Managing the Evolution and Preservation of the Data Web

Γιώργος Παπαστεφανάτος

The Context: Data Web

Data Web becomes a reality by connecting existing isolated data islands

- **Enterprise Intranets**
- **Public Sector Information**
- **Scientific Research**

Linked Open Dat

- Linked Open Data (LOD) is a way of publishing data on the (Semantic) Web that:

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

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

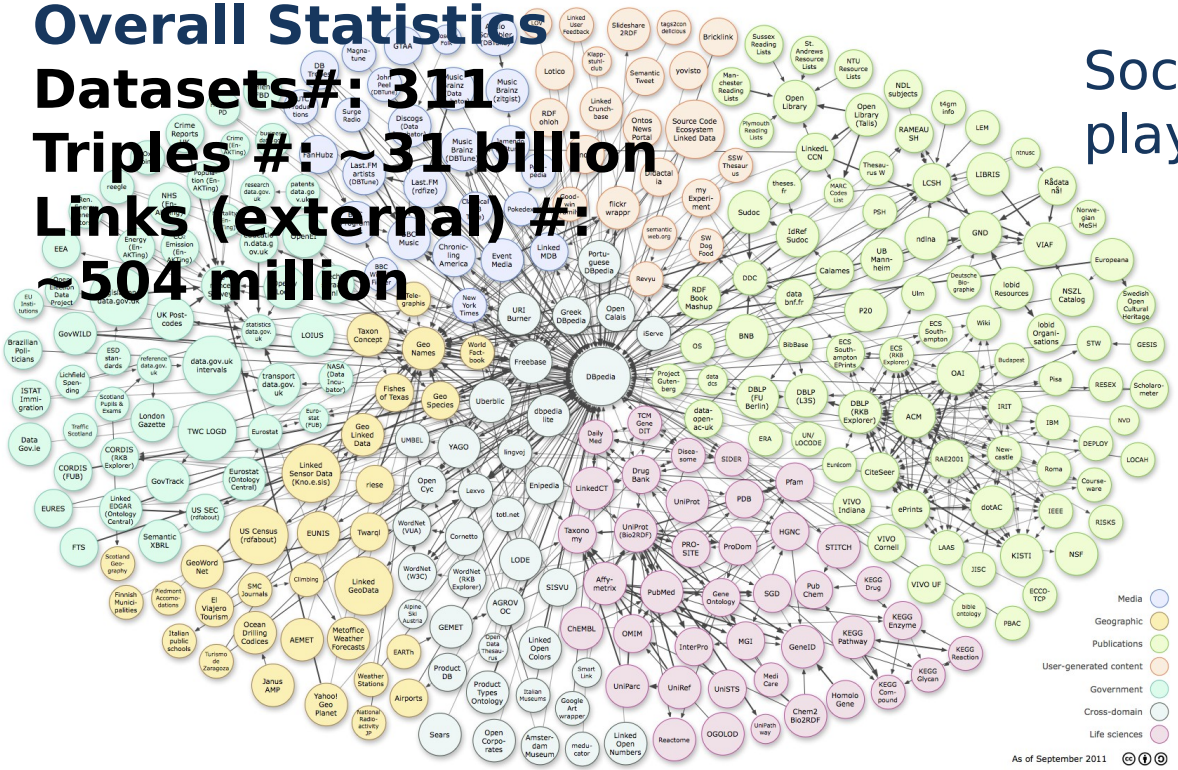
★★★ as (2) plus non-proprietary format (e.g. CSV instead of excel)

★★★★ as (3), plus using open standards from W3C (triple-based data, RDF and SPARQL) to identify things through de-referenceable HTTP URIs, to ensure effective access

★★★★★ as all the above plus establishing links between data of different sources.

THE LOD ECOSYSTEM

Overall Statistics
Datasets#: 311
Triples #: ~31 billion
Links (external) #: ~504 million

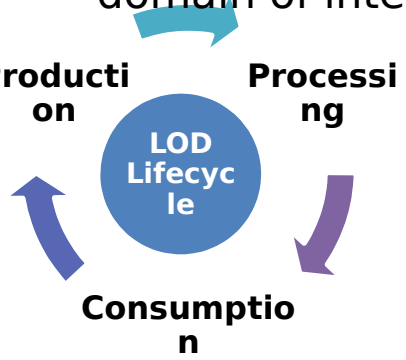


As of September 2011

Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

Social system with several players

- ✓Data producers
- ✓Data consumers
- ✓Data matchmakers
  - ✓ discover publicly available data silos
  - ✓ establishing mappings / correspondences commonly used in a domain of interest



G.Papastefanatos @OpenData Day 2013, Athens

The Challenge: Data Web Evolution

Change at **different granularity levels**

Evolution without notification

Differences are not detected and propagated among data sources

Much data is **impossible to reproduce**

Much data **can only be recovered at enormous costs**

Much data was produced with **heavy human involvement** (aka curated)

Digital preservation is often understood as **“pickling” and “locking away”** individual data sets for future use

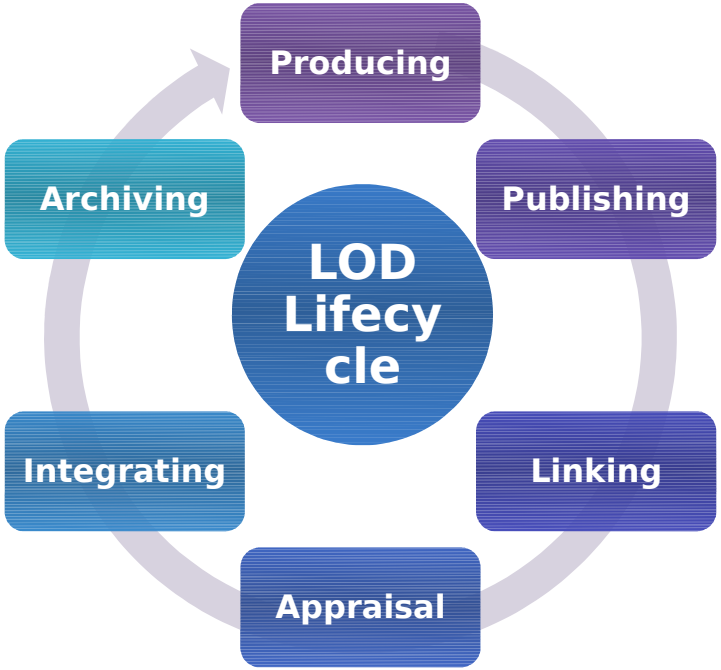
LOD Dynamics: Challenges and Problems

- *crawling and appraisal*
  - how we can *assess* the quality (temporal and spatial) and *change frequency* of LOD datasets in order to be able to decide which and how many *versions* of them deserve to be further preserve?
- *evolution tracking and change management*
  - how can we *monitor*, *model* and *synchronize* changes between LOD datasets?
- *provenance*
  - how can we *monitor the provenance* of LOD datasets when replicated in multiple sources?
- *archiving and citation*
  - How do we *cite* particular versions of a LOD dataset. How will we be able to retrieve a *past* version and not the most recently available?
- *curation*
  - how various data imperfections (e.g., granularity inconsistencies) can be *repaired*?
- *sustainability* problem
  - how can we *spread preservation costs* and ensure long-term access?

Why LOD Dynamics need special treatment?

- *Linked Data are Structured* unlike documents
- *Linked Data are Dynamic* unlike closed settings in which change monitoring is build in a central authority
- *Linked Data are Distributed* and we need preservation techniques based on data replication enhanced with diachronic (temporal and provenance) annotations
- *Linked Data are Uncertain* because publishing and linking can be approximate and uncertain at best (e.g., extracting structured information from text, from social data, employing entity resolution algorithms)

Our goal: Diachronic linked data



- Preservation of LOD lifecycle
- Embed temporal properties and time-aware links at the time of data creation, allowing a one-step production and preservation
- Self-Preserving: Records its own history, its evolution and its usage context
- Mechanisms for
  - Adaptive Focused Crawling
  - Change Detection
  - Multiversion Archiving
  - Longitudinal query capabilities
  - Provenance Support

The DIACHRON Solution

The DIACHRON Solution

Discovers new relevant data from various domains

Collects superimposed information concerning provenance, interpretation, and use of data

Identifies and manages changes within the LOD cloud

Stores and accesses the data

FP7 – IP
Starting : Apr 2013
Duration : 3 years