Keith Alexander (Talis), Richard Cyganiak (DERI),
Michael Hausenblas (DERI) and Jun Zhao (University of Oxford)

Describing Linked Datasets

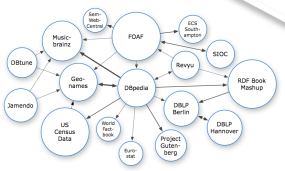
On the Design and Usage of **voiD**,
the 'Vocabulary Of Interlinked Datasets'
Linked Data Workshop at WWW09, 2009-04-20, Madrid, Spain

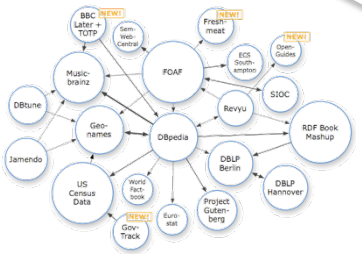
Agenda

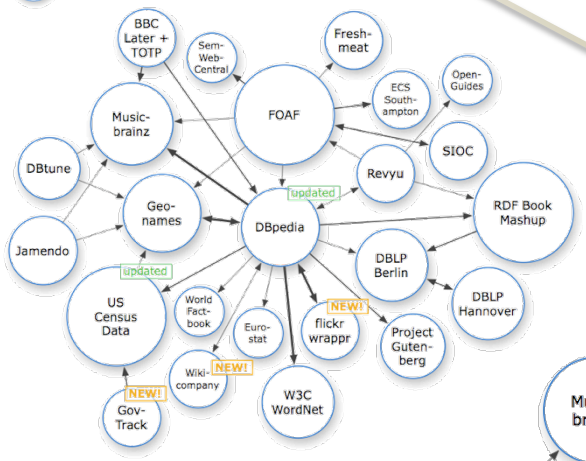
- The Problem
- Our Proposal – voiD
- Applications
- Next Steps

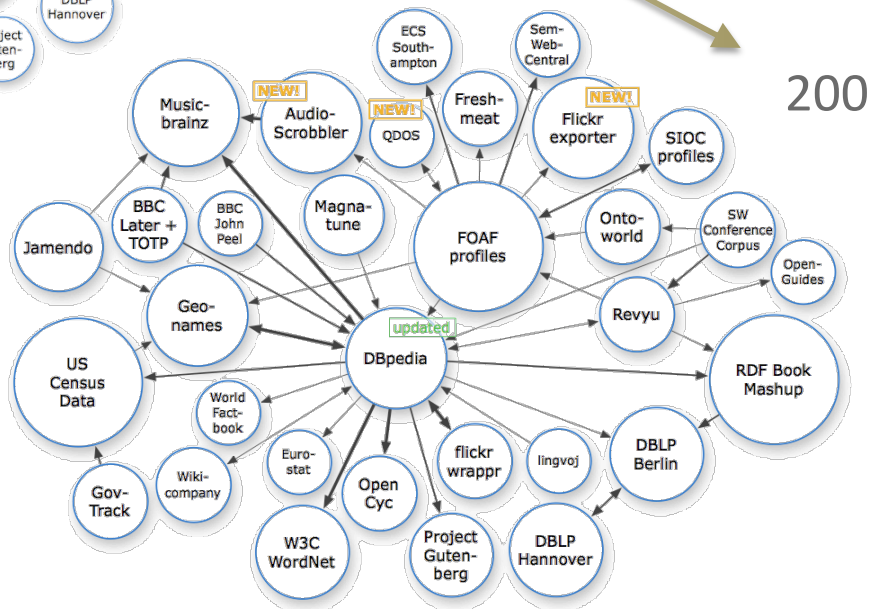
The Problem

2007



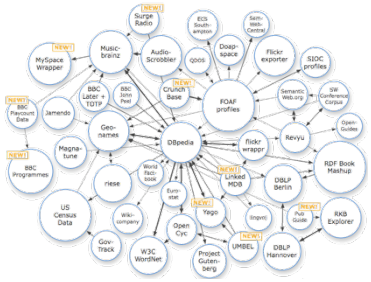






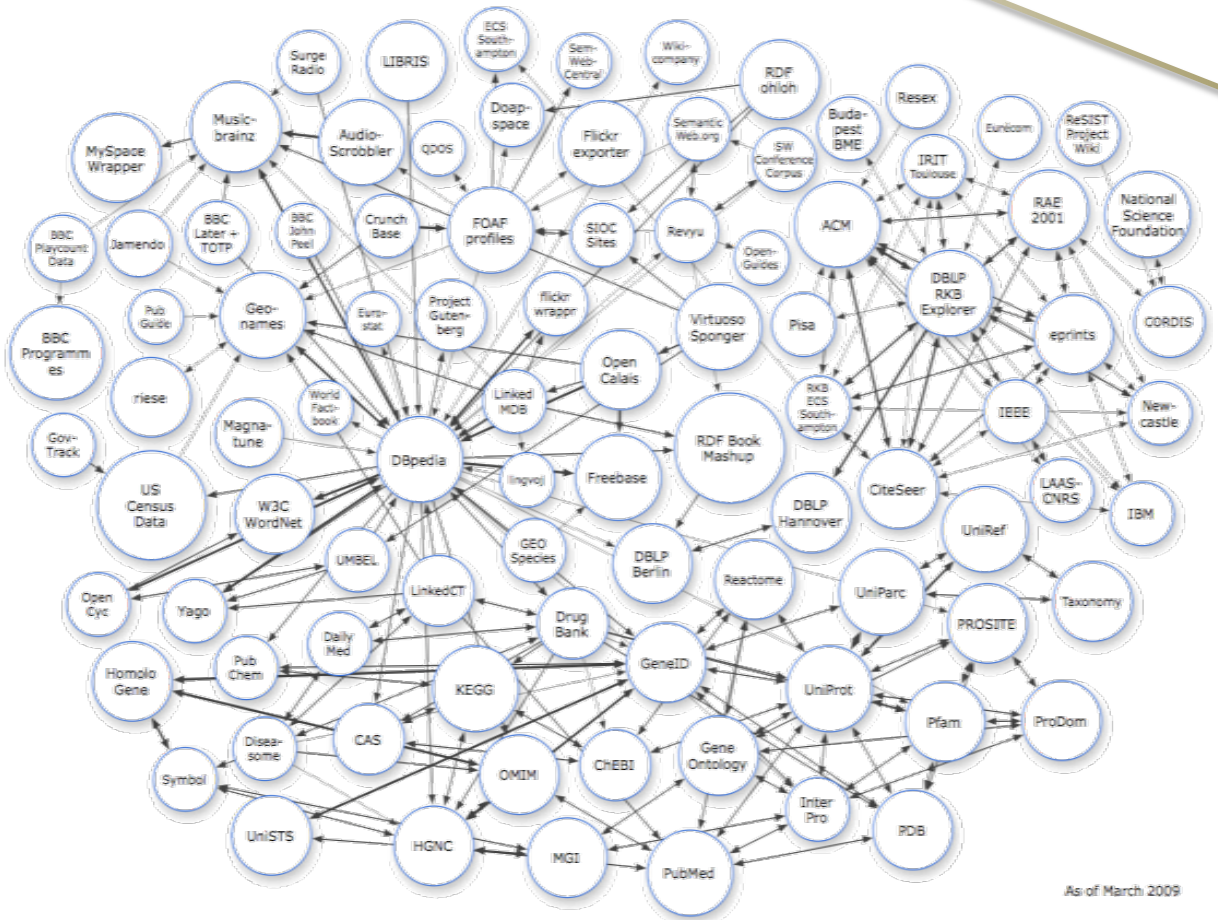
2008

The Problem



2008

2009



The Problem

- The Linking Open Data (LOD) cloud gathers currently roughly the same momentum as the **Web** in the early **1990s**
- How did people deal with the consequences of having a decentralized system, back then?

The Problem


YAHOO!
NEW - COOL - RANDOM
HEAD LINES
YAHOO INFO
ADD URL
Yahoo! Deutschland
CLICK HERE TO VISIT THE STARS
YAHOO! LOS ANGELES
Weekly Picks
Search
Options
Yellow Pages - People Search - City Maps -- News Headlines - Stock Quotes - Sports Scores
Arts - - Humanities, Photography, Architecture, ...
Business and Economy [Xtra!] - - Directory, Investments, Classifieds, ...
Computers and Internet [Xtra!] - - Internet, WWW, Software, Multimedia, ...
Education - - Universities, K-12, Courses, ...
Entertainment [Xtra!] - - TV, Movies, Music, Magazines, ...
Government - - Politics [Xtra!], Agencies, Law, Military, ...
Health [Xtra!] - - Medicine, Drugs, Diseases, Fitness, ...
News [Xtra!] - - World [Xtra!], Daily, Current Events, ...
Recreation and Sports [Xtra!] - - Sports, Games, Travel, Autos, Outdoors, ...
Reference - - Libraries, Dictionaries, Phone Numbers, ...
Regional - - Countries, Regions, U.S. States, ...
Science - - CS, Biology, Astronomy, Engineering, ...
Social Science - - Anthropology, Sociology, Economics, ...

The Problem

- From 2007 on, we have been doing it in the *Yahoo!-catalog-style*: **manually collecting** and **representing** data about the Linking Open Data cloud:
  - In the LOD cloud diagram, we give a qualitative view in form of a visual graph
  - In various ESW Wiki pages we create HTML tables:
    - http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics
    - http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets/LinkStatistics

The Problem

http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets/LinkStatistics

Statistics on links between Data sets

This page collects statistics on links between Data sets that are available as Linked Data.

The link statistics are also used to update the LOD cloud diagram for the main page of the project. So if you publish a dataset yourself or if you know detailed statistics for datasets that you use, please add them to the table and we will include them in the next revision of the LOD cloud.

Links between Data sets

- Outdated CSV version of this table
- Note that **Link count (range)** originally had 3 valid values – "> 100", "> 1,000", "> 100,000" – based on **Link count (actual)**.
  - Now has 5 levels – "> 100", "> 1,000", "> 10,000", "> 100,000", "> 1,000,000" – being updated (and corrected) as **Link count (actual)** is updated. (Linkages fewer than 100 are not counted here nor shown in the Cloud diagrams.)
- Also note – some links (e.g., that for RAE 2001) have been percent-escaped to get around spam-blocks in the ESW Wiki configuration. De-escaping these should allow loading, if they fail when simply clicked in this table.
- re "Type" column –
  moment, so counts
  Equivalent/Broader

Source data
ACM (RKB)
ACM (RKB)
ACM (RKB)
ACM (RKB)
ACM (RKB)
ACM (RKB)
ACM (RKB)
ACM (RKB)

Statistics on Data sets

This page collects statistics on Data sets that are available as Linked Data.

The statistics are also used to update the LOD cloud diagram for the main page of the project. So if you publish a dataset yourself or if you know detailed statistics for datasets that you use, please add them to the table and we will include them in the next revision of the LOD cloud.

Data set sizes

- "Wrapper" denotes any "data set" which is not available as an RDF Dump, for which size cannot be accurately assessed, or for which triples are dynamically produced.
- Note – some links (e.g., those for RAE 2001 and "R e s e x") have been percent-escaped to get around spam-blocks in the ESW Wiki configuration. De-escaping these should allow loading, if they fail when simply clicked in this table.

Data set	Size of the data set (number of triples)	Wrapper?	endpoint?	RDF dump?
ACM (RKB)	13,644,052	N	Y	Y
AudioScrobbler	600,000,000	Y	Y	N
BBC John Peel	277,000	N	Y	Y
BBC Later + TOTP (link not responding - 2009-04-01)	10,000	N	Y	Y
BBC Playcount Data	10,000	N	Y	Y
BBC Programmes	10,000,000	N	Y	Y
Budapest BME (RKB)	42,064	N	Y	Y
CAS	100,000	N	Y	Y
ChEBI	510,866	N	Y	Y
CiteSeer (RKB)	8,294,523	N	Y	Y
CrunchBase	955,676	Y	Y	N
Daily Med	170,000	N	Y	Y

http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics

The Problem

- Currently, only **human comprehensible descriptions** (the LOD cloud, Wiki pages) available
- We **can't automate tasks**, such as
  - Efficient & effective search
  - Selection of dataset (for apps, interlinking targets)
  - Generation of maps, etc.

The Problem

- We **can't apply our tools** and methods we have experiences with, such as editors, engines, stores, etc.
- Even worse, it **doesn't scale**
  - We'd need a Google-style approach that scales like hell and is powerful enough to enable the above mentioned
  - Providing **metadata** about the **LOD cloud** in a **machine-comprehensible** way

Agenda

- ✓ The Problem
- Our Proposal – voiD
- Applications
- Next Steps

Our Proposal - voiD

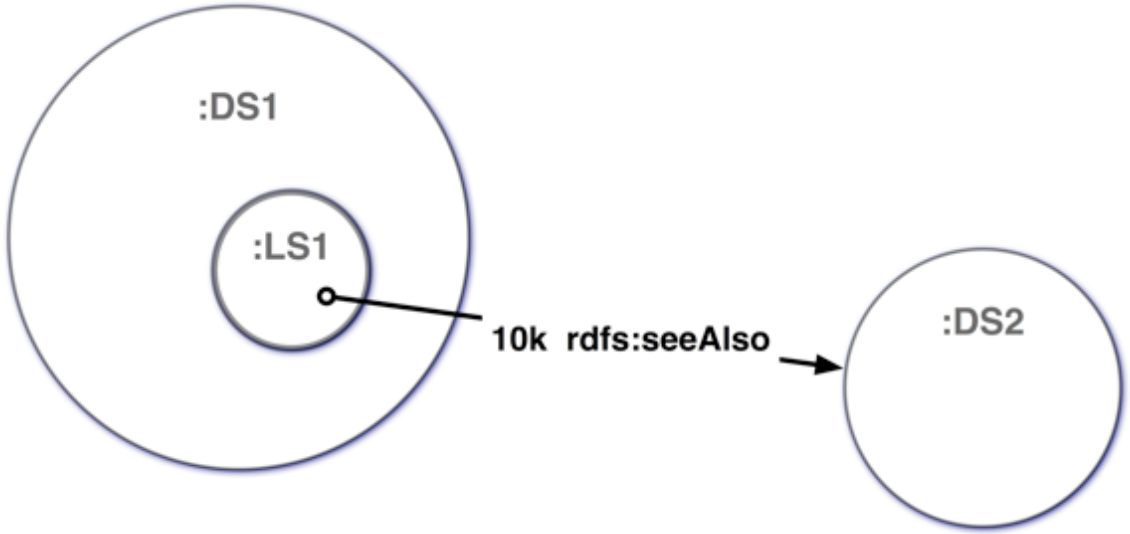
- Solution: providing a formal description of
  - What a dataset is about (topic, technical details)
  - How and under which conditions to access it
  - How the dataset is interlinked with other datasets
    - Qualitative level: type of interlinking
    - Quantitative level: number of links, resources, etc.
  - How to discover the metadata
- **voiD**, the "Vocabulary of Interlinked Datasets" provides precisely this

Our Proposal - voiD

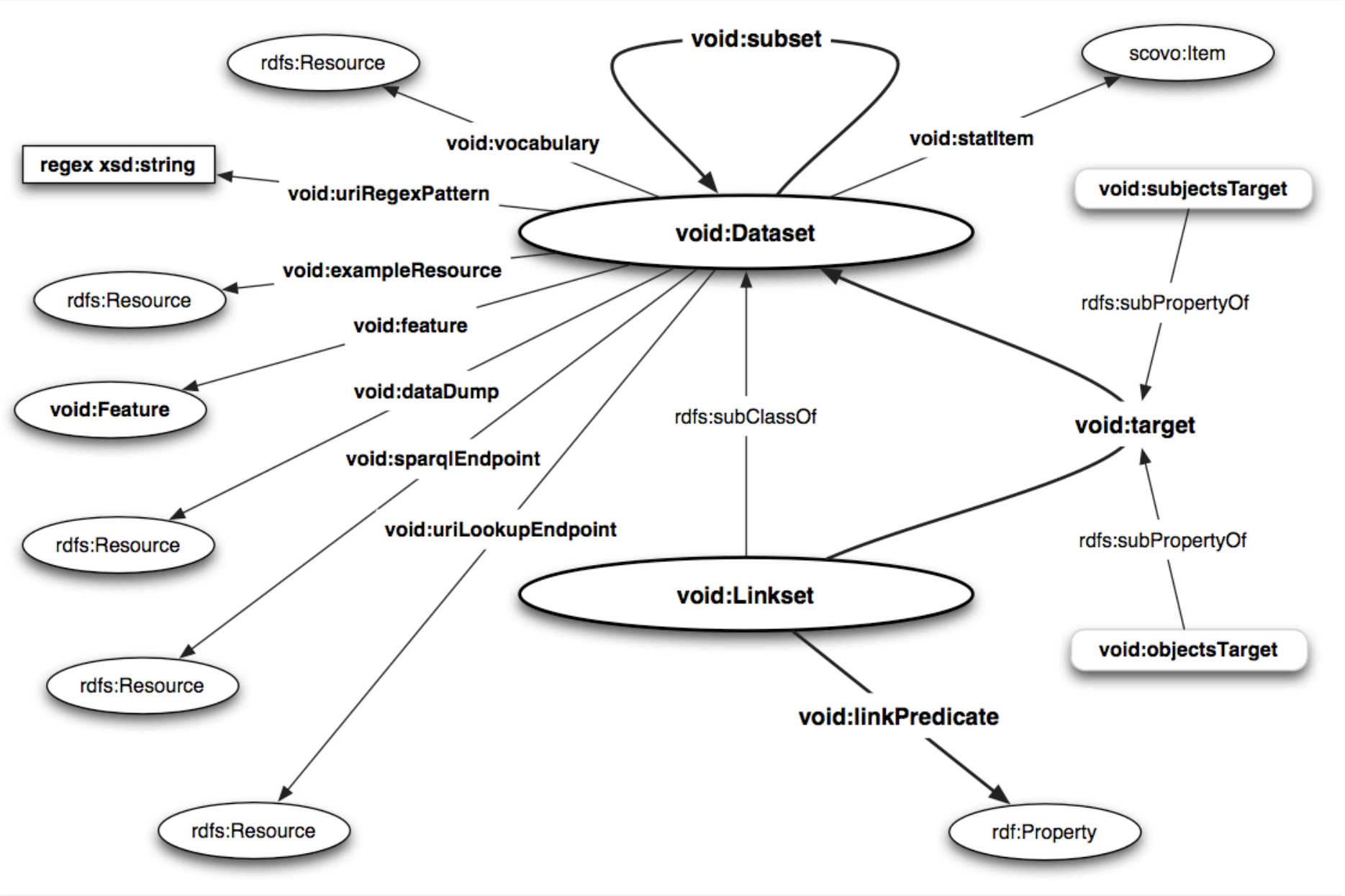
- A **dataset** is a set of RDF triples that are published, maintained or aggregated by a single provider.
- A **dataset** is **authoritative** with respect to a certain URI namespace if it contains information about resources named by URIs in this namespace, and is **published** by the **URI owner** (→URI ownership as of the AWWW1)

Our Proposal - voiD

- A **linkset** LS is a set of RDF triples where for all triples $t_i = \langle s_i, p_i, o_i \rangle \in LS$, the subject is in one dataset, i.e. all s_i are described in DS_1 , and the object is in another dataset, i.e. all o_i are described in DS_2 .



Our Proposal - voiD

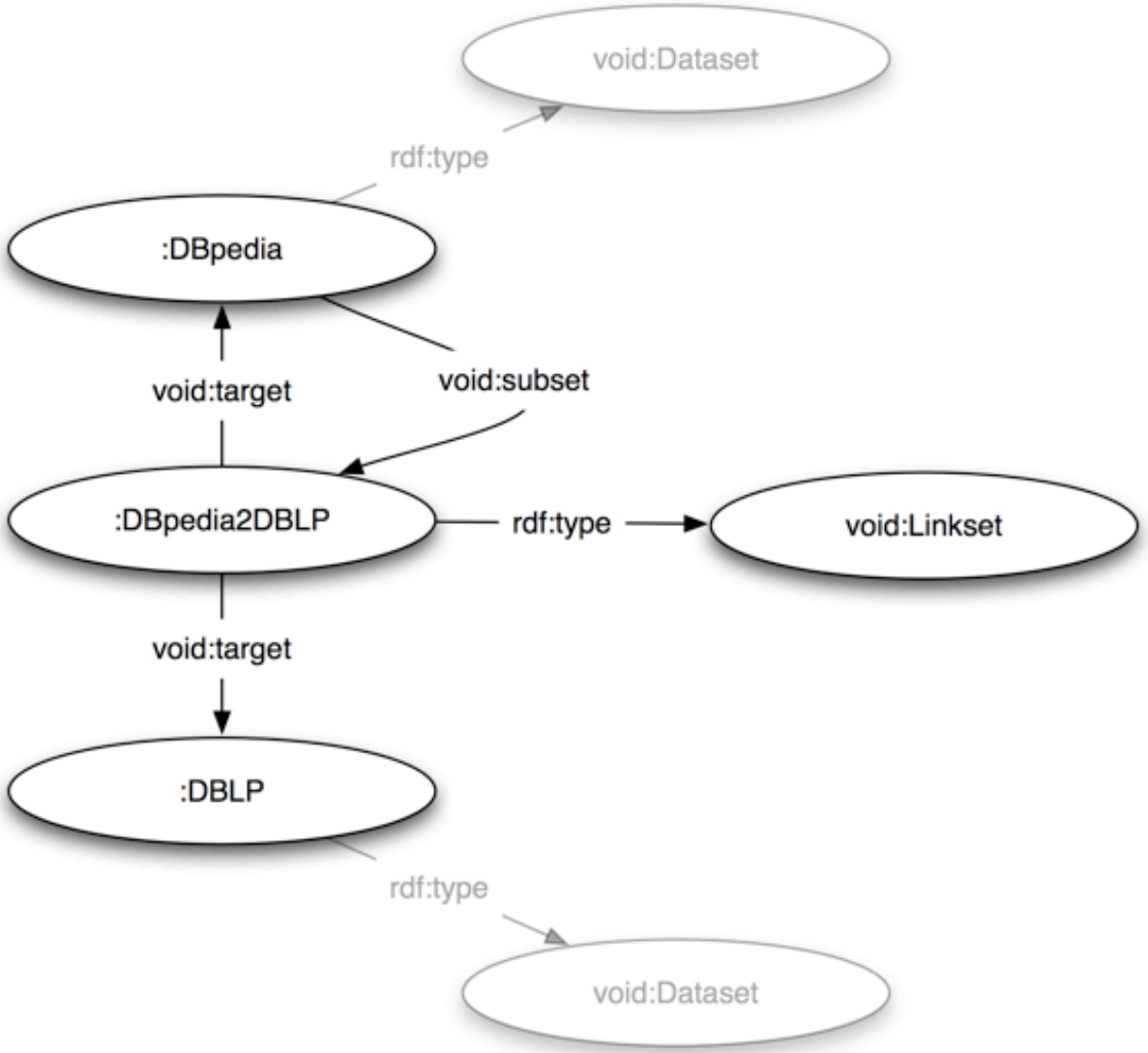


Our Proposal - voiD

voiD offers two orthogonal interlinking types:

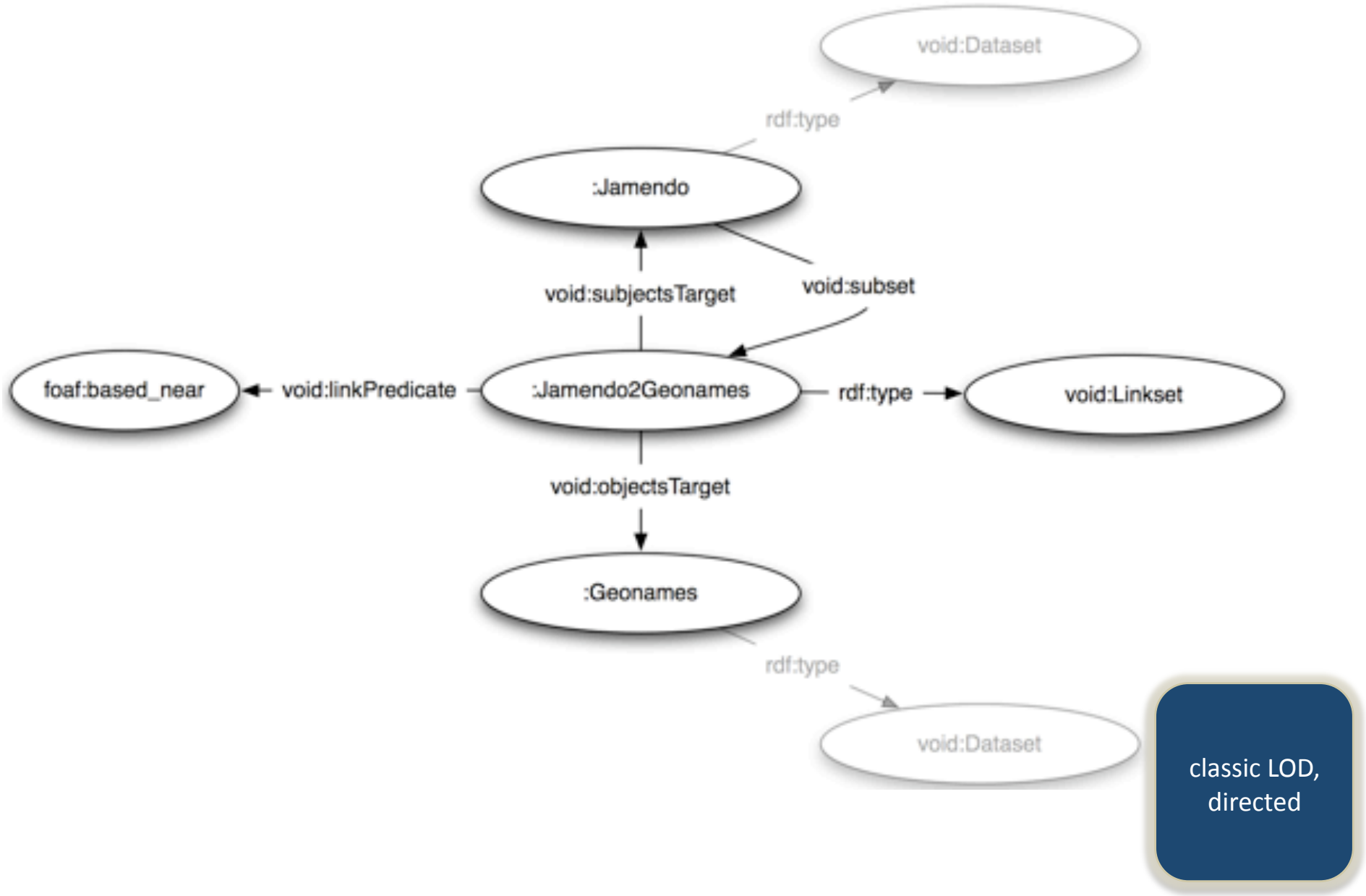
- **classic LOD** vs**. 3rd-party**, differing in where the interlinking statements are kept. In the first case the interlinking triples, i.e. a linkset, are hosted in one of the two involved datasets, while in the latter case there is a third dataset involved that contains the interlinking triples, i.e. the linkset;
- **non-directed** vs. **directed**, which addresses the issue if someone is interested in stating the direction of the interlinking or not (for example with owl:sameAs)

Our Proposal - voiD

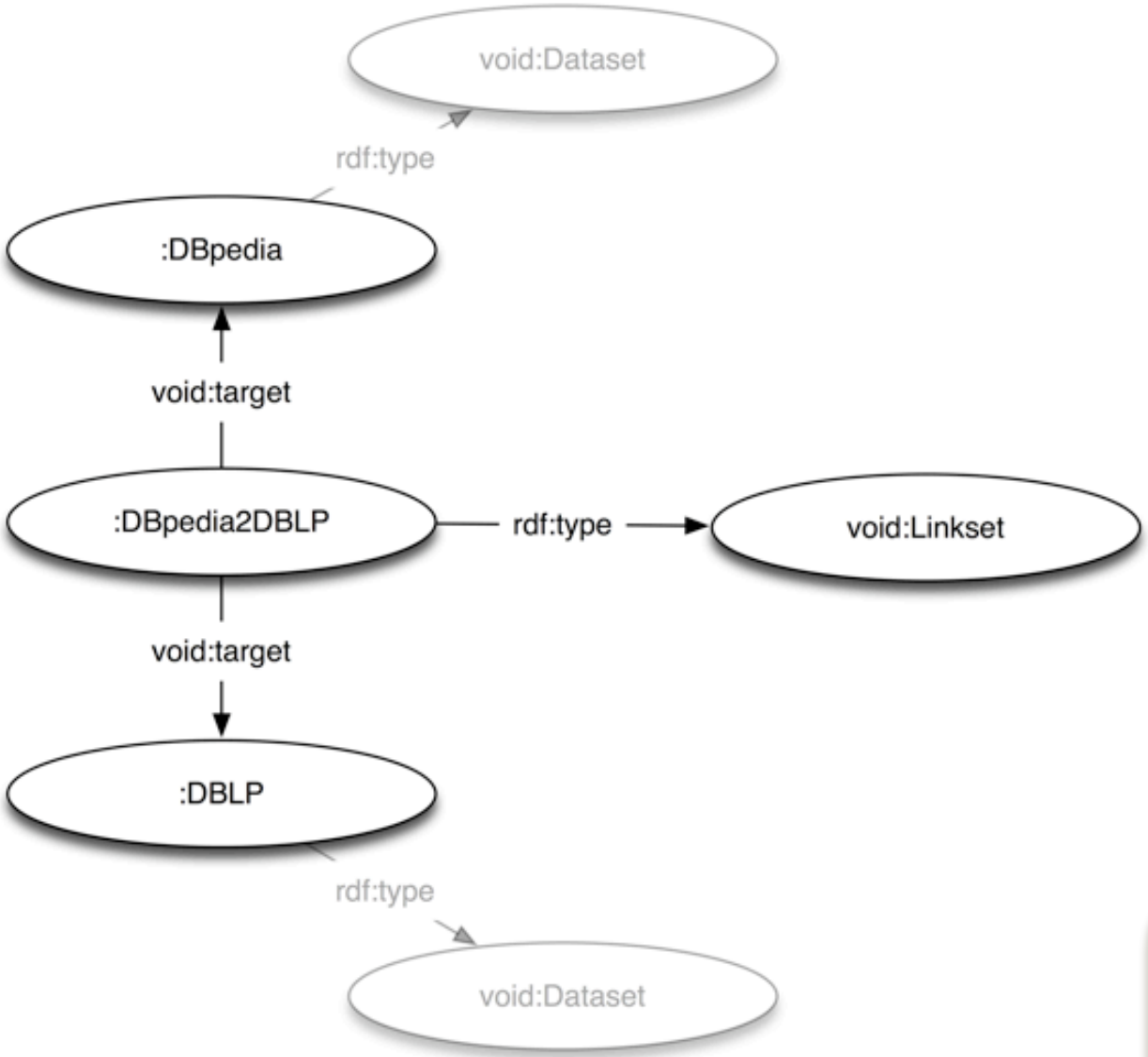


classic LOD, non-directed

Our Proposal - voiD

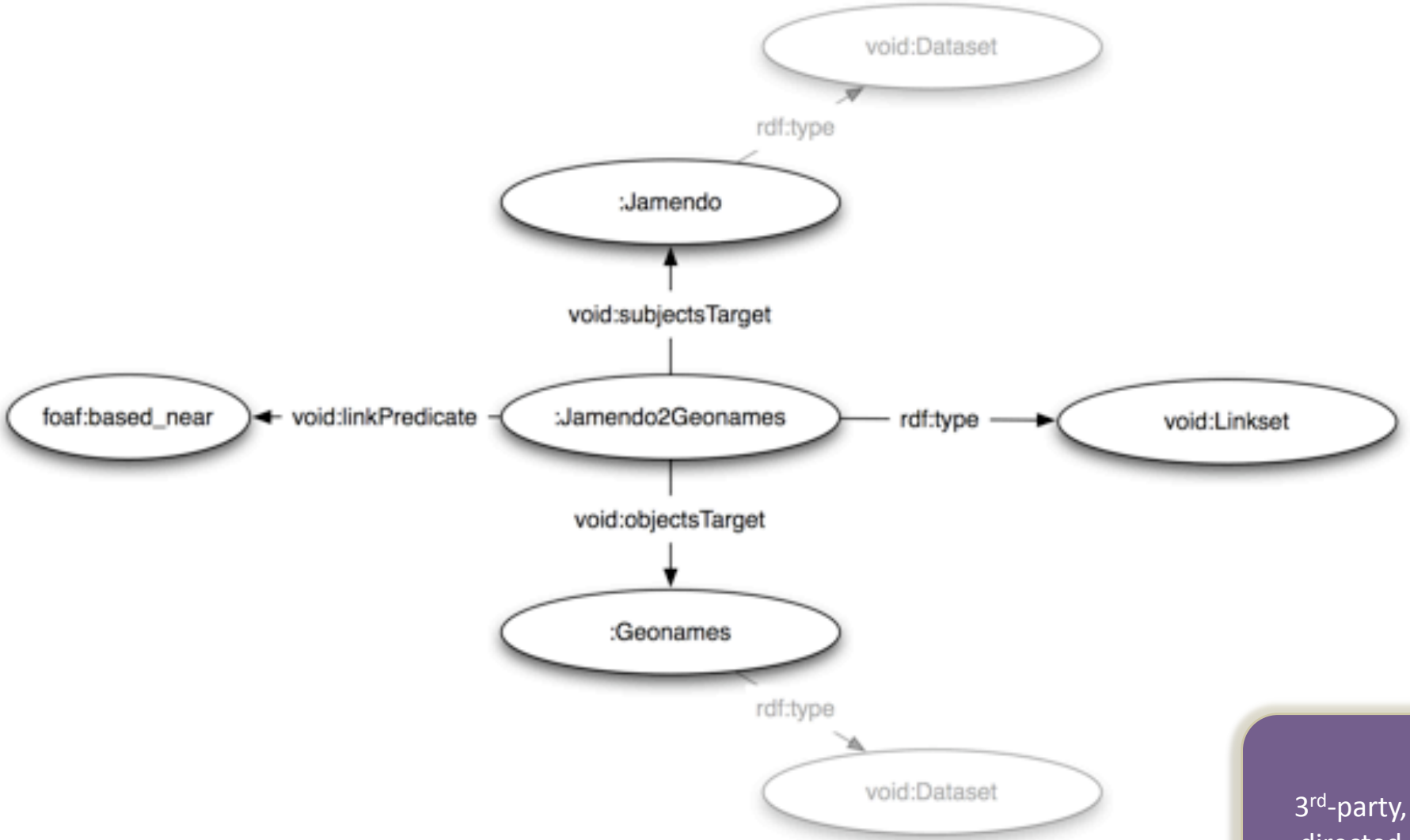


Our Proposal - voiD



3rd-party, non-directed

Our Proposal - voiD



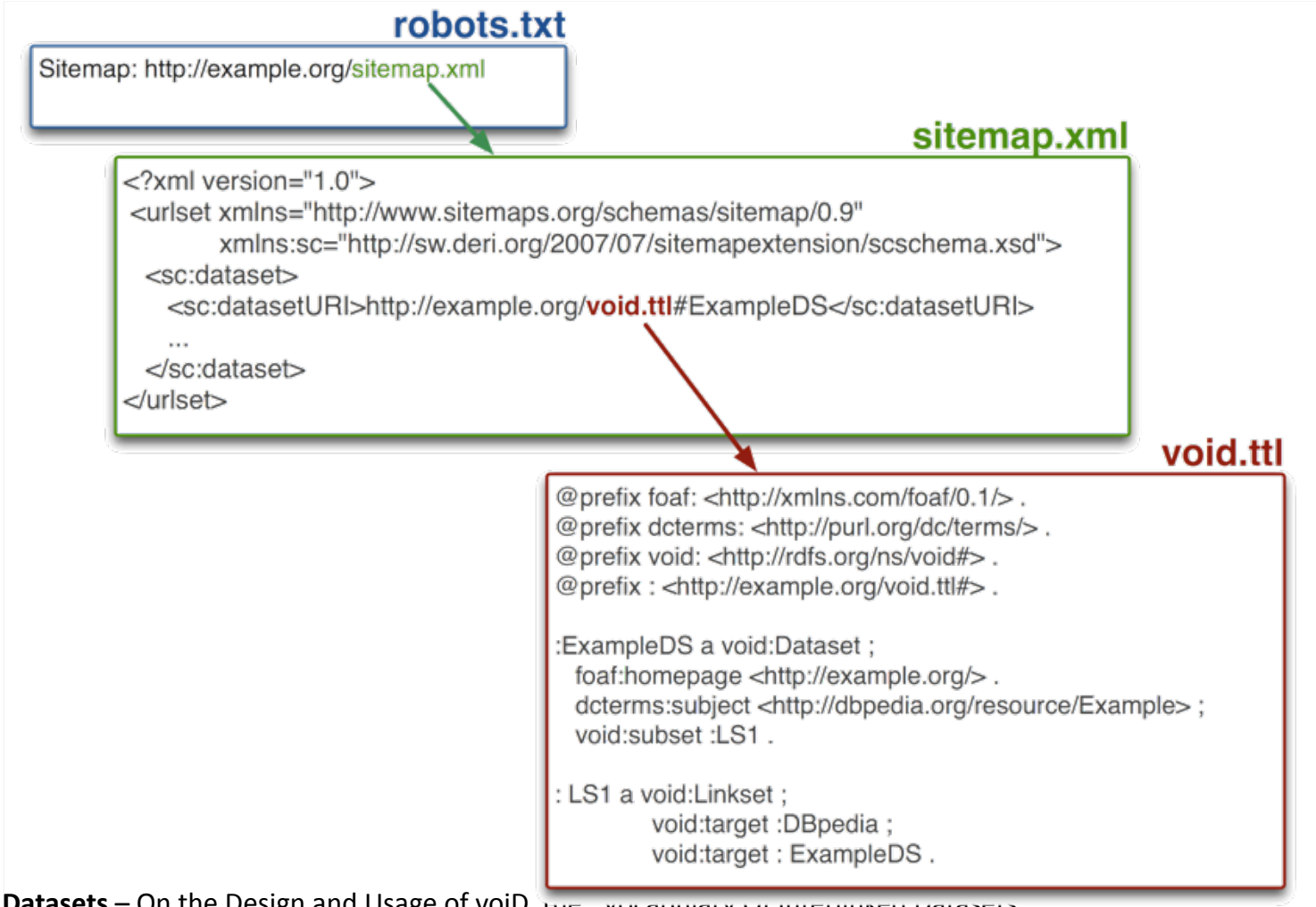
3rd-party, directed

Our Proposal - voiD

- Reusing terms from other vocabularies
  - foaf:homepage/IFP
  - dcterms:subject along with DBpedia URIs http://dbpedia.org/resource/ XXX
  - SCOVO for statistics about triples, links, etc

Our Proposal - voiD

- Publication & discovery via sitemaps and/or backlinks (dcterms:isPartOf)



Our Proposal - voiD

- Once dataset providers have published their voiD description in RDF along with their dataset, one can address the following issues:
  - How to find **some** datasets?
  - How to **efficiently** find a specific dataset?
  - How to **effectively** find datasets?
  - How to **dynamically** select datasets?
  - How to select datasets based on certain **preferences**?

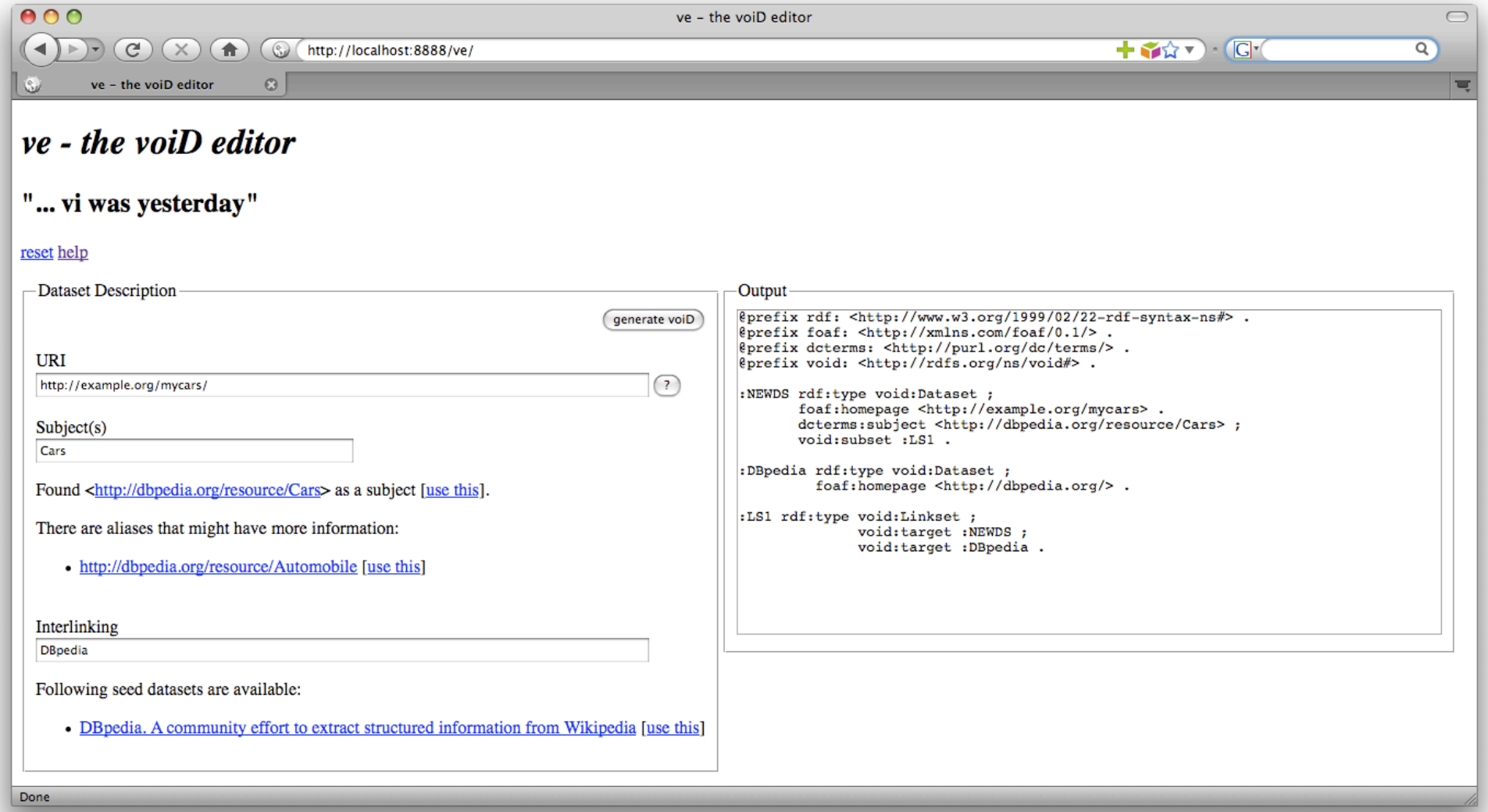
Agenda

- ✓ The Problem
- ✓ Our Proposal – voiD
- Applications
- Next Steps

Applications

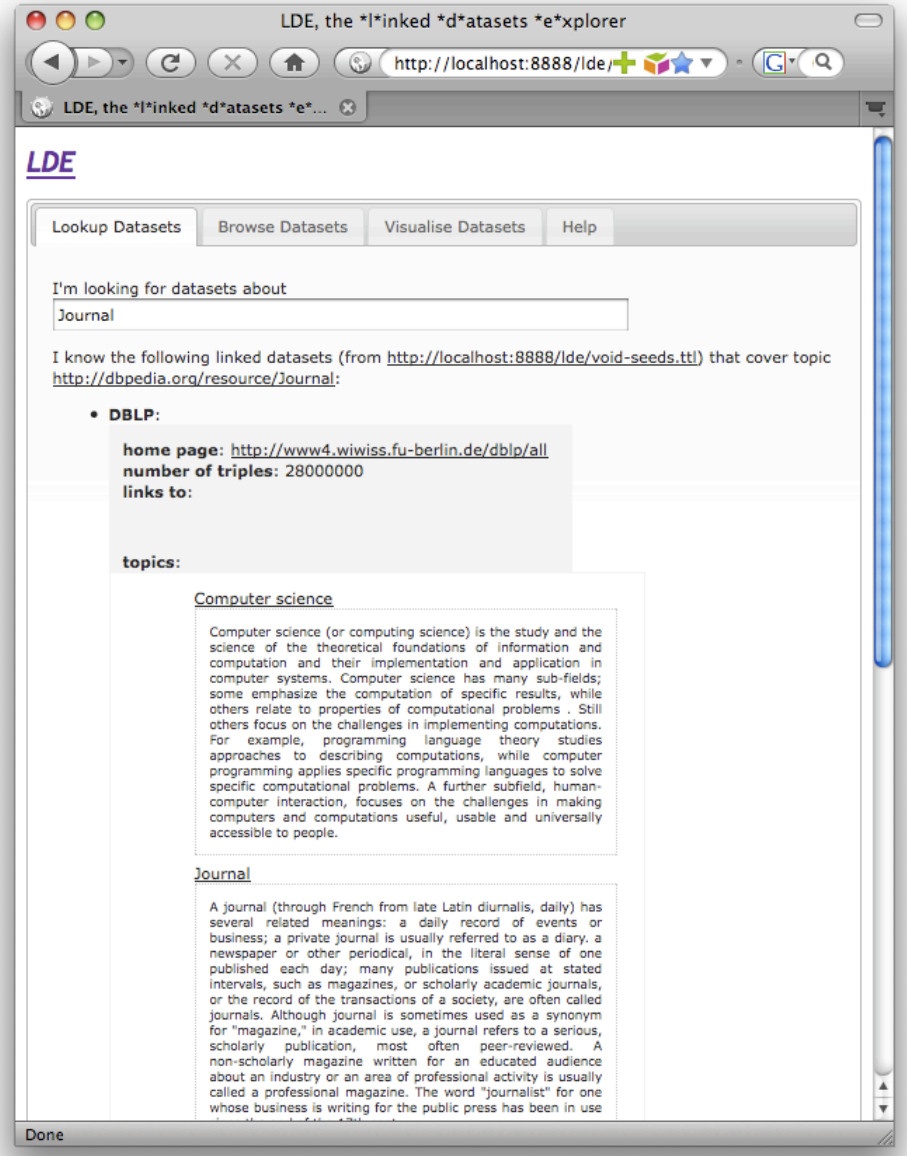
- Generation (ve, liftSSM, NX parser)
- Vocabulary Management (Talis)
- Explorer (RKB, LDE)
- Query Federation (Clarck-Parsia, OpenLink)
- Dataset ranking (→ DING! talk)
- Potential Applications
  - Map of data (Sindice)
  - Dynamic Meshups for Application

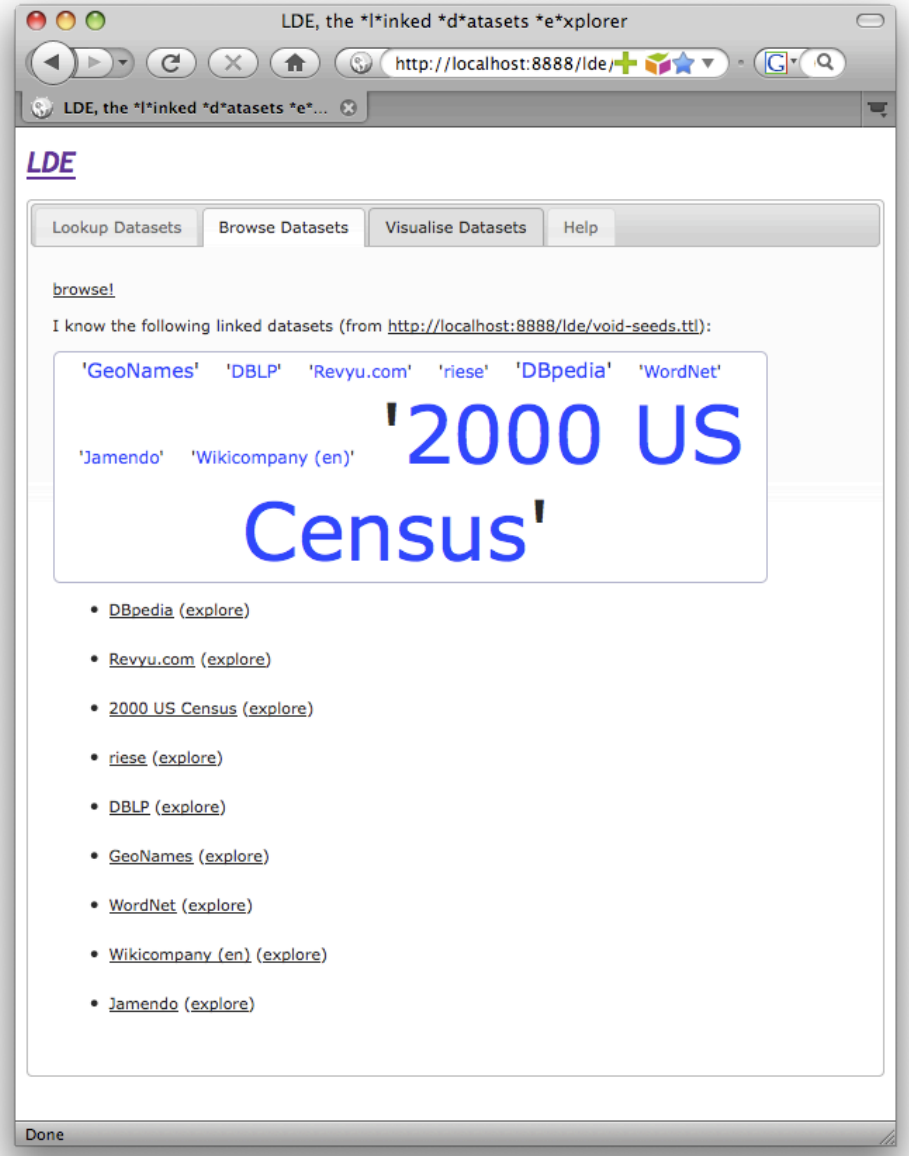
Applications



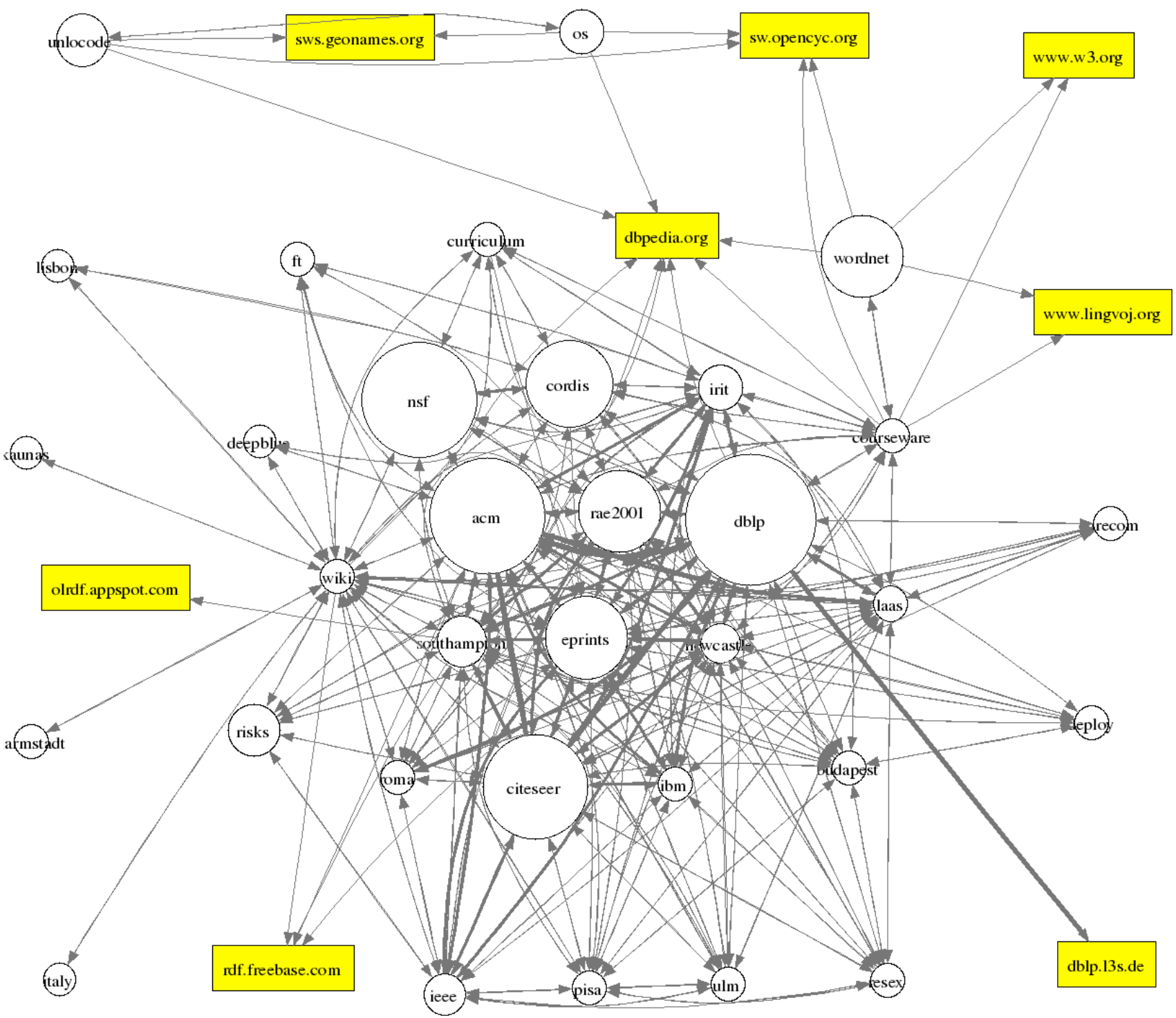
http://ld2sd.deri.org/ve

Applications



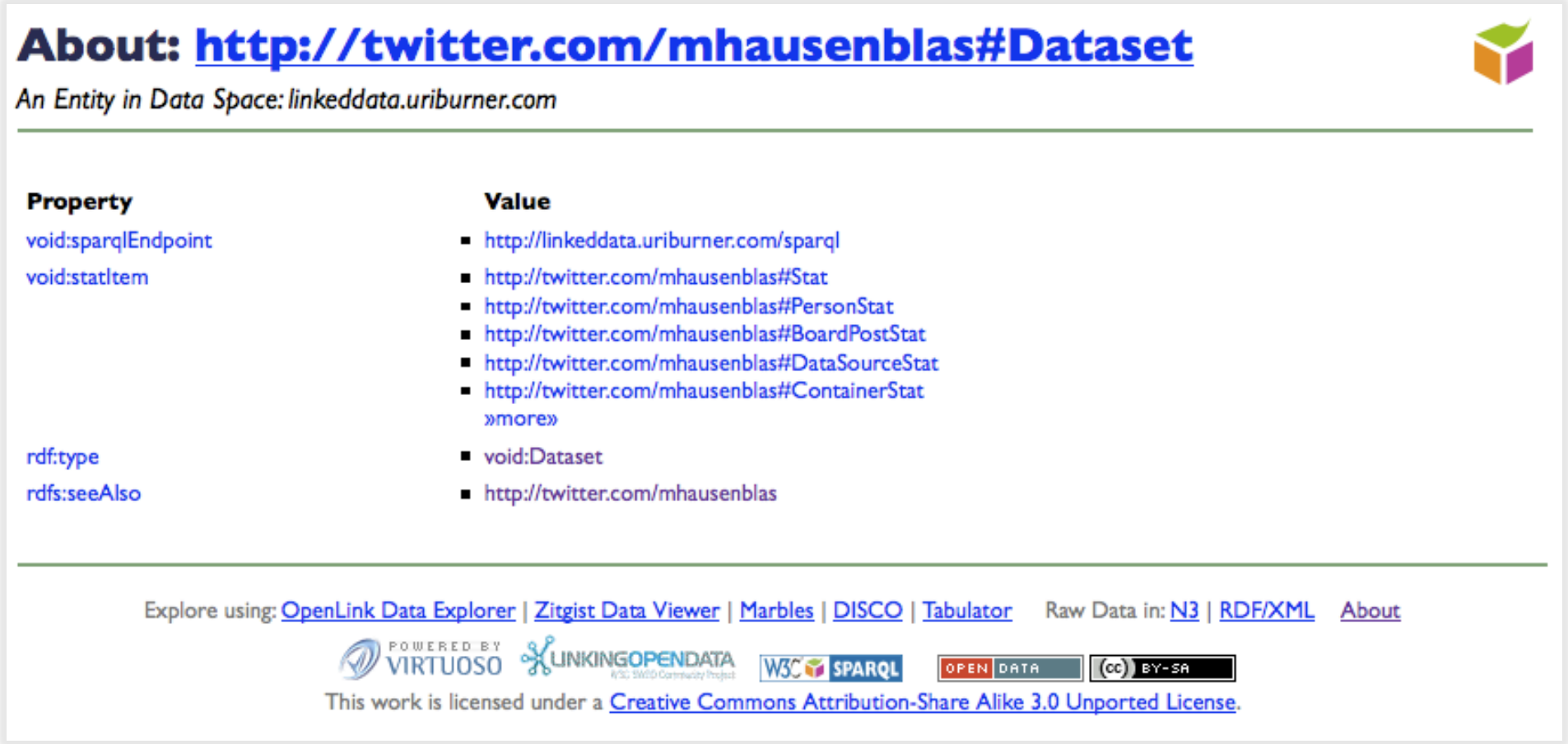


http://ld2sd.deri.org/lde



http://dblp.rkbexplorer.com/models/void.ttl

Applications



About: http://twitter.com/mhausenblas#Dataset

An Entity in Data Space: linkeddata.uriburner.com

Property	Value
void:sparqlEndpoint	▪ http://linkeddata.uriburner.com/sparql
void:statItem	▪ http://twitter.com/mhausenblas#Stat ▪ http://twitter.com/mhausenblas#PersonStat ▪ http://twitter.com/mhausenblas#BoardPostStat ▪ http://twitter.com/mhausenblas#DataSourceStat ▪ http://twitter.com/mhausenblas#ContainerStat »more»
rdf:type	▪ void:Dataset
rdfs:seeAlso	▪ http://twitter.com/mhausenblas

Explore using: OpenLink Data Explorer | Zitgist Data Viewer | Marbles | DISCO | Tabulator Raw Data in: N3 | RDF/XML About

POWERED BY VIRTUOSO LINKINGOPENDATA W3C SPARQL OPEN DATA (cc) BY-SA

This work is licensed under a Creative Commons Attribution-Share Alike 3.0 Unported License.

http://linkeddata.uriburner. com/

Agenda

- ✓ The Problem
- ✓ Our Proposal – voiD
- ✓ Applications
- Next Steps

Next Steps

- voiD 2.0 see issues at http://code.google.com/p/void-impl/issues/list
- statistics module (fix/extend re SCOVO)
- SPARQL endpoints
- provenance, trust (?)
- Assist people in publishing voiD