National Technical University of Athens
School of Electrical and Computer Engineering
Multimedia, Communications & Web Technologies

ΕΘΝΙΚΟ ΚΕΝΤΡΟ ΤΕΚΜΗΡΙΩΣΗΣ
NATIONAL DOCUMENTATION CENTRE

Exposing Bibliographic Information as Linked Open Data using Standards-based Mappings: Methodology and Results

Nikolaos Konstantinou
Nikos Houssos
Anastasia Manta

3rd International Conference on Integrated Information (IC-ININFO'13)
Prague, Czech Republic, September 5-9, 2013

Introduction

- Linked Open Data (LOD) paradigm constantly gaining worldwide acceptance
- Examples in various domains include:
  - Government data
    - http://www.data.gov.uk
  - Financial data
    - http://www.openspending.org
  - News data
    - http://www.guardian.co.uk/data
  - Cultural heritage
    - http://www.europeana.eu
  - *Bibliographic information*
    - http://data.ekt.gr

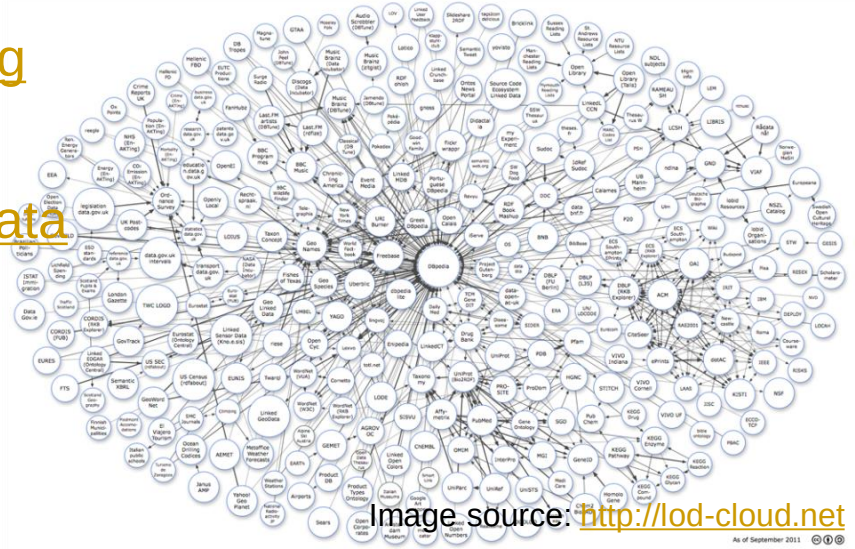


Image source: http://lod-cloud.net

Why Linked Open Data (LOD)?

- Mature technological background
  - W3C Recommendations, i.e. Web standards
    - RDF, OWL, SPARQL, *R2RML*, but also HTTP, XML, etc.
- LOD benefits (indicatively)
  - Integration
    - With data models from other domains
  - Expressiveness
    - In describing information
  - Query answering
    - Graphs: beyond keyword-based searches

The EKT case (1/3)

- National Documentation Centre (EKT)
  - Part of the National Hellenic Research Foundation (NHRF)
  - Mission-critical digital preservation
  - Numerous repositories, maintained by teams of software engineers, librarians and domain experts
  - A living organism is created around these repositories
- *Problem statement:* How to benefit from semantic technologies while:
  - Keeping existing practices unaltered (as possible)
  - Respecting nationwide responsibility
  - Ensuring viability and durability of the result

The EKT case (2/3)

- The national archive of PhD theses (http://phdtheses.ekt.gr)
  - 29,284 theses
  - 21,793 full text records
  - 35,925 downloads from 68 countries
  - 14,742 registered users from 97 countries
  - 173,610 online views

- The Helios repository (http://helios-eie.ekt.gr)
  - 5,735 records by researchers affiliated with the NHRF
  - 1,930 full text records
  - 700 videos

The EKT case (3/3)

- Suggested methodology and approach
  - Maintain LOD repositories side-by-side with existing bibliographic content repositories
  - Respect standards to the maximum degree possible
    - Regarding technologies and vocabularies involved
  - Use open-source tools
    - R2RML Parser
      - Export database contents as RDF
    - Biblio-Transformation-Engine (BTE)
      - Process authority files

The R2RML Parser (1/3)

- An R2RML implementation
- A tool that can export relational database contents as RDF graphs, based on an R2RML mapping document
- See http://www.w3.org/2001/sw/wiki/R2RML_Parser
- R2RML
  - RDB to RDF Mapping Language
  - W3C Recommendation, as of Sept. 2012
  - Reusable mapping definitions
  - Supported by numerous tools
    - db2triples, d2rq, capsenta's ultrawrap, openlink's virtuoso, etc.

The R2RML Parser (2/3)

- Command-line tool
- Fully written in Java
- Open-source (

)
- Publicly available at https://github.com/nkons/r2rml-parser
- Tested against MySQL and PostgreSQL
- Output can be written in RDF/OWL
  - N3, Turtle, N-Triple, TTL, RDF/XML notation
  - Relational database (Jena SDB backend)

The R2RML Parser (3/3)

- Covers most of the R2RML constructs
  - See https://github.com/nkons/r2rml-parser/wiki
- Allows arbitrary SQL queries to be used as logical views (`rr:sqlQuery` construct)
  - Allows SQL functions and function nesting
  - Allows foreign keys
- Limitations
  - No query nesting, union, intersection or difference
  - No multiple graphs from a single execution
    - No support for rr:defaultGraph, rr:graph, rr:graphMap
- Does *not* offer SPARQL-to-SQL translations

The Big Picture

- From DSpace (http://dspace.org) records to RDF

DSpace field	Values
dc.creator	Kollia, Zoe Sarantopoulou, Evangelia Cefalas, Alciviadis Constantinos Kobe, S. Samardzija, Z.
dc.date	2004
dc.format.extent	379-382
dc.identifier.uri	http://hdl.handle.net/10442/7055
dc.language	eng
dc.publisher	Springer
dc.title	Nanometric size control and treatment of historic paper manuscript and prints with laser light at 157 nm
dc.type	Article
dc.subject	Printmaking and Engraving

Resulting RDF snippet in turtle syntax

```
<http://data.ekt.gr/helios/item/10442/7055>
  a  dcterms:BibliographicResource;
  dcterms:creator "Kobe, S." ,
    <http://data.ekt.gr/person/48>,
    <http://data.ekt.gr/person/14>,
    "Samardzija, Z.",
    <http://data.ekt.gr/person/112>;
  dcterms:date "2004";
  dcterms:extent "379-382";
  dcterms:identifier
"http://hdl.handle.net/10442/7055" ;
  dcterms:language
<http://www.lexvo.org/page/iso639-3/eng>;
  dcterms:publisher "Springer";
  dcterms:title
    "Nanometric size control and treatment of
     historic paper manuscript and prints with
     laser light at 157 nm";
  dcterms:type "Article";
  dc.subject

<http://id.loc.gov/authorities/classification/NE1-
NE978>.
```

R2RML Mapping Definition Example

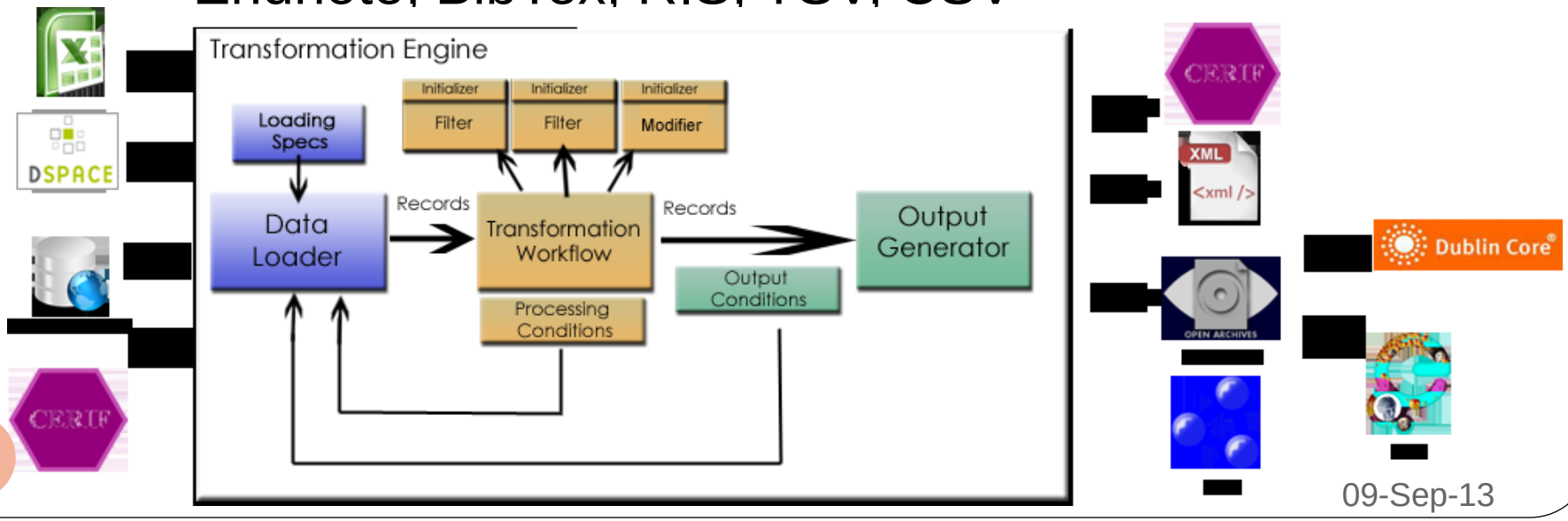
```
@prefix map: <#>.
@prefix rr: <http://www.w3.org/ns/r2rml#>.
@prefix dcterms:
<http://purl.org/dc/terms/>.
map:items
rr:logicalTable <#item-view>;
rr:subjectMap [
rr:template
'http://data.ekt.gr/helios/item/{"handle"}';
rr:class dcterms:BibliographicResource;
].
map:dc-description-abstract
rr:logicalTable <#dc-description-
abstractview>;
rr:subjectMap [ rr:template
'http://data.ekt.gr/helios/item/{"handle"}';
];
rr:predicateObjectMap [
rr:predicate dcterms:abstract;
rr:objectMap [ rr:column '"text_value"' ];
].
```

SQL query

```
<#dc-description-abstract-view>
rr:sqlQuery """
SELECT h.handle AS handle, mv.text_value AS
text_value
FROM handle AS h, item AS i, metadatavalue AS
mv, metadataschemaregistry AS msr,
metadatafieldregistry AS mfr WHERE
i.in_archive=TRUE AND
h.resource_id=i.item_id AND
h.resource_type_id=2 AND
msr.metadata_schema_id=mfr.metadata_schema_id
AND
mfr.metadata_field_id=mv.metadata_field_id AND
mv.text_value is not null AND
i.item_id=mv.item_id AND
msr.namespace =
'http://dublincore.org/documents/dcmi-terms/'
AND
mfr.element='description' AND
mfr.qualifier='abstract' """.
```

Biblio-Transformation-Engine (BTE)

- An open-source java framework
  https://code.google.com/p/biblio-transformation-engine/
- Part of the core DSpace distribution (release 3.0)
- Enables importing Items via basic bibliographic formats
  - Endnote, BibTex, RIS, TSV, CSV



Authority files

- Using BTE, a graph with researcher records is exported

```
<mads>
    <authority lang="en">
        <name><namePart>Sarantopoulou, Evangelia</namePart></name>
    </authority>
    <related lang="gr" type="equivalent">
        <name><namePart>Σαραντοπούλου, Ευαγγελία</namePart></name>
    </related>
    <variant type="other" lang="en">
        <name><namePart>Sarantopoulou, E.</namePart></name>
        <name><namePart>Sarantopoulou, E</namePart></name>
    </variant>
    <variant type="other" lang="gr">
        <name><namePart>Σαραντοπούλου, Ε.</namePart></name>
    </variant>
    <affiliation>
        <organization>ΙΤΕ</organization>
        <email>esarant@eie.gr</email>
        <phone>(+30) 210 7273 840</phone>
        <position>Ερευνήτρια</position>
    </affiliation>
</mads>
```

- Input
  - MADS*-based XML
- Output
  - MADS/RDF
  - Subjects of the form http://data.ekt.gr/persons/{researcher_id}

* Metadata Authority Description Schema: http://www.loc.gov/standards/mads/

The L in LOD

- Open Data is *Linked* when it contains links to other URI's
  - Allows the user to discover more things
- In the EKT case, we linked fields
  - `dc.language` to lexvo.org (language-related concepts)
    - E.g. "`eng`" to http://www.lexvo.org/page/iso639-3/eng
  - `dc.subject` to LCC terms (Library of Congress Classification)
    - E.g. "`Printmaking and Engraving`" to http://id.loc.gov/authorities/classification/NE1-NE978

System Architecture

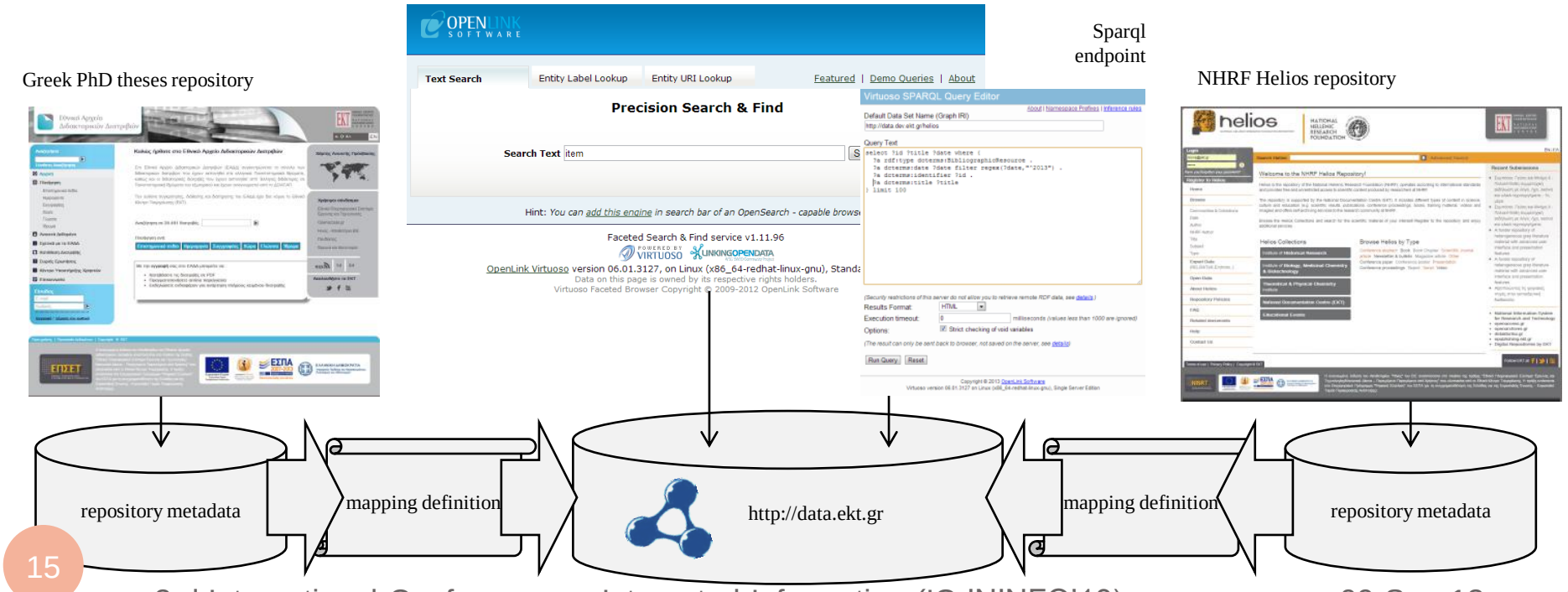
- Virtuoso-backed quadstore
  - Hosts RDF dumps from repository contents
    - Integrated query capabilities
    - Exposes a SPARQL endpoint and a faceted browser

Faceted browsing

Sparql endpoint

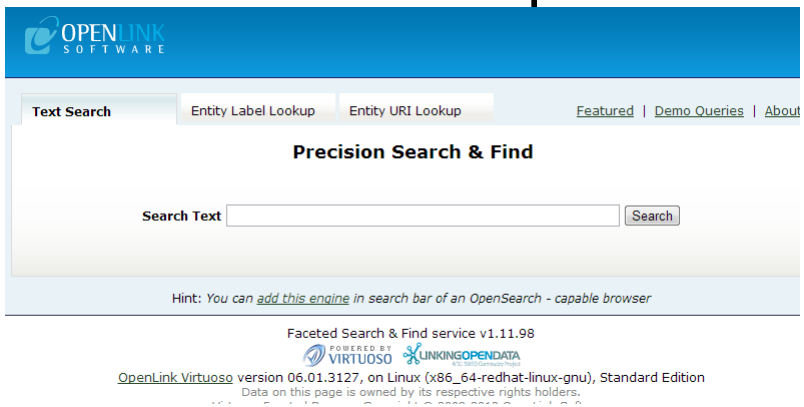
Greek PhD theses repository

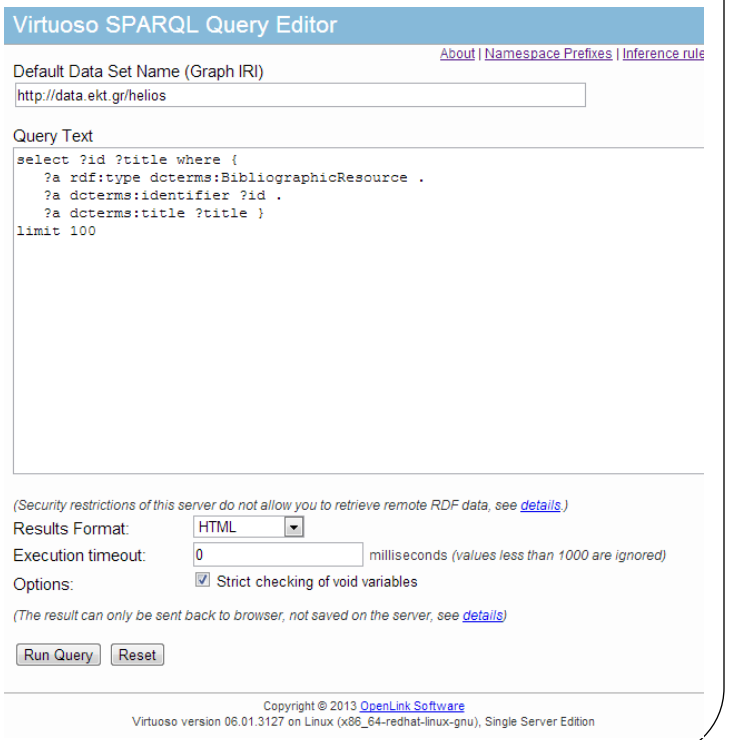
NHRF Helios repository



Virtuoso – data.ekt.gr

- SPARQL endpoint
  - http://data.ekt.gr/sparql
  - Allows arbitrary SPARQL queries on all graphs
  - Results in HTML, JSON, RDF/XML, CSV etc.
    - Allows programmatic access
- Faceted view
  - http://data.ekt.gr/fct
  - Full-text search capabilities





Discussion – Benefits (1/2)

- Semantic annotation
  - Data is unambiguously interpreted and understood by humans and software clients
- Query simplification
  - Complex SQL queries can be mapped to concepts

SPARQL Query: Article abstracts

```
SELECT ?id ?abstract
FROM <http://data.ekt.gr/helios>
FROM <http://data.ekt.gr/phdtheses>
WHERE {
?a rdf:type
dcterms:BibliographicResource .
?a dcterms:identifier ?id .
?a dcterms:abstract ?abstract }
```

SQL Query: Article abstracts

```
SELECT h.handle AS handle, mv.text_value
AS text_value
FROM handle AS h, item AS i, metadatavalue
AS
mv, metadataschemaregistry AS msr,
metadatafieldregistry AS mfr WHERE
i.in_archive=TRUE AND
h.resource_id=i.item_id AND
h.resource_type_id=2 AND
msr.metadata_schema_id=mfr.metadata_schema
_id AND
mfr.metadata_field_id=mv.metadata_field_id
AND
mv.text_value is not null AND
i.item_id=mv.item_id AND
msr.namespace =
'http://dublincore.org/documents/dcmi-
terms/' AND
mfr.element='description' AND
mfr.qualifier='abstract' """.
```

Discussion – Benefits (2/2)

- Increased discoverability
  - Through interconnections to other datasets
- Reduced effort required for schema modifications
  - New concepts can be created without altering the source schema
- Synthesis
  - Integration, fusion, mashups
- Inference
  - Reasoning is possible over the result
- Reusability
  - Third parties can reuse the data

Discussion – Challenges (1/2)

- Multidisciplinarity
  - Computer Science, Library Science
  - Contributions from both domains are required
- The technological barrier
  - No advanced mapping tools exist yet
  - Presence of a technical expert is required
- Result is prone to errors
  - Even after the resulting graph is produced
  - Lack of validation or automation can leave errors or bad practices go unnoticed

Discussion – Challenges (2/2)

- Concept mismatch
  - RDB fields and values may not be exact matches to RDF concepts and instances
  - Identical mappings will not always be present
- Exceptions to general mapping rules
  - Automated curation procedures will apply to the majority but not to all metadata fields and values
  - Post-transformation manual interventions will be required

Synchronous vs. Asynchronous access

- Asynchronous: persistent RDF views
  - Data is exposed periodically
    - RDF graph is materialized
    - Data does not change as frequently as it does in e.g. sensor or social network data
    - More viable option in the case of digital repositories
- Synchronous: transient views
  - Real-time SPARQL-to-SQL translation
    - RDF data is not materialized (as in SQL views)
    - Queries are round-trips to the database
    - Higher cost in terms of computational burden
    - Small benefit (since data does not change frequently)

Conclusions – Future Work

- Conclusions
  - Balance between
    - Experimenting with state-of-the-art technologies
      - Initial investment pays off in numerous ways
    - Carrying the responsibility of maintaining national archives
      - Ensure dataset high value and, most importantly, its viability
- Future work
  - Put more effort in R2RML Parser development
    - Cover more R2RML functionality, offer more related services
  - Improve dataset
    - Quantity: Map and export more database fields, and more datasets as RDF graphs in http://data.ekt.gr
    - Quality: Denser links to other datasets

Thank you!

Questions?