W3C Working Group Note

Best Practices for Publishing Linked Data

W3C Working Group Note 09 January 2014

This version:
http://www.w3.org/TR/2014/NOTE-ld-bp-20140109/
Latest published version:
http://www.w3.org/TR/ld-bp/

Editors:
Bernadette Hyland, 3 Round Stones, Inc.
Ghislain Atemezing, EURECOM
Boris Villazón-Terrazas, iSOCO, Intelligent Software Components S.A.

Abstract

This document sets out a series of best practices designed to facilitate development and delivery of open government data as Linked Open Data. Linked Open Data makes the World Wide Web into a global database, sometimes refered to as the "Web of Data". Using Linked Data Principles, developers can query Linked Data from multiple sources at once and combine it without the need for a single common schema that all data shares. Prior to international data exchange standards for data on the Web, it was time consuming and difficult to build applications using traditional data management techniques. As more open government data is published on the Web, best practices are evolving too. The goal of this document is to compile the most relevant data management practices for the publication and use of of high quality data published by governments around the world as Linked Open Data.

Status of This Document

This section describes the status of this document at the time of its publication. Other documents may supersede this document. A list of current W3C publications and the latest revision of this technical report can be found in the W3C technical reports index at http://www.w3.org/TR/.

This document was published by the Government Linked Data Working Group as a First Public Working Group Note. If you wish to make comments regarding this document, please send them to public-gld-comments@w3.org (subscribe, archives). All comments are welcome. Since the Working Group's charter is ending, the group might not officially respond to comments, but individual members may. As usual, comments are publicly archived, available to both readers and any group updating this document in the future.

Publication as a Working Group Note does not imply endorsement by the W3C Membership. This is a draft document and may be updated, replaced or obsoleted by other documents at any time. It is inappropriate to cite this document as other than work in progress.

This document was produced by a group operating under the 5 February 2004 W3C Patent Policy. W3C maintains a public list of any patent disclosures made in connection with the deliverables of the group; that page also includes instructions for disclosing a patent. An individual who has actual knowledge of a patent which the individual believes contains Essential Claim(s) must disclose the information in accordance with section 6 of the W3C Patent Policy.

Table of Contents

Audience

Readers of this document are expected to be familiar with fundamental Web technologies such as HTML, URIs, and HTTP. The document is targeted at developers, government information management staff, and Web site administrators.

Scope

Linked Data refers to a set of best practices for publishing and interlinking structured data for access by both humans and machines via the use of the RDF (Resource Description Framework) family of standards for data interchange [RDF-CONCEPTS] and SPARQL for query. RDF and Linked Data are not synonyms. Linked Data however could not exist without the consistent underlying data model that we call RDF [RDF-CONCEPTS]. Understanding the basics of RDF is helpful in leveraging Linked Data.

Background

In recent years, governments worldwide have mandated publication of open government content to the public Web for the purpose of facilitating open societies and to support governmental accountability and transparency initiatives. In order to realize the goals of open government initiatives, the W3C Government Linked Data Working Group offers the following guidance to aid in the access and re-use of open government data. Linked Data provides a simple mechanism for combining data from multiple sources across the Web. Linked Data addresses many objectives of open government transparency initiatives through the use international Web standards for the publication, dissemination and reuse of structured data.

Summary of Best Practices

The following best practices are discussed in this document and listed here for convenience.

STEP #1 PREPARE STAKEHOLDERS:
Prepare stakeholders by explaining the process of creating and maintaining Linked Open Data.

STEP #2 SELECT A DATASET:
Select a dataset that provides benefit to others for reuse.

STEP #3 MODEL THE DATA:
Modeling Linked Data involves representing data objects and how they are related in an application-independent way.

STEP #4 SPECIFY AN APPROPRIATE LICENSE:
Specify an appropriate open data license. Data reuse is more likely to occur when there is a clear statement about the origin, ownership and terms related to the use of the published data.

STEP #5 GOOD URIs FOR LINKED DATA:
The core of Linked Data is a well-considered URI naming strategy and implementation plan, based on HTTP URIs. Consideration for naming objects, multilingual support, data change over time and persistence strategy are the building blocks for useful Linked Data.

STEP #6 USE STANDARD VOCABULARIES:
Describe objects with previously defined vocabularies whenever possible. Extend standard vocabularies where necessary, and create vocabularies (only when required) that follow best practices whenever possible.

STEP #7 CONVERT DATA:
Convert data to a Linked Data representation. This is typically done by script or other automated processes.

STEP #8 PROVIDE MACHINE ACCESS TO DATA:
Provide various ways for search engines and other automated processes to

access data using standard Web mechanisms.

STEP #9 ANNOUNCE NEW DATA SETS:
Remember to announce new data sets on an authoritative domain. Importantly, remember that as a Linked Open Data publisher, an implicit social contract is in effect.

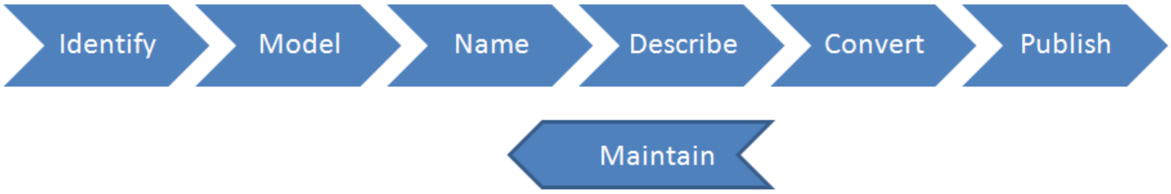
STEP #10 RECOGNIZE THE SOCIAL CONTRACT:
Recognize your responsibility in maintaining data once it is published. Ensure that the dataset(s) remain available where your organization says it will be and is maintained over time.

1. Prepare Stakeholders

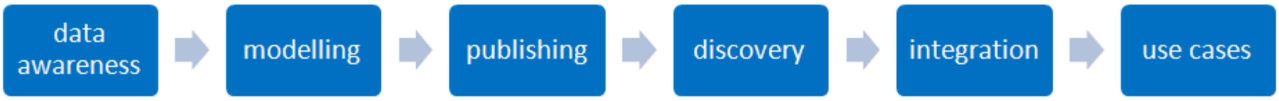
Preparation is crucial for success of an information management project. Sharing with government stakeholders the benefits of data sharing in terms of their agency mission or charter helps ensure success. The concepts of data modeling will be familiar to information management professionals. While the specifics of Linked Open Data may be new to people who are used to traditional information manaagement approaches, they are well-documented in W3C Recommendations, Notes and many peer reviewed publications [WOOD2013], [howto-lodp], [BHYLAND2011], [BVILLAZON]. Linked Data has entered the mainstream and is used by governments around the world, major search engines, international corporations and agile startups.

To help prepare stakeholders, we've included three life cycle models, however it is evident that they all share common (and sometimes overlapping) activities. For example, they all identify the need to specify, model and publish data in standard open Web formats. In essence, they capture the same tasks that are needed in the process, but provide different boundaries between these tasks. One workflow is not better than another, they are simply different ways to visualize a familiar information management process.

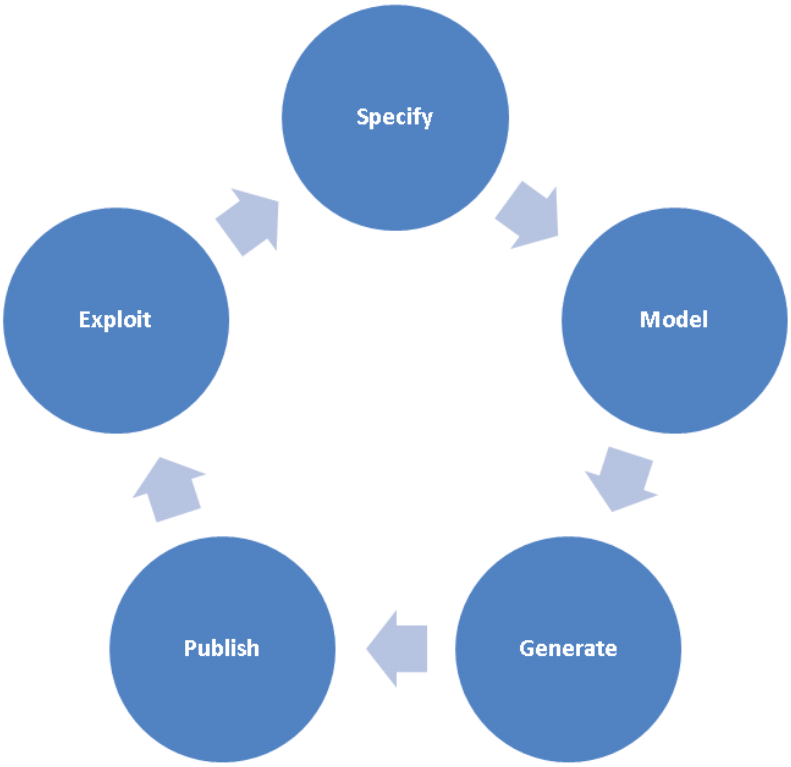
- Hyland et al. [BHYLAND2011] provide a six-step “cookbook” to model, create, publish, and announce government linked data. They highlight the role of the World Wide Web Consortium (W3C) which is currently driving specifications and best practices for the publication of governmental data. Hyland et al. lifecycle consists of the following activities: (1) Identify, (2) Model, (3) Name, (4) Describe, (5) Convert, (6) Publish, and (7) Maintain.



- According to Hausenblas et al. [HAUSENBLAS] existing data management approaches assume control over schema, data and data generation, which is not the case in the Web because it is open, de-centralized environment. Based on their experience in Linked Data publishing and consumption over the past years, they identify involved parties and fundamental phases, which provide for a multitude of so called Linked Data life cycles that consist of the following steps: (1) data awareness, (2) modeling, (3) publishing, (4) discovery, (5) integration, and (6) use cases.



- Villazón-Terrazas et al. propose in [BVILLAZON] a first step to formalize their experience gained in the development of government Linked Data, into a preliminary set of methodological guidelines for generating, publishing and exploiting Linked Government Data. Their life cycle consists of the following activities: (1) Specify, (2) Model, (3) Generate, (4) Publish, and (5) Exploit.



2. Select a Dataset

When publishing a dataset, select data that is uniquely collected or created by your organization. Ideally, this information when combined with other open data provides greater value. Government agencies are in a unique position to collect and curate valuable datasets. Since there is effort and cost associated with modeling, publishing and maintaining any data set as a public service, selection of high value data sets may be guided re-use potential and popularity, among other factors. Data about geographic features, human health, legislation, population and demographics, and the environmental data are just some of the popular open government data sets that have been published as Linked Open Data.

For example, publishing regulated facilities that can then be linked with latitude and longitude allows the facilities to be plotted on a map. That data can then be extended using post codes allowing people to search via post code what facilities are near them on a map view. Facilities data published as extensible Linked Data allows Web developers to rapidly build Web interfaces that are both useful to machines and humans.

3. Model the Data

It is not within scope of this document to treat the Linked Open Data modeling process comprehensively. Rather, we provide guidance on conducting Linked Data modeling and describe a few aspects that differentiate Linked Data modeling from other approaches.

Participants

The modeling process should include participants who represent a broad range of concerns including: the government program or office, the data steward of the originating data source, data standards and policies. For example, if the source data is from a relational database, the modeling meetings may include a database administrator (DBA) and/or data steward. If the organization has a data standards group, include a stakeholder in the modeling effort. A Linked Data subject matter expert should facilitate the modeling process and be capable of explaining Linked Data Principles and the data life cycle (see Prepare Stakeholders). The modeling phase may involve onsite or virtual meetings during which stakeholders specify details about the data, including what the objects mean and how they are related to each other. The Linked Data subject matter expert typically records this information in order complete the remaining steps in the modeling process.

Understanding the Differences

Linked Data modeling involves data going from one model to another. For example, modeling may involve converting a tabular representation of the data to a graph-based representation. Often extracts from relational databases are modeled and converted to Linked Data to more rapidly integrate datasets from different authorities or with other open source datasets. During the data modeling process, stakeholders are encouraged to describe how objects are related. The subject matter expert is recording how various objects are related, using standard vocabularies wherever possible. Best practices for using standard vocabularies are detailed later in this document. In Linked Data, the data schema is represented with the data itself. This mechanism of self-describing data contrasts with the relational approach where external documents (e.g., data dictionaries) and diagrams (e.g., entity relationship diagrams, logical schemas) describe the data.

Linked Data modeling is differentiated through its use of international open Web standards. Linked Data is predicated on the use of international standards for data interchange (e.g., RDFa, JSON-LD, Turtle and RDF/XML) and query SPARQL. Linked Data modeling leverages many of the advances in modern information management, including increased levels of data abstraction. We hope that highlighting some of the differences proves helpful and better informs your efforts to publish open government data.

4. Specify an Appropriate License

It is important to specify who owns data published on the Web and to explicitely connect that license with the data itself. Governmental authorities publishing open data are encouraged to review the relevant guidance for open licenses and copyright. Publishing Linked Open Data makes associating a license that travels with the data itself easier. People are more likely to reuse data when there is a clear, acceptable license associated with it.

A valuable resource for open data publishers may be found on the Creative Commons Web site. Creative Commons develops, supports, and stewards legal and technical infrastructure for digital content publishing.

5. The Role of "Good URIs" for Linked Data

URI Design Principles

The Web makes use of the URI as a single global identification system. The global scope of URIs promotes large-scale "network effects". Therefore, in order to benefit from the value of LD, government and governmental agencies need to identify their resources using URIs. This section provides a set of general principles aimed at helping government stakeholders to define and manage URIs for their resources.

Use HTTP URIs
To benefit from and increase the value of the World Wide Web, governments and agencies SHOULD provide HTTP URIs as identifiers for their resources. There are many benefits to participating in the existing network of URIs, including linking, caching, and indexing by search engines. As stated in [howto-lodp], HTTP URIs enable people to "look-up" or "dereference" a URI in order to access a representation of the resource identified by that URI. To benefit from and increase the value of the World Wide Web, data publishers SHOULD provide URIs as identifiers for their resources.

Provide at least one machine-readable representation of the resource identified by the URI
In order to enable HTTP URIs to be "dereferenced", data publishers have to set up the necessary infrastructure elements (e.g. TCP-based HTTP servers) to serve representations of the resources they want to make available (e.g. a human-readable HTML representation or a machine-readable Turtle). A publisher may supply zero or more representations of the resource identified by that URI. However, there is a clear benefit to data users in providing at least one machine-readable representation. More information about serving different representations of a resource can be found in [COOLURIS].

A URI structure will not contain anything that could change
It is good practice that URIs do not contain anything that could easily change or that is expected to change like session tokens or other state information. URIs should be stable and reliable in order to maximize the

possibilities of reuse that Linked Data brings to users. There must be a balance between making URIs readable and keeping them more stable by removing descriptive information that will likely change. For more information on this see Architecture of the World Wide Web: URI Persistence.

URI Opacity

The Architecture of the World Wide Web [webarch], provides best practices for the treatment of URIs at the time they are resolved by a Web client: *Agents making use of URIs* SHOULD NOT *attempt to infer properties of the referenced resource.* URIs SHOULD be constructed in accordance with the guidance provided in this document to ensure ease of use during development and proper consideration to the guidelines given herein. However, Web clients accessing such URIs SHOULD NOT parse or otherwise read into the meaning of URIs.

URI Policy for Persistence

Defining and documenting a persistent URI policy and implementation plan is vital to the ongoing success and stability of publishing open government data.

The effect of changing or moving resources has the effect of breaking applications dependent upon it. Therefore, government authorities should define a persistence strategy and implementation plan to provide content using the same Web address, even though the resources in question may have moved. Persistent identifiers are used to retain addresses to information resources over the long term. Persistent identifiers are used to uniquely identify objects in the real world and concepts, in addition to information resources.

The choice of a particular URI scheme provides no guarantee that those URIs will be persistent. URI persistence is a matter of policy and commitment on the part of the URI owner. HTTP [RFC2616] has been designed to help manage URI persistence. For example, HTTP redirection (using the 3xx response codes) permits servers to tell an agent that further action needs to be taken by the agent in order to fulfill the request (for example, a new URI is associated with the resource).

The PURL concept allows for generalized URL curation of HTTP URIs on the World Wide Web. PURLs allow third party control over both URL resolution and resource metadata provision. A Persistent URL is an address on the World Wide Web that causes a redirection to another Web resource. If a Web resource changes location (and hence URL), a PURL pointing to it can be updated.

A user of a PURL always uses the same Web address, even though the resource in question may have moved. PURLs may be used by publishers to manage their own information space or by Web users to manage theirs; a PURL service is independent of the publisher of information. PURL services thus allow the management of hyperlink integrity. Hyperlink integrity is a design trade-off of the World Wide Web, but may be partially restored by allowing resource users or third parties to influence where and how a URL resolves.

The Open Source PURLs Project is used widely to run persistent identifier management sites. The Open Source PURLs Project is used by libraries, academic organizations, government agencies and non-government organizations around the world. For example, persistent URLs are used by the United Nations Food and Agriculture Organization (FAO) to provide URIs for major food crops. The National Center for Biomedical Ontology provides persistent URLs to unify and address the terminology used in many existing biomedical databases. The US Government Printing Office also uses persistent URLs to point to documents like the U.S. Budget that are deemed essential to a democratic, transparent government.

Recently, a software project called Permanent Identifiers for the Web emerged to provide a secure, permanent URL re-direction service for Web applications. The service operates in HTTPS-only mode to ensure end-to-end security. This means that it may be used for Linked Data applications that require high levels of security such as those found in the financial, medical, and public infrastructure sectors. A growing group of organizations that have pledged responsibility to ensure the operation of this website over time. Those interested in learning more are encouraged to contact the W3C Permanent Identifier Community Group.

PURLs implement one form of persistent identifier for virtual resources. Other persistent identifier schemes include Digital Object Identifiers (DOIs), Life Sciences Identifiers (LSIDs) and INFO URIs. All persistent identification schemes provide unique identifiers for (possibly changing) virtual resources, but not all schemes provide curation opportunities. Curation of virtual resources has been defined as, “the active involvement of information professionals in the management, including the preservation, of digital data for future use.” [yakel-07] For a persistent identification scheme to provide a curation opportunity for a virtual resource, it must allow real-time resolution of that resource and also allow real-time administration of the identifier.

URI Construction

The following guidance is has been developed by organizations involved in URI strategy and implementation for government agencies:

- Cool URIs for the Semantic Web [COOLURIS]
- Designing URI Sets for the UK Public Sector [uk-govuri]
- Designing URI Sets for the UK Public Sector, a document from the UK Cabinet offices that defines the design considerations on how to URIs can be used to publish public sector reference data;
- Study on Persistent URIs with identification of best practices and recommendations on the topic for the Member States and the European Commission [PURI]
- Towards a NL URI Strategy

General-purpose guidelines exist for the URI designer to consider, including

- Cool URIs for the Semantic Web, which provides guidance on how to use URIs to describe things that are not Web documents;
- Style Guidelines for Naming and Labeling Ontologies in the Multilingual Web (PDF)

Internationalized Resource Identifiers

Stakeholders who are planning to create URIs using characters that go beyond the subset defined in [RFC3986] are encouraged to reference IRIs. Defined in (RFC 3987), IRI is a protocol element that represents a complement to the Uniform Resource Identifier (URI). An IRI is a sequence of characters from the Universal Character Set (Unicode/ISO 10646) that can be therefore used to mint identifiers that use a wider set of characters than the one defined in [RFC3986].

The Internationalized Domain Name or IDN is a standard approach to dealing with multilingual domain names was agreed by the IETF in March 2003.

Internationalized Resource Identifiers use non-ASCII characters in URIs which is relevent to those organizations interested in minting URIs in languages including German, Dutch, Spanish, French and Chinese.

Although there exist some standards focused on enabling the use of international characters in Web identifiers, government stakeholders need to take into account several issues before constructing such internationalized identifiers. This section is not exhaustive and the editors point the interested audience to An Introduction to Multilingual Web Addresses, however some of the most relevant issues are following:

- **Domain Name lookup:** Numerous domain name authorities already offer registration of internationalized domain names. These include providers for top level country domains as `.cn, .jp, .kr` , etc., and global top level domains such as `.info, .org` and `.museum.`
- **Domain names and phishing:** One of the problems associated with IDN support in browsers is that it can facilitate phishing through what are called 'homograph attacks'. Consequently, most browsers that support IDN also put in place some safeguards to protect users from such fraud.
- **Encoding problems:** IRI provides a standard way for creating and handling international identifiers, however the support for IRIs among the various semantic Web technology stacks and libraries is not uniform and may lead to difficulties for applications working with this kind of identifiers. A good reference on this subject can be found in [i18n-web] .

The URI syntax defined in [RFC3986] STD 66 (Uniform Resource Identifier (URI): Generic Syntax) restricts URIs to a small number of characters: basically, just upper and lower case letters of the English alphabet, European numerals and a small number of symbols.

6. Standard Vocabularies

Standardized vocabularies should be reused as much as possible to facilitate inclusion and expansion of the Web of data. The W3C has published several useful vocabularies for Linked Data. For example, the following standard vocabularies help developers to describe basic or more complex relationships for describing data catalogs, organizations, and multidimensional data, such as statistics on the Web. Government publishers are encouraged to use standardized vocabularies rather than reinventing the wheel, wherever possible.

Specifically, Data Catalog Vocabulary (DCAT) [vocab-dcat] is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. By using DCAT to describe datasets in data catalogs, publishers increase discoverability and enable applications easily to consume metadata from multiple catalogs. It further enables decentralized publishing of catalogs and facilitates federated dataset search across sites. Aggregated DCAT metadata can serve as a manifest file to facilitate digital preservation.

Organizational structures and activities are often described by government authorities. The Organization Ontology [vocab-org] supports the publishing of organizational information across a number of domains, as Linked Data. The Organizational Ontology is designed to allow domain-specific extensions to add classification of organizations and roles, as well as extensions to support neighboring information such as organizational activities.

Many government agencies publish statistical information on the public Web. The Data Cube Vocabulary [vocab-data-cube] provides a means to do this using the Resource Description Framework (RDF). The RDF Data Cube Vocabulary makes it possible to discover and identify statistical data artifacts in a uniform way. [CSARVEN] The model underpinning the Data Cube vocabulary is compatible with the cube model that underlies SDMX (Statistical Data and Metadata eXchange), an ISO standard for exchanging and sharing statistical data and metadata among organizations. The Data Cube vocabulary is a core foundation which supports extension vocabularies to enable publication of other aspects of statistical data flows or other multi-dimensional datasets.

How to Find Existing Vocabularies

There are search tools that collect, analyze and index vocabularies and semantic data available online for efficient access. Search tools that use structured data represented as Linked Data include: (Falcons, Watson, Sindice, Semantic Web Search Engine, Swoogle, and Schemapedia).

Others include the LOV directory, Prefix.cc, Bioportal (biological domain) and the European Commission's Joinup platform.

Where to find existing vocabularies in data catalogues
Another way around is to perform search using the previously identified key terms in datasets catalogs. Some of these catalogs provide samples of how the underlying data was modeled and used.

Vocabulary Checklist

This section provides a set of considerations aimed at helping stakeholders review a vocabulary to evaluate its usefulness.

NOTE

It is best practice to use or extend an existing vocabulary before creating a new vocabulary.

A basic vocabulary checklist:

- ensure vocabularies you use are published by a trusted group or organization;
- ensure vocabularies have permanent URIs; and
- confirm the versioning policy.

Vocabularies MUST be documented
A vocabulary MUST be documented. This includes the liberal use of labels and comments, as well as appropriate language tags. The publisher must provide human-readable pages that describe the vocabulary, along with its constituent classes and properties. Preferably, easily comprehensible use-cases should be defined and documented.

Vocabularies SHOULD be self-descriptive
Each property or term in a vocabulary should have a Label, Definition and Comment defined. Self-describing data suggests that information about the encodings used for each representation is provided explicitly within the representation. The ability for Linked Data to describe itself, to place itself in context, contributes to the usefulness of the underlying data.

For example, the widely-used Dublin Core vocabulary (formally `DCMI Metadata Terms`) has a Term Name Contributor which has a:

```
Label: Contributor
Definition: An entity responsible for making contributions to
the resource
Comment: Examples of a Contributor include a person, an
organization, or a service.
```

Vocabularies SHOULD be described in more than one language
Multilingualism should be supported by the vocabulary, i.e. all the elements of the vocabulary should have labels, definitions and comments available in the government's official language(s), e.g. Spanish and at least in English. This is also important as the documentation should suppoly appropriate tags for the language used for the comments or labels.

For example, for the same term Contributor

```
rdfs:label "Contributor"@en, "Colaborador"@es
rdfs:comment "Examples of a Contributor include a person, an
organization, or a service"@en , "Ejemplos de collaborator
incluyen persona, organización o servicio"@es
```

Vocabularies SHOULD be used by other datasets
If the vocabulary is used by other authoritative Linked Open datasets that is helpful. It is in re-use of vocabularies that we achieve the benefits of Linked Open Data. Selected vocabularies from third parties should be already in use by other datasets, as this shows that they are already

established in the LOD community, and thus better candidates for wider adoption and reuse.

For example: An analysis on the use of vocabularies on the Linked Data cloud reveals that FOAF is reused by more than 55 other vocabularies.

Vocabularies SHOULD be accessible for a long period
The vocabulary selected should provide some guarantee of maintenance over a specified period, ideally indefinitely.

Vocabularies SHOULD be published by a trusted group or organization
Although anyone can create a vocabulary, it is always better to check if it is one person, group or authoritative organization that is responsible for publishing and maintaining the vocabulary.

Vocabularies SHOULD have persistent URLs
Persistent access to the server hosting the vocabulary, facilitating reusability is necessary.

Example: The Geo W3C vocabulary [vocab-geo] is one of the most used vocabularies for a basic representation of geometry points (latitute/longitude) and has been around since 2009, always available at the same namespace.

Vocabularies SHOULD provide a versioning policy
The publisher ideally will address compatibility of versions over time. Major changes to the vocabularies should be reflected in the documentation.

Vocabulary Creation

This section provides a set of informative considerations aimed at stakeholders who decide they must develop their own vocabularies.

Define the URI of the vocabulary.
The URI that identifies your vocabulary must be defined. This is strongly related to the Best Practices described in section URI Construction.

For example: If we are minting new vocabulary terms from a particular government, we should define the URI of that particular vocabulary.

URIs for properties with non-literal ranges
What it means: Name all properties as verb senses, so that triples may be actually read; e.g. *hasProperty* .

Vocabularies should be self-descriptive

What it means: Each property or term in a vocabulary should have a Label, Definition and Comment defined. Self-describing data suggests that information about the encodings used for each representation is provided explicitly within the representation. The ability for Linked Data to describe itself, to place itself in context, contributes to the usefulness of the underlying data.

For example, the widely-used Dublin Core vocabulary (formally `DCMI Metadata Terms`) has a Term Name Contributor which has a:

```
Label: Contributor
Definition: An entity responsible for making contributions to
the resource
Comment: Examples of a Contributor include a person, an
organization, or a service.
```

Vocabularies should be described in more than one language

Multilingualism should be supported by the vocabulary, i.e., all the elements of the vocabulary should have labels, definitions and comments available in the government's official language, e.g., Spanish, and at least in English. That is also very important as the documentation should be clear enough with appropriate tag for the language used for the comments or labels.

For example, for the same term `Contributor`

```
rdfs:label "Contributor"@en, "Colaborador"@es
rdfs:comment "Examples of a Contributor include a person, an
organization, or a service"@en , "Ejemplos de collaborator
incluyen persona, organización o servicio"@es
```

Vocabularies should provide a versioning policy

It refers to the mechanism put in place by the publisher to always take care of backward compatibilities of the versions, the ways those changes affected the previous versions. Major changes of the vocabularies should be reflected on the documentation, in both machine or human-readable formats.

Vocabularies should provide documentation

A vocabulary should be well-documented for machine readable (use of labels and comments; tags to language used). Also for human-readable, an extra documentation should be provided by the publisher to better understand the classes and properties, and if possible with some valuable use cases. **Provide human-readable documentation and basic metadata such as creator, publisher, date of creation, last modification, version number.**

Vocabularies should be published following available best practices

Publish your vocabulary on the Web at a stable URI using an open license.. One of the goals is to contribute to the community by sharing the new vocabulary. To this end, it is recommended to follow available recipes for publishing RDF vocabularies e.g. Best Practice Recipes for Publishing RDF Vocabularies [bp-pub].

Using SKOS to Create a Controlled Vocabulary

SKOS, the Simple Knowledge Organization System [SKOS-REFERENCE], is a W3C standard, based on other Semantic Web standards (RDF and OWL), that provides a way to represent controlled vocabularies, taxonomies and thesauri. Specifically, SKOS itself is an OWL ontology and it can be written out in any RDF flavor.

The W3C SKOS standard defines a portable, flexible controlled vocabulary format that is increasingly popular, with the added benefit of a good entry-level step toward the use of Semantic Web technology.

SKOS is appropriate in the following situations:

- There is a need to publish a controlled list of terms or taxonomies having a special meaning for the domain.
- The complexity and formality of an OWL ontology is not appropriate (for example the terms are not themselves entities that will be richly described).

In creating a SKOS vocabulary bear the following good practice in mind:

- Make a clear distinction between the collections of concepts (ConceptScheme) and the different individual concepts.
- Define when possible a different namespace for each `skos:ConceptScheme`
- Structure the concepts in the list using properties `skos:hasTopConcept`, `skos:broader`, `skos:narrower.`
- Consider defining a Class to represent all the skos:Concepts in your controlled list (this can facilitate declaration of properties that will use this list).
- Provide multilingual labels for the terms.

Multilingual Vocabularies

This section is not comprehensive however, is intended to mention some of the issues identified by the Working Group and some of the work performed by others in relation to publishing Linked Data in multiple languages. For more details on the multilingualism on the Web, see the MultilingualWeb-LT Working Group

Multilingual Vocabularies broaden Search
As of the writing of this Note, many of the available Linked Data vocabularies are in English. This may restrict your content from being searched by multilingual search engines and by non-English speakers.

If designing a vocabulary, provide labels and descriptions if possible, in several languages, to make the vocabulary usable by a global audience.

Multilingual vocabularies may be found in the following formats

- As a set of `rdfs:label` in which the language has been restricted (@en, @fr...). Currently, this is the most commonly used approach.
- As `skos:prefLabel` (or `skosxl:Label`), in which the language has also been restricted.
- As a set of monolingual ontologies (ontologies in which labels are expressed in one natural language) in the same domain mapped or aligned to each other (see the example of EuroWordNet, in which wordnets in different natural languages are mapped to each other through the so-called `ILI - inter-lingual-index-`, which consists of a set of concepts common to all categorizations).
- As a set of ontology + lexicon. This is an approach to the representation of linguistic (multilingual) information associated to ontologies. The idea is that the ontology is associated to an external ontology of linguistic descriptions. One of the best exponents in this case is the lemon model, an ontology of linguistic descriptions that is to be related with the concepts and properties in an ontology to provide lexical, terminological, morphosyntactic, etc., information. One of the main advantages of this approach is that semantics and linguistic information are kept separated. One can link several lemon models in different natural languages to the same ontology.
- A list of codes and their corresponding URIs for the representation of language names is published and maintained by the official registration authority of ISO639-2, the US Library of Congress. [ISO-639-1], [ISO-639-2]

NOTE

The current trend is to follow the first approach, i.e. to use at least a `rdfs:label` and `rdfs:comment` for each term in the vocabulary.

7. Convert Data to Linked Data

Now with the ground work in place, the next step is to actually convert a dataset into a Linked Data representation. There is more than one way to convert data including scripts, declarative mapping languages, languages that perform query translation rather then data translation (e.g. R2RML). Regardless of which approach is used, data conversion involves mapping the source data into a set of RDF statements. As data is converted, data is serialized into RDF statements. RDF can be converted into a range of RDF serializations that include:

- RDFa,
- JSON-LD,
- Turtle and N-Triples,
- RDF/XML

Linked Data modelers and developers have certain reasons they prefer to use one RDF serialization over another. No one RDF serialization is better than the other. Benefits of using one over another include simplicity, ease of reading (for a human) and speed of processing.

Provide Basic Metadata

When modeling Linked Data metadata, it is a best practice to include the MIME type, publishing organization and/or agency, creation date, modification date, version, frequency of updates, and contact email address, if this information is available and appropriate to the data. In subsequent sections we outline guidance for the use of vocabularies, as well as, a vocabulary "checklist" to assist in the modeling process.

Link to Other Stuff

As the name suggests, Linked Open Data means the data links to other stuff. Data in isolation is rarely valuable, however, interlinked data is suddenly very valuable. There are many popular datasets, such as DBpedia that provide valuable data, including photos and geographic information. Being able to connect data from a government authority with DBpedia for example, is quick way to show the value of adding content to the Linked Data Cloud.

8. Provide Machine Access to Data

A major benefit of Linked Data is that it provides access to data for machines. Machines can use a variety of methods to read data including, but not limited to:

- Direct URI resolution ("follow your nose"),
- a RESTful API,
- a SPARQL endpoint, and/or
- via file download.

The SPARQL Protocol and RDF Query Language (SPARQL) defines a query language for RDF data, analogous to the Structured Query Language (SQL) for relational databases. SPARQL is to RDF data what SQL is to a relational database. For more information, see the SPARQL 1.1 Overview [sparql11-overview].

A SPARQL endpoint is a a service that accepts SPARQL queries and returns answers to them as SPARQL result sets. It is a best practice for datasets providers to give the URL of their SPARQL endpoint to allow access to their data programmatically or through a Web interface. A list of SPARQL endpoints monitoring the availability, performance, interoperability and discoverability of SPARQL Endpoints is published by the Open Knowledge Foundation.

9. Announce to the Public

It is not within scope of this document to discuss domain name issues and data hosting however, it is a vital part of the publication process. Hosting Linked Open Data may require involvement with agency system security staff and require planning that often takes considerable time and experise for compliance, so involve stakeholders early and schedule accordingly.

Now you're ready to point people to authoritative open government data. Be sure the datasets are available via an authoritative domain. Using an authoritative domain increases the perception of trusted content. Authoritative data that is regularly updated on a government domain is critical to re-use of authoritative datasets.

The following checklist is intended to help organizations realize the benefits of publishing open government data, as well as, communicate to the public that you are serious about providing this data over time.

- Use multiple channels including mailing lists, blogs and newsletters to announce a newly published data set;
- Publish a description for each published dataset using [vocab-dcat] or [void] vocabulary;
- Define the frequency of data updates (as metadata);
- Associate an appropriate license;
- Plan and implement a persistence strategy;
- Ensure data is accurate to the greatest degree possible;
- Provide a form for people to report problematic data and give feedback;
- Provide a contact email address (alias) for those responsible for curating and publishing the data; and
- Ensure staff have the necessary training to respond in a timely manner to feedback.

10. Social Contract of a Linked Data Publisher

Government publishers of Linked Open Data are entering into a sort of "social contract" with users of their data. Publishers must recognize their responsibility in maintaining data once it is published. Key to both access and reuse is ensuring that the dataset(s) your organization publishes remains available where you say it will be and is maintained over time.

Giving due consideration to your organization's URI strategy should be one of the first activities your team undertakes as they prepare a Linked Open Data strategy. Authoritative data requires the permanence and resolution of HTTP URIs. If publishers move or remove data that was published to the Web, third party applications or mashups may break. This is considered rude for obvious reasons and is the basis for the Linked Data "social contract." A good way to prevent causing HTTP 404s is for your organization to implement a persistence strategy. Below we provide an introduction to the best practice of defining a persistence strategy and implementation plan.

Stability Properties

It is beyond the scope of this document to comprehensively treat issues related to data stability over time on the Web. Mention is included such that readers may consider data stability in the context of a given agency and region. There are characteristics that influence the stability or longevity of useful open government data. Many of these properties are not unique to government Linked Open Data, yet they influence data cost and therefore data value.

As a final note related to the importance of stability. The W3C prepares to celebrate its 20th anniversary and the Web turns 25 years old in 2014. Perhaps surprisingly, the first Web page cannot be found. A team at CERN is looking into restoring it, however at the time of the writing of this document, it has not yet been found.[GBRUMFIEL] Thus, the Government Linked Data Working Group wished to reference the importance of *data stability* as the vast majority of government data is quickly available *only* in digital form. As stewards and supporters of open government data, it is encumbant upon us all to pursue the methods and tools to support responsible data stability on the Web over time. Thanks for your interest in this topic and please join us in helping evolve the Web of Data into the 21st Century and beyond!

A. Acknowledgments

The editors wish to gratefully acknowledge the considerable contributions to the Linked Data Best Practices document by the following people: Dave Reynolds, (Epimorphics,UK), Phil Archer, (W3C / ERCIM, UK), Makx Dekkers, (Independent Consultant, Spain), John Erickson (Rensselaer Polytechnic Institute, USA), João Paulo Almeida , (Federal University of Espírito Santo, Brazil), Tom Heath , (Open Data Institute, UK), Thomas Baker , (Dublin Core Metadata Initiative, US) Sarven Capadisli, (UK) Bernard Vatant (Mondeca, France), Michael Pendleton (U.S. Environmental Protection Agency, USA), Biplav Srivastava (IBM India), Daniel Vila (Ontology Engineering Group, Universidad Politécnica de Madrid, UPM, Spain), Martín Álvarez Espinar (CTIC-Centro Tecnológico, Spain), David Wood (3 Round Stones, USA), Michael Hausenblas (MapR, USA), our working group co-chair, Hadley Beeman (UK LinkedGov, UK), and Sandro Hawke (W3C/MIT). Please accept our apologies in advance if we've inadvertantly omitted your name as many people provided valuable feedback and were instrumental in the production of this best practices publication.

Thank you, grazie, gracias, obrigado, merci, धन्यवाद.

This document has been produced by the Government Linked Data Working Group, and its contents reflect extensive discussion within the Working Group as a whole.

B. References

B.1 Informative references

[BHYLAND2011]
Bernadette Hyland; David Wood. *The Joy of Data - Cookbook for Publishing Linked Government Data on the Web*. URL: http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook

[BVILLAZON]
Boris Villazón-Terrazas; et al.. *Methodological Guidelines for Publishing Government Linked Data*. URL: http://link.springer.com/chapter/10.1007/978-1-4614-1767-5_2

[COOLURIS]
Leo Sauermann; Richard Cyganiak. *Cool URIs for the Semantic Web*. 3 December 2008. W3C Note. URL: http://www.w3.org/TR/cooluris

[CSARVEN]
Sarven Capadisli. *Towards Linked Statistical Data Analysis*. URL: http://csarven.ca/linked-statistical-data-analysis

[GBRUMFIEL]
Geoff Brumfiel. *The First Web Page, Amazingly, Is Lost*. URL: http://www.npr.org/2013/05/22/185788651/the-first-web-page-amazingly-is-lost

[HAUSENBLAS]
Michael Hausenblas; Richard Cygankiak. *Linked Data Life cycles*, formerly at http://linked-data-life-cycles.info/.

[ISO-639-1]
U.S. Library of Congress. *ISO 639-1: Codes for the Representation of Names of Languages - Part 1: Two letter codes for languages*. URL: http://id.loc.gov/vocabulary/iso639-1.html

[ISO-639-2]
U.S. Library of Congress. *ISO 639-2: Codes for the Representation of Names of Languages - Part 2: Alpha-3 Code for the Names of Lanuguages*. URL: http://id.loc.gov/vocabulary/iso639-2.html

[PURI]
Phil Archer; et al.. *Study on Persistent URIs*. URL: http://philarcher.org/diary/2013/uripersistence/#recs

[RDF-CONCEPTS]
Graham Klyne; Jeremy Carroll. *Resource Description Framework (RDF): Concepts and Abstract Syntax*. 10 February 2004. W3C Recommendation. URL: http://www.w3.org/TR/rdf-concepts/

[RFC2616]
R. Fielding et al. *Hypertext Transfer Protocol - HTTP/1.1*. June 1999. RFC. URL: http://www.ietf.org/rfc/rfc2616.txt

[RFC3986]
T. Berners-Lee; R. Fielding; L. Masinter. *Uniform Resource Identifier (URI): Generic Syntax (RFC 3986)*. January 2005. RFC. URL: http://www.ietf.org/rfc/rfc3986.txt

[SKOS-REFERENCE]
Alistair Miles; Sean Bechhofer. *SKOS Simple Knowledge Organization System Reference*. 18 August 2009. W3C Recommendation. URL:

http://www.w3.org/TR/skos-reference

[WOOD2013]

Wood, D.; Zaidman, M.; Ruth, L.. *Linked Data: Structured Data on the Web*. URL: http://www.manning.com/dwood/

[bp-pub]

Diego Berrueta; Jon Phipps. *Best Practice Recipes for Publishing RDF Vocabularies*. W3C Working Group Note. URL: http://www.w3.org/TR/swbp-vocab-pub/

[howto-lodp]

Christian Bizer; Richard Cyganiak; Tom Heath. *How to Publish Linked Data on the Web*. URL: http://linkeddata.org/docs/how-to-publish

[i18n-web]

S. Auer; M. Weidl; J. Lehmann; Amrapali J. Zaveri; Key-Sun Choi. *I18n of Semantic Web Applications*. URL: http://svn.aksw.org/papers/2010/ISWC_I18n/public.pdf

[sparql11-overview]

The W3C SPARQL Working Group. *SPARQL 1.1 Overview*. 21 March 2013. W3C Recommendation. URL: http://www.w3.org/TR/sparql11-overview/

[uk-govuri]

Cabinet Office GOV.UK. *Designing URI sets for the UK public sector*. URL: https://www.gov.uk/government/publications/designing-uri-sets-for-the-uk-public-sector/

[vocab-data-cube]

Richard Cyganiak; Dave Reynolds. *The RDF Data Cube Vocabulary*. 17 December 2013. W3C Proposed Recommendation. URL: http://www.w3.org/TR/vocab-data-cube/

[vocab-dcat]

Fadi Maali; John Erickson. *Data Catalog Vocabulary (DCAT)*. 17 December 2013. W3C Proposed Recommendation. URL: http://www.w3.org/TR/vocab-dcat/

[vocab-geo]

Dan Brickley; Tim Berners-Lee. *Basic Geo (WGS84 lat/long) Vocabulary*. URL: http://www.w3.org/2003/01/geo/

[vocab-org]

Dave Reynolds. *The Organization Ontology*. 17 December 2013. W3C Proposed Recommendation. URL: http://www.w3.org/TR/vocab-org/

[void]

Keith Alexander; Richard Cyganiak; Michael Hausenblas; Jun Zhao. *Describing Linked Datasets with the VoID Vocabulary*. 3 March 2011. W3C Note. URL: http://www.w3.org/TR/void/

[webarch]

Ian Jacobs; Norman Walsh. *Architecture of the World Wide Web, Volume One*. 15 December 2004. W3C Recommendation. URL: http://www.w3.org/TR/webarch/

[yakel-07]

Elizabeth Yakel. *Digital curation*. URL: http://dx.doi.org/10.1108/10650750710831466