Linked Data

Angelica Lo Duca
IIT-CNR
angelica.loduca@iit.cnr.it

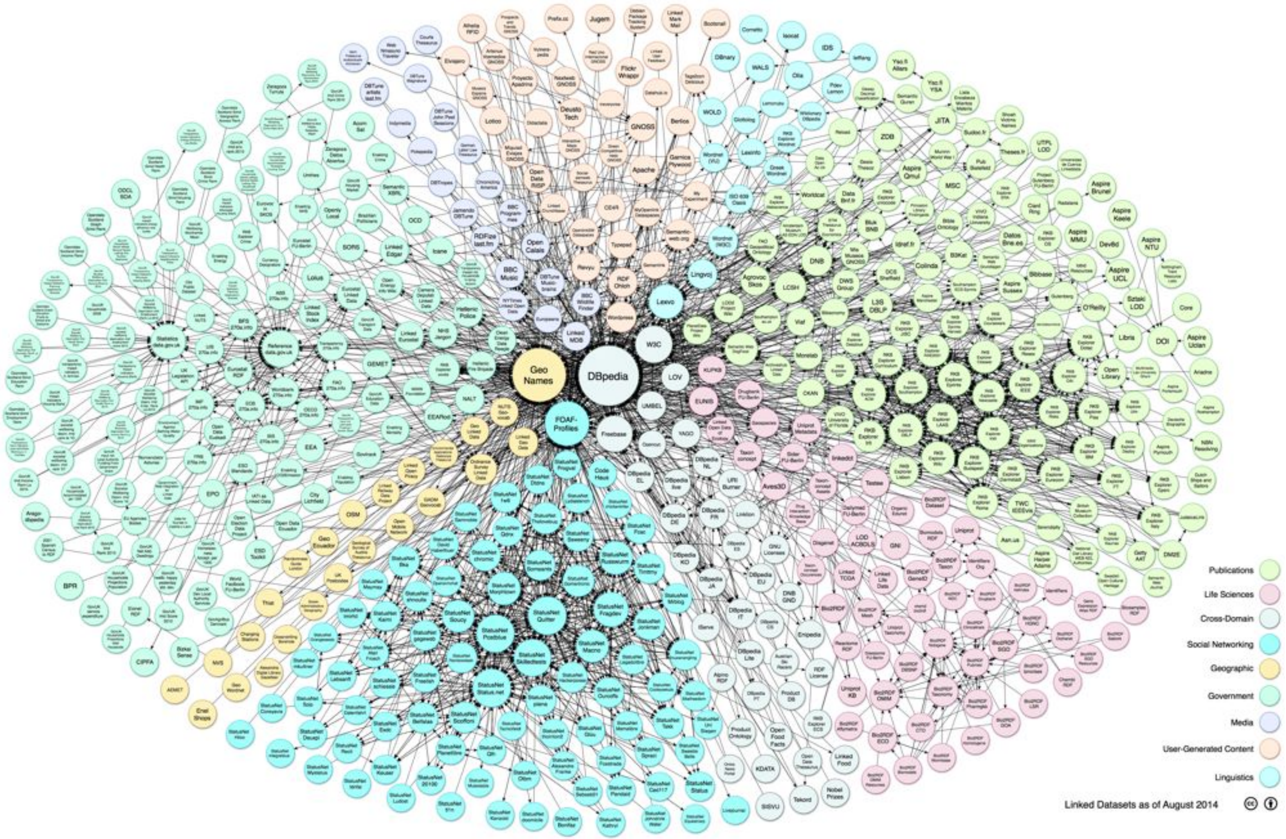
Linked Open Data:
a paradigm for the
Semantic Web

Linked Data are a series of ***best practices*** to connect **structured data** through the Web.

Three questions

- *data access* - easy way for data reusage.
- *data discovery* among a multitude of relevant datasets.
- *data integration* among a large number of data sources previously unknown.

The Linked Data Cloud



Linked Datasets as of August 2014

Existing Linked Data nodes

- http://datahub.io/
  - web site which allows the creation, publication and search of datasets
- http://sparqles.okfn.org
  - to see the list and the status of all SPARQL endpoints maintained by datahub.io

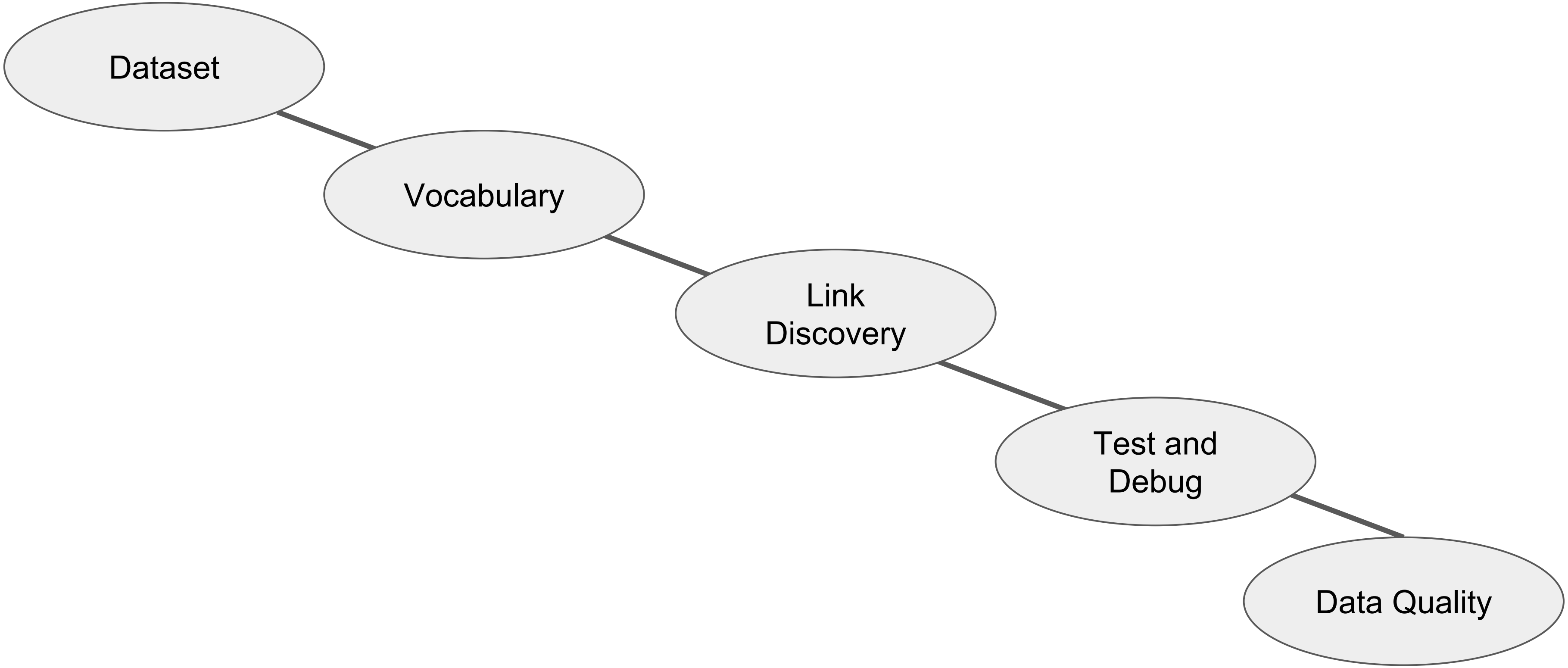
Four principles

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

Kinds of Links

- **Relationship Links** point at related things in other data sources.
- **Identity Links** point at URI aliases used by other data sources to identify the same real-world object or abstract concept.
- **Vocabulary Links** point from data to the definitions of the vocabulary terms that are used to represent the data.

Publish Linked Data



Dataset

Vocabulary

Link Discovery

Test and Debug

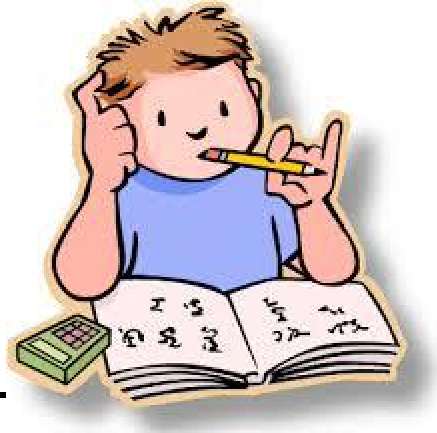
Data Quality

Assignment 1

Think about a topic of your interest (tourism, cultural heritage, books, health, …) and imagine that you have a dataset containing many records of that topic (e.g. hotels, books, patients of a hospital, …)

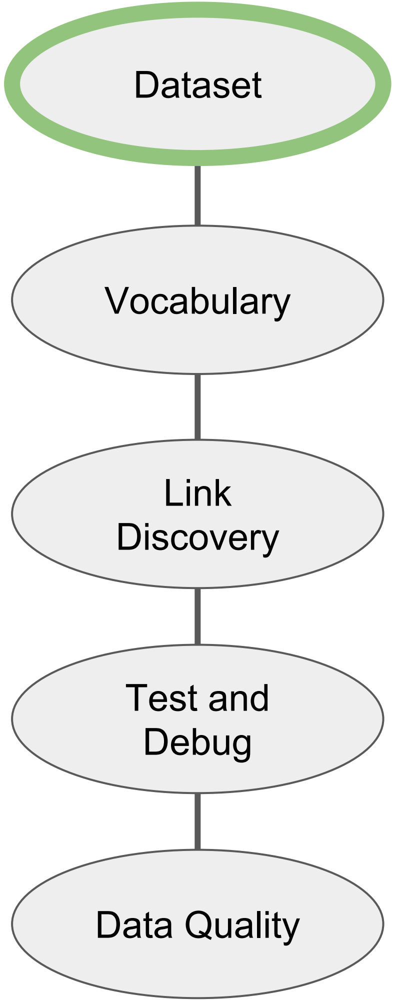
Example: dataset of hotels

Name	**Address**	**City**	**Stars**
Hotel Bologna	via Mazzini 3	Pisa	4
Flowers Hotel	via Rosi 2	Milano	3
...	...	...	...



Dataset

- Describe the dataset
  - use VOID ontology
- Provenance metadata
- License



VOID (Vocabulary of Interlinked Datasets)

- Provides classes and properties to describe a dataset
- A dataset is modelled as an instance of the **void:Dataset** class.
    - The void:Dataset instance is a single RDF resource that represents the entire dataset

```
@prefix void: <http://rdfs.org/ns/void#> .
@prefix : <#> .

:DBpedia a void:Dataset .
```

VOID - Linkset

- VoID also allows the description of ***RDF links*** between datasets.
  - An RDF link is an RDF triple whose subject and object are described in different datasets.
- A ***linkset*** is a collection of RDF links between two datasets.
- A linkset is modelled as an instance of the void:Linkset class.
  - void:Linkset is a subclass of void:Dataset.

```
:DBpedia_Geonames a void:Linkset ;
     void:target :DBpedia;
     void:target :Geonames;
     void:subset :DBpedia;
     void:triples 252000;
     void:linkPredicate owl:sameAs .
```

VOID - General dataset metadata

Term	Purpose
dcterms:title	The name of the dataset.
dcterms:description	A textual description of the dataset.
dcterms:creator	An entity primarily responsible for creating the dataset.
dcterms:publisher	An entity responsible for making the dataset available.
dcterms:contributor	An entity responsible for making contributions to the dataset.
dcterms:source	A related resource from which the dataset is derived.
dcterms:created	Date of creation of the dataset.
dcterms:modified	Date on which the dataset was changed.

```
:DBpedia a void:Dataset;
    dcterms:title "DBPedia";
    dcterms:description "RDF data extracted
from Wikipedia";
    dcterms:contributor :FU_Berlin;
    dcterms:contributor :University_Leipzig;
    dcterms:contributor :OpenLink_Software;
    dcterms:contributor :DBpedia_community;
    dcterms:source
<http://dbpedia.org/resource/Wikipedia>;
    dcterms:modified "2008-11-17"^^xsd:date;
    .
:FU_Berlin a foaf:Organization;
    rdfs:label "Freie Universität Berlin";
    foaf:homepage
<http://www.fu-berlin.de/>;
    .
 # Similar descriptions of the other
contributors go here
```

VOID - License

- The **dcterms:license** property should be used to to point to the license under which a dataset has been published.
    a. Public Domain Dedication and License (PDDL) — places the data(base) in the public do (waiving all rights)

    b. Open Data Commons Attribution (ODC-By) — free to share, create, adapt dat attribute any public use of the database

    c. Open Database License (ODC-ODbL) — free to share, create, adapt data but attribute any public use of the database, redistribute data under the same licer a-like), keep redistributed data open

    d. CC0 1.0 Universal — copy, modify, distribute and perform the work, even for c purposes, all without asking permission

VOID - Dataset Subject

- The **dcterms:subject** property should be used to tag a dataset with a topic.
- For the general case, use a DBpedia resource URI (http://dbpedia.org/resource/XXX) to categorise a dataset
  - XXX stands for the thing which best describes the main topic of what the dataset is about.

```
:DBLP a void:Dataset;
    dcterms:subject <http://dbpedia.org/resource/Computer_science>;
    dcterms:subject <http://dbpedia.org/resource/Journal>;
    dcterms:subject <http://dbpedia.org/resource/Proceedings>;
    .
```

DBLP is a computer science bibliographical database.

VOID - Access Metadata

- SPARQL endpoint
    - `void:sparqlEndpoint <http://dbpedia.org/sparql>;`
      `.`
- RDF data dumps
    - `void:dataDump <http://data.nytimes.com/people.rdf>;`

VOID - Structural Metadata

- Example resources
  - `void:exampleResource <http://dbpedia.org/resource/Berlin> ;`
- Pattern for resource URIs
  - `void:uriSpace "http://dbpedia.org/resource/";`
- Vocabularies used in the dataset
  - `void:vocabulary <http://xmlns.com/foaf/0.1/>;`

VOID - Statistics about dataset

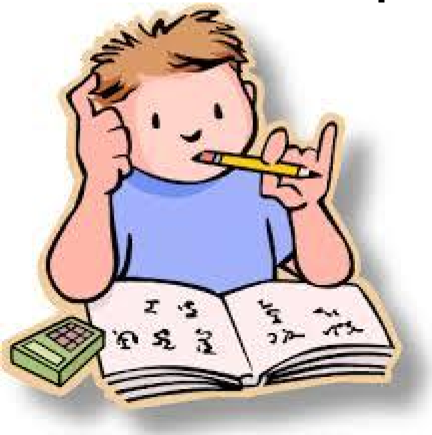
Property	Purpose
void:triples	The total number of triples contained in the dataset.
void:entities	The total number of entities that are described in the dataset.
void:classes	The total number of distinct classes in the dataset.
void:properties	The total number of distinct properties in the dataset.
void:distinctSubjects	The total number of distinct subjects in the dataset.
void:distinctObjects	The total number of distinct objects in the dataset.

VOID - Publish the void file

- Publish a Turtle file named `void.ttl` in the root directory of the site, with a local “hash URI” for the dataset, yielding a dataset URI such as `http://example.com/void.ttl#MyDataset.`

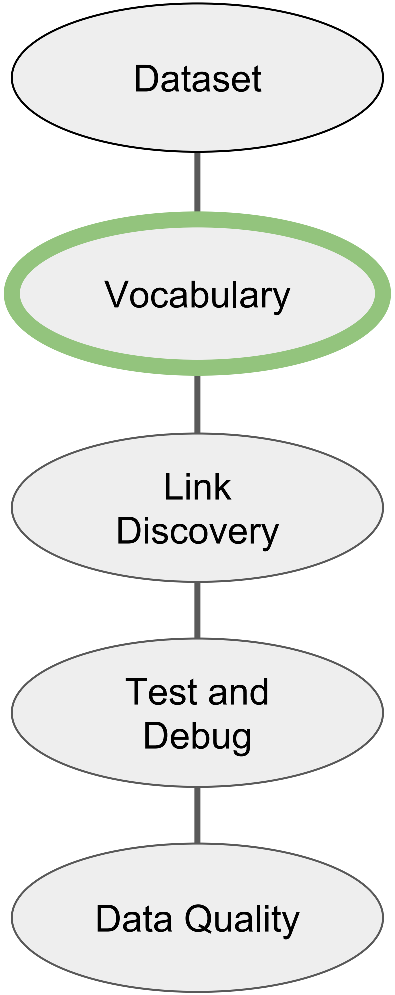
Assignment 2

Describe your dataset in VOID.



Vocabulary

- Choose the vocabularies to describe data
    - RDF Schema
    - OWL
    - SKOS
    - ...



The Web Ontology Language (OWL)

- OWL extends the expressivity of RDFS with additional modeling primitives.
- For example, OWL defines the primitives **owl:equivalentClass** and **owl:equivalentProperty**.

Simple Knowledge Organization System (SKOS)

- SKOS is a vocabulary for expressing **conceptual hierarchies**, often referred to as taxonomies, while RDFS and OWL provide vocabularies for describing **conceptual models** in terms of classes and their properties.

Reusing existing terms

If suitable terms can be found in existing vocabularies, these should be reused to describe data wherever possible, rather than reinvented.

Some common vocabularies

- The **Dublin Core Metadata Initiative (DCMI) Metadata Terms** vocabulary defines general metadata attributes such as *title*, *creator*, *date* and *subject*.
- The **Friend-of-a-Friend (FOAF)** vocabulary defines terms for describing persons, their activities and their relations to other people and objects.
- The **Semantically-Interlinked Online Communities (SIOC)** vocabulary (pronounced *"shock"*) is designed for describing aspects of online community sites, such as users, posts and forums.
- The **Description of a Project (DOAP)** vocabulary(pronounced *"dope"*) defines terms for describing software projects, particularly those that are Open Source.
- The **Music Ontology** defines terms for describing various aspects related to music, such as artists, albums, tracks, performances and arrangements.
- The **Programmes Ontology** defines terms for describing programmes such as TV and radio broadcasts.
- The **Good Relations Ontology** defines terms for describing products, services and other aspects relevant to e-commerce applications.
- The **Creative Commons (CC)** schema defines terms for describing copyright licenses in RDF.
- The **Bibliographic Ontology (BIBO)** provides concepts and properties for describing citations and bibliographic references (i.e., quotes, books, articles, etc.).
- The **OAI Object Reuse and Exchange** vocabulary is used by various library and publication data sources to represent resource aggregations such as different editions of a document or its internal structure.
- The **Review Vocabulary** provides a vocabulary for representing reviews and ratings, as are often applied to products and services.
- The **Basic Geo (WGS84)** vocabulary defines terms such as *lat* and *long* for describing geographically-located things.

How to select a vocabulary

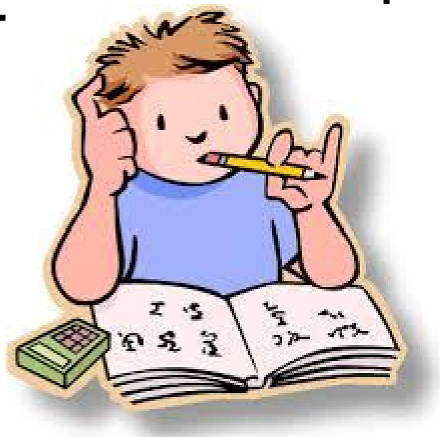
1. **Usage and uptake** – is the vocabulary in widespread usage? Will using this vocabulary make a data set more or less accessible to existing Linked Data applications?
2. **Maintenance and governance** – is the vocabulary actively maintained according to a clear governance process? When, and on what basis, are updates made?
3. **Coverage** – does the vocabulary cover enough of the data set to justify adopting its terms and *ontological commitments*?
4. **Expressivity** – is the degree of expressivity in the vocabulary appropriate to the data set and application scenario? Is it too expressive, or not expressive enough?

How to define a new vocabulary

- Supplement **existing vocabularies** rather than reinventing their terms.
- Only define **new terms** in a namespace that you control.
- **Use terms from RDFS and OWL to relate new terms** to those in existing vocabularies.
- **Document** each new term with human-friendly labels and comments

Assignment 3

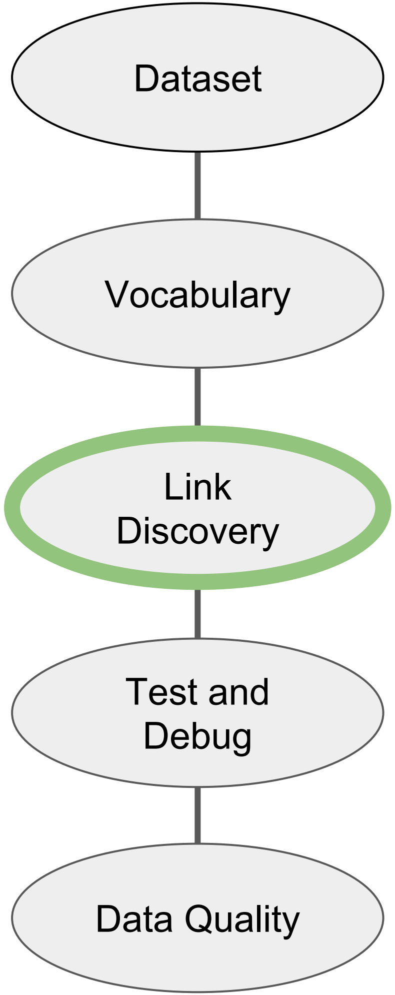
Choose a vocabulary for your imaginary dataset.



Link Discovery

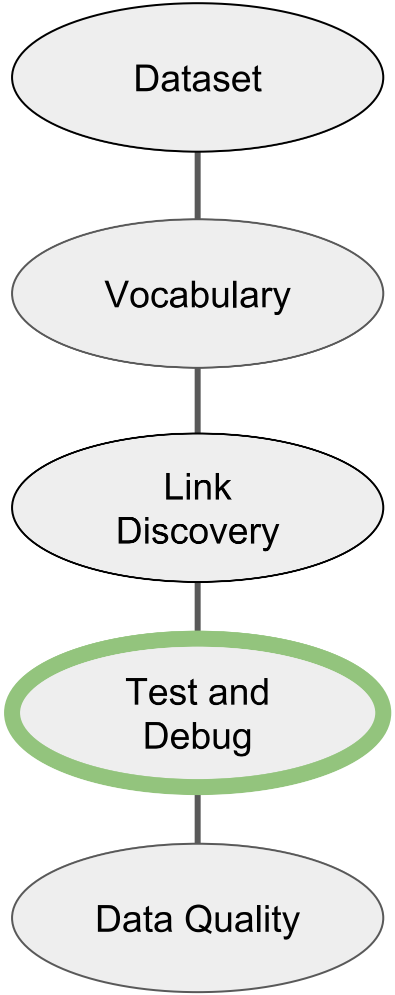
- Establish internal and external links
  - *internal links* connect pairs of nodes, both belonging to the same dataset
  - *external links* connect pairs of nodes, one belonging to our dataset and the other to an external one

Link Discovery will be discussed later



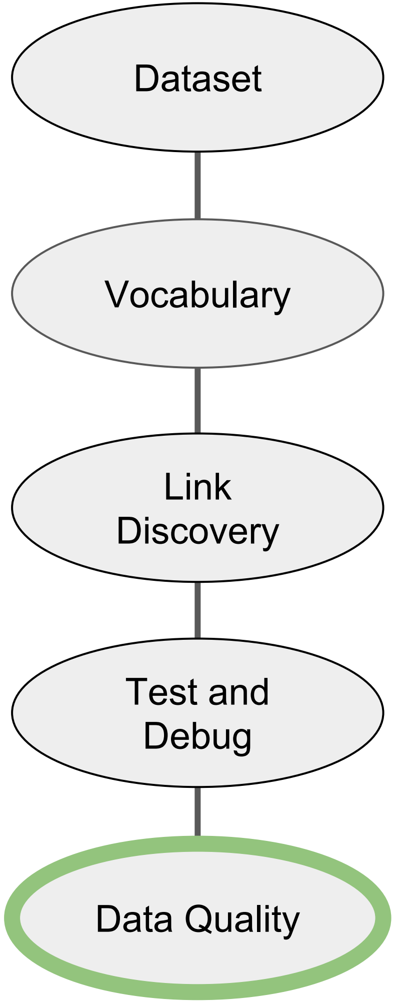
Test and Debug

- Check the syntax of RDF triples
  - W3C RDF Validator can check RDF/XML for syntactic correctness
- Check the infrastructure
  - RDF:Alerts



Data Quality

- Does your data set links to other data sets?
- Do you provide provenance metadata?
- Do you provide licensing metadata?
- Do you use terms from widely deployed vocabularies? Are the URIs of proprietary vocabulary terms dereferenceable?
- Do you map proprietary vocabulary terms to other vocabularies?
- Do you provide data set-level metadata?
- Do you refer to additional access methods?



Five Star Linked Data

* Data available on the web (in whatever format), but with an open licence

** Available as machine-readable structured data (e.g. Excel instead of image scan of a table)

*** All the above, plus: Use non-proprietary data format (e.g. CSV instead of Excel)

**** All the above, plus: Use open standards from W3C (e.g. HTTP URIs) to identify things, so that people can point at your stuff

***** All the above, plus: Link your data to other people’s data to provide context

How to publish Linked Data

- Serving Linked Data as Static RDF/XML Files
- Serving Linked Data as RDF Embedded in HTML Files
- Serving Linked Data from Relational Databases
- Serving Linked Data from RDF Triple Stores
- Serving Linked Data by Wrapping Existing Application or Web APIs

Consuming Linked Data

- Two basic types of generic Linked Data applications:
  - Linked Data browsers
    - allow users to navigate between data sources by following RDF links.
  - Linked Data search engines
    - crawl Linked Data from the Web by following RDF links, and provide query capabilities over aggregated data.

References

- VOID - https://www.w3.org/TR/void/
- Dean Allemang and Jim Hendler. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Morgan Kaufmann, 2008.
- Tom Heath and Christian Bizer (2011) Linked Data: Evolving the Web into a Global Data Space(1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.