European Commission

OPEN DATA SUPPORT

Training Module 2.3

Design & Manage Persistent URIs

Presentation metadata

Open Data Support is funded by the European Commission under SMART 2012/0107 'Lot 2: Provision of services for the Publication, Access and Reuse of Open Public Data across the European Union, through existing open data portals'(Contract No. 30-CE-0530965/00-17).

This presentation has been created by PwC

Authors:
Nikolaos Loutas, Michiel De Keyzer and Stijn Goedertier

Disclaimers

1. The views expressed in this presentation are purely those of the authors and may not, in any circumstances, be interpreted as stating an official position of the European Commission.
The European Commission does not guarantee the accuracy of the information included in this presentation, nor does it accept any responsibility for any use thereof.
Reference herein to any specific products, specifications, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favouring by the European Commission.
All care has been taken by the author to ensure that s/he has obtained, where necessary, permission to use any parts of manuscripts including illustrations, maps, and graphs, on which intellectual property rights already exist from the titular holder(s) of such rights or from her/his or their legal representative.

2. This presentation has been carefully compiled by PwC, but no representation is made or warranty given (either express or implied) as to the completeness or accuracy of the information it contains. PwC is not liable for the information in this presentation or any decision or consequence based on the use of it.. PwC will not be liable for any damages arising from the use of the information contained in this presentation. The information contained in this presentation is of a general nature and is solely for guidance on matters of general interest. This presentation is not a substitute for professional advice on any particular matter. No reader should act on the basis of any matter contained in this publication without considering appropriate professional advice.

Learning Objectives

By the end of this training module you should have an understanding of:

- What is a Uniform Resource Identifier (URI) is.
- Why is URI persistence important.
- How to design and manage persistent URIs for data resources.

Content

This module contains …

- An introduction to Uniform Resource Identifiers (URI).
- A set of design principles for building persistent URIs.
- Service requirements for persistent URIs.

Uniform Resource Identifiers (URIs)

As common identifiers for things, e.g. people, buildings, locations, information resources…

What is a URI?

A URI is

"a compact sequence of characters that identifies an abstract or physical resource"

[TBL et al, 2005].

- "**compact**" means that the string must contain no white-space padding;
- "**abstract or physical**" means that the URI may refer to a real-world object (or thing), e.g. a person, a building or even abstract ideas like a service, or to a Web document.

For example…

A country, e.g. Belgium

- http://publications.europa.eu/resource/authority/country/BEL

An organisation, e.g. the Publications Office of the EU

- http://publications.europa.eu/resource/authority/corporate-body/PUBL

A dataset, e.g. Country Name Authority List

- http://publications.europa.eu/resource/authority/country/

Key principles

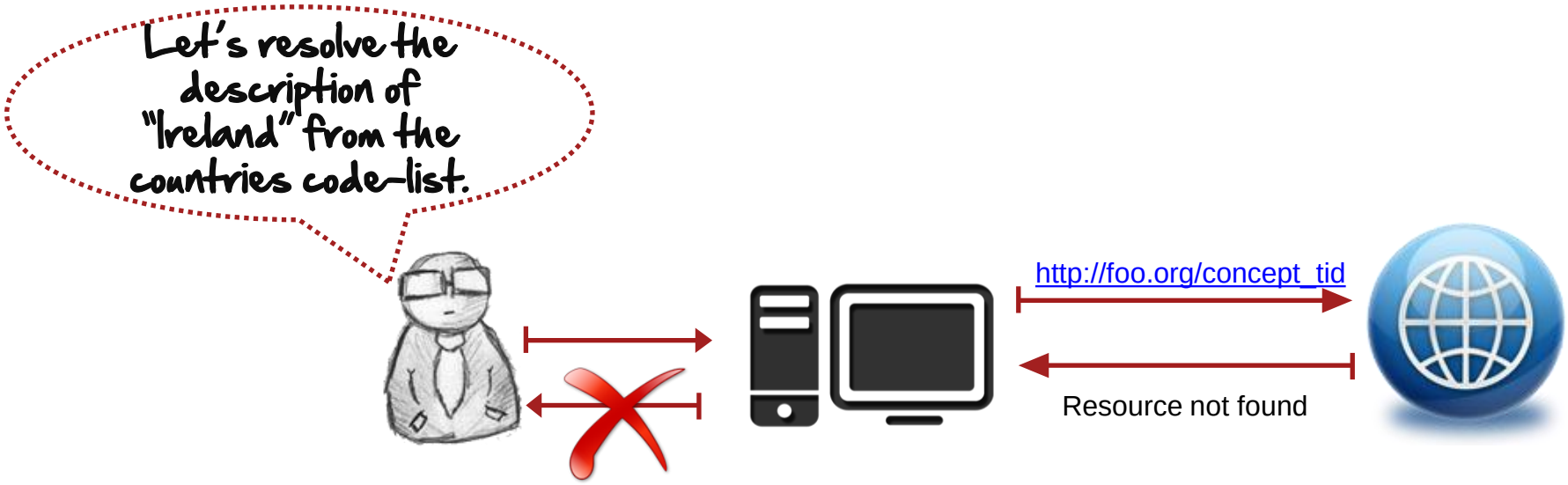
- **Persistent ,** i.e. a URI permanently assigned to a particular resource. It is stable and does not change or vanish over time.
- **Dereferencable,** i.e. a user agent can make a request to that URI over the Internet and receive a meaningful response back.
  - If the user agent is a *Web browser*, then what comes back should be a human readable HTML document.
  - If the user agent is an *RDF client* then RDF should be returned *from the same URI*.
- **Unambiguous,** i.e. here should be no confusion between identifiers for Web documents and identifiers for other resources.
  - There should be a different URI for referencing the author of a Web page and the Web page itself.

Key assumptions

- In order to create and manage URIs, one **should be the owner** of the respective Internet domain and have administrator's rights on it.
- For government domains, it is very likely that this is **managed by a central agency**. So please check with your colleagues before starting.
- Persistent and dereferencable URIs must be supported by **a trusted underlying Web infrastructure**. Such an infrastructure may be available in house in your organisation or may be provided by a different organisation – e.g. as a shared resource. So please check with your IT colleagues before starting.

What if a URI is not dereferencable and/or persistent?

Imagine the following situation…



Designing persistent URIs for datasets

10 Dos and Dont's

Follow a generic URI format

`http://{domain}/{type}/{concept}/{reference}`

- `{domain}` is a combination of the host and the relevant sector.
- `{type}` should be one of a small number of possible values that declare the type of resource that is being identified. Typical examples include:
  - `'id'` or `'item'` for real-world objects;
  - `'doc'` for documents that describe those objects;
  - `'def'` for concepts;
  - `'set'` for datasets;
  - a string specific to the context.
- `{concept}` might be a collection, the type of real-world object identified or the name of the concept scheme;
- `{reference}` is a specific item, term or concept.

Mint URIs reusing existing identifiers

- Existing identifiers of resources, e.g. database keys, should be incorporated into the URI.
  - Reuse identifiers that themselves are likely to be persistent.
  - Reuse standard identifiers rather than internal system-specific codes.

- For example, if the identifier of a company in a national business register is a string like AB123456, then the URI for this company could be:

  `http://businessdata.gov/id/company/AB123456`

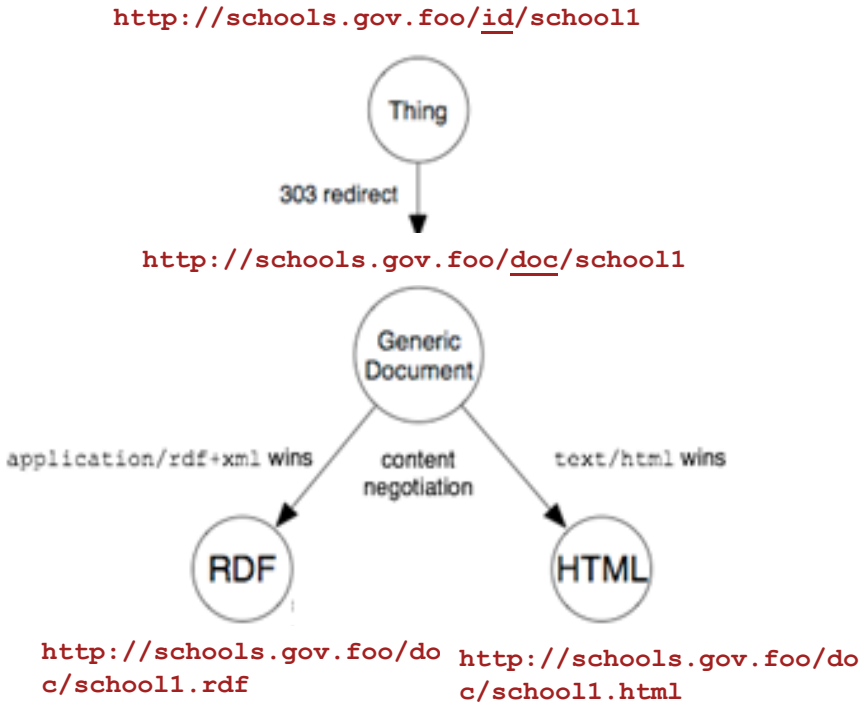
Implement 303 URIs for real-world resources

- As **no suitable representation** exists **for a real-world resources** (i.e. a non-document resources such as a person, business, location…) it is useful to be directed to a Web document which holds information about that resources.
- **Avoids ambiguity** between the real-world resource and the document that represents it.
- For example, if a government decides to create 303 URIs to represent primary schools, the result may be:
  - `http://schools.gov.foo/id/school1`
  - `http://schools.gov.foo/id/school2`

See also:
Cool URIs for the Semantic Web. http://www.w3.org/TR/cooluris

Dereferencing 303 URIs and content-negotiation

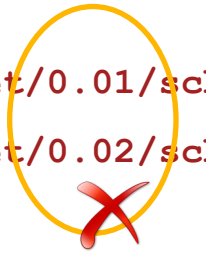
- When dereferenced, the URIs of these resources should respond with **HTTP 303** to a document that describes the object.
- The Web server needs to be configured to **redirect**:
  - from `http://schools.gov.foo/id/school1`
  - To `http://schools.gov.foo/doc/school1`
- A **URI re-write rule** is in place, typically replacing the URI {type} of 'id' with 'doc'
- **Different representations** are possible, e.g. RDF, XML, HTML…



See also:
Cool URIs for the Semantic Web.
http://www.w3.org/TR/cooluris

Avoid including version numbers in the URIs

- Datasets, concept schemes, ontologies, taxonomies and vocabularies are released in successive versions following iterative cycles of change/update.
- The URIs should remain **stable between versions**.
  - Version numbers and status information **should not be included** in the URI.
- For example, imaging two consecutive releases, v0.01 and v0.02 of the schools dataset. If version information was included in the URI, then the URI of the dataset has to change every time a new release is out.
  - `http://schools.gov.foo/set/0.01/schools`
  - `http://schools.gov.foo/set/0.02/schools`



Avoid using auto-increment when minting new URIs

- Simply incrementing a counter when creating URIs for a large dataset may be simple, but can result in serious problems.
  - What happens if the dataset is updated and URIs have to be assigned again. How can we ensure that the sequence will be the same?

Does this mean I should never do it?

The use of auto-increment in URIs may be considered, when:

- the process will never be repeated;
- the process can be repeated to create exactly the same URIs for the same input data with new URIs minted only for new items.

Avoid using query strings

- A query string (e.g. '?param=value') is text appended at the end of a URL that contains data to be passed to Web applications, e.g. search parameters to look up terms in a database.
  - Query strings are not persistent as they **rely on particular implementations**. Therefore, they should be avoided from URIs.

- For example, imagine that the URI of a company published by a national business register (NBR) was

  `http://businessdata.gov/NBR/id/company?id="AB123456"` 

  instead of

  `http://businessdata.gov/NBR/id/company/AB123456`

Avoid including information about ownership

- A persistent URI template should **not include the name of the organisation or project** that minted the URI.

- For example, imagine that the URI of a company published by a national business register (NBR) was

  `http://businessdata.gov/NBR/id/company/AB123456` 

- After a couple of years NBR is renamed to national company register (NCR). Hence all URIs have to be updated.
- In this case a URI designed for persistence would be

  `http://businessdata.gov/id/company/AB123456`

Avoid using file extensions

- File extensions reveal the file type of specific document.
- The use of file extensions should be avoided in persistent URIs.
- For example, the URI of a dataset containing the list of schools in a Member State would rather be
  - `http://data.gov.foo/set/schools`

  than
  - `http://data.gov.foo/set/schools.csv`
- The file extension can be part of the document's metadata.
  - e.g. dcat:mediaType in the Data Catalogue Vocabulary of W3C for describing datasets.

Serving persistent URIs for data resources

Use a dedicated service

- A dedicated, trusted service that is **independent of the data originator** has to be put in place.
- **Easy to be transferred** and run by someone else if necessary.
  - Dublin Core uses purl.org
  - data.gov.uk and publications.europa.eu are all also independent of a specific government department
- Not necessary to adopt a single service for multiple data providers.
  - Higher risk as this would be a single point of failure, but
  - Easier to manage and more cost-efficient.

Conclusions

A URI is “a compact sequence of characters that identifies an abstract or physical resource”.

Follow the pattern

e.g. http://{domain}/{type}/{concept}/{reference}

Re-use existing identifiers

e.g. http://education.data.gov.uk/id/school/123456

Link multiple representations

e.g. http://data.example.org/doc/foo/bar.html

e.g. http://data.example.org/doc/foo/bar.rdf

Implement 303 redirects for real-world objects

e.g. http://www.example.com/id/alice_brown

Use a dedicated service

i.e. independent of the data originator

10 rules for persistent URIs

Avoid stating ownership

e.g. http://education.data.gov.uk/ministryofeducation/id/school/123456

Avoid version numbers

e.g. http://education.data.gov.uk/doc/school/v1/123456

Avoid using auto-increment

e.g. http://education.data.gov.uk/id/school1/123456

e.g. http://education.data.gov.uk/id/school1/123457

Avoid query strings

e.g. http://education.data.gov.uk/doc/school?id=123456

Avoid file extensions

http://education.data.gov.uk/doc/schools/123456.csv

See also:
10 Rules for Persistent URIs. https://joinup.ec.europa.eu/node/53858

Group questions

http://www.visualpharm.com

Does your country have a national URI policy? If so, which are the key principles?

http://www.visualpharm.com

Does your country have in place a dedicated service for URI persistency? If so, which organisation is managing this service? If not, why?

Take also the online test here!

Thank you!

...and now YOUR questions?

References

Slide 6:

- T. Berners-Lee, R. Fielding and L. Masinter (2005) "Uniform Resource Identifier (URI): Generic Syntax". http://tools.ietf.org/html/rfc3986

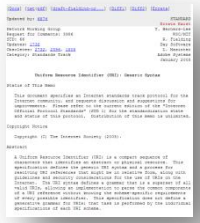
Slides 11-22:

- UK Government, CTO Council, Designing URI sets of the UK Public Sector. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/60975/designing-URI-sets-uk-public-sector.pdf
- EC ISA Programme, Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC. https://joinup.ec.europa.eu/community/semic/document/10-rules-persistent-uris

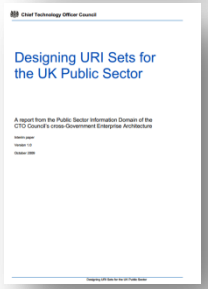
Slides 14-15:

- Cool URIs for the Semantic Web, http://www.w3.org/TR/cooluris

Further reading (1/2)



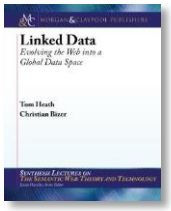
T. Berners-Lee, R. Fielding and L. Masinter (2005) "Uniform Resource Identifier (URI): Generic Syntax". http://tools.ietf.org/html/rfc3986



UK Government, CTO Council, Designing URI sets of the UK Public Sector. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/60975/designing-URI-sets-uk-public-sector.pdf

EC ISA Programme, Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC. https://joinup.ec.europa.eu/community/semic/document/10-rules-persistent-uris

Further reading (2/2)



Linked Data: Evolving the Web into a Global Data Space. Tom Heath and Christian Bizer.

http://linkeddatabook.com/editions/1.0/

Related projects and initiatives

LOD2 FP7 project, http://lod2.eu

W3C Cool URIs for the Semantic Web

- http://www.w3.org/TR/cooluris
- http://www.w3.org/wiki/GoodURIs

URI Design Principles: Creating Unique URIs for Government Linked Data, http://logd.tw.rpi.edu/instance-hub-uri-design

Publications Office of the European Commission, http://publications.europa.eu

Data.gov.uk, http://data.gov.uk/linked-data

Be part of our team…

Find us on

Open Data Support
http://www.slideshare.net/OpenDataSupport

Open Data Support
http://goo.gl/y9ZZI

Join us on

http://www.opendatasupport.eu

Follow us

@OpenDataSupport

Contact us

contact@opendatasupport.eu