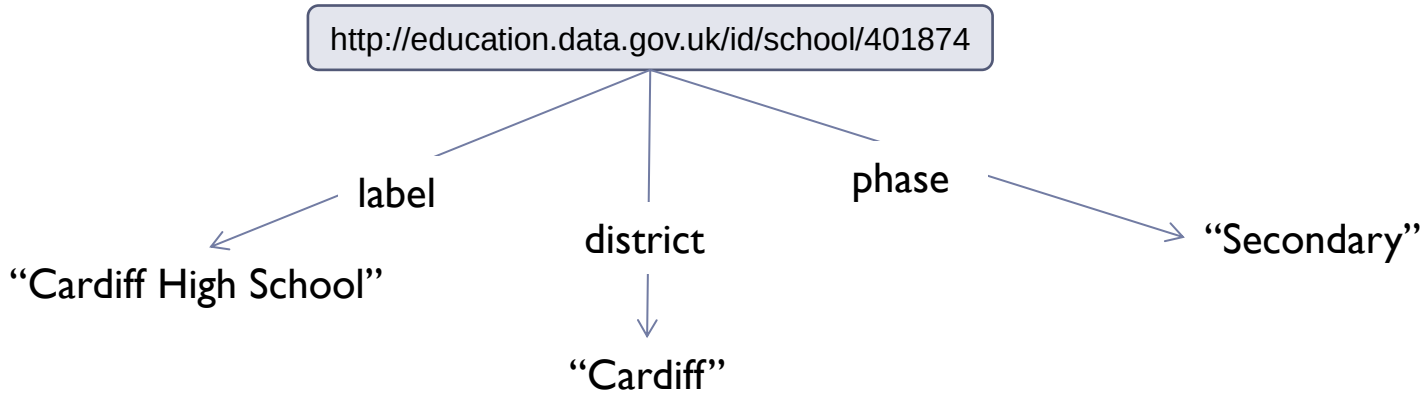
- is a global network of *things*
- each identified by a URI
- fetching a URI gives a set of statements
- *things* connected by typed links
- open, anyone can say anything about anything else

Linked data is “data you can click on”

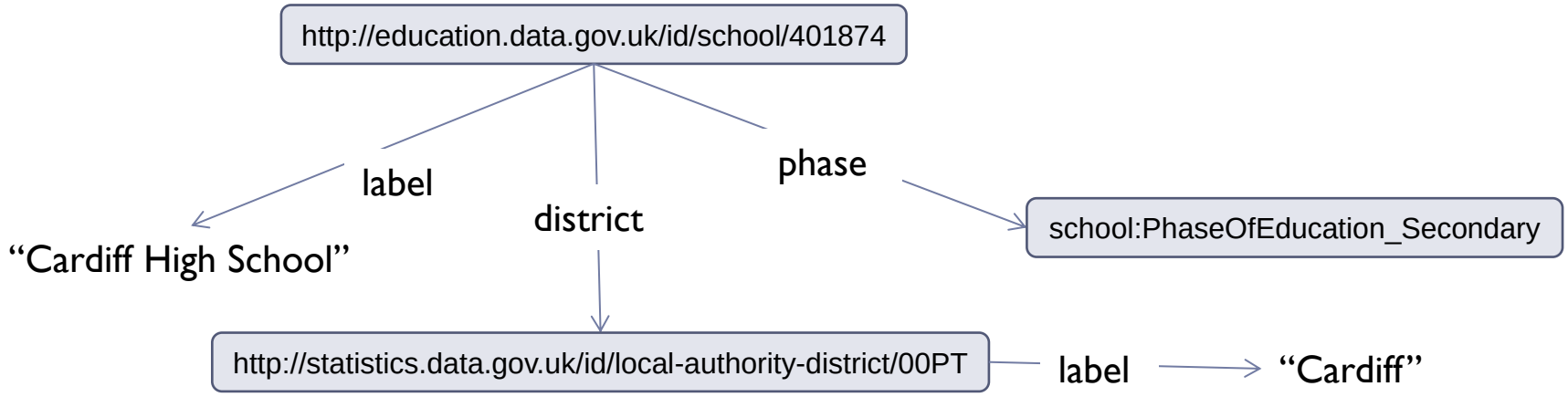
Example schools information

http://education.data.gov.uk/id/school/401874

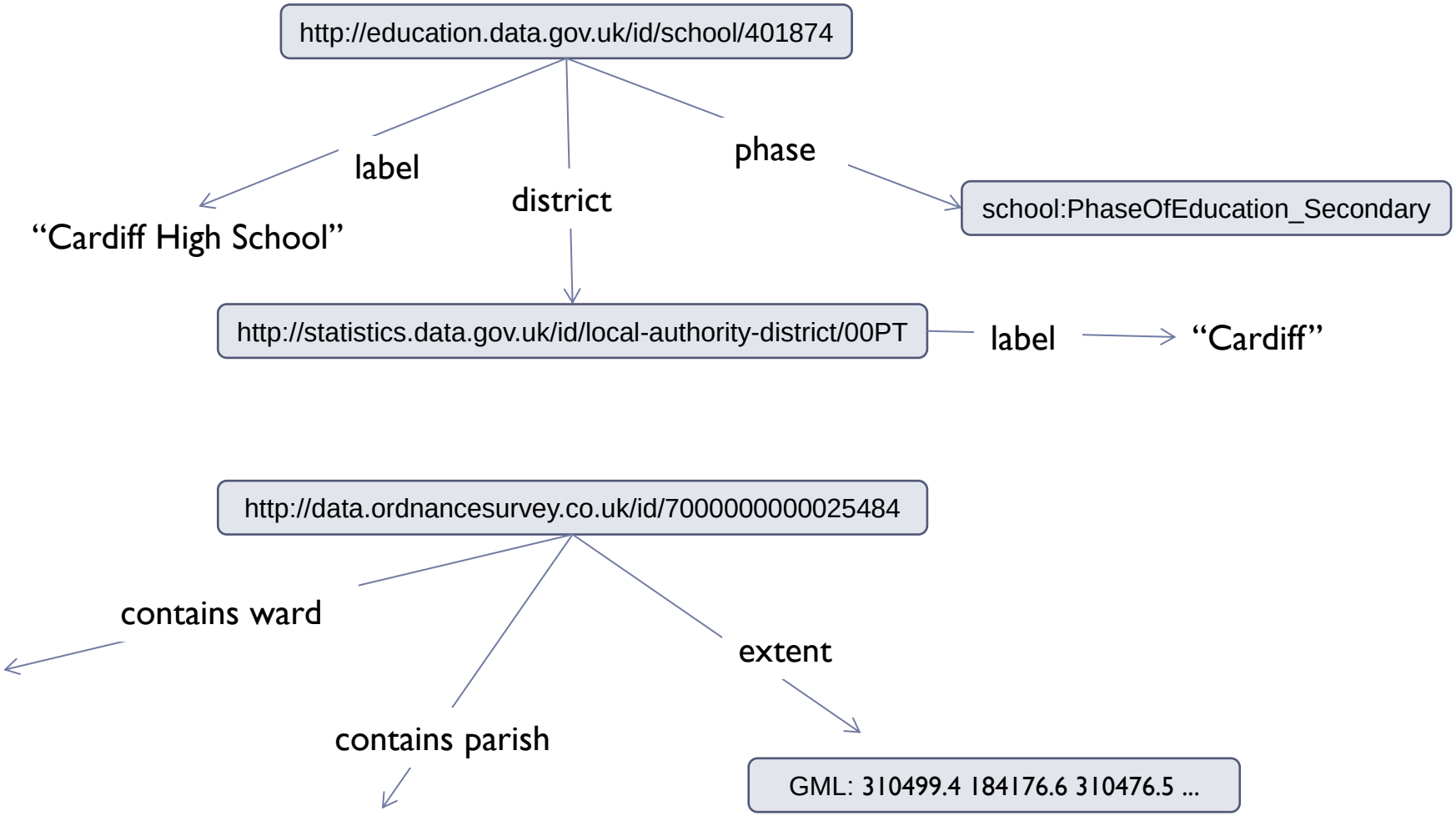
Example schools information



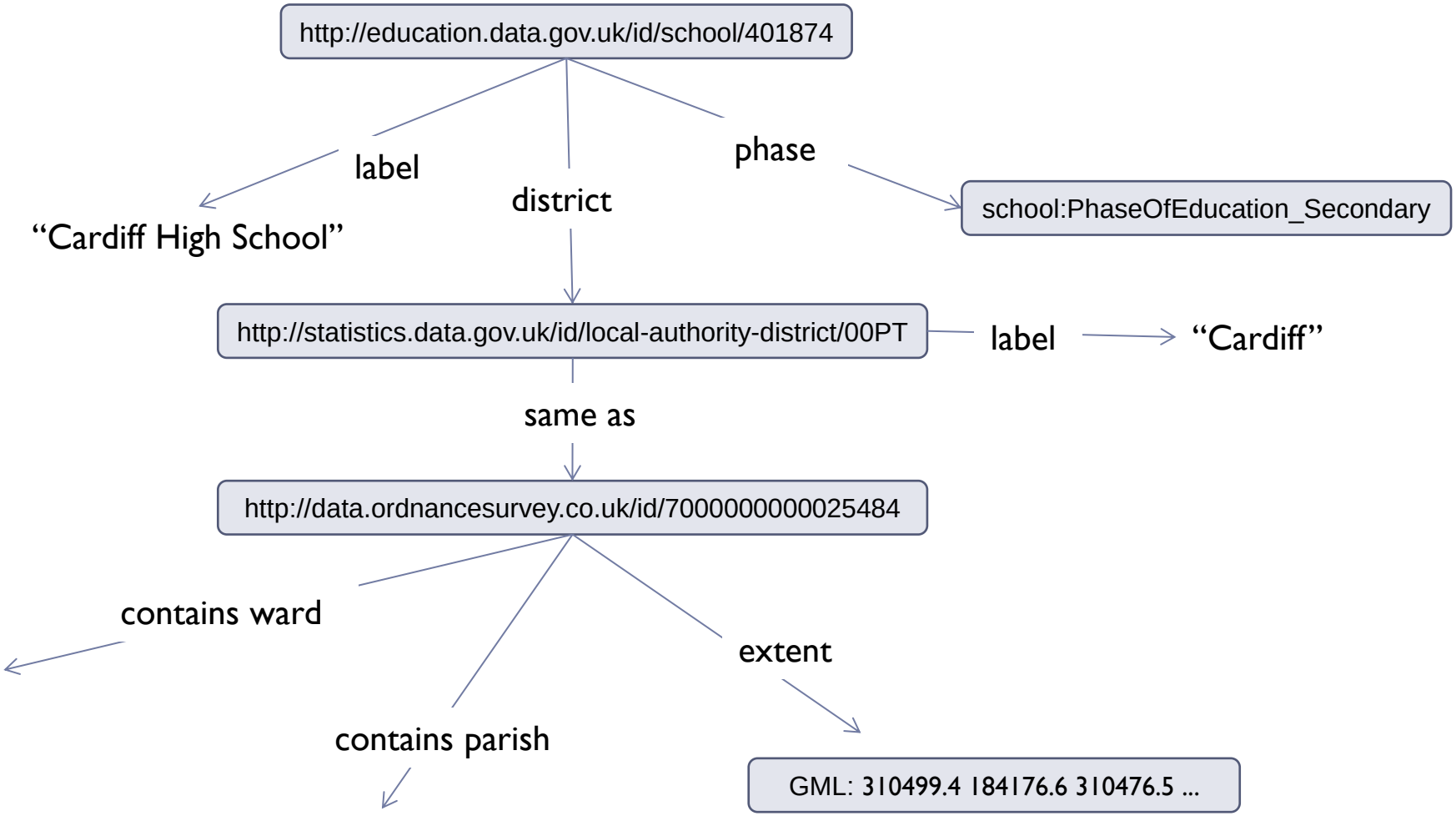
Example schools information



Example schools information



Example schools information



Linked data principles

- Use URIs as names for things
- Use HTTP URIs so that people can look up those names
- When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
- Include links to other URIs, so that they can discover more things

Pattern of application of semantic web stack

Why?

- self-describing
- everything is addressable
  - annotation, put in context
- APIs not just data dumps
- particularly good for graph structured data
  - but OK for regular data too
- open standards
- integration and comparison
  - global identifiers, vocabulary reuse
  - connected web rather than silos
  - decentralized

Who (illustrative, not exhaustive)

Environment Agency

Ordnance Survey

Companies House

Land Registry

legislation.gov.uk

The National Archives

Communities and Local Government

CabinetOffice

nature publishing group npg

BBC

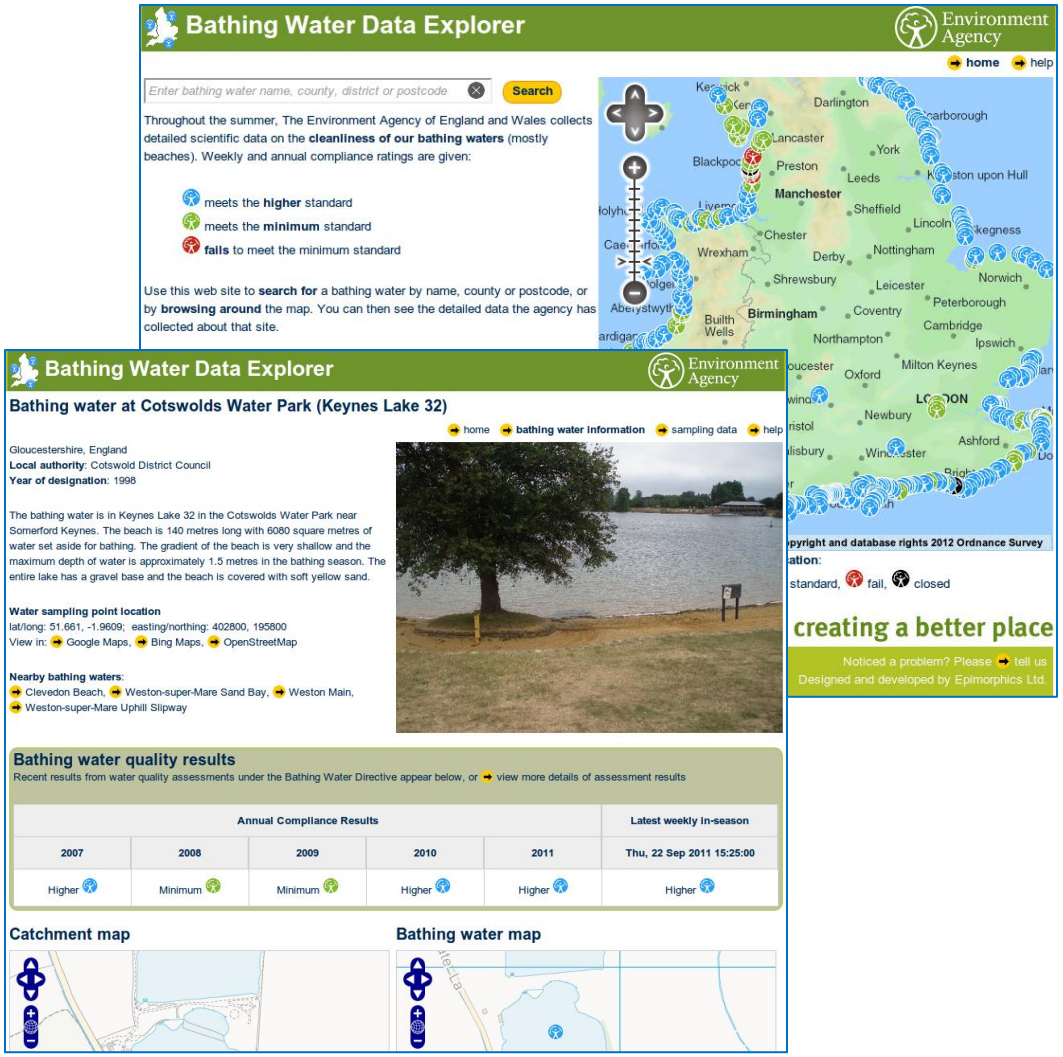
PEARSON

LIBRARY OF CONGRESS

BRITISH LIBRARY

Environment Agency

- monitoring of bathing water quality
- static pilot
- live pilot
  - historic annual assessments
  - weekly assessments
- operational system
  - additional data feeds
  - live update
  - integrated API
  - data explorer



Epimorphics Ltd
Linked data solutions

The problem for today

"decentralized, distributed, graph"

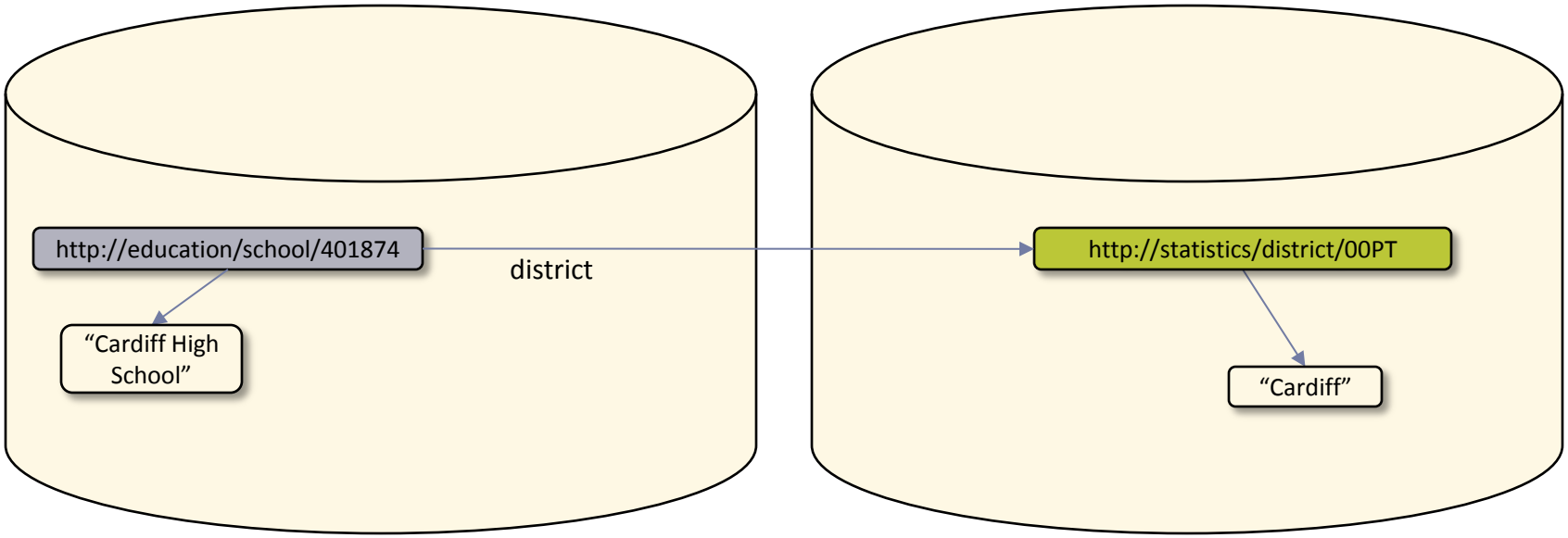
=>

"internet of dependencies"?

- silos have benefits
  - self contained – verify, dump
  - controllable governance – no external changes
- web of data sounds fragile
  - linked resources stops resolving - 404
  - query endpoints disappear, timeout – 408, 500
  - data at the end changes

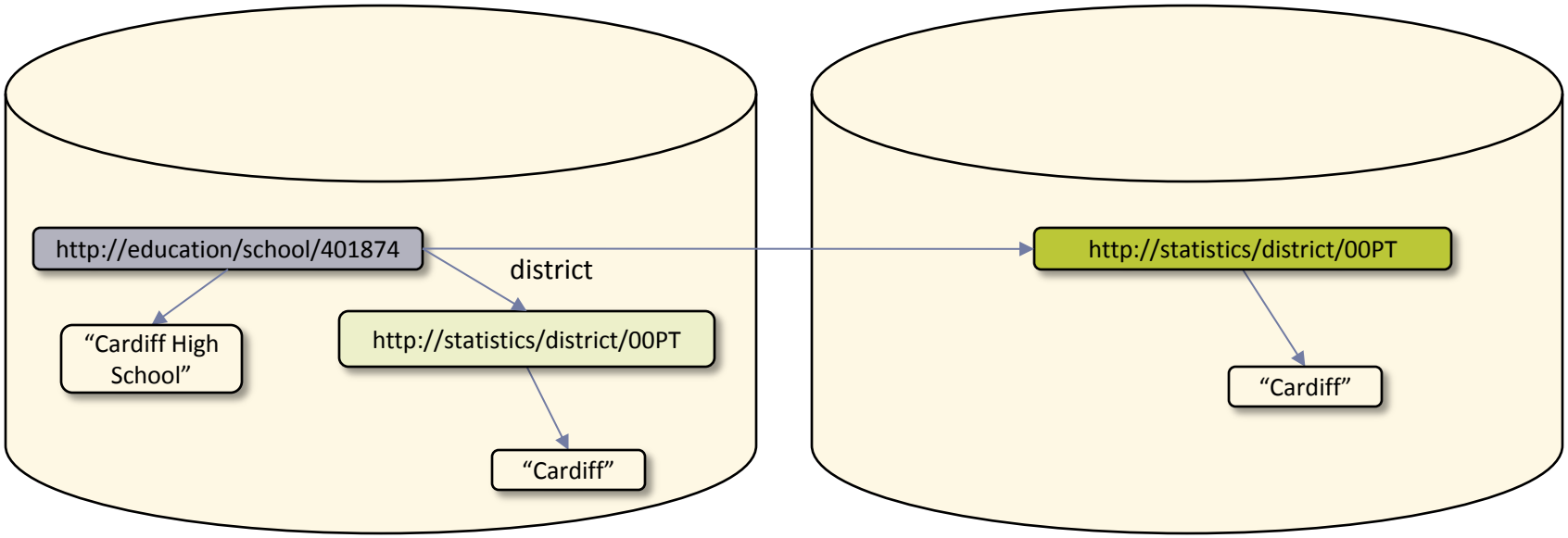
Approach 1: coalesce the graph

- value of using http URIs is that you *can* dereference them
- doesn't mean you have to



Approach 1: coalesce the graph

- RDF store ("triple" store) can hold arbitrary graphs
- so can include copy of the key information you rely on

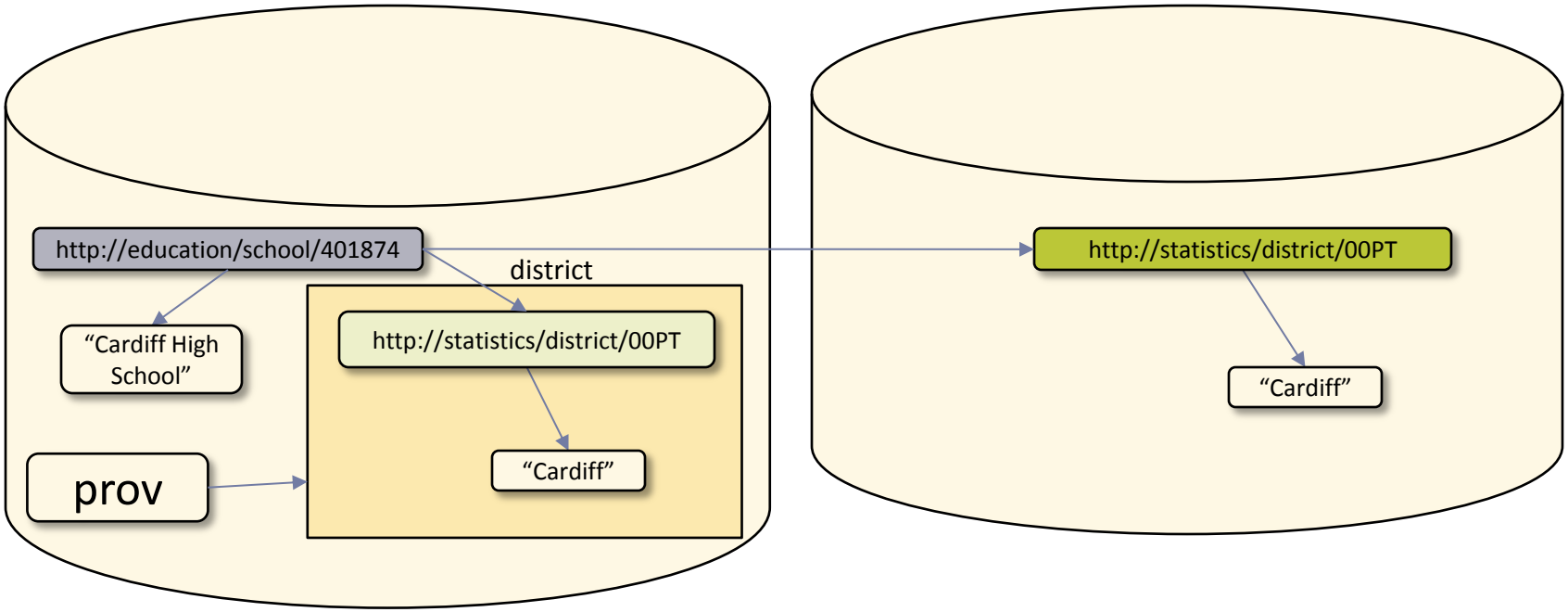


Approach 1: coalesce the graph

- higher performance access
  - avoid distributed queries
- stability
  - queries work even if the other sources are off-line
  - makes explicit what you thought the URI referred to
- the URI still points to the authoritative source
- comes at cost
  - have to maintain up to date ("cool URIs", but ...)
  - storage cost (how many steps away do you go?)
  - if clients rely on this view may miss out on third party links
- commonly used for:
  - vocabulary terms
  - key reference data, sufficient to do useful local queries

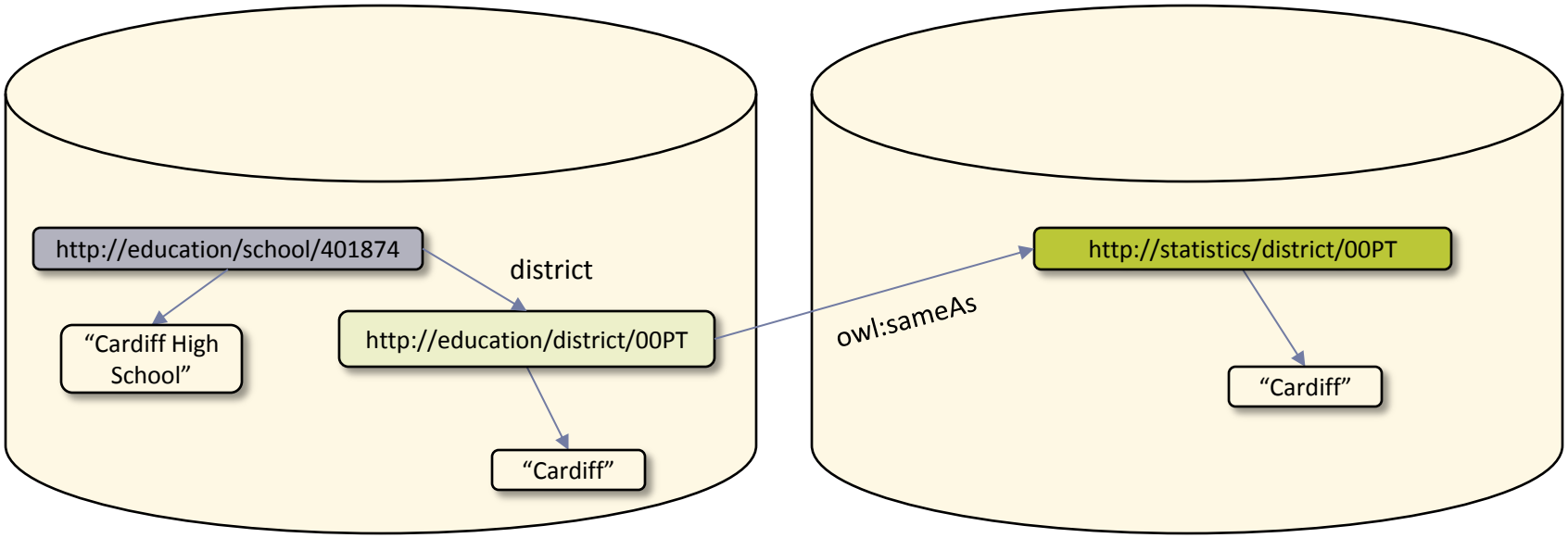
Approach 1b: named graphs

- can partition store in graphs also identified by URIs
- annotate graph URI with statements such as provenance PROV-O ontology now a recommendation



Approach 2: link sets

- create local identifiers for the concepts you refer to
- link them to external sources, e.g. via owl:sameAs



Approach 2: Link sets

- you are in charge of what your URIs mean
- can change the link without changing the data and queries
  - incorrect choice of external URI
  - external URI changes (shouldn't but ...)
- still able to integrate data via the external URIs
- more work
  - create and maintain these reference identifiers
  - maintain link sets
- more complex to query
- clients may miss out on third party links
- commonly used for:
  - cross data set links that may be fragile (e.g. automated matching)
  - bootstrapping identifiers before authoritative source is ready

Approach 3: URI architecture

- "cool URIs don't change"
  - stability – avoid technology and organizational dependency
  - manageability – domain structure for ease delegation
- old data.gov.uk guidance

```
http://{sector}.data.gov.uk/{type}[/{concept}/{reference}]*
```

- revised data.gov.uk guidance (in progress)

```
http://{sector}.data.gov.uk{/collection*}/{type}[/{concept}/{ref}]*
http://{domain}{/collection*}/{type}[/{concept}/{reference}]*
```

- ease of handover
  - DNS domain, proxy
- assurances of stability
  - data quality metadata

Approach 4: governance and registries

- reference data
  - code lists, vocabulary terms, entities
  - key to interoperability
  - high value use of Linked Data
- Linked Data registry
  - Tooling to enable organizations to create and manage authoritative lists reference of identifiers as resolvable URIs

Registry functions

1. Manage lists of identifiers
   - lists of things (*entities*) identified by URIs described by RDF
   - metadata on status, governance, submitters
   - lifecycle support, with history and versioning
   - workflow is external
2. Repository
   - optionally store entity descriptions
   - easy to create a new, resolvable URI
3. Namespace management
   - delegate parts of namespace for others to run
   - key to both scaling and stability

Registry status

- Activity sponsored by UKGovLD working group
  - open specification
  - open source implementation
  - not a single central instance
- Further info
  - Design notes and API details

    https://github.com/der/ukl-registry-poc/wiki
  - Proof of concept deployment

    http://ukgovld-registry.dnsalias.net/
  - Summary presentation

    http://www.slideshare.net/der42/ukgovld-registryintro

Summary

- Linked Data brings interesting set of capabilities
- does imply distributed and thus fragile information space
- but provides several mechanisms to mitigate this:
  - RDF graph data model
    - independent of whether URIs resolve
    - flexible trade-off in how you rely on the network
  - named graphs
  - indirection with owl:sameAs
  - web architecture for delegation, hand over

Thanks