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
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
The Problem

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
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

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
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/Statistics)
    • [http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets/LinkStatistics](http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets/LinkStatistics)
The Problem

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

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
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
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** $\mathcal{LS}$ is a set of RDF triples where for all triples $t_i = \langle s_i, p_i, o_i \rangle \in \mathcal{LS}$, the subject is in one dataset, i.e. all $s_i$ are described in $\mathcal{DS}_1$, and the object is in another dataset, i.e. all $o_i$ are described in $\mathcal{DS}_2$. 

![Diagram showing two datasets connected by a linkset]
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
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

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
Our Proposal - voiD
Our Proposal - voiD

3rd-party, non-directed
Our Proposal - voiD
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)

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
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?
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
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

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
Applications

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

http://ld2sd.deri.org/lde
Describing Linked Datasets – On the Design and Usage of voiD, the "Vocabulary Of Interlinked Datasets"
Applications

http://linkeddata.uriburner.com/

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
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