D7.1.3 - Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC

Deliverable

JOINING UP GOVERNMENTS

Document Metadata

Property	Value
Release date	*17/12/2012*
Status	*Acceptance*
Version	*0.16*
Authors	*Phil Archer – W3C/ERCIM* *Stijn Goedertier – PwC EU Services* *Nikolaos Loutas – PwC EU Services*
Reviewed by	*João Rodrigues Frade – PwC EU Services* *Giorgios Georgiannakis – European Commission* *Antonio Maccioni – Agenzia per l'Italia Digitale* *Priit Parmakson – Estonian Information Systems Authority* *Peter Kranz - eGov Consultant, Sweden* *Andrea Perego – EC JRC*
Approved by	

Document History

Version	Date	Description	Action
0.01	*01/11/2012*	*For internal review (TOC +)*	*Review*
0.02	*06/11/2012*	*TOC for acceptance*	*Acceptance*
0.03	07/11/2012	*TOC for acceptance*	*Acceptance*
0.04	*20/11/2012*	*Revised TOC, CELLAR case study*	*Update*
0.05	*21/11/2012*	*Added ANDS case study, minor edits elsewhere*	*Update*
0.06	*22/11/2012*	*Added DCMI case study*	*Update*
0.08	*30/11/2012*	*Content complete, just needs todying up*	*Update*
0.09	*30/11/2012*	*Updates all over the document*	*Update*
0.10	*03/12/2012*	*Very minor updates, native speaker corrections etc.*	*Update*
0.11	*03/12/2012*	*Minor updates to Europeana section*	*Update*
0.12	*03/12/2012*	*Updates all over the document*	*Update*
0.13	*03/12/2012*	*Updates and restructuring*	*Update*
0.14-0.15	*04/12/2012*	*Minor changes ref: Spain, Europeana, DCMI – delivered for acceptance*	*Acceptance*
0.16	*17/12/2012*	*Minor updates following feedback of the contributors*	*Acceptance*

This report was prepared for the ISA programme by:

PwC EU Services

Disclaimer:

Table of Contents

List of Tables

List of Figures

10 rules for persistent URIs

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

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

1 Introduction

This document explores best practices on the publication of Uniform Resource Identifiers (URI) sets, both in terms of format and of their design rules and management. The first two elements are linked in that a well-designed URI is more likely to persist than a badly designed one. Management issues are independent to URI design itself.

Why is URI persistence important?
When a book is published, if nowhere else, it should still be found in national libraries many years in the future. When a patent is lodged or a work copyrighted, that creates a legal status that can be referred to reliably both now and in the future. Books, patents and legal documents are matters of record and what is sought here is the equivalent for identifiers that lie at the heart of Web-based interoperable data.

The recent development of open data and the desire to increase its interoperability have lead to an increased reliance on URIs as identifiers for a wide variety of concepts; everything from languages to buildings, government departments to currencies. Against this demand there is a natural human reluctance to depend on the Web - a system that is seen as being 'new.' This reluctance is entirely understandable given that the RFC that defines the URI syntax is only 14 years old[i]. Furthermore, de-referenceable[ii] URIs depend on the provision of an online service, one that cannot be maintained without some agency funding the relevant server infrastructure. Such funding is itself ultimately dependent on a decision that the cost is less than the benefit, a balance that is very much subject to change in either direction over time.

Why are persistent and well-formed URIs important when public administrations exchange data?
During data exchange, there is a need for common identifiers for the resources (classes, properties, individuals, real world entities) exchanged. Even before the evolution of linked data, Peristeras et al. (2008)[iii] emphasised in their work the need for common identifiers to support cross-border public service provision. Nowadays, the use of URIs as means of assigning unique, global identifiers to resources can provide an effective solution to this. By referring to the same URI, different agents (let them be human or machines) can easily reason that they are referring to the same resource, regardless of how this resource is modelled in national/regional/local information systems. This practically means the use of persistent and well-formed URIs, can help EU Member States to overcome semantic interoperability conflicts and provide to their citizens and business cross-border public services, thus supporting the Single Market and the mobility of people, information and goods in the EU.

1.1 Objectives

The specific objectives for the study are to:

- survey current practice for publishing URI sets;
- identify the issues related to URI design rules and management for maximum stability;
- consider the use of URIs in multilingual data;
- identify the technical issues relevant to URI design and persistence;
- codify best practice for designing and publishing stable URIs;

1.2 Scope

The stability of URIs depends on the way in which a given organisation prepares and manages them. In this context, there is an underlying dependency on the Internet infrastructure, specifically the Domain Name System (DNS). At the present time, this system underpins the entirety of the World Wide Web that itself underpins huge volumes of information exchange. Set against a historical timeline measured in centuries, the DNS system will undoubtedly evolve and be replaced. However, we cannot predict when and how such evolution will happen.

There are many types of URI. ISBN numbers seen on the back of books, for example, can be rendered as URIs (e.g. isbn:978-0-575-08360-8). Digital Object Identifiers (DOIs) can also be rendered as URIs and so on. The best known example though is of course HTTP URIs, those that begin with http://. These are de-referenceable. HTTP URIs can be put in a browser's address bar to return more information about the resource. Other URI schemes may not be de-referenceable in the same way which is the key difference between RDF (which supports the use of any URI scheme) and Linked Data, which depends on HTTP. This document focuses entirely on HTTP URIs. However, it is noteworthy that services such as handle.net provide HTTP URIs for other persistent identifier schemes, such as DOI[iv] and ARK[v]. These allow non-HTTP based identifiers to be appended to a common HTTP URI prefix and thus make them de-referenceable. Whist noting the existence of such systems, particularly in the ANDS Case Study (section 3.2), this document will focus entirely on HTTP URIs that act as complete identifiers without the need to refer to separate identifier schemes. URI schemes such as mailto: and isbn: are not considered, again, as they are not de-referenceable.

1.3 Intended audience

This study intends to reach out to government officials and CIOs of governments as well as private companies, technology consultants and the research community, who:

- need a comprehensive set of good practices on how to design and manage persistent URIs (which constitute one of the building blocks of any Linked Data initiative);
- are interested in finding out the URI persistence policies of EU agencies, national governments, and major standardization bodies and initiatives.

1.4 Structure

This document is structured as follows:

- Chapter 2 provides a short introduction to URIs and collects a number of issues related to the technical infrastructure that underlies URIs;

- Chapter 3 shows a number of case studies where URI management and persistence have been subject to a policy (as opposed to merely ad-hoc design). In these cases URIs design has been discussed and agreed. In addition, and to contrast, we look at some cases where URI design has been carried out in the absence of a clear policy. The content of this chapter is organised in the following categories, based on the nature of the stakeholders that were surveyed:
    - EU Agencies and Services (section 3.1);
    - EU Member States (section 3.2);
    - Standardisation bodies and initiatives (section 3.3); and
    - Others (section 3.4).

- Chapter 4 presents a distillation of the available information as a set of best practices that can and should be followed by publishers of URI sets. These must be:
    - consistent with the HTTP protocol;
    - consistent with the semantics of URIs;
    - consistent with the architecture of the World Wide Web;
    - consistent with common practice (unless there is a direct conflict with the above);
    - amenable to long term management;
    - implementable.

2 Beginner's guide to URIs

URIs are a manifestation of a technical architecture - the World Wide Web - that is an application of a deeper system, the Internet. These technical foundations are important when designing identifiers that are intended to be used and re-used by persons unknown, into and perhaps beyond the foreseeable future. As noted in section 1.2, this document takes the persistence of the Web and the DNS system as a given but it is important to note that the architecture of the Web is well defined and based on a set of principles that are themselves built for long term persistence, including the evolution of relevant technologies[vi].

It is within that technical framework that the following sub sections discuss the steps that can be taken to maximise, if not ensure, the long term stability of a URI.

> A Uniform Resource Identifier (URI) is a compact sequence of characters that identifies an abstract or physical resource. (...)The following example URIs illustrate several URI schemes and variations in their common syntax components:
>
> - ftp://ftp.is.co.za/rfc/rfc1808.txt
> - http://www.ietf.org/rfc/rfc2396.txt
> - ldap://[2001:db8::7]/c=GB?objectClass?one
> - mailto:John.Doe@example.com
> - news:comp.infosystems.www.servers.unix
>
> A URI can be further classified as a locator, a name, or both. The term "Uniform Resource Locator" (URL) refers to the subset of URIs that, in addition to identifying a resource, provide a means of locating the resource by describing its primary access mechanism (e.g., its network "location"). The term "Uniform Resource Name" (URN) has been used historically to refer to both URIs under the "urn" scheme [RFC2141], which are required to remain globally unique and persistent even when the resource ceases to exist or becomes unavailable, and to any other URI with the properties of a name.
> Source: RFC 3986 - Uniform Resource Identifier (URI): Generic Syntax

2.1 Detailed aims

The aim when publishing URI sets is that a given URI can be de-referenced. That is, 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*. In order for this to happen, it is important to consider the technology behind this, namely HTTP, and how this is implemented on a server.

2.2 Server capability and URI opacity

In any discussion of URI persistence it is necessary to understand two facets of the discussion:

- URIs are dumb strings, i.e. they carry no meaning except to identify a resource. For clarity, a URI such as http://example.com/document.pdf does **not** convey that there is a PDF available at that location. It would be perfectly conformant, contrary to what one would expect, for this URI to return a CSS stylesheet.
- Servers are smart and flexible – they can be configured to do a great deal more than return a static file and such configuration means that a single URI might de-reference to different resources in future.

These facets of HTTP mean that we can immediately say that persistent URIs should not include file extensions or technologies.

Many URI sets will be published and de-referenced programmatically and this will be done using a particular technology. 15 years ago it would probably have been done using Perl, 10 year ago it would be done with PHP, today it might be with Python, Ruby, ASP.Net or any number of alternatives. Even something as seemingly stable as .html should be avoided. A document might be published today in HTML but in 20 years time, maybe HTML8 will be so different that the file extension .html8 becomes common and some important documents might get updated accordingly. File extensions often (although not necessarily) reveal the technology used to create the resource and few things change as rapidly as technology.

It follows that query strings should always be avoided too. So, something like `http://example.com/getId.aspx?id=7` is almost guaranteed to be ephemeral. Better to establish a URI such as `http://example.com/id/7` and let the server deconstruct it and return the relevant data through whatever technology is in use at the time, which can be updated as required with no change to the URI.

2.3 Content negotiation

Another aspect of server configuration and the design of the HTTP protocol itself is content negotiation. As we have seen, a URI such as `http://example.com/id/7` includes no information about the nature of the resource itself. It might be machine readable data in any number of formats, it may be a human readable document that is available as HTML and PDF in any number of human languages. A properly configured server can receive a request for a simple URI like that and return the correct representation of the resource based on the detail in the request. A user agent, be it a browser or something looking for data to process, will include information about what kinds of formats it can handle and the human languages its user can process. It is this data within the HTTP request that determines what the response will be.
New representations of the resource may become available and these can be added to the resources available to the server with no change to the URI which continues to identify the specific resource.

2.4 Minimal information

Tim Berners-Lee first addressed the issue of URI stability in his 1998 paper “Cool URIs don't change”[vii]. Essentially, the advice is to *include* data that will not change within the URI and to *leave out* anything that will. Very little is guaranteed never to change. Perhaps the only exception is the date of creation of a new resource such as a document. It may change status, author, owner, title etc. but its first creation date is something that can be included in a URI and this might be useful in some circumstances, but if it can be left out, then it should be. Where a URI identifies a previously-existing resource then its creation date cannot be known with sufficient certainty to include it in the identifier.

It may be considered that the subject of a document is something that will never change but this is not so. New drafts of the document may use a new title, a policy of using a particular taxonomy to describe subjects may be brought in and so on. So even something as 'stable' as a subject should not be included in any URI designed for long term survival.

2.5 Latest versions

A class of URI that deserves special mention is that of 'latest version.' Such a URI should be very stable but what it returns might vary frequently. The W3C publishing system gives a good example of this.

The URI `http://www.w3.org/TR/vocab-org/` *always* points to the latest version of the Organisation Ontology. At the time of writing, the latest version is the one published on 23 October 2012 which has its own stable URI of `http://www.w3.org/TR/2012/WD-vocab-org-20121023/` and that is the document returned from the short URI for now. When a new version is published, it will have its own identifier and the latest version URI will return that document when de-referenced.

`http://www.w3.org/TR/vocab-org/` is a persistent URI and, like a news portal's home page, is guaranteed to return the latest information, even when specific information has gone out of date.

3 Case Studies

This chapter sets out a series of case studies. The focus is very much on public sector use of URIs, particularly linked data. A variety of online sources were used including academic papers and official guidelines, and these were augmented by direct e-mail exchanges and communications.

3.1 EU Agencies and Services

This section reports on the persistent URI policies of EU agencies and services.

3.1.1 Publication Office – CELLAR project

The Publications Office (OP) of the European Union began its CELLAR project in 2010 with the vision to make all the metadata and digital content it manages available at a single place in a harmonised and standardised way in order to:

- guarantee that citizens have better **access** to law and publications of the EU;
- encourage and facilitate reuse of content and metadata by professionals and experts;
- preserve content and metadata by making it accessible over time.

From a presentation given to the European Information Association by the head of the Enterprise Architecture Unit, Peter Schmitz, in March 2011[viii], we can see that URI-based technologies are an integral part of CELLAR and that long term preservation was recognised as being important from the outset. When the project began, the OP was able to draw on in-house skills and expertise but to realise the project it was necessary to engage contractors. At the time of writing, initial data is being loaded into CELLAR ahead of its formal launch and further data will be added over the coming months.

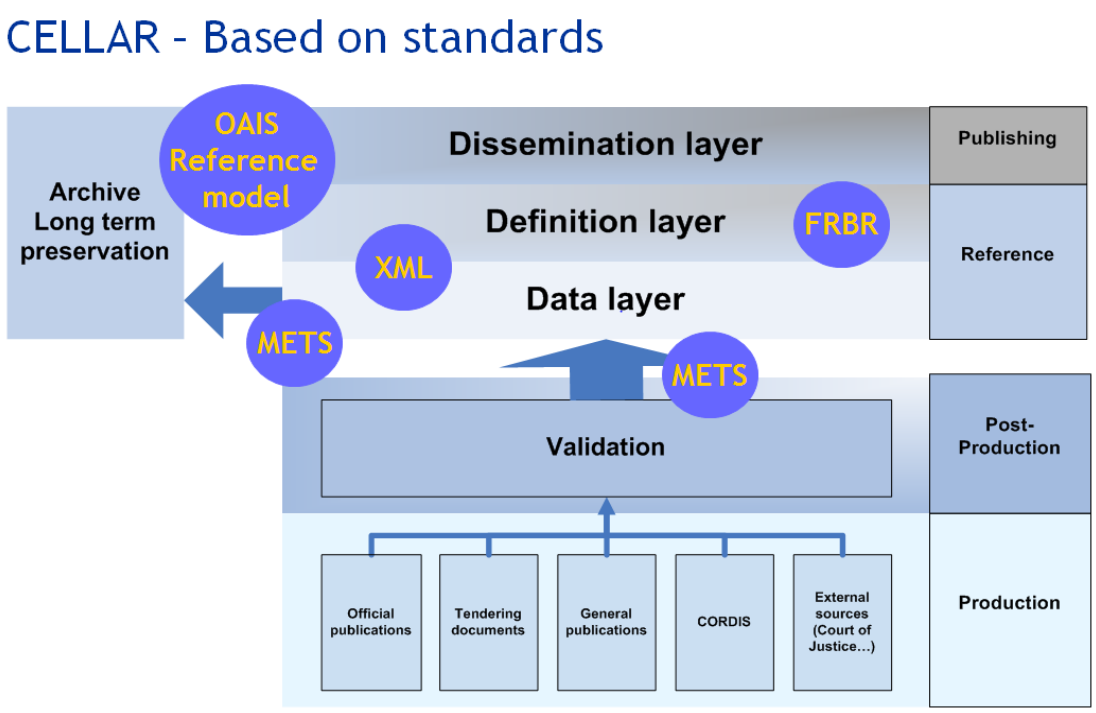

Figure 1 - Slide from Peter Schmitz's presentation to the EIA, March 2011 showing the semantic technologies used in CELLAR

Publishing information and committing to its long term preservation and stability has always been an important aspect of the OP's work. This organisation maintains several *de facto* schemas such as the EUR-Lex schema smartAPI that is already more than 10 years old and that can be used as a URI set. However, its use is no longer encouraged and, for reasons that will discussed in later sections, URIs such as

```
http://eur-lex.europa.eu/smartapi/cgi/sga_doc?smartapi!celexplus!prod!CELEXnumdoc&lg=en&numdoc=308R1008
```

must considered brittle since they depend on a particular implementation. The adoption of linked data has put greater emphasis on the importance of URIs as identifiers and it is this change that has lead to the development of CELLAR.

3.1.1.1 URI format

URIs need to be resolvable so that further information can be extracted from them and for information to be available in multiple languages and multiple formats. The URIs themselves are designed carefully for long term management and stability. Every URI begins with the same pattern:

```
http://publications.europa.eu/{type}/{subtype}
```

where there are just three possible values for `{type}`:

1. `resource`, for content and metadata resources;
2. `ontology`, for schemas;
3. `webapi`, for Web API services.

For example, editions of the Official Journal all have URIs beginning with:

`http://publications.europa.eu/resource/oj/`

where `'oj'` acts as the subtype. Named Authority Lists begin with

`http://publications.europa.eu/resource/authority/`

Beyond the second path component, the structure depends on the specific case. OJ editions all have identifers based on their year of publication and edition within that year. Each edition of the OJ is available in multiple languages and these are included in the URI; so the the first edition of the OJ from 1952, in German, is identified by:

`http://publications.europa.eu/resource/oj/JOP_1952_001_R.DEU.`

The entry for English in the Named Authority List for languages is

`http://publications.europa.eu/resource/authority/language/ENG.`

These URIs are inherently stable since any new Named Authority List can be added, with the list name as the third path element, and the specific entry in that list as the fourth. The name of a publication will always appear as the second path element and so on. The Publications Office consistently uses ISO 693-3 3 character language codes as the level of detail provided is the right one for OP's multilingual environment.

The OP makes extensive use of content negotiation (section 2.3) and language negotiation. Many items published by the OP are 'works' within the FRBR[ix] sense of the word. Each work has its own URI and CELLAR returns a specific manifestation of that work based on the HTTP Request headers.

CELLAR makes use of its HTTP server's native support for content negotiation for the first of these but not the second. That is, the server inspects the Accept header in the HTTP request and returns HTML (for humans), RDF or XML. (All HTTP requests include information about the device making the request. In the case of a Web browser, this will include the type of browser, operating system and language preferences). One of these three formats is always returned. HTTP is less deterministic for languages so that if the request header specifies a language in which the particular work is not available, the server response can vary between implementations. The canonical response is either 'No Acceptable Variant' or 'Multiple Choices' - neither of which may be helpful for some users and so CELLAR uses its own software to always return a representation of the work. Each manifestation of a work, that is, a particular version of the requested resource in a specific language and specific data format, has its own URI and this can of course be accessed directly.

3.1.1.2 URI design rules and management

As noted, the Publications office has designed its URIs to survive for the long term. The *intention* is clear: that URIs will persist and will continue to mean the same thing over the long term. However, no organisation can be certain about its future and a commitment to maintain a service indefinitely is very hard for anyone to make. The recent European Council Decision to support European Legislation Identifiers (ELI)[x] was widely welcomed. However, Member States have not translated that into a firm commitment to guarantee the stability of ELIs over the long term. Similarly, like any organisation, the OP itself is subject to political change[xi] and it is clearly possible that the office might be reorganised, re-named, merged or split. Such eventualities cannot be foreseen or guaranteed against.

> The best guarantee of persistence is usefulness. The OP is making its best effort to create a stable URI set, designed, managed and published with longevity in mind. The OP has a track record of maintaining URIs for more than 10 years and cannot today see any reason why the URIs embodied in CELLAR will not persist. The subdomain, publications.europa.eu, is as stable as any can be and was chosen deliberately for that reason. Even if the name of the institution were to change, the subdomain is sufficiently generic that it could easily survive. This would not be the case if, for example, the name CELLAR, i.e. the project that created it, had been used as the subdomain.

The systems behind the various publications handled by the OP vary depending on the type of publication itself. The OJ is developed as a separate document whereas the Named Authority Tables (reference tables used throughout the European Institutions) are managed directly by the OP and edited using Microsoft Excel. From there an XML document is created and this becomes the Master file. Various scripts are then used to generate the HTML and RDF versions of the tables. Importantly, the output of each of those scripts is a static file so that the resolution of each URI is not dependent on a dynamic process, this adds stability to the system.

3.1.2 EC Informal Working Group on Persistent URIs

Recently, several Directorate Generals of the Commission have formed an informal working group on persistent URIs. This group was 'officially' launched in February 2012[xii]. Its objective is to propose a solution on how to proceed with the compilation of guidelines that can be used to create consistent URIs and a common URI assignment policy. The aim of such guidelines would be to define a working, scalable and performant URI model, that can be rolled out Commission-wide, to uniquely identify each physical item (e.g. data objects like 'toxic substance' or a NUTS region), each abstract concept (e.g. 'governance', styles, map layers) and each core vocabulary and dataset[xiii].

The WG focuses on three areas, which are reflected in the recommendations of chapter 4:

- the governance and maintenance of the URIs,
- the universal use of persistent URIs and a URI model,
- the application of a common naming scheme across Directorate Generals.

N.B.: Please note that the URI format, design rules and management are an initial proposal still under review (valid at the time that this report was compiled). It is expected to evolve through the discussions of the Working Group.

3.1.2.1 URI format

According to the WG, a URI model should follow the following form:

```
{URI Root}/{Resource Path}/{ID}/{String}/{Options}
```

Where:

- **URI Root** provides information about the URI scheme and provider – in most other cases studied this is usually referred to as Base URI;
- **Resource path** is a hierarchical organisation to specify the scope of the URI (this resource path captures a certain context and knowledge for the resource, as is the case for example with the EuroVoc thesaurus);
- **ID string**, to be used in case of large numbers of resources of the same type, corresponds to the actual textual format of the identity of the resource. The identity is assigned from an authority. Examples of such cases can be found in domains where a large number of same type resources exist, as is the case with legal entities and bank account numbers;
- **Options** is a way to address different services related to a URI (e.g. resources, services, SPARQL endpoints).

Only the URI Root and Resource Path are necessary for all resources. The URI root can have the following form: `http://{Europa Home URI}`. The standard URI Root can be:

`http://ec.europa.eu/URI/.`

The Resource Path is a hierarchical scope definition for the URI and can have the form:

`{Policy Definition}/{Sub-domain Definition}*/{entity}`

For example http://ec.europa.eu/URI/health/indicators/echi is made of the following components:

- Europa Home URI = ec.europa.eu
- Policy Definition = health[1]
- Sub-domain Definition = indicators
- Entity = echi

[1] It is very likely that the Working Group will decide to remove the `policy definition` from the URI as this information is volatile and thus subject to change.

A URI structure with a fixed part and loosely defined part is sufficient to specify entities as well as instances in the case of integral resources (such as big datasets, collections and queries on them). The options part remains to be further defined but models like Open Data Protocol can serve to establish specific guidelines at a later stage.

3.1.2.2 URI design rules and management

The WG published also a set of recommendations to be considered when EU agencies and services design URIs.

- Define a common URI assignment policy using one root, e.g. http://ec.europa.eu/open-data;
- Include each policy area as a domain prefix name using a managed restricted list, referenced and linked to a standard taxonomy, e.g. EuroVoc[xiv];
- Make provision so that DGs that have already created URIs with different parameters can retrospectively link to a subsequently agreed common URI policy. This will ensure a uniform application of the guidelines across the institution, increasing the possibilities of linking data;
- Take account existing best practice and guidelines from W3C[xv] and advice from national administrations or other international organisations with known experience of persistent URIs;
- Incorporate the essentials of existing best practice and guidelines, which will ensure that URIs:
    - identify each resource;
    - are permanent and stable;
    - are manageable;
    - are unique;
    - are clear, concise and short;
    - are explicitly linked with each other;
    - are user-friendly (i.e. human readable), i.e. `uropa.eu/health/guidelines` not `europa.eu/health/12355647987654654 98;`
    - are consistent in format and structure, e.g. `europa.eu/health/…` not `europa.eu/healthy/;`
    - do not contain keywords;
    - do not contain file extensions;
    - have no 'www' (use a 301 redirection in case its presumed);
    - are lowercase only;
    - do not contain accents or spaces;
    - replace special characters !"£$%^&*() with hyphens or underscores.

- Decide on the scope of the assignment policy and the guidelines and on the use of ontologies in which each resource will be uniquely identified.
    - In which cases will external ontologies be used for wider data coverage? Which instances are at the core of the European Commission and how many controlled vocabularies are defined?

- Will one DG manage all URIs for both published and other data?

An open question remains: will one DG become the single, central authority for managing the attribution of URI sets or will there be a set of authorities depending on the policy content?

3.1.3 Eurostat Linked Data

Under the LATC Support Action[xvi], a team at DERI[xvii] created a linked data version of the Eurostat data[xviii]. This is a flagship sophisticated system that gets top marks in the 5 Stars of Linked Open Data[xix] devised by Tim Berners-Lee. As Figure 2 shows, 5 star data is published in RDF and linked to other data sets, all under an open licence. In addition to meeting the 5 star criteria, the data is updated each week and a policy is in place[xx] that makes a best effort to ensure that the system persists beyond the life of the project. The Eurostat Linked Data project not only created the data but also produced a number of freely available tools and was the basis of DERI's Sarven Capadisli's MSc thesis[xxi].

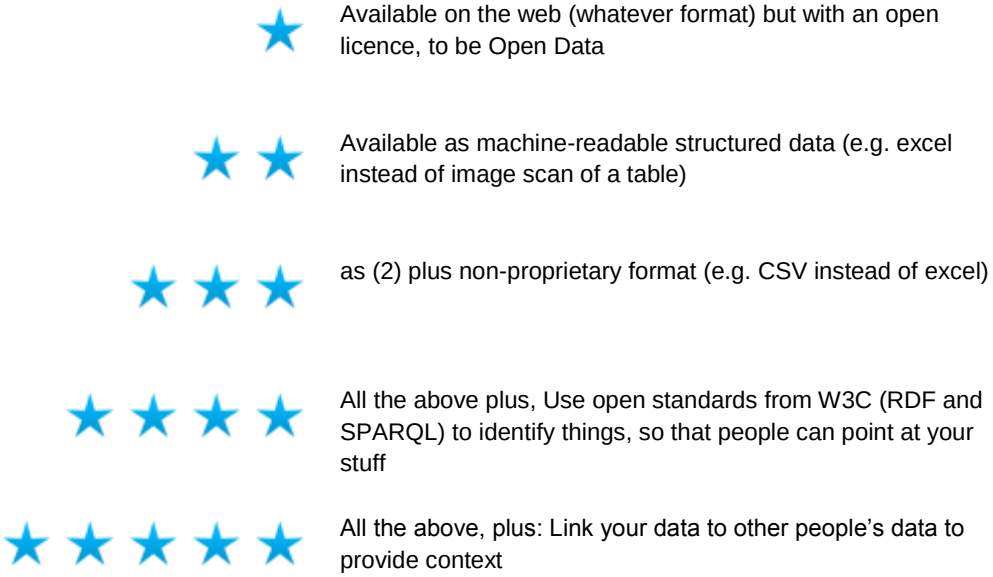

Figure 2 - The 5 Stars of Open Linked Data

3.1.3.1 URI format

The base URI proposed for Eurostat's data is:

`http://eurostat.linked-statistics.org/`

Capadisli and Hausenblas say that they decided to keep the same file name for the metadata and the actual dataset containing observation values as they appear in the original data and distinguish them by using `dsd` and data in the URI pattern. The code lists shared among all datasets are provided by using `dic` in the URI pattern. Hence the following URI patterns are defined for:

Metadata: http://eurostat.linked-statistics.org/dsd/{id}

where id is one of the dataset's metadata file.

Datasets: http://eurostat.linked-statistics.org/data/{id}

where id is the filename of the dataset containing observation values.

Code lists: http://eurostat.linked-statistics.org/dic/{id}

where id is the filename of dictionary.

Observations:http://eurostat.linked-statistics.org/data/{dataset}#{dimension1},{dimensionN}

where the order of dimension values in the URI space depends on the order of dimension values present in the dataset.

3.1.3.2 URI design rules and management

The Eurostat Linked Data project also is a rare example of a multilingual dataset. Although the language used to mint a URI may be apparent, technically, all URIs are dumb strings and therefore language-neutral (see section 2.2). Any number of labels may be attached to URIs in any language as set out by Jose Emilio Labra Gayo in Dublin at the Linked Open Data and MultilingualWeb workshop in June 2012[xxii]. The Eurostat Linked Data Project home page includes some sample SPARQL queries. Queries like this are executed behind the scenes when URIs such as http://eunis.eea.europa.eu/species/124054/linkeddata are de-referenced.

The Web page seen in a browser offers a human reader a number of options for navigating around the data (note the tabs such as 'General Information' in Figure 3). When the same URI is de-referenced by a machine that accepts RDF, all the data is returned, including the vernacular names. The careful design and publication of persistent URIs does **not** imply the publication of multilingual 5 star linked data. Persistent URIs are also an important component in less sophisticated systems that target interoperability. However, if public administrations are to enjoy the maximum benefit of this powerful technology, for their own internal data management as much as reporting to the outside world, then the Eurostat Linked Data offers a show case for what can be achieved and how to achieve it.

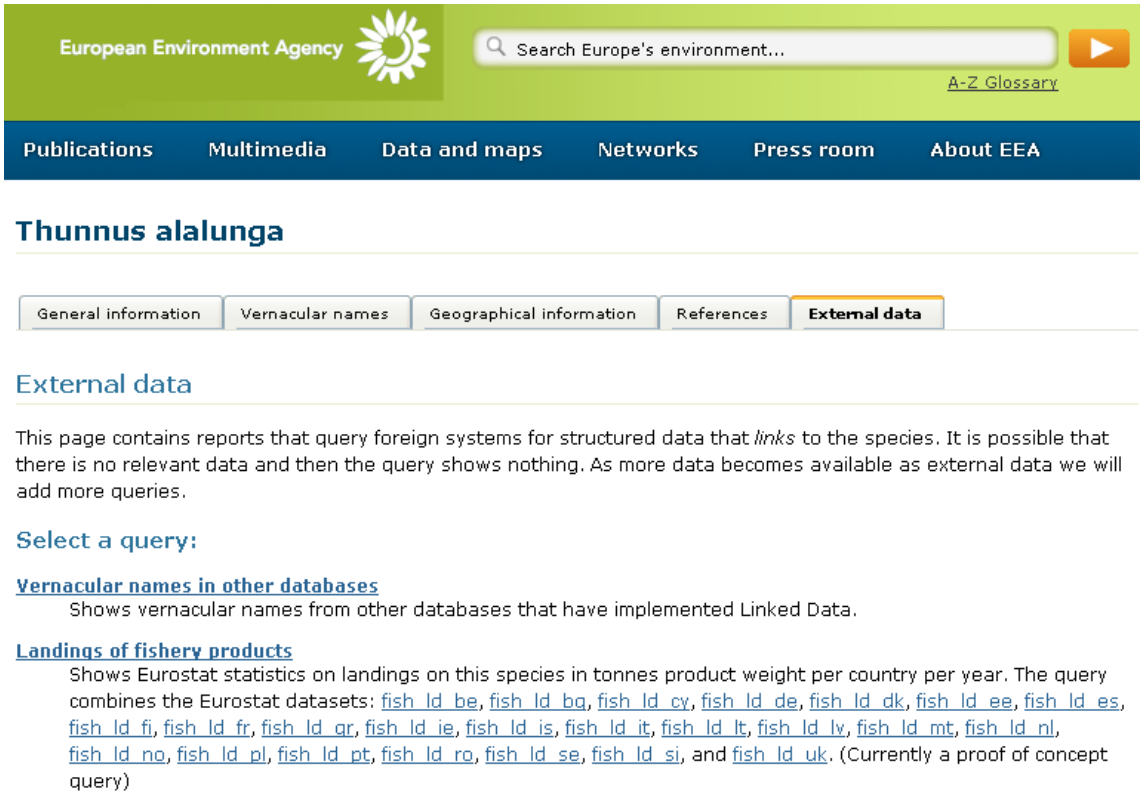

Figure 3 - The EEA Linked Data concerning Thunnus alalunga as seen in a Web browser at http://eunis.eea.europa.eu/species/124054/linkeddata

Table 1 - The EEA Linked Data concerning Thunnus alalunga returned to a user agent that accepts RDF at http://eunis.eea.europa.eu/species/124054/linkeddata

```
<SpeciesSynonym rdf:about="species/124054">
 <speciesCode>124054</speciesCode>
 <foaf:isPrimaryTopicOf rdf:resource="124054/general"/>
 <binomialName>Thunnus alalunga</binomialName>
 <validName
rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean">true</validName>
 <eunisPrimaryName rdf:resource="species/124054"/>
 <taxonomicRank>Species</taxonomicRank>
 <taxonomy rdf:resource="taxonomy/2201"/>
 <dwc:scientificNameAuthorship>(Bonnaterre,
1788)</dwc:scientificNameAuthorship>
 <dwc:scientificName>Thunnus alalunga</dwc:scientificName>
 <rdfs:label>Thunnus alalunga (Bonnaterre, 1788)</rdfs:label>
 <dwc:genus>Thunnus</dwc:genus>
 <speciesGroup rdf:resource="speciesgroup/2"/>
 <dwc:nameAccordingToID rdf:resource="references/1785"/>
 <ignoreOnNameMatch
rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean">false</ignoreOnNameM
atch>
</SpeciesSynonym>
<rdf:Description rdf:about="species/124054">
```

```
</rdf:Description>
<rdf:Description rdf:about="species/124054">
 <dwc:vernacularName xml:lang="sq">Ton pendgjate</dwc:vernacularName>
 <dwc:vernacularName xml:lang="de">Thun</dwc:vernacularName>
 <dwc:vernacularName xml:lang="de">Thunfisch</dwc:vernacularName>
 <dwc:vernacularName xml:lang="de">Weißer thun</dwc:vernacularName>
 <dwc:vernacularName xml:lang="da">Albacore</dwc:vernacularName>
 <dwc:vernacularName xml:lang="da">Hvid tun</dwc:vernacularName>
 <dwc:vernacularName xml:lang="da">Tun</dwc:vernacularName>
 <dwc:vernacularName xml:lang="es">Albacora</dwc:vernacularName>
… More
```

3.2 EU Member States

3.2.1 UK Government

The UK was an early adopter of open data and of linked data in particular. Its data portal, data.gov.uk, went online on 30th September 2009, 4 months after the US portal at data.gov. An important difference between the two is that the British portal has already from its inception put greater emphasis on the added value of linked data.

In fact, the UK portal has always given prominence to linked data as Figure 4 makes clear with its reference to SPARQL (these days the link is to 'Linked Data'). The personal involvement of Tim Berners-Lee and Nigel Shadbolt is an important factor here as both are staunch advocates of linked data and therefore the use of stable URIs.

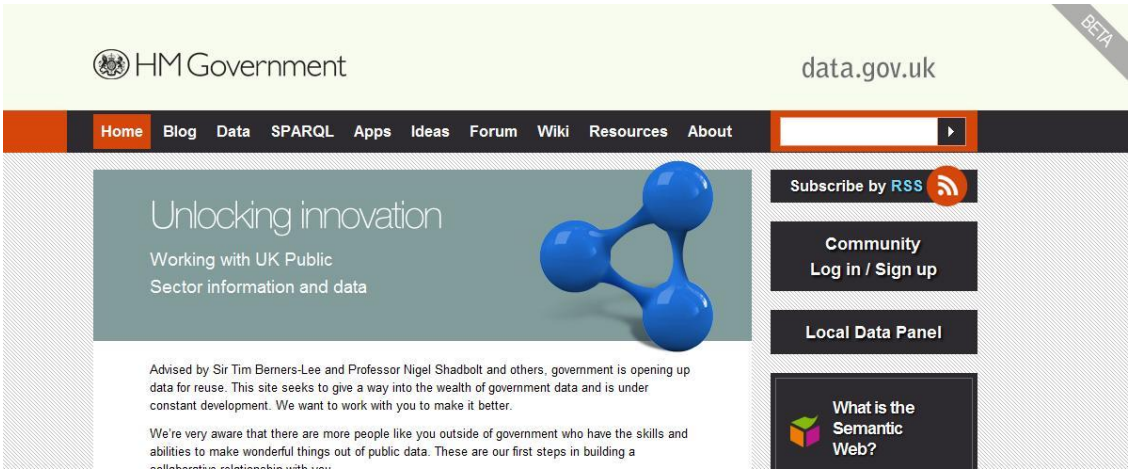

Figure 4 - Screenshot of data.gov.uk taken 4 April 2010. Source: Wikipedia[xxiii]

It was against this background that a group within the UK government wrote a document called *Designing URI Sets for the UK Public Sector*[xxiv], published in October 2009, less than 2 weeks after data.gov.uk went online. The document covers many design issues (see next section) to ensure persistence, these are copied below[xxv].

General

1. When considering the domain to root a URI set in:
   - the publisher will require content control of the sub-domain that it ultimately resolves to;
   - the domain will have appropriate service-levels and scalability for resilience and performance.

Requirements for URI sets that are promoted for re-use

2. In addition, where a URI set is promoted for re-use, the following considerations apply to find a balance for central and federated components:
   - flexibility and readability;
   - administrative burden;
   - infrastructure costs.

3. In particular, the domain will:
   - expect to be maintained in perpetuity;
   - not contain the name of the department or agency currently defining and naming a concept, as that may be re-assigned;
   - support a direct response, or redirect to department/agency servers;
   - ensure that concepts do not collide;
   - require the minimum of central administration and infrastructure costs;
   - be scalable for throughput, performance, resilience.

4. The choice of domain should provide the confidence to the consumer, that the URI set has met minimum quality criteria, including implementing these design considerations. In other words, the domain itself should convey an assurance of quality and longevity.

5. Due to the drive to rationalise websites and also to separate presentation of data from its location, UK public sector URIs will be based around the **`data.gov.uk`** domain, split by sectors as sub-domains. When looking up a URI, the data.gov.uk servers either provide the response themselves, or DNS is used to redirect enquiries to the appropriate department or agency server.

6. A sector is NOT a department name. Sectors should be understandable by the public, rather than reflecting how government is currently organised. New departments taking over all or part of a sector are required to maintain the URI sets. The community of practice will provide further information about the use of sectors which are likely to be aligned to other initiatives such as Directgov.

7. Using ‘education’ as an example sector for a URI set promoted for re-use, gives:

   `http://education.data.gov.uk`

   This:

- shows that the set is a part of the education sector;
- puts it in the data.gov.uk collection of UK public sector URIs promoted for re-use;
- can be redirected using DNS to a departmental server for the content;
- is from the data.gov.uk domain and therefore not confused with a presentation website.

The first point highlights that a URI data set requires a resilient technical infrastructure. The first point under item 3 is stark: "expect to be maintained in perpetuity." That is a bold statement and goes well beyond what anyone can reasonably predict, however, the key word in that bullet point is 'expect.' What the guidelines are highlighting is the need to *think* for the long term and to show the *intention* that the URIs will persist.

The `data.gov.uk` domain is put forward as the long term domain to use although it has not been possible to identify a published URI persistence policy for the domain. Using that as the upper domain name means that sectors can then be defined as sub-domains rather than using a public-facing Web site that will be on a domain name that reflects the departmental name. Departments come and go - anecdotally, the average lifespan of a UK government department is 4 years. A case that might have been in the mind of the authors of *Designing URI Sets for the UK Public Sector* was the Department for Children, Families and Schools known before June 2007 and since May 2010 as the Department for Education.

The UK Government now has adopted a policy that explicitly recognises the value of URIs as identifiers[xxvi] and it is clear from the policy that government URI data sets are expected to persist over the long term.

3.2.1.1 URI format

Designing URI Sets for the UK Public Sector suggests that URIs should include the following components:

- A **concept**: a word or string to capture the essence of the real-world 'Thing' that the set names, e.g. school.
- A **reference**: a string that is used by the set publisher to identify an individual instance of concept. The reference should match the way that it is used in normal use. In some circumstances, a name may be appropriate as the Reference, e.g. 'England'. Where the name may change, or becomes overly verbose, a code may be more appropriate. For example:
    - road/M5 (roads rarely change their names);
    - school/123 (a specific school).

  URI **type**, for example one of:
    - id – where the URI identifies a non information resource (a real world object);
    - doc – where the URI identifies a document, including one that describes a non information resource;
    - def – a concept definition;

- set – a data set.

These elements can be combined into a general URI pattern thus:

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

where `{domain}` is a combination of the host (e.g. data.gov.uk, europa.eu etc.) and the relevant sector ('transport', 'education' etc. in the case of the UK, 'resource', 'ontology' or WebAPI' in the case of the Publications Office). It is a matter of choice whether the sector is defined as a sub-domain of the host or as the first component of the path, so that both `http://transport.data.gov.uk` of the UK government and the example shown before, of the Publications Office, `http://publications/europa.eu/resource` are both valid values for the `{domain}` variable.

The `{type}` element will vary between different use cases. At data.gov.uk it will be one of the types listed immediately above; for Dublin Core (section 3.3.1) it will be one of `'terms'`, `'dcmitype'` etc. and for CELLAR (section 3.1.2) it will be something like `'oj'` or `'authority.'` In Europeana (section 3.4.3) it will be simply `'item'` meaning that the URI identifies an item in a collection.

The `{concept}` and `{reference}` elements are also found in different forms across many examples but there will be more variation depending on the specific case. For Dublin Core (section 3.3.1) terms it is enough to simply add the relevant term on to:

```
http://purl.org/dc/terms/{term}
```

The UK example already given shows how a URI such as:

`http://transport.data.gov.uk/id/road/B3178`

is minted (the B3178 is a road in Devon) and in Europeana we see URIs such as the one below `http://data.europeana.eu/item/00000/E2AAA3C6DF09F9FAA6F951FC4C4A9CC80B5D4154` where `'00000'` identifies the collection (this can be thought of as the `{concept}`) and the long string identifies the specific item within that collection (the `{reference}`).
A final design feature captured in *Designing URI Sets for the UK Public Sector* and repeated elsewhere is the addition of a file extension at the end of a URI that reflects the likely media type. For example, pasting `http://transport.data.gov.uk/id/road/B3178` (the URI

for the B1378) into a Web browser gives a 303 redirect to the document that describes that road which is at `http://transport.data.gov.uk/doc/road/B3178` (note the substitution of `'doc'` for `'id'`). The data shown at that page is available in 7 different formats: HTML, CSV, JSON, RDF/XML, Text, Turtle and XML.

The particular version returned when `http://transport.data.gov.uk/doc/road/B3178` is de-referenced will be decided through content negotiation (section 2.3) but it is possible to refer to a specific representation of the data by appending the relevant file extension so that appending '.ttl' (http://transport.data.gov.uk/doc/road/B3178.ttl) for example will return *exactly* the same data as seen as an HTML page in a regular Web browser but encoded as RDF and serialised in Turtle. Importantly, every representation of the document includes links to all the others. A similar set up can be seen, for example, at Open Corporates[xxvii].
This kind of functionality, and the ability to resolve well-designed URIs at different levels of the tree, is provided by the Linked Data API[xxviii]. Although adopted across data.gov.uk, it is not a recognised standard and is not used as widely as the designed principles on which it is based.

3.2.1.2 URI design rules and management

To persist, a URI must be designed to persist. In summary: that which is subject to change should not be included in a URI. That which is permanent should be included. Notes on what to leave out of URIs are discussed in section 2.2 later on. This section focuses on what to include.

The UK Government's *Designing URI Sets for the UK Public Sector* paper again provides a key reference point. It was expanded upon shortly after its publication in an influential series of blog posts[xxix] by Jeni Tennison[xxx], then lead developer at legislation.gov.uk and now Technical Director at the Open Data Institute. The ideas have been extended substantially by Leigh Dodds and Ian Davis in their book *Linked Data Patterns*[xxxi]. Like Jeni Tennison, Leigh Dodds and Ian Davis were directly involved in the development of *Designing URI Sets for the UK Public Sector* and although their book is more recent (it was published in May 2012) and goes into considerably more depth, the thinking behind it is the same.

The core ideas in all these documents stem originally from a W3C Semantic Web Interest Group Note *Cool URIs for the Semantic Web*[xxxii] which in turn is a development of Tim Berners-Lee's original *Cool URIs Don't Change* document from 1998[xxxiii]. The design principles have been collected, re-phrased and discussed in other publications, such as the American Federal data portal, data.gov[xxxiv], and Tom Heath and Chris Bizer's book *Linked Data: Evolving the Web into a Global Data Space*[xxxv] but the core ideas have not changed and can be seen repeated everywhere with little variation.

As discussed in previous sections, the choice of domain name is a critical one: the domain name itself must be one that is expected to persist for as far into the future as anyone can reasonably see. In many cases, this means establishing or using a service *specifically established* to provide that stability.

Designing other URI components requires an understanding of what a given URI actually refers

to: what it identifies and what it does not. The use of HTTP URIs as identifiers is extremely powerful since it allows computers to look up information about a given 'thing' and to recognise that wherever the same URI is used, it identifies the same 'thing.' However, there is a useful distinction to be made between URIs that identify real world things as opposed to concepts and between a single piece of data and a set of data. There is a more fundamental difference between a real world object, such as a school or a person, and a document that describes that object. This difference is at the heart of a discussion that has been going on for more than a decade under the arcane title of HttpRange-14[xxxvi]. W3C's Sandro Hawke gathered many references to the discussion when it first became 'old' - in 2003[xxxvii]. The W3C Technical Architecture Group (TAG) - the permanent Working Group that oversees the overall direction of Web technologies - is *still* grappling with the issue[xxxviii]. An article by Mike Bergman from January 2012 "*Give Me a Sign: What Do Things Mean on the Semantic Web?"* offers a philosophical discussion of the issue[xxxix].
In a nutshell, if a URI identifies a physical object that cannot be transmitted over the Internet, what should be returned instead when the URI is de-referenced?

The usual solution - that is, the solution agreed by the TAG in 2005[xl] - is that where a URI refers to an information resource, i.e. something that can be represented as a stream of bytes, the data should be returned. Where the URI identifies a non information resource (like a building or a person) then the HTTP server should offer a 303 'Other' re-direct[xli] to a different URI that identifies a document that describes the object. This solution has many adherents as it uses existing HTTP infrastructure and is faithful to the semantics. However, the additional round trip to the server to fetch the document describing the thing you asked about in the first place, and the fact that you very often can't copy and paste a URI from a browser window into another document (because it's changed), are things that many developers and practitioners are unhappy with - hence the endless discussion. Be that as it may, the resolution is embodied in all the URI design patterns seen in the course of compiling this document.

3.2.2 Estonia

Estonia, a country known for its advanced use of e-Government services, has recently published its *Framework of Websites version 1.0*[xlii] document, part of the *Interoperability Framework of the State Information System version 3.0*[xliii]. These documents do not address URIs in the sense discussed in this document, i.e. as identifiers used in data sets.

Priit Parmakson of the Estonian Information Systems Authority provided us with a comprehensive overview of initiatives in Estonia that deal with persistent URI design rules and management.

3.2.2.1 URI format

Although there is no official URI policy, Estonian's *Framework of Websites* document does, however, offer guidelines on Web site construction. All public administrations in Estonia are required to publish a Web site and within it there are certain sections that must be present at defined URLs. For example, the contact information must be at `http://{domain}/kontakt`,

the news must be at `http://{domain}/uudised` and so on. We can draw from these examples the following URI model:

```
{domain}/{type}/
```

The above pattern is not always used as there are examples of the use of URIs as identifiers within datasets. More information is provided in the next section.

3.2.2.2 URI design rules and management

The Estonian Land Cadastre[xliv], operated by the Estonian Land Board (Eesti Maa-amet[xlv]) has a public interface where cadastral units can be accessed by permalinks in the form `http://xgis.maaamet.ee/ky/FindKYByT.asp?txtCU=72704:004:0430`. This is an interesting case akin to ANDS (section **Error! Reference source not found.**) as the online service is being used to resolve a non-URI identifier. In this case, 72704:004:0430 is the permanent identifier. The bulk of the URI (`http://xgis.maaamet.ee/ky/FindKYByT.asp?txtCU=`) is clearly not designed for persistance because:

- it includes the technology used - ASP - which was superseded by ASP.NET several years ago and so it already looks 'old;'
- it includes a query string `(?txtCU=)` and thus the URI reveals a lot about the system behind the URI (it's looking up a value in a database).

The replacement either of the database system or of ASP would almost certainly require a change in the URI and so is unlikely to persist. We understand that URI persistence may not have been a documented requirement when this system was developed. It may also be that in this case the need for persistence is understood as not too relevant.

A similar situation applies to the Estonian National Place Names Register (Riigi kohanimeregister[xlvi]). Owned by the Ministry of Interior Affairs, and operated by the Estonian Land Board, the register has a publicly accessible service (avalik liides[xlvii]) where a comprehensive database of Estonian place names can be accessed by URLs in the form:

`http://xgis.maaamet.ee/knravalik/knr?obj_id=3375`.

Here again we can surmise that that the actual identifier is the number 3375 and that a query is being made to a relational database rather than data being provided directly through a URI such as `http://xgis.maaamet.ee/knravalik/knr/3375` which would fit the pattern used elsewhere.

The Estonian State Gazette (Riigi Teataja[xlviii]), the national database of legislation, owned by the Ministry of Justice has made all Estonian law referenceable and de-referenceable, down to

paragraph level. For example, the Public Information Act (Avaliku teabe seadus) is accessible at `https://www.riigiteataja.ee/akt/122032011010?leiaKehtiv`.

In this instance, the query string `(?leiaKehtiv)` is not being used as part of the identifier as such. Instead, it is an instruction to return the latest version. Although not directly inline with the recommended practice of assigning a stable URI that always points to the latest version with different URIs for each draft (section 2.5) it is closer in spirit than the other examples. It is also noteworthy that individual paragraphs within the legislation can be accessed directly using a fragment identifier so that to refer to § 431 one can use:

`https://www.riigiteataja.ee/akt/122032011010?leiaKehtiv#para43b1`

There are several stable URI patterns in use in Estonia. Authentic descriptions of all public sector information systems, for example the Riiklik ehitisregister (Estonian Register of Buildings) is identified by `https://riha.eesti.ee/riha/main/inf/riiklik_ehitisregister` and an ontology such as the Ontology of Health Insurance is identified as `https://riha.eesti.ee/riha/onto/ravikindlustus/2008/r2` with appended individual terms so that the role of an Assistance Doctor (and abiarist) is identified within the relevant ontology as:

`http://riha.eesti.ee/riha/onto/toohoivejasotsiaalkysimused/ravikindlustus/2008/r1/abiarst`

> Although these URIs have obviously been carefully designed to persist, they are not fully conformant with the practices examined so far as they include the version number/date of the ontology, a practice which may create persistence problems (see section 3.3.1 for an example of where this goes wrong).

Estonia is already perfecting their URIs. An example of this is the requirement to indentify objects with URIs in the Estonian Open Data Guidelines[xlix]. In the summer 2012, a programme of open data projects was launched with a budget of € 1.3 million. The 8-10 projects now under way are dealing with the establishment of URI schemes for different types of objects. For example, the national sports person database already has convenient URI scheme in use `http://www.spordiinfo.ee/esbl/biograafia/Ilmar_Ruus`. This is exactly in line with persistent URI design discussed earlier, having the basic structure of :

> `http://{domain}/{type}/{concept}/{reference}` (section 3.2.1.1).

The introduction of URIs as identifiers in the Estonian public sector should be viewed in the larger context of naming and identification practices which are recognised as being extremely important for interoperability and data quality. One current large programme, where identifier issues play major part, is Estonia's Computerised Census 2020. This involves close linking of

data from a number of base registers and several projects are underway that review and improve identifier systems in these registers.

3.2.3 Italy – The case of Agenzia per l'Italia Digitale

The Agenzia per l'Italia Digitale[l] has recently published a set of guidelines for achieving semantic interoperability in the public sector through Linked Open Data[li]. As part of their Linked Open Government Data roadmap, the agency is planning to gradually open-up data about administration, public contracts, geo-data and taxonomies, people with key positions in the public sector, public services and organisational units.

We have interviewed Dr. Giorgia Lodi and Mr. Antonio Maccioni to find out how the agency is dealing with persistent URI design rules and management for their LOGD. Mr. Maccioni said that the agency's URI strategy was inspired by the following two reference initiatives:

- Cool URIs for the Semantic Web; and
- Section 4.1 of Linked Data: Evolving the Web into a Global Data Space.

This is another indication of the great impact that these two initiatives have on URI design, which makes them important references.

3.2.3.1 URI format

Dr. Lodi, Mr. Maccioni and their team have specified URI structures for classes, properties, instances and individuals. The base URI for all resources is http://spcdata.digitpa.gov.it.

The URIs for classes must follow the following structure:

```
http://spcdata.digitpa.gov.it /{concept name}
```

Different values for {concept name} are possible depending on the type of the class, e.g. administration, organisational unit or public contract.

3.2.3.2 URI design rules and management

Unlike the suggestion of the UK guidelines, the Italian team decided not to use the government sector in the URI and refer directly to the concept name. According to Mr. Maccioni, removing the public administration's hierarchical structure from the URIs structure will increase their reusability across different sectors, which then fosters the vision of a unique global knowledge space known as the semantic web.

The URIs for properties must follow the following structure:

```
http://spcdata.digitpa.gov.it/{property name}
```

while the URIs for instance must be structured as follows:

```
http://spcdata.digitpa.gov.it/{concept name}/{natural key}
```

In this case, `{concept name}` refers to the class to which the instance belongs to and {natural key} denotes a unique alphanumeric identifier of the instance, e.g. the legal identifier of a public administration agency or a post code (depending on the nature of the instance).

A key design decision of the team is that all terms in a URI will be in Italian. This is expected to limit the complexity of the URIs and make them more understandable and easier to interpret especially among a non-technical audience. Of course, this is pursued by avoiding duplicate URIs while reusing external ontologies and vocabularies. In fact, when third party classes are used, class names are translated into Italian to be used as concept names. For instance, if the class `<http://purl.org/goodrelations/v1#ProductOrService>` is used for describing products (i.e. defined by natural keys prod-id1, prod-id2, etc.), the URIs of the products will be formed as follows:

```
<http://spcdata.digitpa.gov.it/Prodotto/prod-id1>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://purl.org/goodrelations/v1#ProductOrService> .

<http://spcdata.digitpa.gov.it/Prodotto/prod-id2>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://purl.org/goodrelations/v1#ProductOrService> ."
```

Finally, the URI structure for individuals is as follows:

```
http://spcdata.digitpa.gov.it/{individual name}
```

where `{individual name}` can for example be the name of a dataset.

Mr. Maccioni said that currently their persistent URI infrastructure does not support content negotiation but that this is part of their future plans.

3.2.4 Member States with no URI persistence policy

Despite not having in place yet an official policy on persistent URIs, a number of EU member States are aware of the importance of URI persistence and plan to create such a policy in the near future.

3.2.4.1 Sweden

Peter Krantz provided us with a comprehensive overview of initiatives in Sweden that deal with persistent URI design rules and management. Sweden has seen a small number of linked data projects such as Swedish Open Cultural Heritage (SOCH) K-samsök[lii] and one at the National Library of Sweden described in *Making a Library Catalogue Part of the Semantic Web*[liii]. The latter uses URIs that are similar to some of the other initiatives:

`http://libris.kb.se/resource/bib/{number}`

for bibliographic records and for authority records:

`http://libris.kb.se/resource/auth/{number}.`

Notice the type of either `'bib'` or `'auth'`. Dereferencing an example URI of `http://libris.kb.se/bib/7771917` gives either a Web page or RDF data depending on the user agent used (i.e. via content negotiation, see section 2.3). Although there are no (known) official Swedish government guidelines to refer to, Sveriges Domstolar, the Swedish Courts, have published guidance on publishing URIs for legal information[liv]. This paper repeats much of the advice we have already seen and cites many of the same references documents.

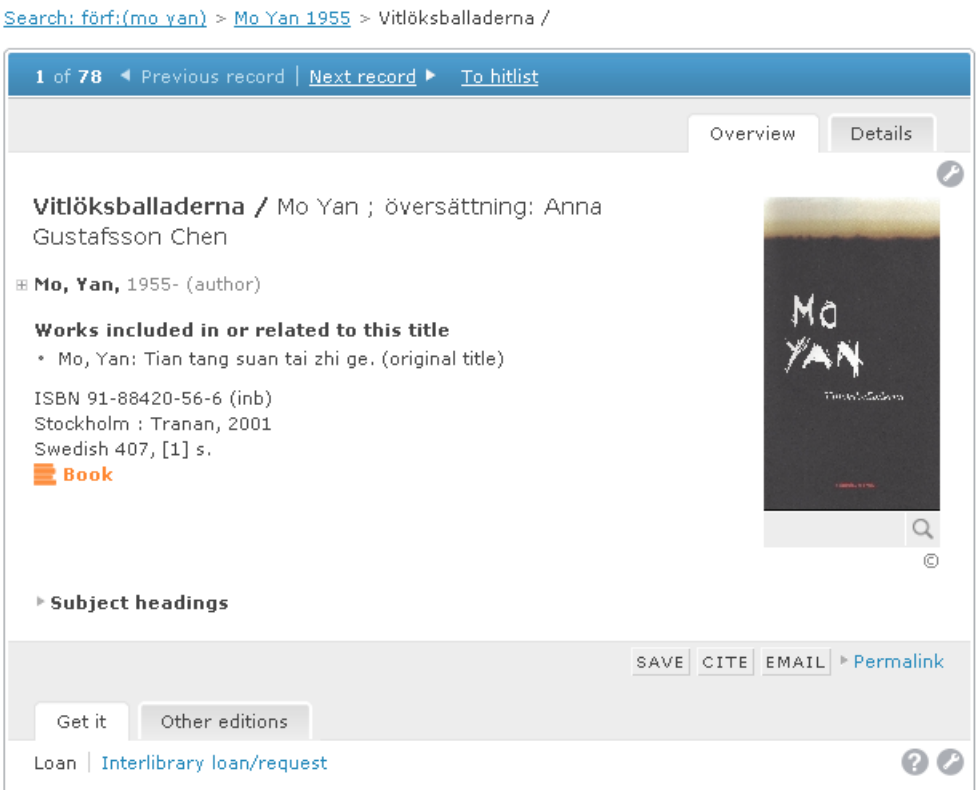

Figure 5 - Part of the HTML page returned from the National Library of Sweden's LIBRIS service at http://libris.kb.se/bib/7771917

Table 2 - Some of the RDF data returned from http://libris.kb.se/bib/7771917

```
<rdf:Description rdf:about="http://libris.kb.se/resource/bib/7771917">
    <dc:identifier rdf:resource="URN:ISBN:9188420566"/>
    <dc:description xml:lang="sv">Li:S</dc:description>
    <dc:publisher>Tranan</dc:publisher>
    <rdfs:isDefinedBy
rdf:resource="http://data.libris.kb.se/open/bib/7771917.rdf"/>
    <dc:title xml:lang="sv">Vitlöksballaderna</dc:title>
    <dc:description xml:lang="sv">Första svenska uppl.
2001</dc:description>
    <dc:date>2001</dc:date>
    <dc:creator
rdf:resource="http://libris.kb.se/resource/auth/205835"/>
    <bibo:isbn10>9188420566</bibo:isbn10>
    <dc:language
rdf:resource="http://purl.org/NET/marccodes/languages/chi#lang"/>
    <dc:creator>Mo, Yan, 1955-</dc:creator>
    <dc:creator>Yan Mo</dc:creator>
    <dc:language
rdf:resource="http://purl.org/NET/marccodes/languages/swe#lang"/>
    <dc:type>text</dc:type>
    <rdf:type rdf:resource="http://purl.org/ontology/bibo/Book"/>
    <dc:description xml:lang="sv">NB: En annan version av orig. titel,
se 740</dc:description>
    <rda:placeOfPublication
rdf:resource="http://purl.org/NET/marccodes/countries/sw#location"/>
  </rdf:Description>
```

3.2.4.2 Greece

Thodoris Papadopoulos of the Ministry of Administrative Reform and eGovernance, Greece provided us with a comprehensive overview of initiatives in Greece that deal with persistent URI design rules and management.

Greece does not yet have a formal policy on persistent URIs, nor any official open data published as linked data. Like Estonia, it does have some rules on Web site URLs. However, Greece still lacks Linked Data building blocks. For example, the list of Official Government Categories[lv] and the Unique Lexical Identifiers (Latin identification description strings) for most of the Greek Public Organizations[lvi] are published online but the URI in use is not designed for persistence:

`http://www.ermis.gov.gr/portal/page/portal/ermis/egcl?p_topic=perivallon_kai_fysikoi_poroi` (the Environment and Natural Resources public services).

> The above URI is tightly coupled to the particular portal implementation. This means that the URI will be impacted once the portal changes. With proper URI design, this should not happen as this change will impact all client applications using the portal.

However, the query string 'perivallon_kai_fysikoi_poroi' is perhaps a component of a future persistent URI scheme. It is noteworthy that the Greek authorities recognise the value in adopting a standards-based approach, an attitude that is being encouraged by non-governmental activities such as publicspending.gr. There is a willingness to participate in, for example, the ISA Programme and recognition that the design of URIs should take into account the following elements:

1. Longevity of URIs;
2. Human & Machine Readable;
3. Inclusion of formal Sectoral parts in URIs (e.g Health , Defence etc);
4. Control Scheme over new URIs (Central / Hierarchical / or per Organization);
5. Multilingual URIs;
6. Technology Abstraction (Content Negotiation).

These are the factors highlighted in *Designing URI Sets for the UK Public Sector* which again was cited in research for this document.

3.2.4.3 Finland

Finland does not have at the moment a national policy on persistent URIs. However, according to Ms. Kauhanen-Simanainen, Ministerial Advisor at the Finnish ministry of Finance, Finland is planning to develop common guidelines for this in the near future.

The National Library of Finland recognises that URIs should always be persistent, both when publishing linked data and when using links to refer to documents and other online resources. To ensure this, they argue that persistent identifiers, such as URNs, should be used as URIs[lvii]. In this case, mapping from a persistent identifier to current location(s) would be maintained centrally in a resolution service. On the contrary, adopting a Linked Data approach, the Finnish Research project FinnONTO, run by the Aalto University[lviii], suggests the use of HTTP URIs versus that of URNs and DOIs. The project has also delivered a tool for managing URIs, which is available through the ONKI service[lix].

3.2.4.4 Others

The authors understand that the subject of URI strategy is being discussed by the relevant authorities in the **Netherlands** but, as yet, nothing has been finalised or published. It is hoped that this document will prove useful in those discussions. In **Denmark**, all data is published in a format in which it is originally used which at the moment doesn't include any linked data. Therefore the subject of URI persistence has not yet been discussed. Likewise the **Czech Republic** has not yet considered the issue. A policy is under development by CTIC[lx] on behalf of the government in **Spain** and follows the UK pattern closely.

3.3 Standardisation bodies and other initiatives

This section reports on the persistent URI policies of standardisation bodies and other similar initiatives.

3.3.1 Dublin Core Metadata Initiative

The Dublin Core Metadata Initiative has a very clear set of policies around URI design and persistence, set out on its Web site[lxi]. It publishes 4 vocabularies, each one with its own namespace.

3.3.1.1 URI format

The URI pattern is shown below:

```
http://purl.org/dc/{vocabulary}/
```

More information is shown in Table 3. Individual terms within those vocabulary follow the final / character. In common with recognised best practice, terms are written in camel case and classes begin with uppercase letters, properties with lower case.

Namespace	Notes
http://purl.org/dc/terms/	All DCMI properties, classes and encoding schemes. This is best known and widely used namespace
http://purl.org/dc/dcmitype/	Classes in the DCMI Type Vocabulary
http://purl.org/dc/dcam/	Terms used in the DCMI Abstract Model
http://purl.org/dc/elements/1.1/	The Dublin Core Metadata Element Set, Version 1.1. This was the namespace of the original 15 elements.

Table 3 - The DCMI namespace

3.3.1.2 URI design rules and management

Two things stand out from the table above. First of all it is interesting to note that the original DC Elements namespace included the version number. A version number also occurs in the (equally widely used) FOAF namespace (`http://xmlns.com/foaf/0.1/`) despite the specification now being on version 0.98[lxii]. Both FOAF and DCMI date from the earliest days of the Semantic Web and so the inclusion of version numbers is an indication of this. At the time, the assumption was that as new versions of the vocabulary came out, new namespaces would need to be declared so that the specific semantics of any term could be updated without affecting previous versions that were already in use. That sounds eminently sensible on paper.

However, the implication is that an application built today should use today's namespace - i.e. the latest one available - whilst older applications and data would be based on whatever was current at the time, even if the two versions of the vocabulary included the same terms. One might have seen modern day applications using:

```
http://purl.org/dc/elements/1.5/title
```

whilst older ones would use

```
http://purl.org/dc/elements/1.1/title
```

even though the meaning of `'title'` has not changed between the versions.

DCMI's move to a namespace without a version number avoided this unhelpful URI proliferation happening, as has FOAF's continued use of '0.1' in its namespace long after it has lost its meaning. Once a term is defined in the Dublin Core namespace, it is subject to extremely strict change control. Quoting from the policy:

Changes of definitions … will be reflected in the affected DCMI recommendation and/or DCMI term declaration. If, in the judgment of the DCMI Directorate, such changes of meaning are likely to have substantial impact on either machine processing of DCMI terms or the functional semantics of the terms, then these changes will be reflected in a change of URI for the DCMI term or terms in question. The URIs for any new DCMI namespaces resulting from such changes will conform to the DCMI namespace URI pattern defined above.

In other words, unless the change in definition is trivial and is very unlikely to affect running applications, a new definition means a new term with its own URI and the old one will persist.

The second thing to notice from Table 3 is DCMI's use of persistent URLs (purls) at `purl.org`.

Some historical context is useful here. OCLC, the Online Computer Library Center, is both the organisation behind `purl.org` and was the host from 1995 to 2008. In that sense one can think of Dublin Core and `purl.org` as having a common ancestry in OCLC. Whether DCMI would have used `purl.org` without that common ancestry is unknowable. The idea behind it is that it provides stable, persistent URIs that can be used even if the resources they ultimately resolve to move. The rationale for using purl.org is that it is generally easier to hand on a service, like purl.org, than it is an organisational domain such as dublincore.org. If required, the purl.org service is easier to be taken on by another organisation. The use of purl.org therefore can be seen as a clear indication of the intention that the URIs will persist for as long into the future as it can reasonably be foreseen. The question of whether DCMI needed to use purl.org arises since its own Web site, dublincore.org, is itself designed for long term stability.

That said, the separation of namespace and definition that purl.org provides does allow DCMI to move things around on their site, publish updated schemas at new URIs and so on without affecting the rock solid stability of URIs like http://prul.org/dc/terms/creator on which so many applications depend. .

`purl.org` itself is used by many organisations and therefore a lot of people regard it as important. It is that diversity of interest that is the best guarantor of `purl.org`'s continued existence, even if ownership changes in future. Nevertheless, it is a service like any other and as its usage continues to grow, so does the cost of running it.

3.3.2 W3C

The W3C operates a strict policy with regard to the creation of its URIs. Certain sections of w3.org are very tightly controlled and there is a policy stating that once a URI has been minted, it should never cease to exist. Team members and document editors are generally authorised to publish and edit documents but do not have sufficient privileges to delete them.

3.3.2.1 URI format

Formally published documents, particularly documents that are evolving into standards, are all published at:

`http://www.w3.org/TR/{shortname}`

where `{shortname}` is something easy to remember like 'mobile-bp' (Mobile Web Best Practices) or 'vocab-org' for the Organisation Ontology. This will be the 'latest version URI' - i.e. the specific document returned will change as the document evolves (see section 2.5).

3.3.2.2 URI design rules and management

Individual documents have URIs in the form: `http://www.w3.org/TR/{status}-{shortname}-{yyyymmdd}` and so include an indication of the status of the document and the date on which it was published. Other parts of the `w3.org` namespace are covered by *URIs for W3C Namespaces*[lxiii]. It is noteworthy that the policy strongly encourages the publication of documents in 'dated space', i.e. `http://www.w3.org/{year}/{month}/`.

On a typical working day, 50 documents are created with URIs of this form, mostly minutes of meetings which are preserved in multiple formats and mostly with filenames beginning with the day of the month, for example the minutes of the eGov Interest Group meeting held on 19th October are at `http://www.w3.org/2012/10/19-egov-minutes.html`. This simple use

of the date and group name as part of the URI helps ensure uniqueness with little effort and is a big aid to persistence. The creation of documents outside dated space and the special area reserved for namespace documents (`http://www.w3.org/ns/{shortname}`) is rare and requires specific authorisation from W3C management.

As well as operating a well-defined policy of URI creation, W3C also has a published policy on URI persistence[lxiv]. The policy takes the form of a pledge by the three host organisations (MIT, ERCIM and Keio University):

- The hosts will ensure that persistent resources continue to be available throughout the life of the Consortium;
- Where a persistent resource is modified, a change history will be archived though the archive will *not* necessarily be available publicly;
- Should the W3C be disbanded, then any Web site will be granted the right to make a copy (at a different URI) of all public persistent resources so long as they are not modified and are preserved in their entirety and made available free of charge, and provided the same persistence policy is applied to these "historical mirrors." In such event, the original `http://www.w3.org` web site will be handed over for management to another organization only if that organization pledges to this policy or one considered more persistent.

The policy was written personally by Tim Berners-Lee' who sums it up as: "The intent is to set an example by reducing the failure of links due to clumsy management or inadequate commitment to information persistence, and to provide a stable reference base of information about W3C-related topics as a service to the community."

This combination of careful URI management and a stated persistence policy, including provision for what should happen should the organisation cease to exist, means that the community is right to have great confidence that resources on w3.org will persist for a long time. Although not covered by the persistence policy, the W3C mailing list archives are also managed and maintained for the long term. The very first archived mail from 28th October 1991, less than 3 months after the original WWW software was released, is still available online[lxv].

3.4 Others

This section reports on the persistent URI policies of pioneers around the globe.

3.4.1 Data.gov

In November 2010, Data.gov announced that some of its datasets would be available also as Linked Data[lxvi].

3.4.1.1 URI format

The URI template of data.gov, which is in line with that of data.gov.uk (in fact it was designed based on the latter), is:

```
http://' BASE '/' 'id' '/' ORG '/' CATEGORY ( '/' TOKEN )+
```

where `id` is used for identifying non information resources; `ORG is` a short token for representing the agency, government, or organization that controls the identfier space; and `CATEGORY` and `TOKEN` identify the specific instance.

For example, in the case of US Government Agencies, the suggested URI template is:

```
http://BASE/id/us/fed/agency/NAME/SUBNAME
```

In that case the URI for the National Ocenaic and Atomsphairic Administration would be:

```
http://BASE/id/us/fed/agency/Commerce/National_Oceanic_and_Atmospheric
_Administration
```

3.4.1.2 URI design rules and management

Jim Hendler's team at RPI[lxvii] undertook the task to design the URIs for data.gov. Citing from their website[lxviii], the data.gov URIs are:

- **easily re-hosted**, meaning that the BASE URI can easily be transformed from one namespace to another, to facilitate the buy-in from government agencies.
- **Concise**, and
- **Cross-domain**, spanning from governmental agencies and zip codes to congressional districts and EPA[lxix] facilities.

3.4.2 Australian National Data Service (ANDS)

In its own words, "ANDS is building the Australian Research Data Commons: a cohesive collection of research resources from all research institutions, to make better use of Australia's research data outputs."

3.4.2.1 URI format

It is an effort to manage its research output for easier discovery and re-use and, as part of this, it runs a service called 'Identify My Data' that acts as a Handle Service. On request from an authorised user (typically a researcher in an Australian scientific institute), it issues a Handle[lxx] of the form:

```
102.100.100/nn
```

where the sequence 102.100.100 is made up of '102' (Australia) dot '100' (e-research) dot '100' (ANDS), followed by a sequential number (*nn*). It is for the individual researcher to then

associate that Handle with metadata about the resource which will typically include its location on the Web. Handles issued by ANDS can be resolved using handle.net, for example, `http://hdl.handle.net/102.100.100/15`.

3.4.2.2 URI design rules and management

The documentation of this system is spread across three increasingly detailed documents:

- Persistent identifiers (awareness level)[lxxi] provides a non-technical high level view of the topic of persistent identifiers.
- Persistent Identifiers (working level)[lxxii] provides a good deal of the background to the subject and surveys the various options for creating and maintaining persistent identifiers.
- Persistent Identifiers (expert level)[lxxiii] is the detailed technical documentation.

The important aspect for the current discussion is that this is a dedicated service established to issue persistent identifiers to the Australian research community. ANDS is committed to maintain the Persistent Identifiers for *at least* twenty years[lxxiv] and is organised in such a way that it could readily continue well beyond that time frame. The two-part nature of the service means that the researcher him/herself also needs to maintain the data and to inform the 'Identify My Data' service of any changes (which can be done programmatically). They can move, update or delete their data but this is independent of the ANDS service.

The up side of this arrangement is that the identifiers are very likely to persist in terms of their structure, uniqueness and meaning, even when a given researcher moves on to new topics. The downside is that by shifting some of the responsibility for maintenance to an outside agency, researchers may be less motivated to maintain their data despite this being an important aspect of the system and so it could be argued that the problem of link rot[lxxv] is not really solved.

> To ameliorate the URI policy explained above, users of 'Identify My Data' are strongly encouraged to provide *authority metadata*, i.e. information about the identifier itself, who has/had responsibility for its creation and the object it identifies. ANDS recommends that contact information is provided and tied to a role, not an individual, to increase the likelihood of future discovery of the responsible researcher.

The final line of the policy statement is illuminating:

"Because authority metadata is used when things go wrong, its availability should not be reliant on external systems: failure to access an external system may be why things have gone wrong to begin with. Contact data should therefore be stored directly in the identifier record, rather than linked through some external database."

This makes perfect sense but it emphasises the fact that the 'Identify My Data' service is

designed and managed for a non linked data environment. Indeed, the term linked data doesn't occur anywhere in the relevant ANDS documentation. This is a service for assigning and managing persistent identifiers for resources that exist elsewhere, rather than a service for managing URIs as first class citizens of the datasphere.

3.4.3 Europeana

Europeana, a comprehensive and growing portal to Europe's cultural heritage collections, was launched in November 2008. At that time it only used URIs as identifiers for the *records about* the millions of items held in cultural heritage collections across Europe. A pilot project that was described in a paper at the 2011 Dublin Core Conference[lxxvi] and released to the public in February 2012 made linked open data available about 2.4 million items.

3.4.3.1 URI format

The original record URIs were of the form:

```
http://www.europeana.eu/resolve/record/{collectionID}/{itemID}
```

and to create identifiers for the objects themselves, the project decided to use the related pattern:

```
http://data.europeana.eu/item/{collectionID}/{itemID}
```

As noted in section 3.2.1.1, this in line with the UK's *Designing URI Sets for the UK Public Sector* which is again cited as a reference.

3.4.3.2 URI design rules and management

However, during the pilot project a problem came to light. Europeana assigns URIs to records as those records are ingested from the relevant cultural heritage institutions. Individual items are assigned a hexadecimal number at the time of ingestion, a number that is incremented as each new item is added. This makes perfect sense as it is a largely automated system handling millions of records. However, during the linked open data pilot, the data from some collections was re-ingested and this generated a different set of identifiers. Projects and applications that had depended on the original URIs had to update their systems. The PATHS Project[lxxvii] as one such.

> It is evident that automated URI minting systems like this must have a means of avoiding such faux-pas in future.

There are only two realistic ways of doing this:

1. advise original data providers to include identification fields that are themselves persistent and that can therefore be referred to when the secondary system ingests the data;
2. perform some data matching, i.e. find resources where most data is identical and then use the 'old' URI. Sadly this is error-prone.

The second option is most likely to apply to Europeana where efforts are being made to define a persistence policy but even so, some sort of DOI or ARK-based system is likely to be necessary to perform the matching process reliably.

The lack of integration between the production system and the Linked Data pilot at Europeana is a continuing source of resistance to the wholesale use of persistent URIs since the URIs in the production system are inherently ephemeral.

3.4.4 Wikipedia: Avowedly non persistent URIs

A discussion of persistent URIs should include at least one example that is avowedly not persistent. Wikipedia provides just such an example.

3.4.4.1 URI format

A page name in Wikipedia may begin with a namespace prefix – a string (ending with a colon) which the MediaWiki software recognizes as placing a page in a particular namespace[lxxviii]. If the page name does not begin with any of the recognised prefixes, then it is considered to be in the main namespace. A full page name therefore takes one of the following forms:

- BaseName (for pages in the main namespace);
- NamespacePrefix:BaseName (for pages in any other namespace).

3.4.4.2 URI Design

URIs are created when new pages are created in Wikipedia. This process is automatic and is based on the page title or label. Since labels are inherently ambiguous, a commonly encountered feature of Wikipedia is its disambiguation pages such as:

`http://en.wikipedia.org/wiki/Europe_(disambiguation)`

At the time of writing, there are 18 pages all about 'Europe' - everything from the physical continent via several references from Greek Mythology to the national anthem of the republic of Kosovo. As the Wikimedia Foundation's notes on the subject[lxxix] make clear, the issue is well understood:

"Note that the URLs for pages in Wikipedia are not persistent (although they are quite persistent, see e.g. Siorpaes/Hepp research on this issue). But still, they can change meaning: if a Kevin Smith should become US president some day, he will most certainly replace the author Kevin Smith from his place as the main topic of the Kevin Smith article. Also changes in a name - e.g. when a person marries or becomes pope - lead to changes of the URL. There will be adequate redirects and disambiguations in most cases, but although these are easy to follow and disambiguate for humans, this is not necessarily true for machines."

Wikipedia is a resource created by humans for humans. The derived linked data version, DBpedia, was a critical aspect of the development of linked data and has brought the issue of URI persistence into focus as shown further down the page from which the quote was taken. In short: URIs need to become more persistent, especially now that the Wiki Media Foundation has begin its new WikiData project is underway. This project marks a move from the human-centric service to one that both humans and machines can use, making URI management all the more critical.

4 Recommended URI design and management principles

Having reviewed several case studies from the public sector and the key technical considerations, it is now possible to derive a set of best practices. The foundations of the best practices have not changed since the earliest days of the Web, however, experience has allowed their refinement and evolution. Table 4 offers a guide to the sources that document this evolution.

Status	Title	Authors and Date
Background	Cool URIs don't change	Tim Berners-Lee, 1998
	Cool URIs for the Semantic Web	Leo Saurman, Richard Cyganiak, 2008
	Linked Data	Tim Berners-Lee, 2009
Key Source	Designing URI Sets for the UK Public Sector	UK Chief Technology Officer Council October 2009
Expansion	Creating Linked Data	Jeni Tennison, 2009
	Linked Data: Evolving the Web into a Global Data Space	Tom Heath & Christian Bizer, 2011
	Linked Data Patterns	Leigh Dodds & Ian Davis, 2012
	Best Practices for Multilingual Linked Open Data	Jose Emilio Labra Gayo, 2012
Detail	Statistical Linked Dataspaces	Sarven Capadisli, 2012

Table 4 - Published sources of information related to URI persistence

Figure 6 below summarises the 10 DOs and DON'Ts of persistent URI design rules and management, which are detailed in the remainder of this chapter.

Figure 6 - The 10 Dos and DONTs for persistent URIs

4.1 Recommended URI format

The recommended pattern for a URI designed for persistence is:

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

This comes originally from *Designing URI Sets for the UK Public Sector* and has been repeated successfully with little variation in many different scenarios. A full explanation is provided in section 3.2.1.1 and can be summarised as:

- `{domain}` is a combination of the host and the relevant sector. It is a matter of choice whether the sector is defined as a sub-domain of the host or as the first component of the path.
- `{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, such as 'authority' (Publications Office, 3.1) or 'dcterms' (DCMI, 3.3.1).
- `{concept}` might be a collection (Europeana 3.4.3), the type of real world object identified (e.g. road, 3.2.1.1), the name of the concept scheme (e.g. 'language' 3.1);
- `{reference}` is a specific item, term or concept.

4.2 Recommended URI design principles

4.2.1 Avoid stating ownership

The URI template above does not include the name of the organisation or project that minted the URI. This makes it much less susceptible to change should the project end or the organisation be merged or renamed.

4.2.2 Avoid version numbers

Although concept schemes, ontologies, taxonomies and vocabularies are likely to go through iterative cycles of change, version numbers and status information should not be included in the URIs. Rather, the URIs should remain stable between versions and new ones minted for new terms. URIs may be deprecated and their use discouraged but they should nevertheless be maintained both in terms of the actual URI and the resource they identify.

4.2.3 Re-use existing identifiers

Where resources are already uniquely identified, those identifiers should be incorporated into the URI. For example, if schools are assigned integer identifiers, the URI for the school with identifier 123456, could be:

```
http://education.data.example/id/school/123456
```

Caution: when re-using an identifier, it is essential to re-use them without changing the original semantics. For example, a URI for a vehicle *licence* is not an identifier for the *vehicle* itself. Furthermore, it is important to only re-use identifiers that themselves are likely to be persistent.

4.2.4 Avoid using auto-increment

Minting new URIs for large datasets will need to be automated and the process must be guaranteed to produce unique identifiers. One way to do this might be to simply increment a counter as each new URI is minted. Imagine that in the example of the previous section the integer URIs given to schools were given based on such a counter. In that case, the following could be possible URIs for two different schools.

```
http://education.data.gov.uk/id/school/123456
http://education.data.gov.uk/id/school/123457
```

Although this approach is perfectly feasible, we would recommend it *only if* one of the following is true:

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

4.2.5 Avoid query strings

Query strings (e.g. `'?param=value'`) are usually used in URIs as keys to look up terms in a database. This is brittle since it often relies on a particular implementation. We recommend conforming to the URI template of section 0 and where necessary, user server configuration to interpret URIs that conform to the template.

4.2.6 Avoid file extensions

For similar reasons to the previous point, avoid file extensions in persistent URIs, particularly those that stem from the technology used such as .php or .py.

4.3 Design and build for multiple formats

A persistent URI should identify a conceptual resource. Where that resource is an information resource - that is, something that can be transmitted as a stream of bytes - then different user agents should be able to access it in different formats. In particular, humans and machines should be able to access it in formats appropriate to their different needs. Typically this will mean that a resource such as

`http://data.example.org/doc/foo/bar`

can be returned in at least HTML and some serialisation of RDF.

Those specific representations of the resource should have their own URI and this should follow a predictable pattern, the simplest of which is to add the relevant file extension, i.e.

`http://data.example.org/doc/foo/bar.html` and
`http://data.example.org/doc/foo/bar.rdf`

This does not break the best practice set out in 4.2.6 since these are not the persistent URIs (http://data.example.org/doc/foo/bar is).

4.3.1 Link multiple representations

Multiple representations of the same resource should all link to each other using a suitable method. In HTML use a link element with the ‘rel’ value of 'alternate', in RDF use dcterms:hasFormat etc.

4.4 Implement 303 redirects for real-world objects

When de-referenced, URIs that identify real world objects that cannot be transmitted as a series of bytes (such as buildings, places and people) should redirect using HTTP response code 303 to a document that describes the object. This should be done in a consistent manner that can be written as a URI re-write rule, typically replacing the URI `{type}` of 'id' with 'doc.' See section 3.2.1.2 for details.

4.5 Use a dedicated service

Without exception, all the use cases discussed in section 3 where a policy of URI persistence has been adopted, have used a dedicated service that is independent of the data originator. The Australian National Data Service uses a handle resolver, Dublin Core uses purl.org, services, data.gov.uk and publications.europa.eu are all also independent of a specific government department and could readily be transferred and run by someone else if necessary. This does not imply that a single service should be adopted for multiple data providers. On the contrary - distribution is a key advantage of the Web. It simply means that the provision of persistent URIs should be independent of the data originator.

Multiple, small scale services that are easily transferable, rather than a single point of failure, plus a continued demand for the service, are the greatest guarantors of persistence.

5 References

[i] http://www.ietf.org/rfc/rfc2396.txt
[ii] A URI that can be looked up directly with a user agent, such as a Web browser
[iii] Peristeras V., Loutas N., Goudos S., Tarabanis K.: A Conceptual Analysis of Semantic Conflicts in Pan-European E-Government Services. In Journal of Information Science, vol. 34 (6), pp. 877-891, 2008
[iv] http://www.doi.org/
[v] http://tools.ietf.org/html/draft-kunze-ark-17
[vi] http://www.w3.org/TR/webarch/
[vii] http://www.w3.org/Provider/Style/URI.html
[viii] www.eia.org.uk/2010conf/Talk-Schmitz.ppt
[ix] http://www.ifla.org/publications/functional-requirements-for-bibliographic-records
[x] http://europa.eu/rapid/press-release_IP-12-1040_en.htm
[xi] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:168:0041:01:EN:HTML
[xii] https://webgate.ec.europa.eu/CITnet/confluence/display/PURI/Home
[xiii] European Commission, Proposal from the informal Working Group on Persistent URIs, November 2012.
[xiv] http://eurovoc.europa.eu/drupal/?q=download/subject_oriented&cl=en
[xv] http://www.w3.org/TR/cooluris/
[xvi] http://latc-project.eu/
[xvii] http://www.deri.ie/
[xviii] http://eurostat.linked-statistics.org/
[xix] http://www.w3.org/DesignIssues/LinkedData.html
[xx] http://latc-project.eu/node/111
[xxi] http://csarven.ca/statistical-linked-dataspaces
[xxii] http://www.multilingualweb.eu/en/documents/dublin-workshop/dublin-program
[xxiii] http://en.wikipedia.org/wiki/File:Screenshot_data.gov.uk_homepage_april_2010.jpg
[xxiv] http://www.cabinetoffice.gov.uk/sites/default/files/resources/designing-URI-sets-uk-public-sector.pdf
[xxv] Choosing the right domain for URIs, p5 onwards
[xxvi] http://www.cabinetoffice.gov.uk/sites/default/files/resources/CM8353_acc.pdf (p 25 and Annex A)
[xxvii] e.g. http://opencorporates.com/companies/gb/04285910
[xxviii] http://data.gov.uk/blog/guest-post-developers-guide-linked-data-apis-jeni-tennison
[xxix] http://www.jenitennison.com/blog/node/136
[xxx] http://theodi.org/people/jeni
[xxxi] http://patterns.dataincubator.org/book/
[xxxii] http://www.w3.org/TR/cooluris/
[xxxiii] http://www.w3.org/Provider/Style/URI
[xxxiv] http://www.data.gov/communities/node/116/forums/topic/207
[xxxv] http://linkeddatabook.com/editions/1.0/

[xxxvi] http://www.w3.org/2001/tag/issues.html#httpRange-14
[xxxvii] http://www.w3.org/wiki/HttpRange14Webography
[xxxviii] http://www.w3.org/2001/tag/awwsw/issue57/latest/
[xxxix] http://www.mkbergman.com/994/give-me-a-sign-what-do-things-mean-on-the-semantic-web/
[xl] http://lists.w3.org/Archives/Public/www-tag/2005Jun/0039.html
[xli] http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html#sec10.3.4
[xlii] http://www.riso.ee/et/koosvoime/web-framework.odt
[xliii] http://www.riso.ee/et/koosvoime/interoperability-framework.odt
[xliv] https://riha.eesti.ee/riha/main/inf/maakataster
[xlv] www.maaamet.ee
[xlvi] https://riha.eesti.ee/riha/main/inf/riigi_kohanimeregister
[xlvii] http://geoportaal.maaamet.ee/est/Teenused/Kohanimeregistri-teenus-p133.html
[xlviii] https://riha.eesti.ee/riha/main/inf/riigi_teataja
[xlix] http://opendata.riik.ee/juhendid/juhend-avaandmete-avaldamiseks-ning-avaandemete-portaali-kasutamiseks Ministry of Economic Affairs and Communications, 2011, in Estonian
[l] http://www.digitpa.gov.it/
[li] http://www.digitpa.gov.it/sites/default/files/allegati_tec/CdC-SPC-GdL6-InteroperabilitaSemOpenData_v2.0_0.pdf
[lii] http://www.ksamsok.se/
[liii] http://www.kb.se/dokument/Libris/artiklar/Project%20report-final.pdf
[liv] http://dev.lagrummet.se/dokumentation/system/uri-principer.pdf
[lv] http://www.ermis.gov.gr/
[lvi] http://diavgeia.gov.gr/en
[lvii] http://www.nationallibrary.fi/publishers/urn.html
[lviii] http://www.seco.tkk.fi/projects/finnonto/
[lix] http://onki.fi/
[lx] http://www.fundacionctic.org/
[lxi] http://dublincore.org/documents/dcmi-namespace/
[lxii] http://xmlns.com/foaf/spec/
[lxiii] http://www.w3.org/2005/07/13-nsuri
[lxiv] http://www.w3.org/Consortium/Persistence
[lxv] http://philarcher.org/diary/2011/20yearsofmlarchives/
[lxvi] http://www.data.gov/communities/node/116/forums/topic/207
[lxvii] http://logd.tw.rpi.edu/
[lxviii] http://logd.tw.rpi.edu/instance-hub-uri-design
[lxix] http://www.epa.gov/
[lxx] http://www.ietf.org/rfc/rfc3650.txt
[lxxi] http://ands.org.au/guides/persistent-identifiers-awareness.html
[lxxii] http://ands.org.au/guides/persistent-identifiers-working.html
[lxxiii] http://ands.org.au/guides/persistent-identifiers-expert.html
[lxxiv] http://www.ands.org.au/services/pid-policy.html, see section 4.6
[lxxv] http://en.wikipedia.org/wiki/Link_rot

[lxxvi] http://dcpapers.dublincore.org/pubs/article/download/3625/1851
[lxxvii] http://paths-project.eu/
[lxxviii] http://en.wikipedia.org/wiki/Wikipedia:Page_name#Namespace_and_base_name
[lxxix] http://meta.wikimedia.org/wiki/Wikidata/Notes/URI_scheme

Acknowledgements

The authors would like to thank the following people for their valuable contribution to this study (in alphabetical order):

Martin Alvarez-Espinar (CTIC, Spain), Adam Arndt (Danish Agency for Digitisation), Sarven Capadisli (DERI, NUI Galway), Makx Dekkers (formerly of DCMI), Giorgos Georgiannakis (DG SANCO, European Commission), Michael Hausenblas (DERI, NUI Galway), Josef Hruška (Ministry of Interior, Czech Republic), Aftab Iqbal (DERI, NUI Galway), Antoine Isaac (Europeana), Anne Kauhanen-Simanainen (Ministry of Finance, Finland), Peter Krantz (eGov Consultant, Sweden), Giorgia Lodi (Agenzia per l'Italia Digitale, Italy), Antonio Maccioni (Agenzia per l'Italia Digitale, Italy), Thodoris Papadopoulos (Ministry of Administrative Reform and eGovernance, Greece), Priit Parmakson (Estonian Information Systems Authority), Paul Suijkerbuilk (data.overheid.nl, The Netherlands).