Summer School LDA

Libraries in the digital age: linked data technologies for a global knowledge sharing

Pula (Cagliari), 29th August – 1st September 2016

Linked Open Data

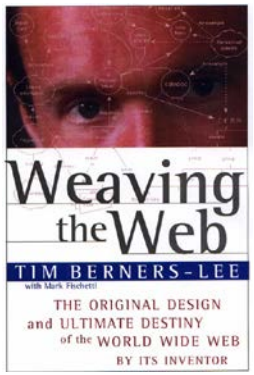
Oreste Signore

(W3C Italy)

Slides a: http://www.orestesignore.eu/education/lda/slides/lod.pdf

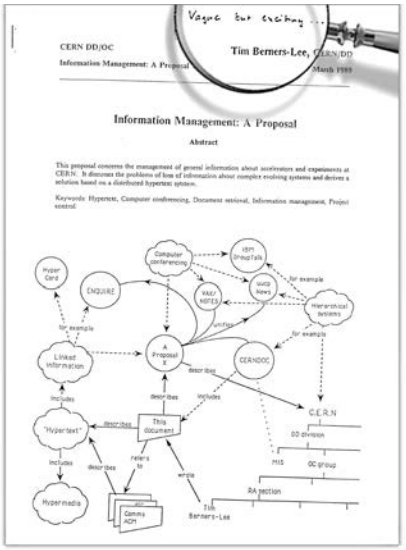
Talk layout

- The birth of Linked Open Data (LOD)
- Linked Open Data
  - ✓ benefits, principles, levels
- Web of Data & Semantic Web
  - ✓ Data integration
  - ✓ RDF (Resource Description Framework)
- One step forward: ontology
- Conclusion



Once upon a time…

- 1970(?) A boy was talking with his father:
  - ✓ How to make a computer intuitive, able to complete connections as the brain did
- 1980, while at CERN:
  - ✓ *Suppose all the information stored on computers everywhere were linked. Suppose I could program my computer to create a space in which anything could be linked to anything… There would be a single, global information space.*
- 1989 Vague but exiciting
- …and there was the Web…
- 1994
  - ✓ "The very first *International World Wide Web Conference*, at CERN, Geneva, Switzerland, in September 1994" http://www.w3.org/Talks/WWW94Tim/
- 1999 Semantic Web Activity in W3C (now: Data Activity)
- 2007 LOD (W3C Linking Open Data project)

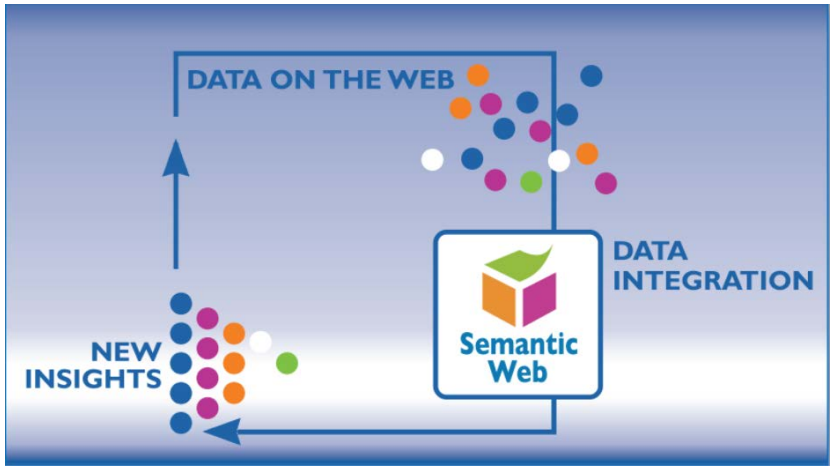


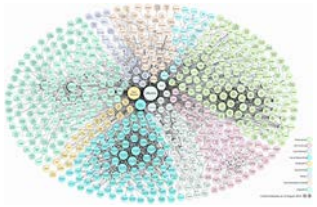
Web architecture

- Decentralization
- Basics
  - ✓ URI
    - The most fundamental innovation of the Web
    - Can address everything (resources, concepts)
  - ✓ HTTP
    - Format negotiation
    - Protocol to fetch resources
  - ✓ HTML
    - Structuring documents
- RDF (Resource Description Framework)
  - ✓ will be for the Semantic Web what HTML has been for the Web

Web of Data and Semantic Web

- **Semantic Web**
  - ✓ **Extends Web principles from documents to data**
  - ✓ **Creates the "*Web of Data*"**
- **Data (and not only data) can be**
  - ✓ **shared and reused in the Web**
- **RDF**
  - ✓ **Resource Description Framework**
  - ✓ **gives the abstraction layer to integrate data on the Web**



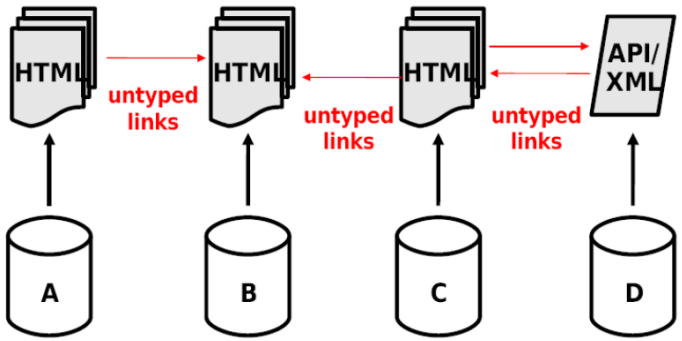


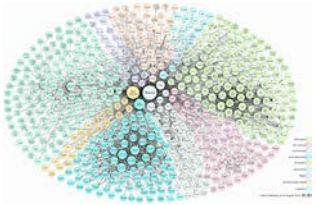
Linked Data

- A term used to describe a recommended best practice for *exposing*, *sharing*, and *connecting* pieces of data, information, and knowledge on the Semantic Web using URIs and RDF
  - ✓ (quoted in Wikipedia)
- See also:
  - ✓ http://linkeddata.org/
  - ✓ http://www.w3.org/standards/semanticweb/data

LOD: the benefits (1)

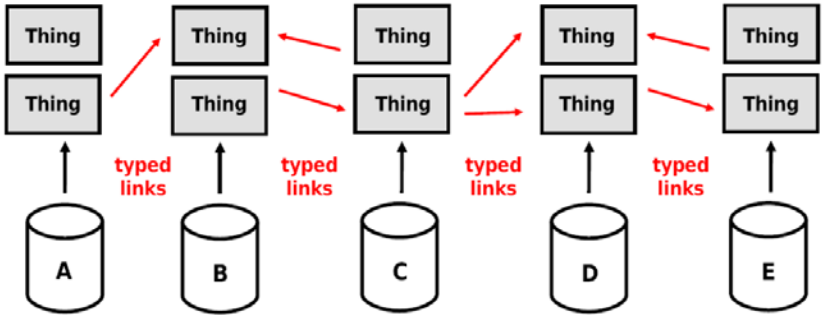
- From the Web of Documents …
  - ✓ A global filesystem
  - ✓ Documents are the primary objects
  - ✓ (Fairly structured) documents connected by untyped links
  - ✓ Implicit semantics of content and links
  - ✓ Designed for human consumption
  - ✓ Simplicity … but disconnected data





LOD: the benefits (cont.)

- … to the **Web of Data**
  - ✓ A global **database**
  - ✓ Primary objects: **Things** (or description of things)
  - ✓ **Typed links** between things (including documents)
  - ✓ High degree of **structure** in (description of) things
  - ✓ **Explicit semantics** of content and links
  - ✓ Designed for
    - **Machines** (first)
    - **Humans** (later)



LOD: the principles

Web of things in the world, described by data on the Web

- **What does LOD mean?**
  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.**

Tim Berners-Lee 2007

http://www.w3.org/DesignIssues/LinkedData.html

LOD: principle 1

Use URIs as names for things

- **URI identify:**
  - ✓ **Documents and digital contents available on the Web**
  - ✓ **Real objects and abstract concepts**
- **Only *HTTP URI*, not other schemas like URN or DOI, because:**
  - ✓ **Provide a simple way to create globally unique names in a decentralized fashion, as every owner of a domain name, or delegate of the domain name owner, may create new URI references**
  - ✓ **They serve not just as a name but also as a means of accessing information describing the identified entity**

LOD: principle 2

Use HTTP URIs so that people can look up those names

- **HTTP is the universal protocol to access Web resources**
- **All HTTP URI must be "*dereferenceable*"**
- **When URIs identify real objects, it's essential distinguish objects from documents that describe them**

LOD: principle 3

When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)

- **Use a single data model to publish data on the Web: *RDF***
- **RDF data model is very simple and strictly coherent with Web architecture**

LOD: principle 4

Include links to other URIs, so that they can discover more things

- **Links (named *RDF links*) are "*typed*"**
- **Set *RDF links* towards other data sources on the Web**
  - ✓ **An external RDF link (having p and/or o defined in an external dataset) allows to access data on remote servers**
  - ✓ **The process is repeated in cascade**
  - ✓ ***External RDF links* are the glue that connects data islands into a global, interconnected data space**

The LOD five levels

★ **On the web**

Available on the web (whatever format) *but with an open licence, to be Open Data*

★★ **Machine-readable data**

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

★★★ **Non-proprietary format**

as (2) plus non-proprietary format (e.g. CSV instead of excel)

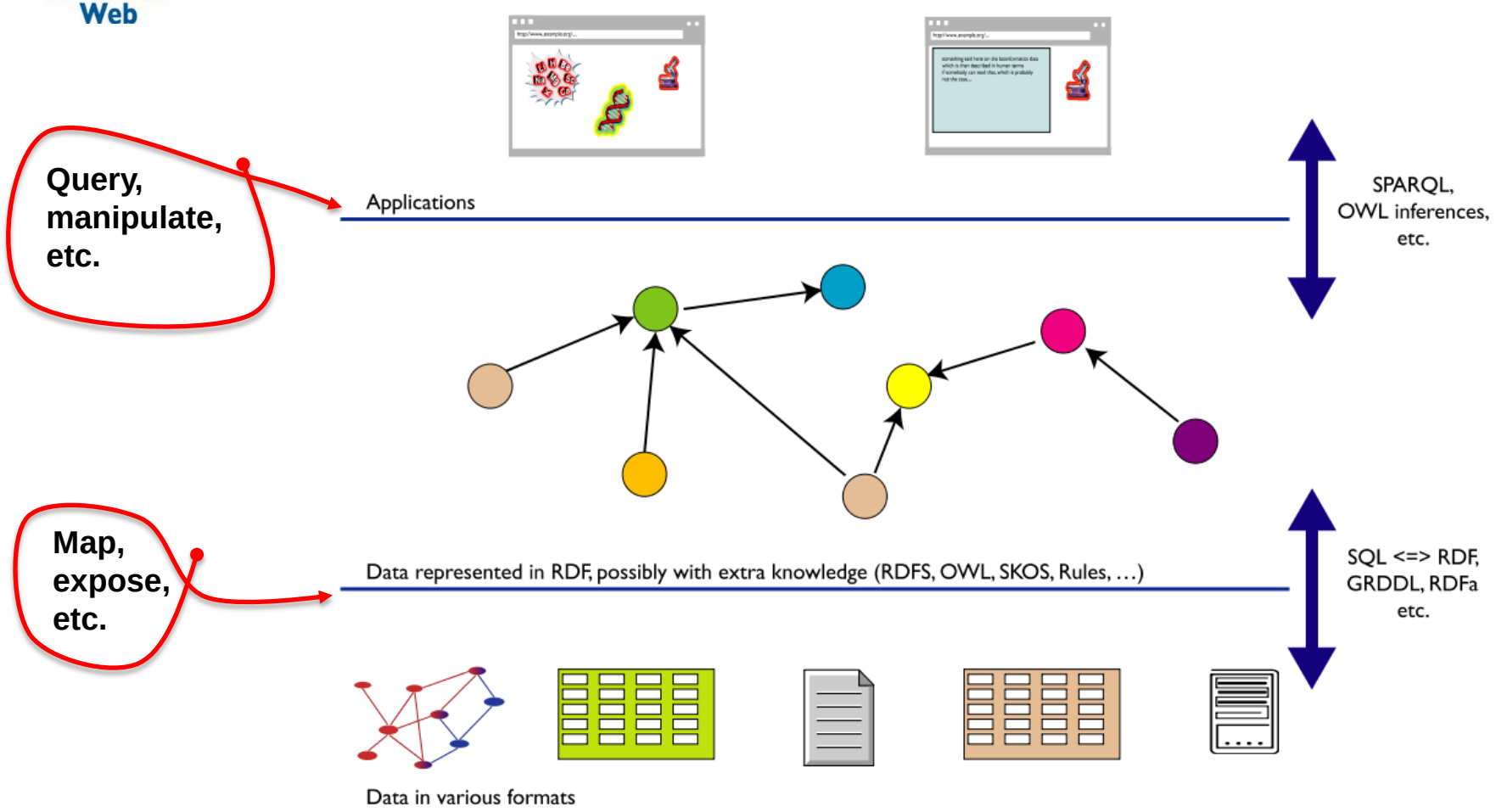
★★★★ **RDF standards**

All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff

★★★★★ **Linked RDF**

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

SW and Data Integration

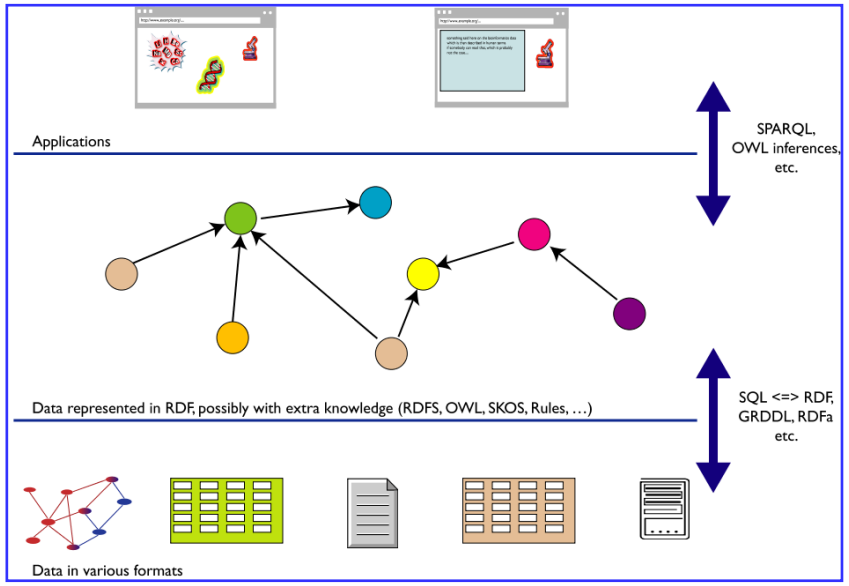


No need to put all your data in RDF!

SW and Data Integration: some advantages

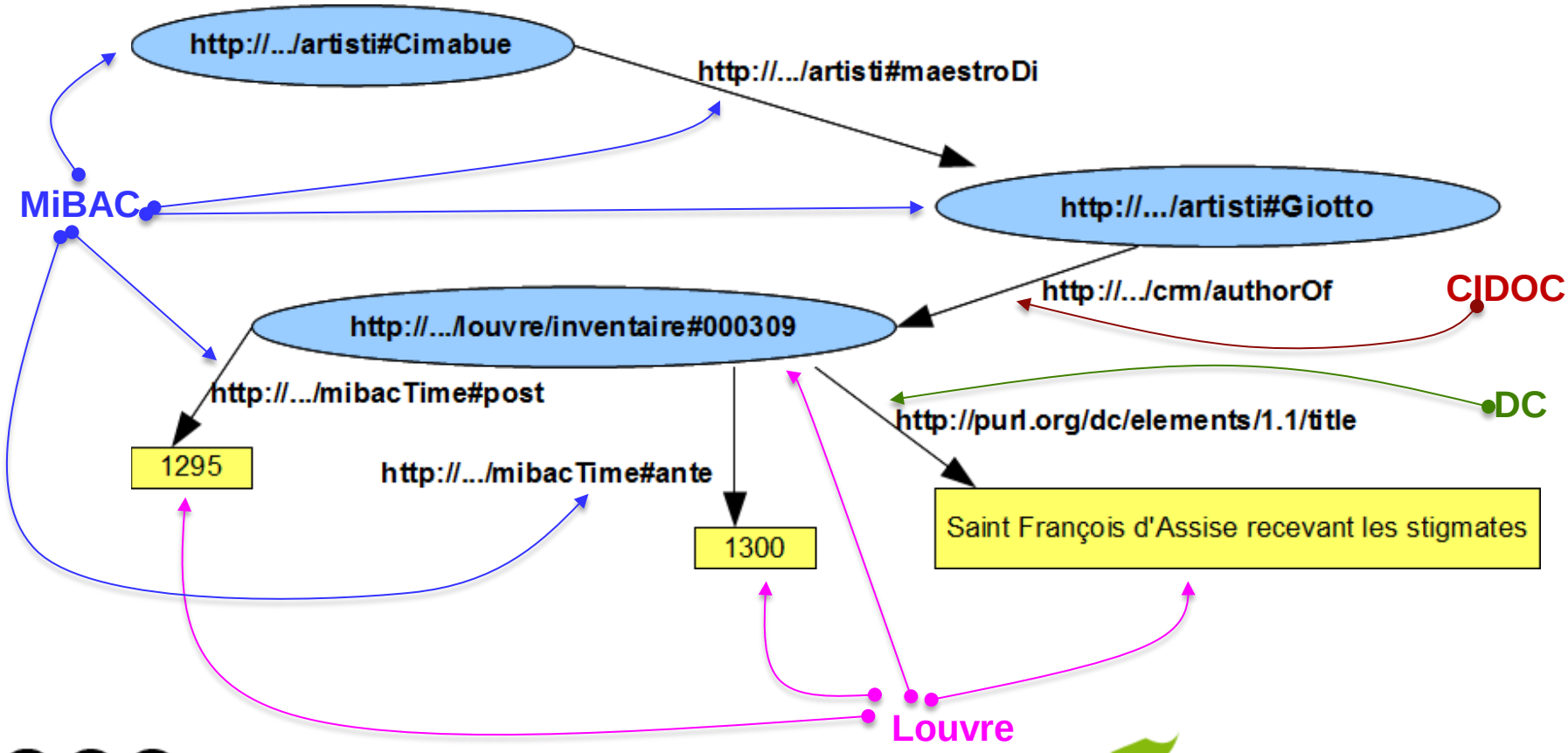
- Representation as a graph
  - ✓ independent of the actual structure of the data
- Changes to the format of the local database, etc.
  - ✓ have no influence on the general level
  - ✓ affect only the level of the step of exporting data (schema independence)
- You can
  - ✓ add new data
  - ✓ add more connections

  seamlessly, regardless of the structure of other data sources



A RDF graph (annotated)

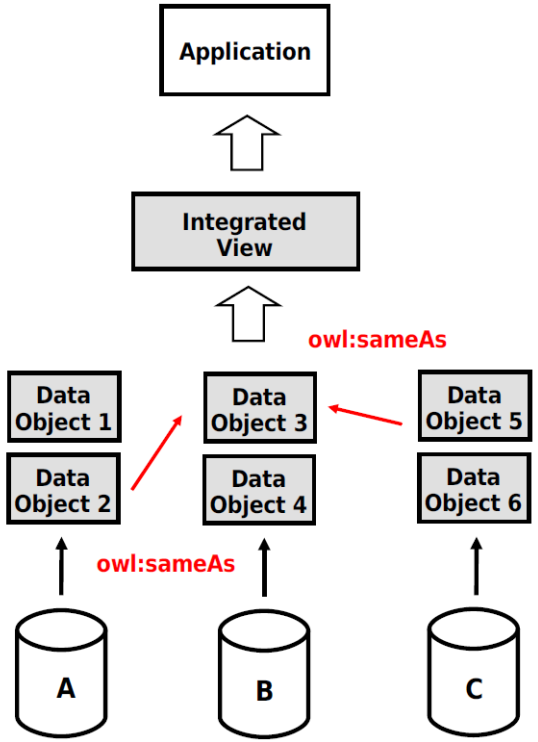
...a set of s-p-o (subject-predicate-object) triples



Reconciling differences

- For classes:
  - ✓ owl:equivalentClass: two classes have the same individuals
- For properties:
  - ✓ owl:equivalentProperty
- For individuals:
  - ✓ owl:sameAs: two URIs refer to the same concept (“individual”)
- owl:sameAs
  - ✓ is a main mechanism of “linking”
  - ✓

```
<http://louvre.fr/Michel-Ange>
    owl:sameAs  <http://mibac.it/Michelangelo> ;
```



Up to 7th level

- Providing 5-star Linked Data is just the beginning.
- To actually make use of the datasets, consumers need:
  - ✓ more support in getting to know and access them
  - ✓ a better grasp of their quality and provenance.
- Extend the model with two additional stars

Levels 6 and 7

★★★★★

Schema and documentation

Provide your data with a schema and documentation so that people can understand and re-use your data easily

★★★★★★

Validation and provenance

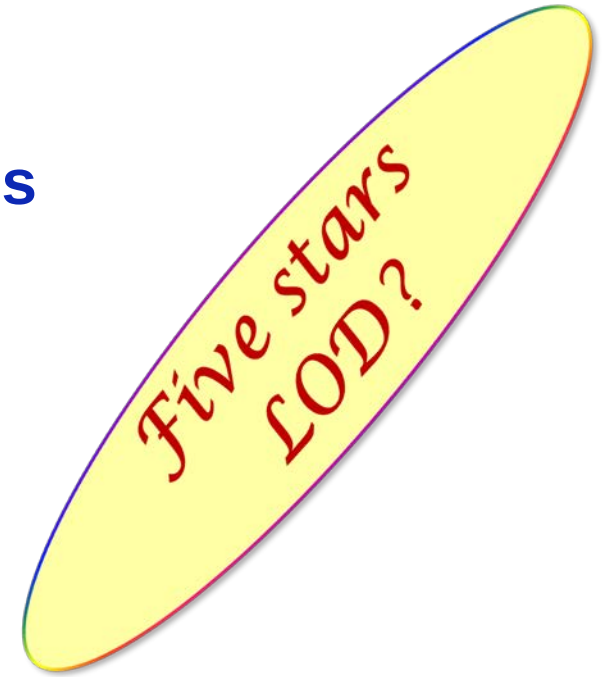
Validate your data and denote its provenance so that people can trust the quality of your data

References:

- ✓ http://www.ldf.fi/
- ✓ http://www.seco.tkk.fi/publications/2014/hyvonen-et-al-ldf-2014.pdf

Work done?

- The ontology (intension):
  - ✓ Models concepts and relationships
  - ✓ Supports multilinguality
  - ✓ Can be referenced by everybody
- Data (extension):
  - ✓ Available as RDF
  - ✓ Can be queried via SPARQL
  - ✓ Can be linked by everyone from everywhere
- No more a single information silo!



Nobody's perfect!

- Is the ontology a **shared** ontology?
- Does it make **reference** to well established ontologies?

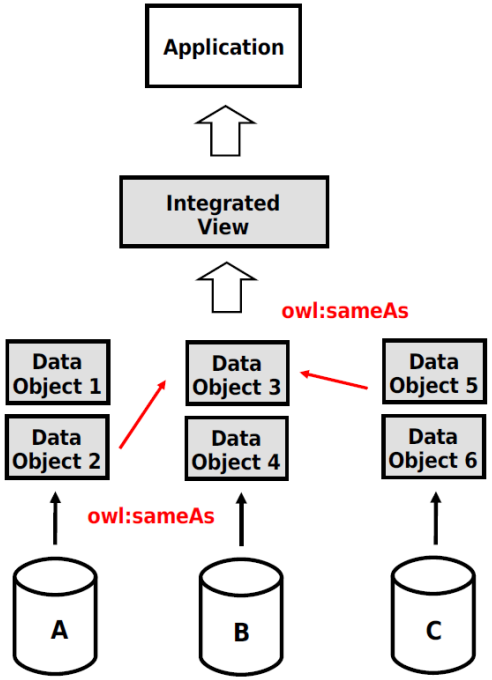
Building ontologies: a methodology (or a rule of thumb?)

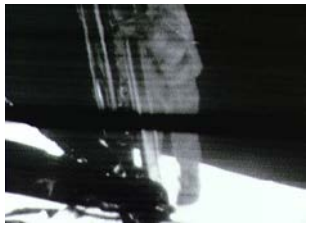
- Analyze and model your "world of interest"
- Check existing ontologies:
  - ✓ does one fits perfectly?
  - ✓ extend one with your own concepts?
  - ✓ combine several existing ontologies?
  - ✓ full import or just refer some class/properties?
- Based on my own experience:
  - ✓ creating your own ontology is easier, but less effective
  - ✓ using/combining/extending existing ontologies is harder, but more effective
  - ✓ keep intensional and extensional components separated

Content of this slide does not necessary reflect the W3C position

Ready to start?

- User requirements
  - ✓ Integrated view of information
- Data fusion: some well known problems
  - ✓ Schema mapping
  - ✓ Conflict resolution: inconsistencies
  - ✓ Trust / Information quality
- Reuse issues
  - ✓ Licences
- Implementation issues
  - ✓ How to publish
  - ✓ Platforms
- Aim: five (or seven?)star dataset, rich and shared ontology. However:
  - ✓ The best is the enemy of the good.
  - ✓ The important is to start, even with raw data
  - ✓ *"One small step for man. One giant leap for mankind."*





References

- Linked Data (Tim Berners-Lee)
- Tim Berners-Lee on the next Web (presentazione a TED2009, con sottotitoli in varie lingue)
- http://esw.w3.org/LinkedData (Wiki W3C)
- http://linkeddata.org/
- Linked Data - The Story So Far (Bizer, Heath,Berners-Lee) - preprint
- Tom Heath, Christian Bizer: Linked Data: Evolving the Web into a Global Data Space

Conclusion

- LOD have been part of the Web since its inception
- The main benefit is to share and improve knowledge
- RDF is the basis
- SW technologies are crucial
- Share ontologies (intension)!
- Keep data decentralized (extension)!
- START NOW

?
Questions

Thank you for your attention!

Slides at: http://www.orestesignore.eu/education/lda/slides/lod.pdf