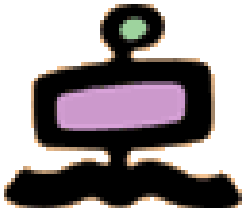
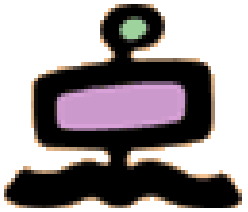
Corso di Biblioteche Digitali



- Vittore Casarosa
  - casarosa@isti.cnr.it
  - Office: 050 621 3115
  - Mobile: 348 397 2168
  - Skype: vittore1201

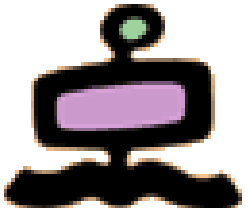
- “Ricevimento” at the end of the lessons or by appointment
- Final assessment
  - 70% oral examination
  - 30% project (development of a small digital library))
- Reference material:
  - Ian Witten, David Bainbridge, David Nichols, How to build a Digital Library, Morgan Kaufmann, 2010, ISBN 978-0-12-374857-7 (Second edition)
  - Material provided by the teacher
- **http://cloudone.isti.cnr.it/casarosa/BDG/**

Modules



- Computer Fundamentals and Networking
- A conceptual model for Digital Libraries
- Bibliographic records and metadata
- Information Retrieval and Search Engines
- Knowledge representation
- Digital Libraries and the Web
- Hands-on laboratory: the Greenstone system

Parallel evolution



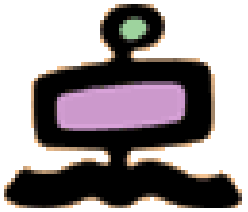
Libraries

- Description (documents)
  - Bibliographic records
  - MARC
- Interoperability
  - Z39-50
- Conceptual model (classes)
  - FRBR – LRM for Works, Expr., Manif.
- Information Retrieval
  - Full text (catalogue and documents)

The Web

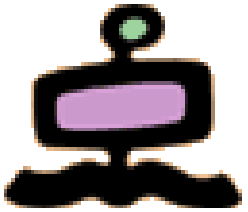
- Description (instances)
  - Metadata
  - Dublin Core
- Interoperability
  - OAI-PMH
- Conceptual model (classes)
  - **RDF and RDF Schema for all resources (ontologies)**
- Information Retrieval
  - Full text (web pages and resources)

Knowledge representation



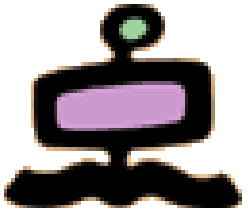
- FRBR: Functional Requirements for Bibliographic Records
  - LRM: Library Reference Model
- RDF: Resource Description Framework
- RDF Schema
  - LOD: Linked Open Data

New requirements from the Web



- Increase in the amount of information available on-line (data bases, repositories, the Web, etc)
- Increase in the variety of resources available on-line (text, sound, images, video, 3D, etc)
- Need to better describe how to find and how to access resources on the Web
  - Dublin Core
  - RDF syntax and RDF Schema (to define ontologies)
  - LOD: Linked Open Data

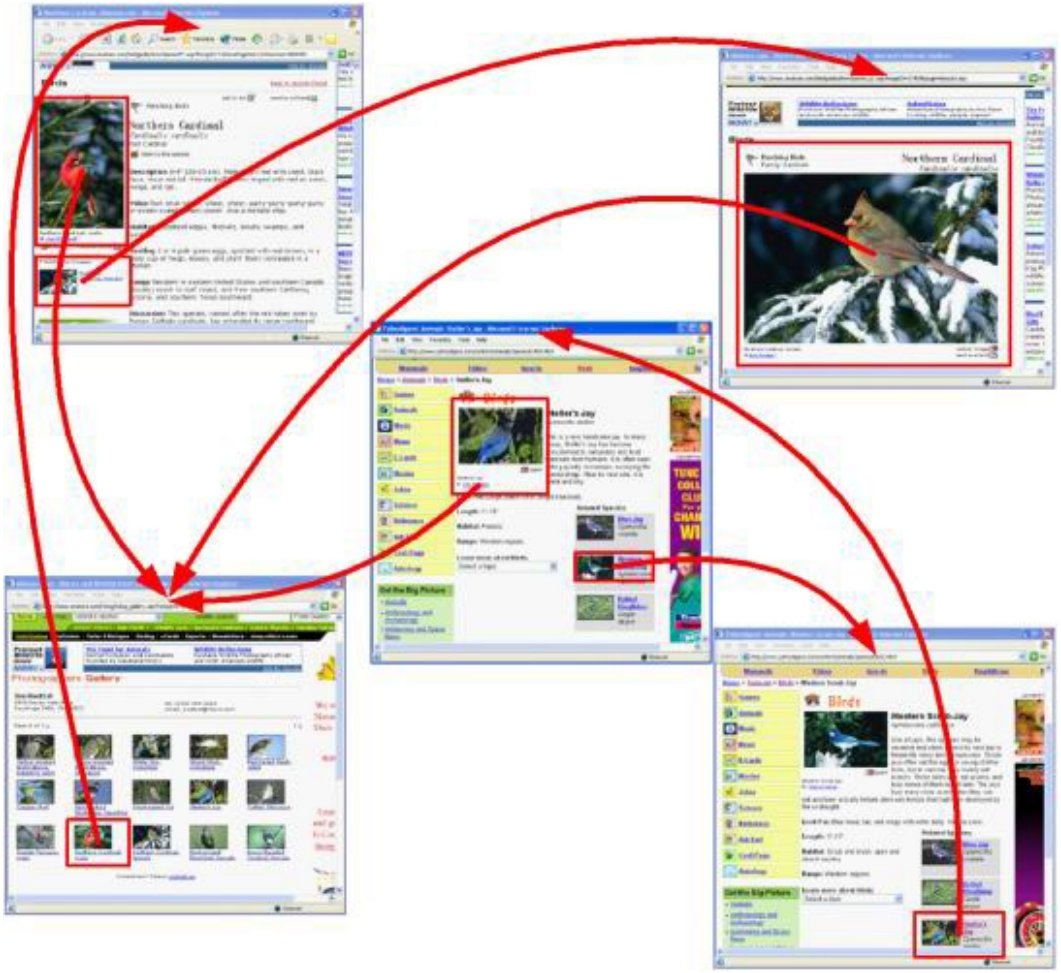
“I have a dream”



“I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A ‘Semantic Web’, which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The ‘intelligent agents’ people have touted for ages will finally materialize”

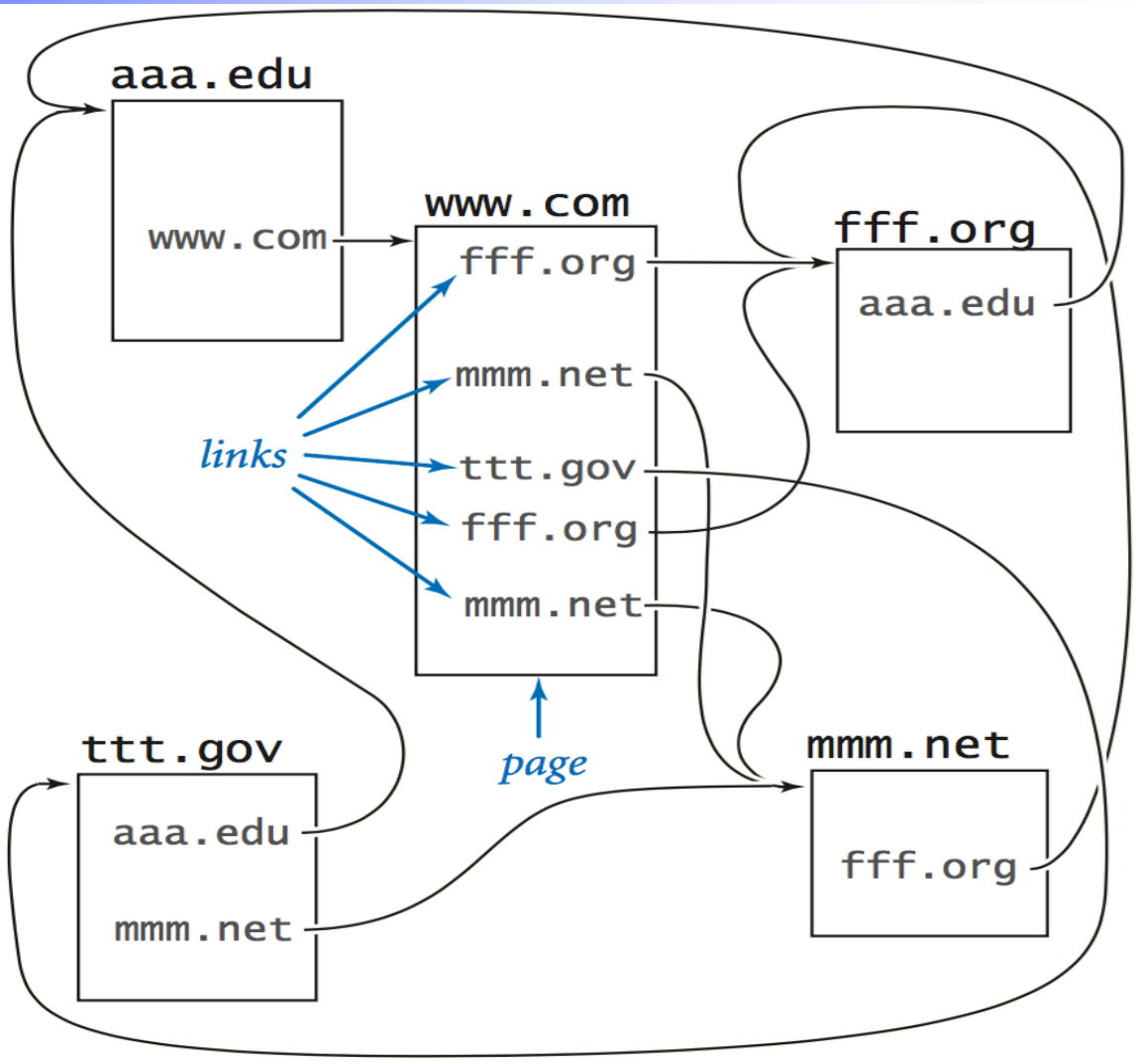
Tim Berners-Lee, 1999

The World Wide Web

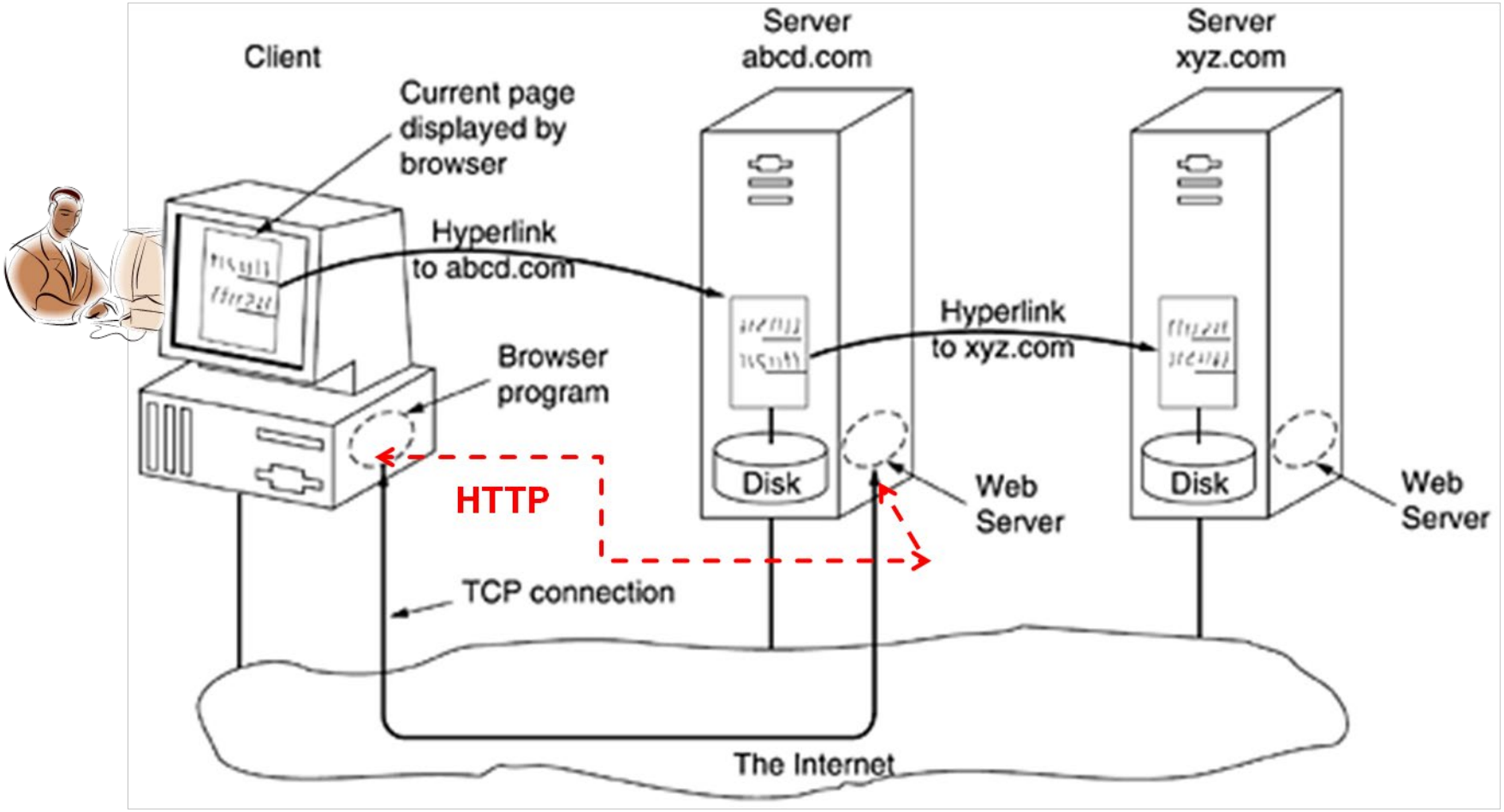


The hyperlinks

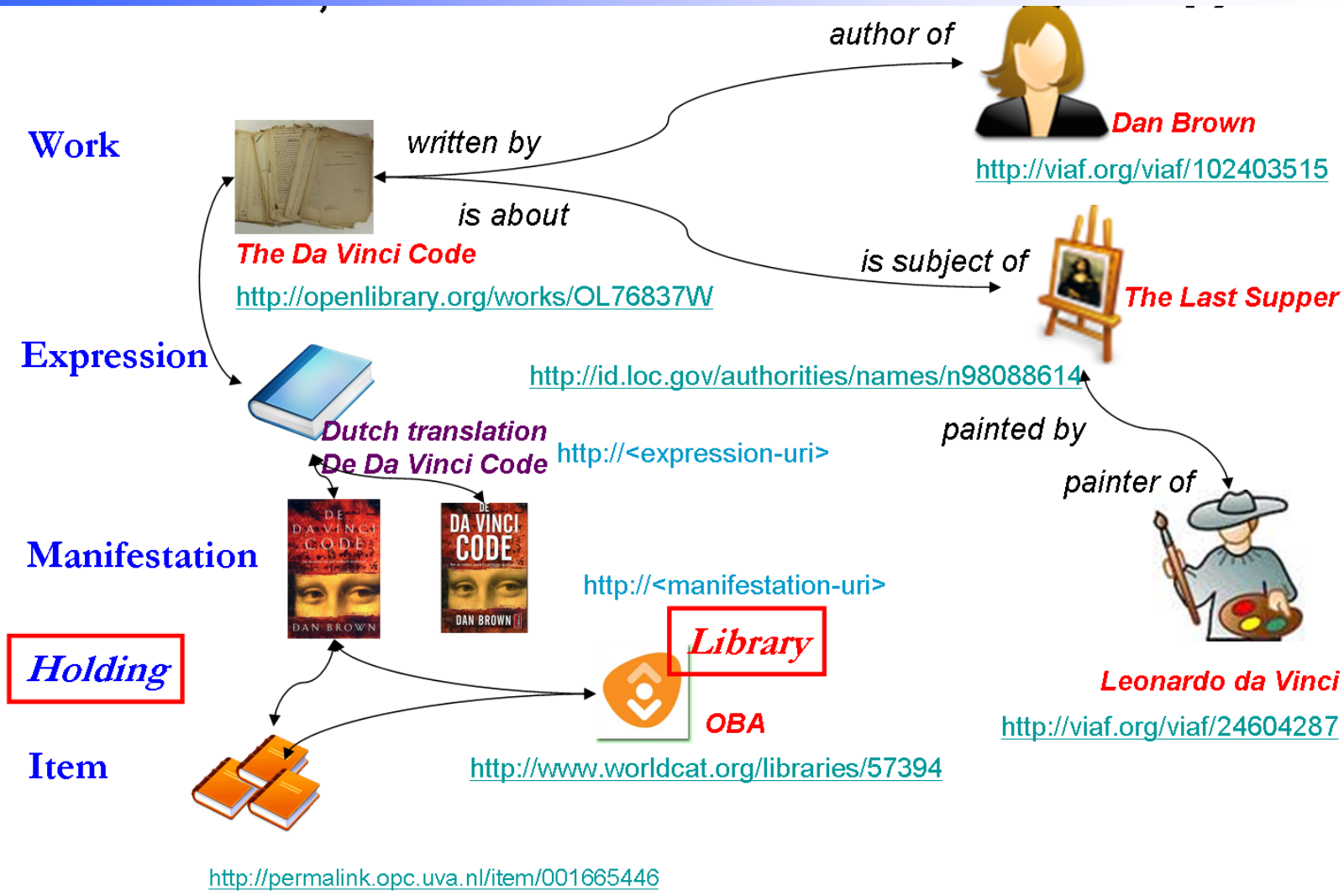
A link is made of two parts:
the visible text or image (the “anchor text”) and the link to the resource (typically a web page) to be accessed (brought into your own PC) when clicking on the visible part



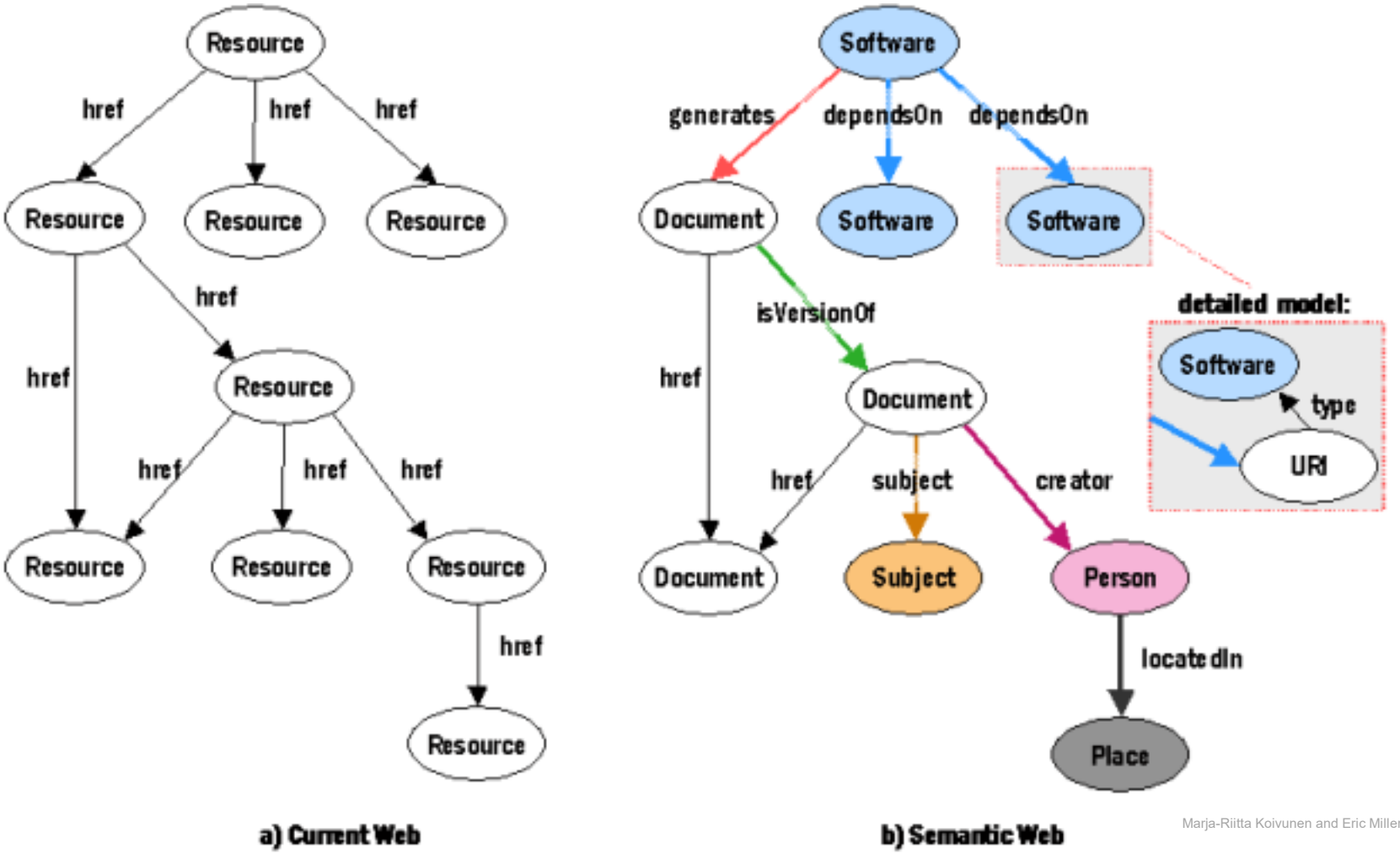
The Web architecture



The Web of Linked Data

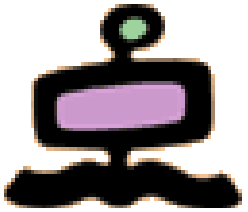


The Semantic Web



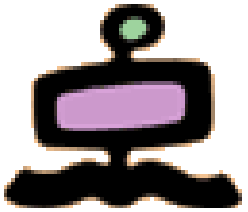
Marja-Riitta Koivunen and Eric Miller w3.org

The Semantic Web



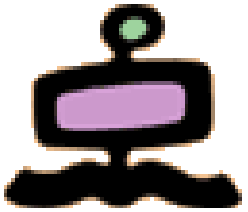
- The whole idea of the Semantic Web is to make available (for use in the Web) resources (or resource descriptions) whose “meaning” is understandable by a computer
- This is accomplished by providing descriptions of resources in a “formal way”, so that these descriptions can be “understood” by a computer (i.e. a program running in a computer)
- The first step in approaching this formal description is to define exactly the “portion of the universe” that we want to describe, and then define a “conceptual model” of it

What is a Conceptual Model ?



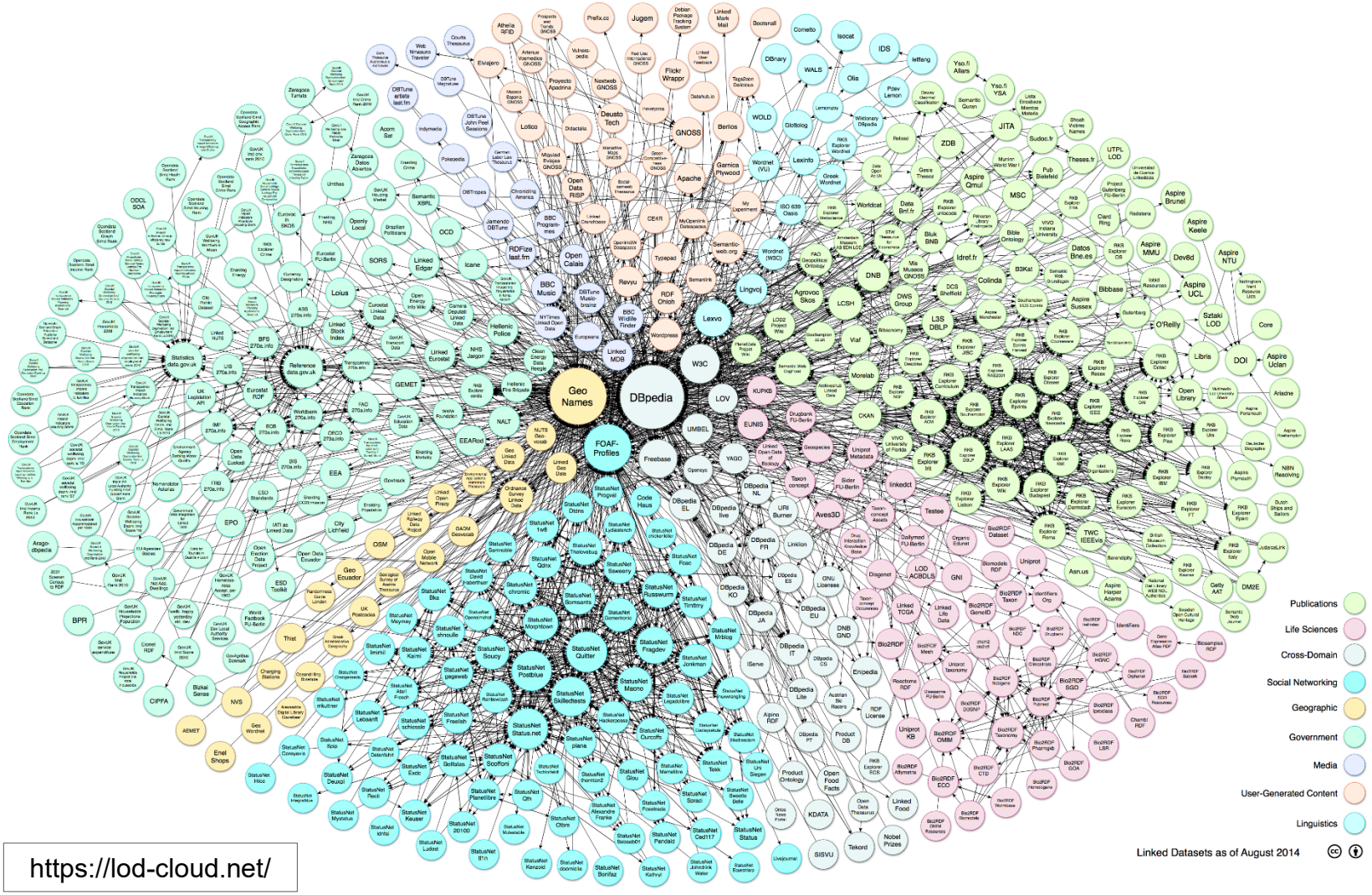
- A conceptual model is an **abstract framework** for identifying and understanding significant relationships among the entities of interest in some “part of the universe” and for the development of consistent standards or specifications supporting that environment
- A conceptual model is based on **a small number of unifying concepts** (entities), their **properties** (attributes) and their **relationships**
- A conceptual model **is not directly tied to any standards, technologies or other concrete implementation details**, but it does seek to provide a common semantics that can be used unambiguously across and between different implementations

LOD and the Semantic Web



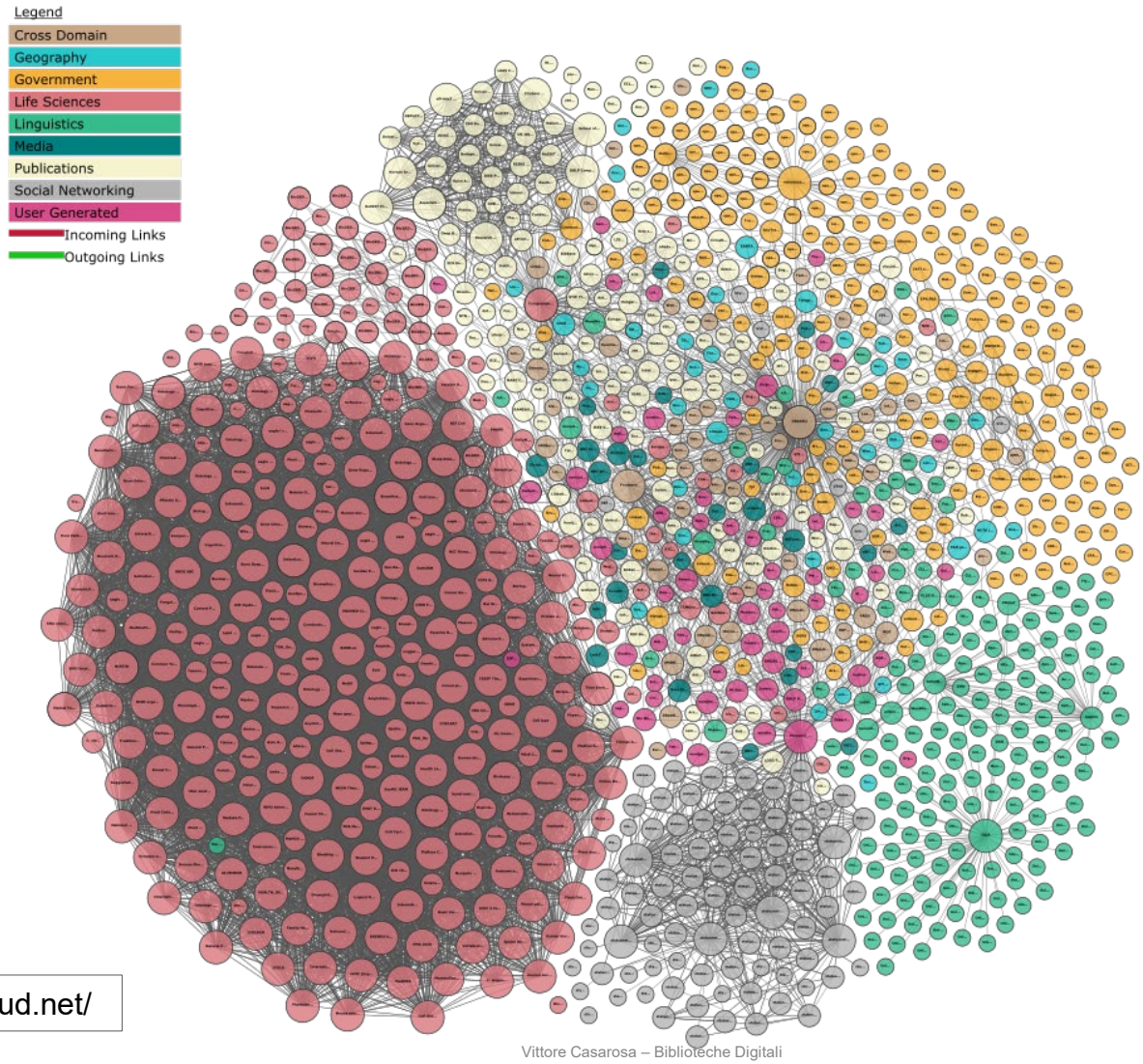
- The main formalism used today for describing resources is RDF – Resource Description Framework
- The RDF descriptions are based on RDF schemas (often called vocabularies or ontologies), which are also described in RDF (they are the “conceptual models”)
- One of the main initiatives in the Semantic Web is “Linked Open Data” (LOD), where the resources (or their descriptions) to be made freely available on the Web must be described in RDF and must be linked one to another with “typed links” (i.e. RDF predicates)
- The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web
- An increasing number of data providers over the last years have contributed to the creation of a global data space containing billions of statements (RDF triples)

LOD in the Web (August 2014)



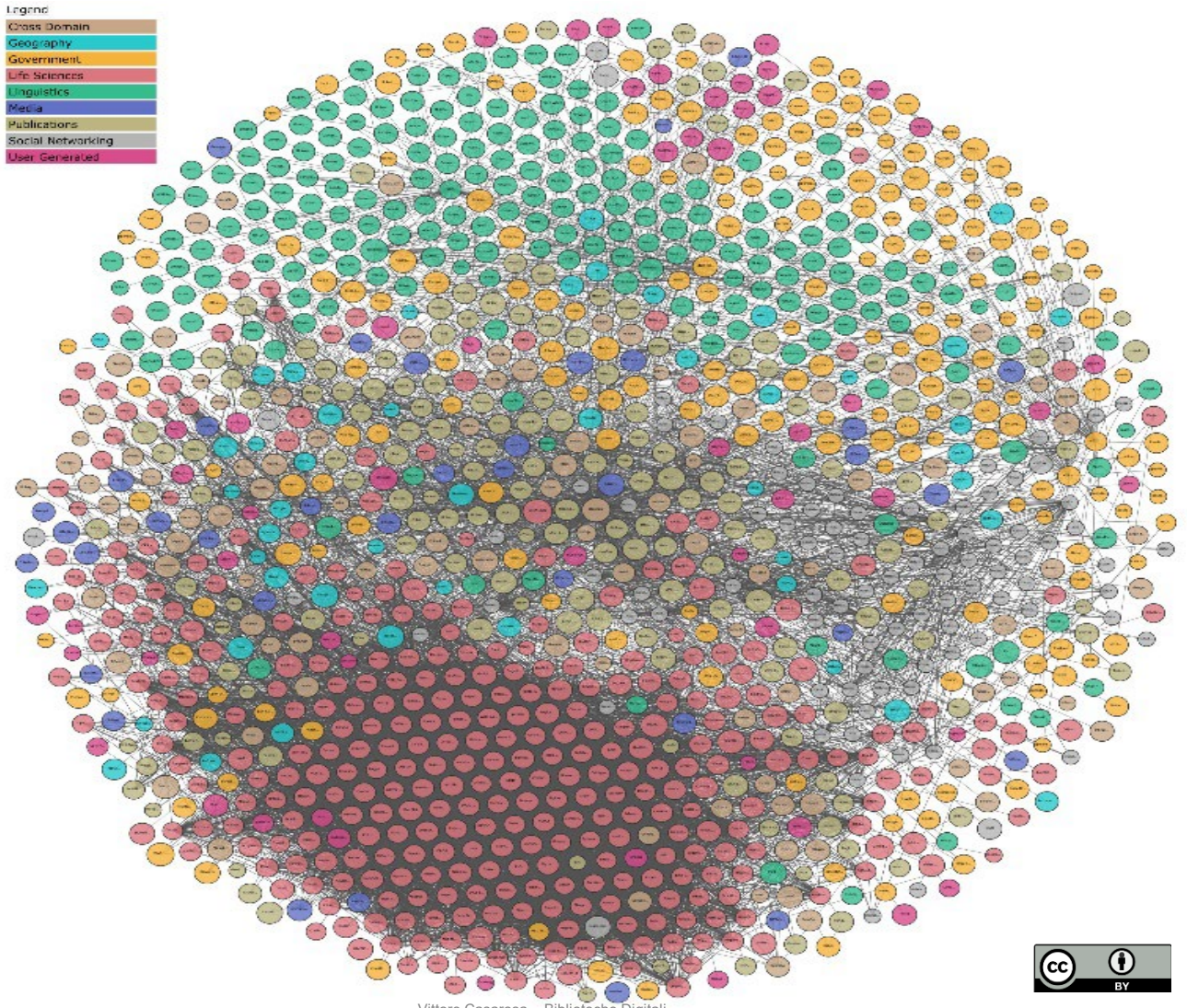
https://lod-cloud.net/

LOD in the Web (September 2017)



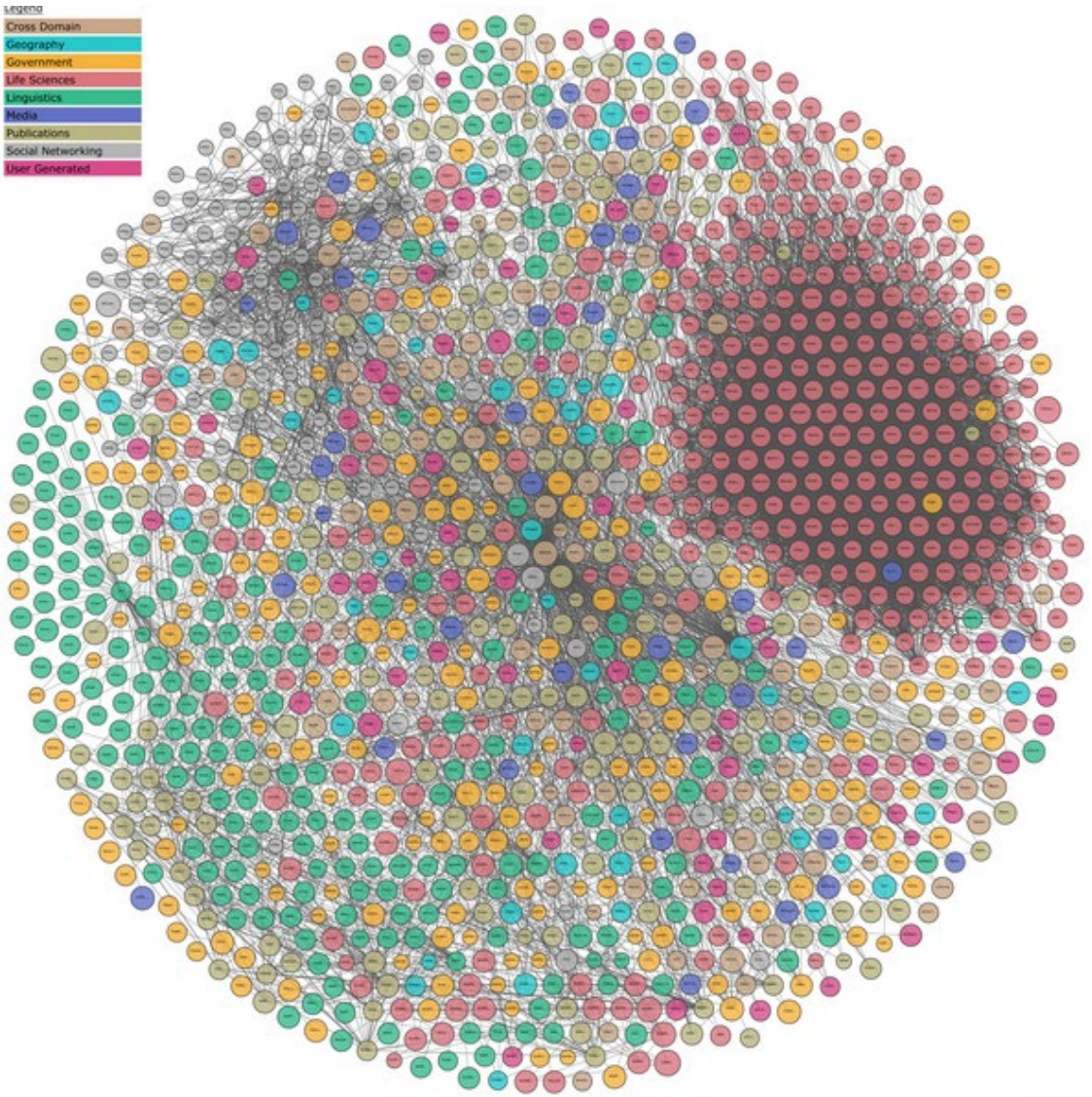
https://lod-cloud.net/

LOD in the Web (May 2020)



https://lod-cloud.net/

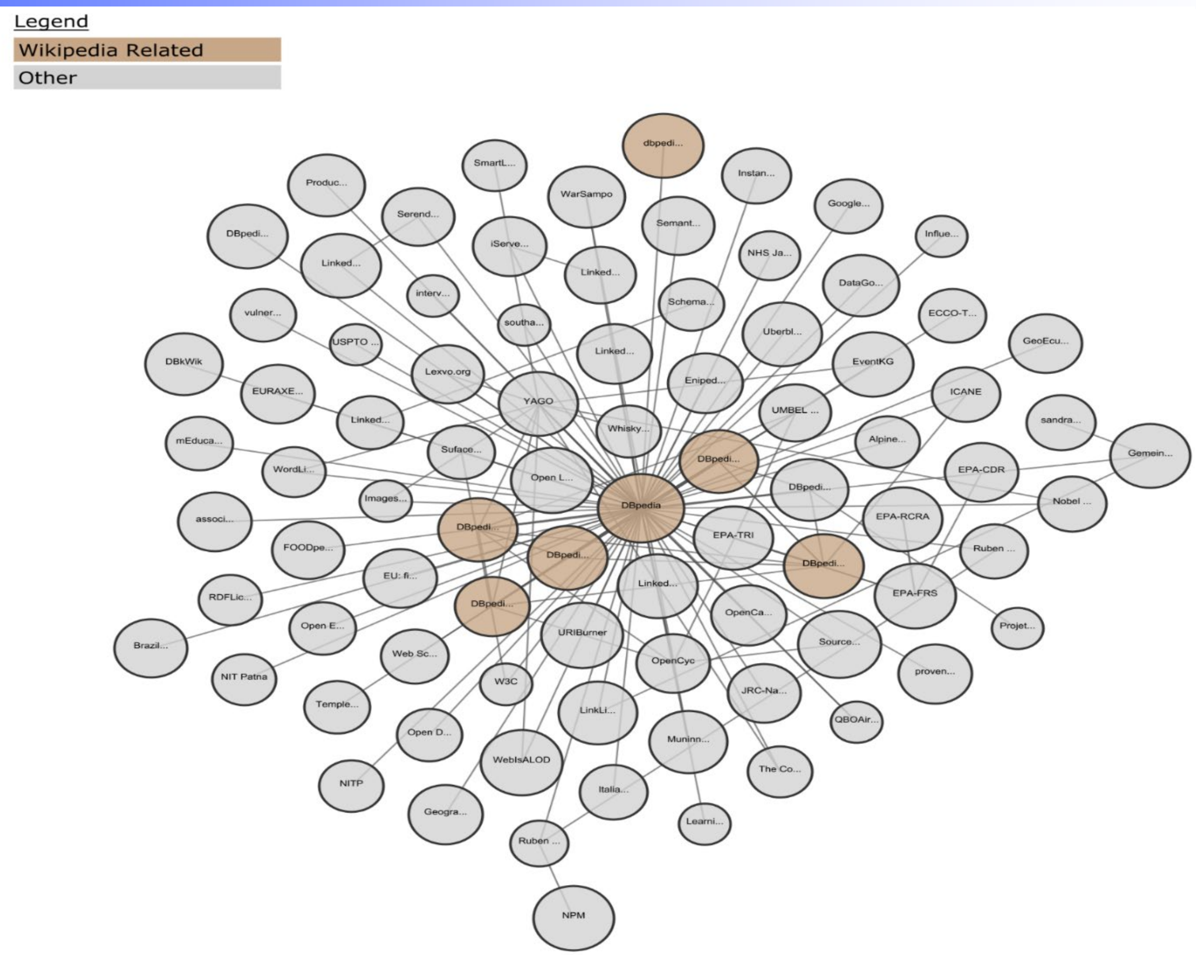
LOD in the Web (Nov 2024)



The LOD cloud currently contains 1349 datasets

https://lod-cloud.net/

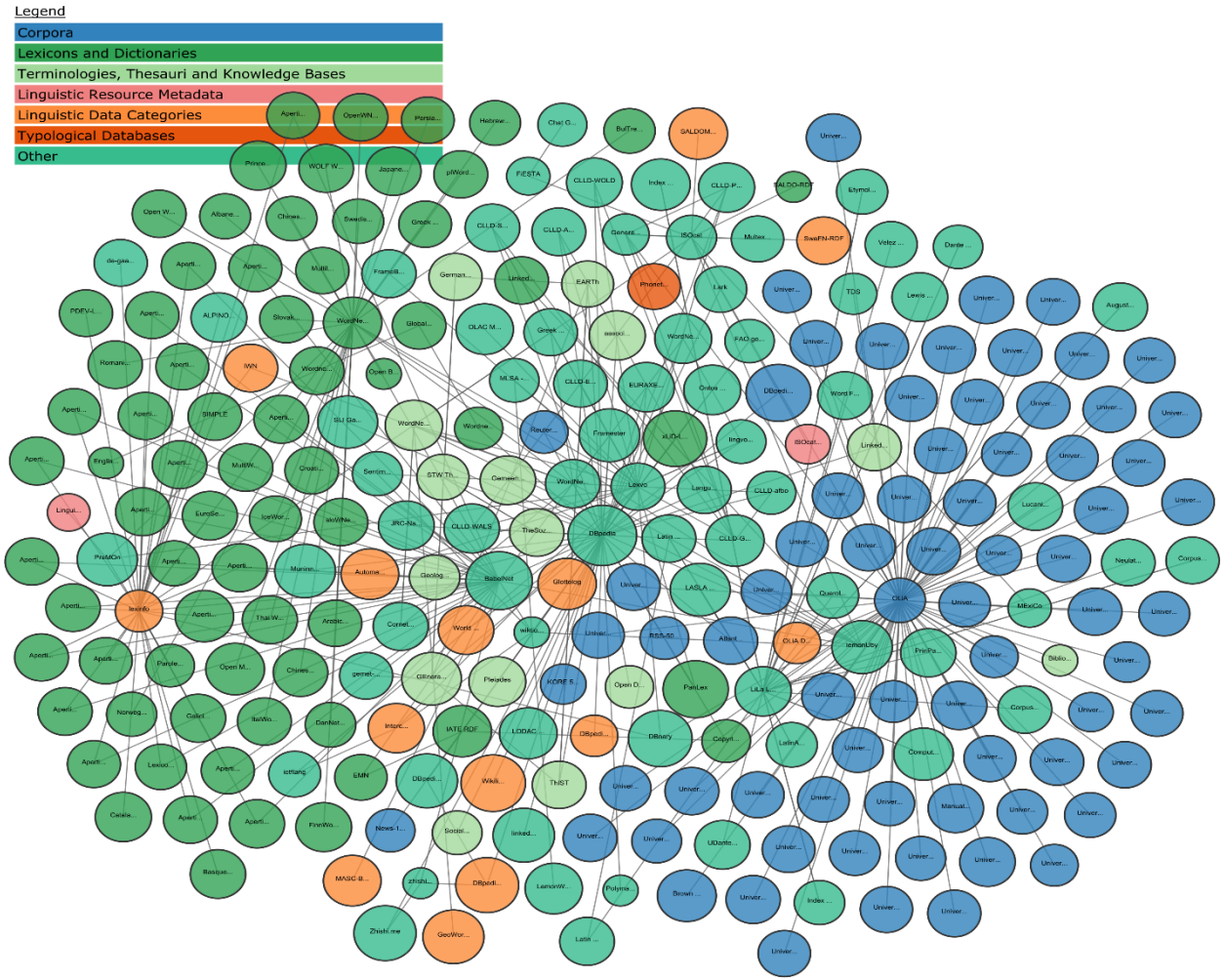
LOD sub-domains



https://lod-cloud.net/

The Cross-Domain Linked Open Data Cloud from lod-cloud.net

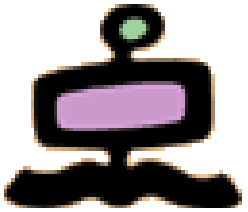
Linked Open Data for Linguistics



https://lod-cloud.net/

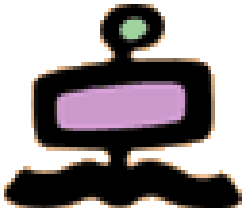
The Linguistic Linked Open Data Cloud from lod-cloud.net

Rules for publishing Linked Data sets



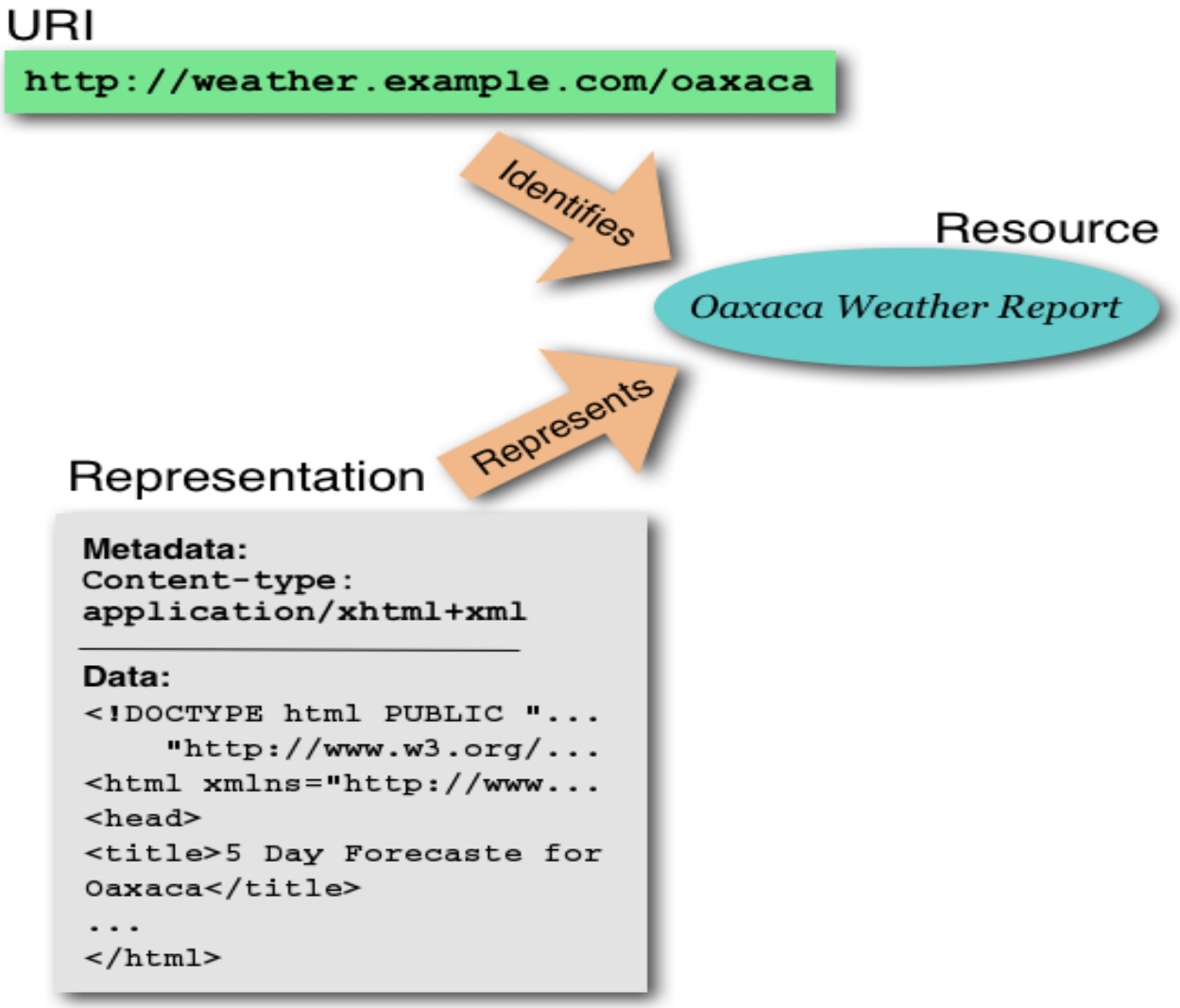
- Publish data according to the Linked Data principles
- There must be resolvable http:// (or https://) URIs.
- They must resolve (with or without content negotiation) to RDF data in one of the popular RDF formats (RDFa, RDF/XML, Turtle, N-Triples, JSON).
- The dataset must contain at least 1000 triples
- The dataset must be connected via RDF links to a dataset that is already in the diagram; we arbitrarily require at least 50 links.
- Access to the entire dataset must be possible via RDF crawling, via an RDF dump, or via a SPARQL endpoint.

Linked Data principles

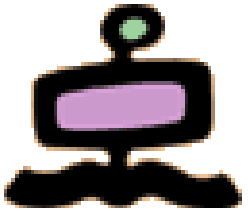


- Use URIs as names for things
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (e.g. HTML, RDF*, SPARQL)
- Include links to other URIs, so that they can discover more things.

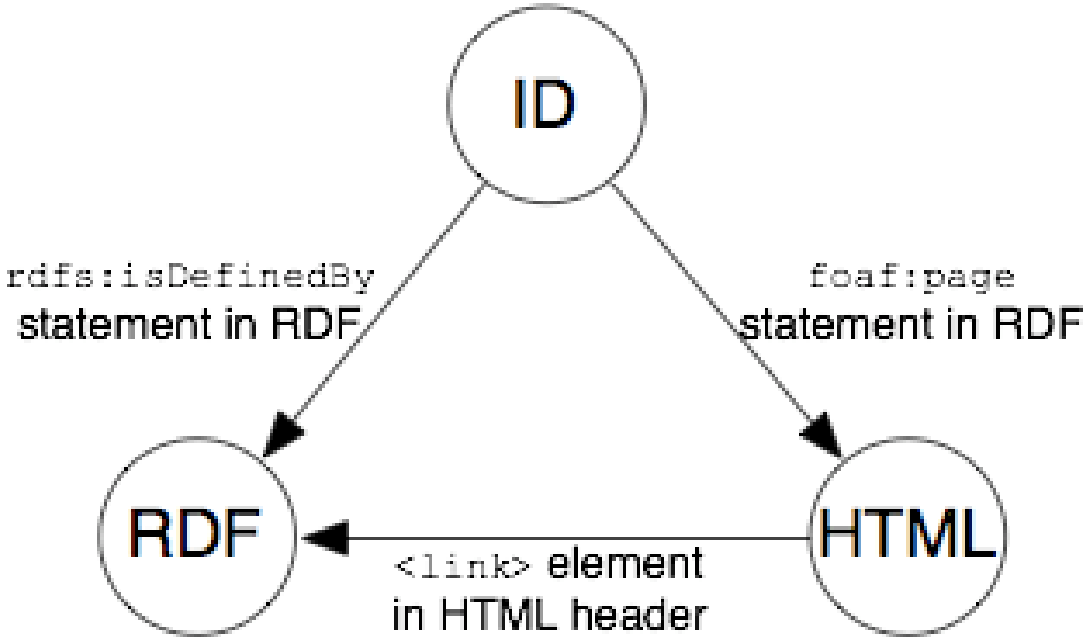
Representation of resources



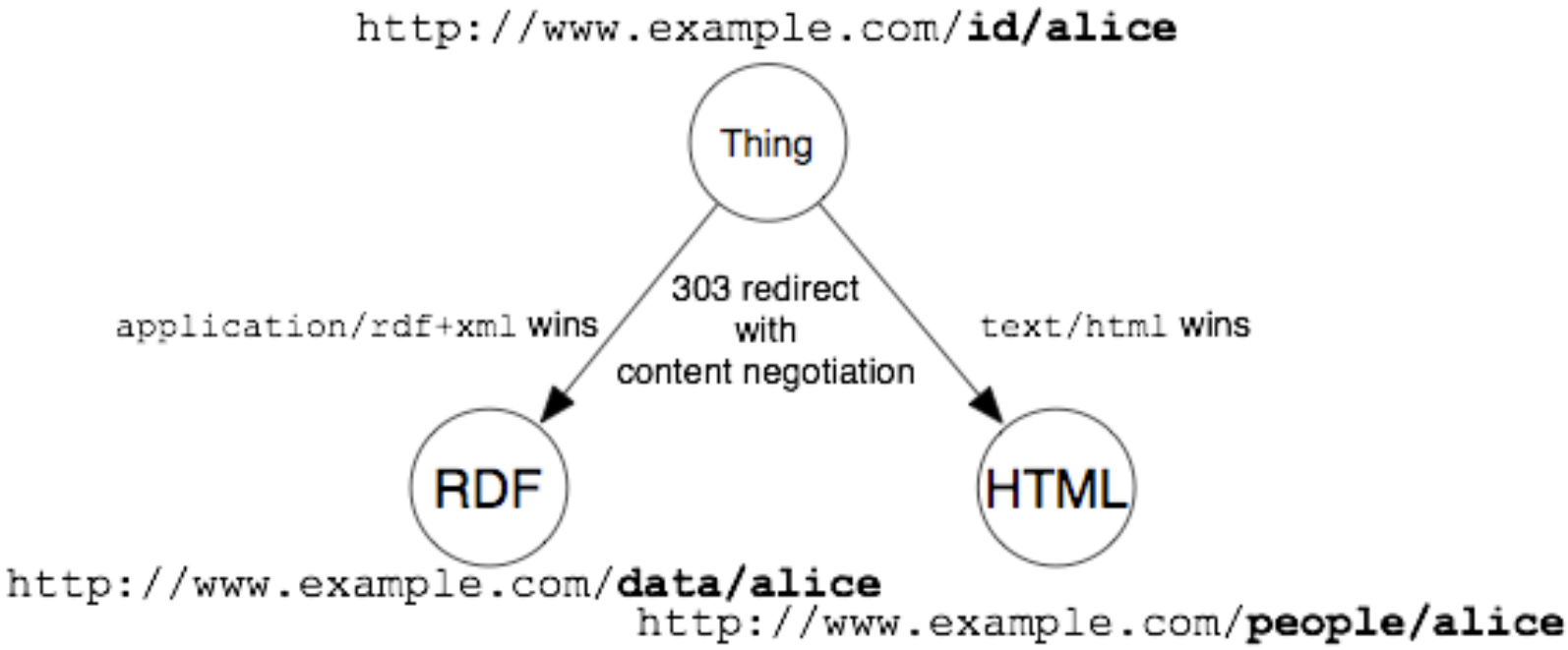
Three URIs for a resource



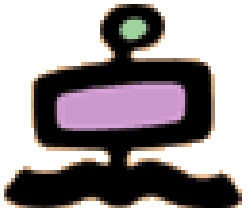
http://www.example.com/id/alice	Identifier for Alice, the person
http://www.example.com/people/alice	Alice's homepage
http://www.example.com/data/alice	RDF document with description of Alice



Dereferencing a URI

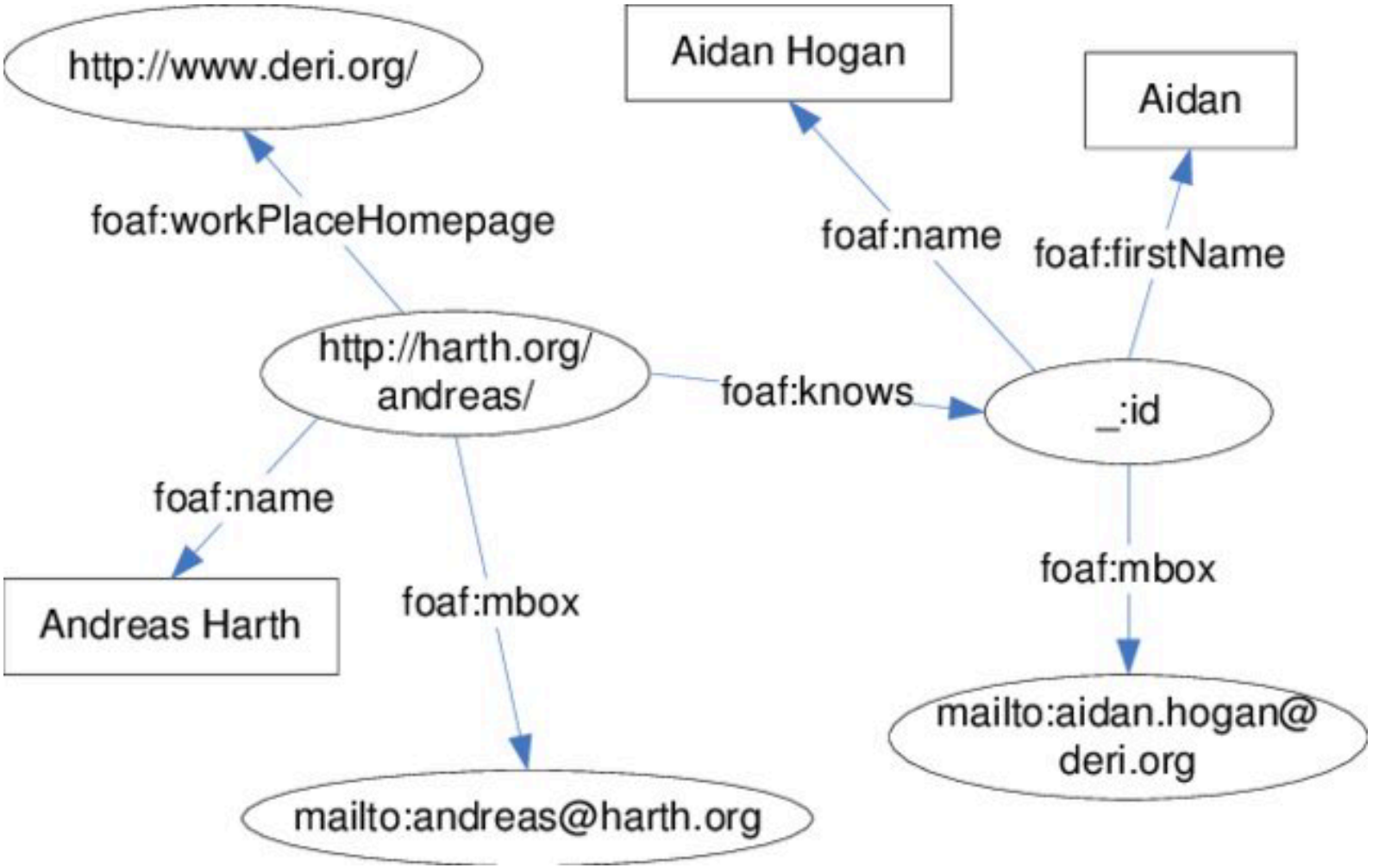


Using “de facto standards”



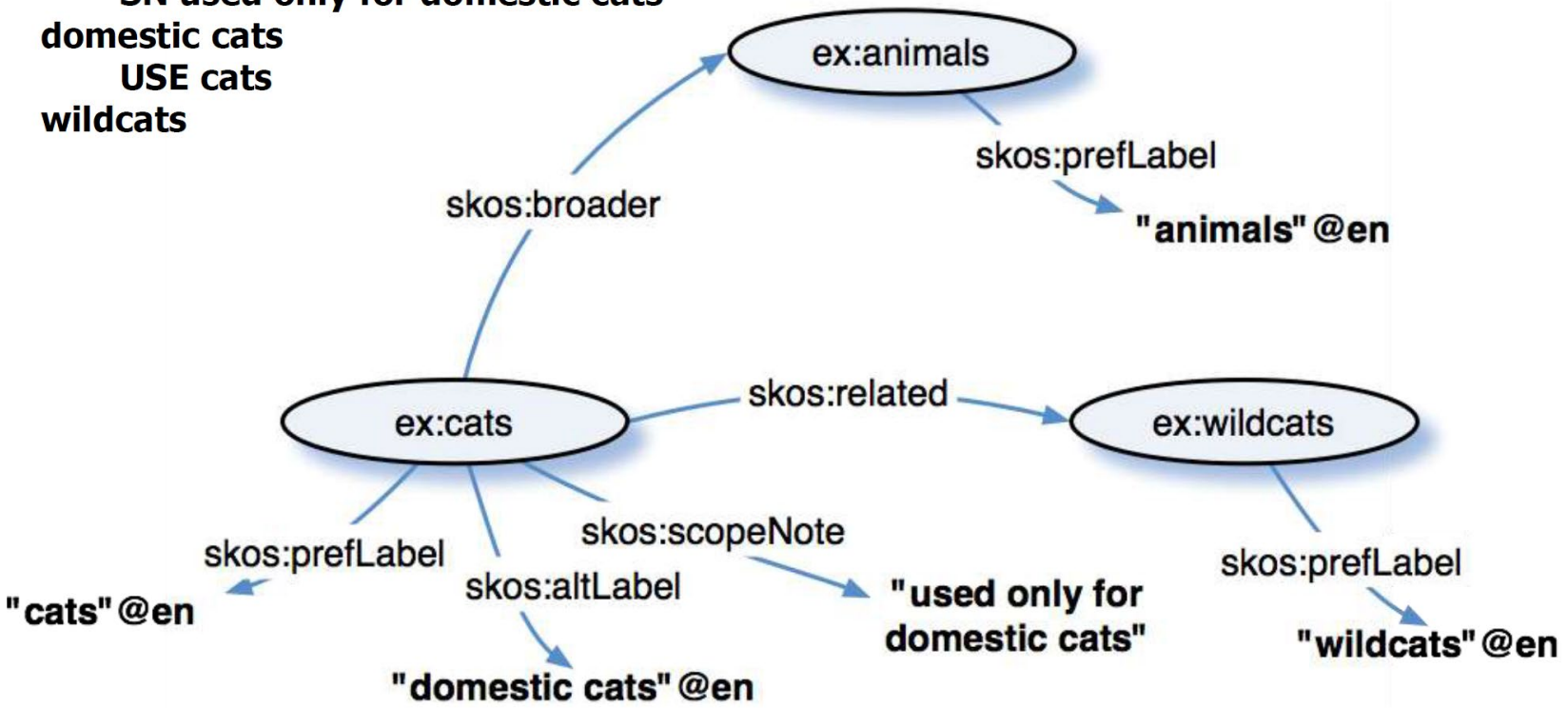
- Different communities have specific preferences on the vocabularies they prefer to use for publishing data on the Web.
- The Web of Data is therefore open to arbitrary vocabularies being used in parallel.
- Despite this general openness, it is considered good practice to reuse terms from well-known RDF vocabularies such as FOAF, SKOS, DOAP, vCard, Dublin Core, or Good Relations wherever possible in order to make it easier for client applications to process Linked Data.
- Only if these vocabularies do not provide the required terms should data publishers define new, data source-specific terminology

Friend of a Friend
foaf example

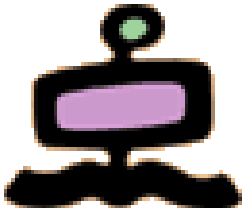


Simple Knowledge Organization System – SKOS example

animals
cats
UF domestic cats
RT wildcats
BT animals
SN used only for domestic cats
domestic cats
USE cats
wildcats

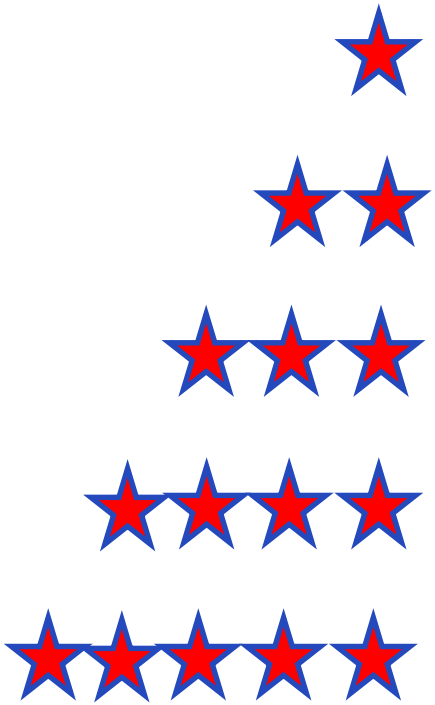


Datasets



- Billions of Linked Data facts are published on the Web with an open license.
- A well-known dataset is DBpedia.
  - Data is extracted automatically from Wikipedia infoboxes.
  - Like Wikipedia, it exists in several different languages.
  - Its quality is acceptable for many queries.
- Another well-known dataset is Wikidata, which is a manually curated alternative.
  - It has its own data model on top of RDF.
  - It grows fast and might overtake DBpedia
  - It is one of the Wikimedia projects
- You can find many other datasets on Datahub.io

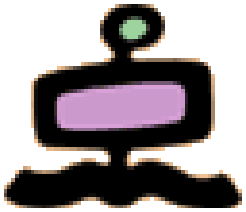
Five Star Open Data



- make your stuff available on the Web (whatever format) under an open license
- make it available as structured data (e.g., Excel instead of image scan of a table)
- use non-proprietary formats (e.g., CSV instead of Excel)
- use URIs to denote things, so that people can point at your stuff
- link your data to other data to provide context

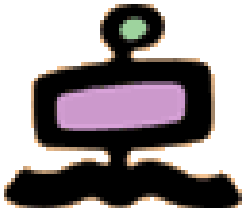
Usually Open Data is available under a CC-BY-SA license. This means you can include it in any other work (Creative Commons) under the condition that you give proper attribution (created BY). If you create derivative works (such as modified or extended versions of the Open Data), then you must also license them as CC-BY-SA (Share Alike).

Creative Commons



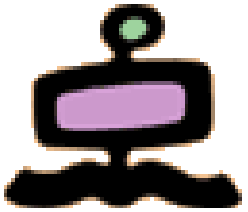
Icon	Description	Acronym	Attribution Required	Allows Remix culture	Allows commercial use	Allows Free Cultural Works	Meets 'Open Definition'
PUBLIC DOMAIN	Freeing content globally without restrictions	CC0	No	Yes	Yes	Yes	Yes
BY	Attribution alone	BY	Yes	Yes	Yes	Yes	Yes
BY SA	Attribution + ShareAlike	BY-SA	Yes	Yes	Yes	Yes	Yes
BY NC	Attribution + Noncommercial	BY-NC	Yes	Yes	No	No	No
BY NC SA	Attribution + Noncommercial + ShareAlike	BY-NC-SA	Yes	Yes	No	No	No
BY ND	Attribution + NoDerivatives	BY-ND	Yes	No	Yes	No	No
BY NC ND	Attribution + Noncommercial + NoDerivatives	BY-NC-ND	Yes	No	No	No	No

LOD in the libraries



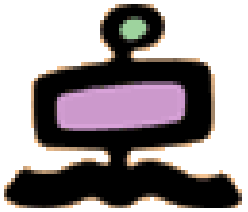
- Memory institutions are key players in providing knowledge:
  - this is their mission
  - their knowledge is trusted and of high quality
- Nowadays, knowledge is shared on the web
  - human consumable knowledge is expressed in natural languages and shared via HTML documents
  - machine consumable knowledge is expressed in RDF and shared through Linked Data
- Memory institutions have a key role to play in Linked Data
- Libraries, in particular, can offer their knowledge to the rest of the world by:
  - encoding it in RDF
  - using standard vocabularies for classes and properties
  - using well-known URIs for naming resources such as people, places, times, concepts, events
  - providing URIs for their own resources so that other institutions can use them

Methodology



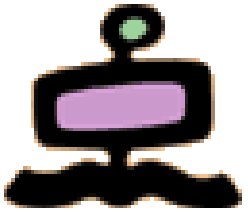
- The steps form the basis for different workflows that can be used to publish Linked Data, depending on purpose, data and context
- Data of interest:
  - knowledge organization systems (classification schemes, thesauri)
  - authority files
  - digital contents and their descriptions
  - catalogue data including circulation data sets.
- All these datasets should have links within themselves and should establish outgoing links to many other web resources, in order to attract many incoming links
- “Web Centric Cataloguing”

15 steps for publishing library data (e.g. the catalog)



1. Motivation
2. Management approval
3. Sorting out the legal and financial issues
4. Assessment of skills & data available
5. Tools assessment and evaluation
6. Dataset analysis
7. URI assignment
8. Vocabulary Modeling
9. Generation of RDF Data
10. Curating the data and outgoing links
11. Describing the data set
12. Evaluating the Dataset
13. Publishing
14. Incoming links
15. Maintenance

Conclusions



- Adopting linked data technologies allows libraries to
  - improve their presence where today's information is sought (i.e. the web)
  - improve the services offered to their users
  - promote innovative use of the data that the libraries held
- A small numbers of libraries (and even less archives and museums) have embraced the Linked Data paradigm
- Awareness is raising and knowledge is coming