WEB SCIENCE BRASIL

Brazilian Institute for Web Science Research

Research on the Web of Data at the INCT for Web Science

Marco A. Casanova

CNPq

Conselho Nacional de Desenvolvimento Científico e Tecnológico

FAPERJ

Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro

Topics

- INCT for Web Science
- Web of Data
- Web of Data at the INCT for Web Science
- Future work

Topics

- INCT for Web Science
  - Web Science
  - Mission
  - Research Areas
  - Partners
- Web of Data
- Web of Data at the INCT for Web Science
- Conclusions

INCT for Web Science

Brazilian Institute for Web Science Research

- Web Science
  - A new domain where the Web is the subject of study
  - *"The science that investigates problems associated with decentralized information systems, encompassing people, software and hardware, and their multiple and complex interactions"*

BERNERS-LEE, T., HALL, W., HENDLER, J.A., O'HARA, K., SHADBOLT, N., WEITZNER, J. "A Framework for Web Science". Found. and Trends in Web Science. Now Publishers Inc, 2006.

INCT for Web Science

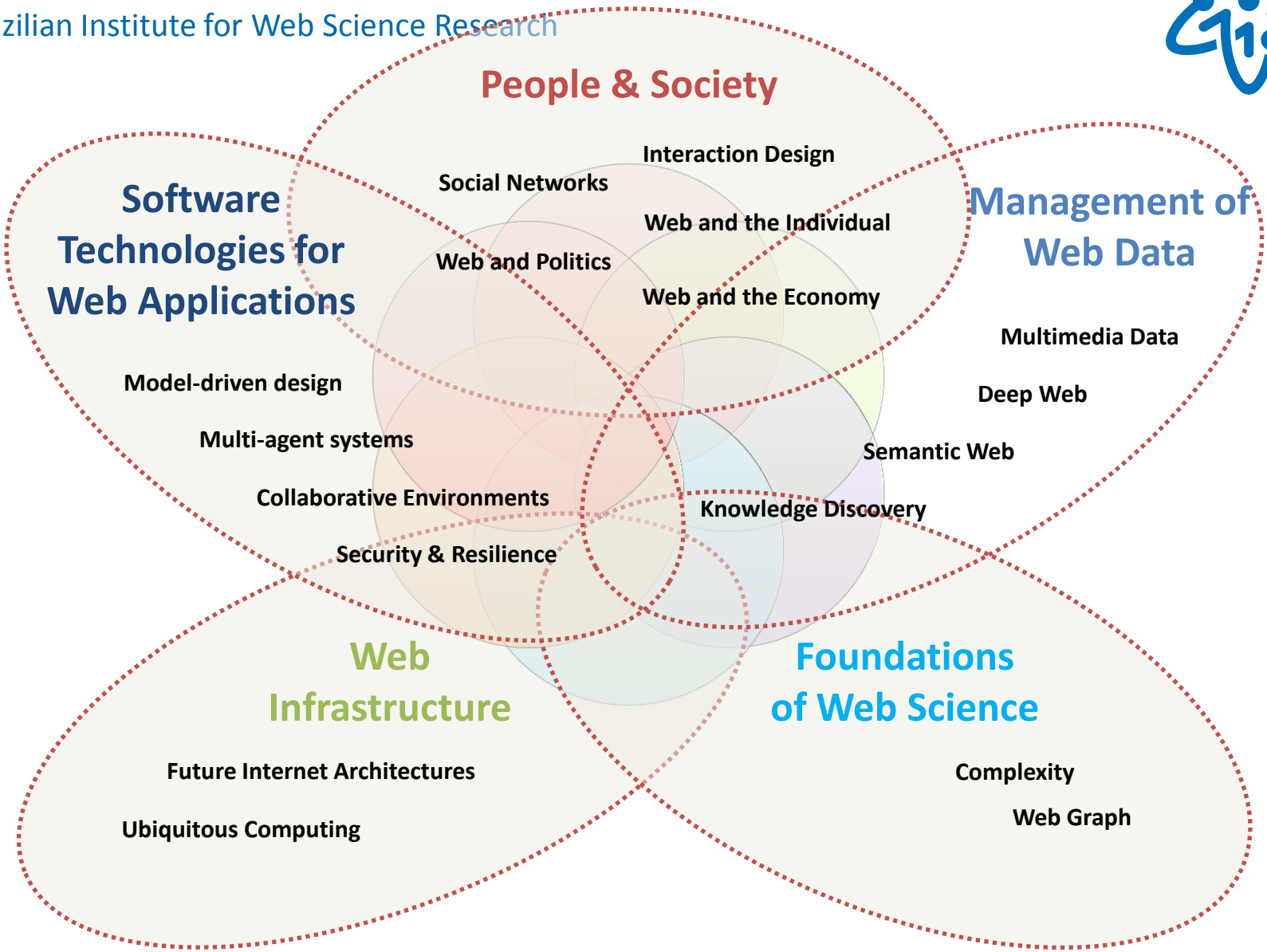
Brazilian Institute for Web Science Research

- Mission
  - To advance scientific research in themes related to the new discipline of Web Science
  - To educate professionals and researchers, to guarantee innovation and to promote interdisciplinary cooperation
  - To transfer knowledge to all sectors of society through comprehensive educational programs
  - To transfer knowledge to industry and to the services sector through specific programs, including residencies in the Institute

MACULAN, N., LUCENA, C.J.P. Brazilian Institute for Web Science Research. MCC46/08. Dept. Informatics, PUC-Rio.

INCT for Web Science

Brazilian Institute for Web Science Research



INCT for Web Science

Brazilian Institute for Web Science Research

- Brazilian Partners
  - PUC-Rio
  - UFRJ
  - UNICAMP
  - RNP
  - UNIRio
  - UFF
  - UERJ
  - UENF
  - UFRN
  - UFC
- International Partners
  - DERI
  - L3S
  - LIP6
  - LERO
  - U. Waterloo
- Member of the Web Science Trust Network of Labs

INCT for Web Science

Brazilian Institute for Web Science Research

- Web of Data
  - Group
    - PUC-Rio (ALF, MAC, KKB)
    - UFC (VMPV, JAFM)
    - UFF (JVF, LAPPL)
  - Additional funding
    - Various CNPq grants
    - CAPES/PROCAD

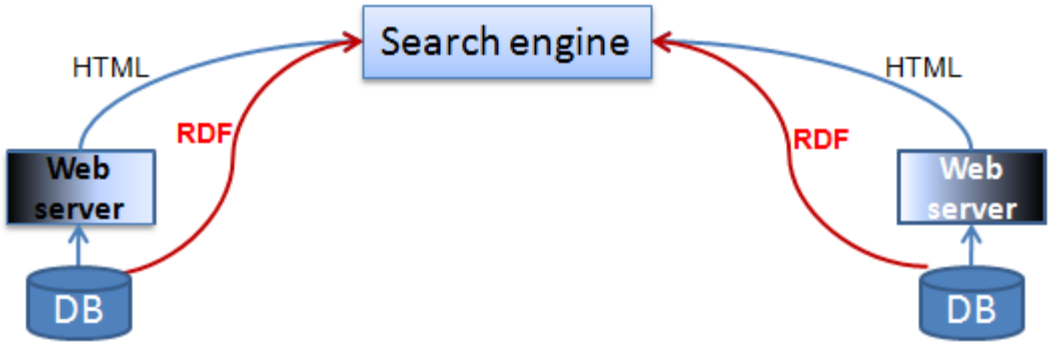
Topics

- INCT for Web Science
- Web of Data
  - Motivation
  - Key concepts and technologies
  - Linked Data
  - Open Government Data
- Web of Data at the INCT for Web Science
- Conclusions

Motivation

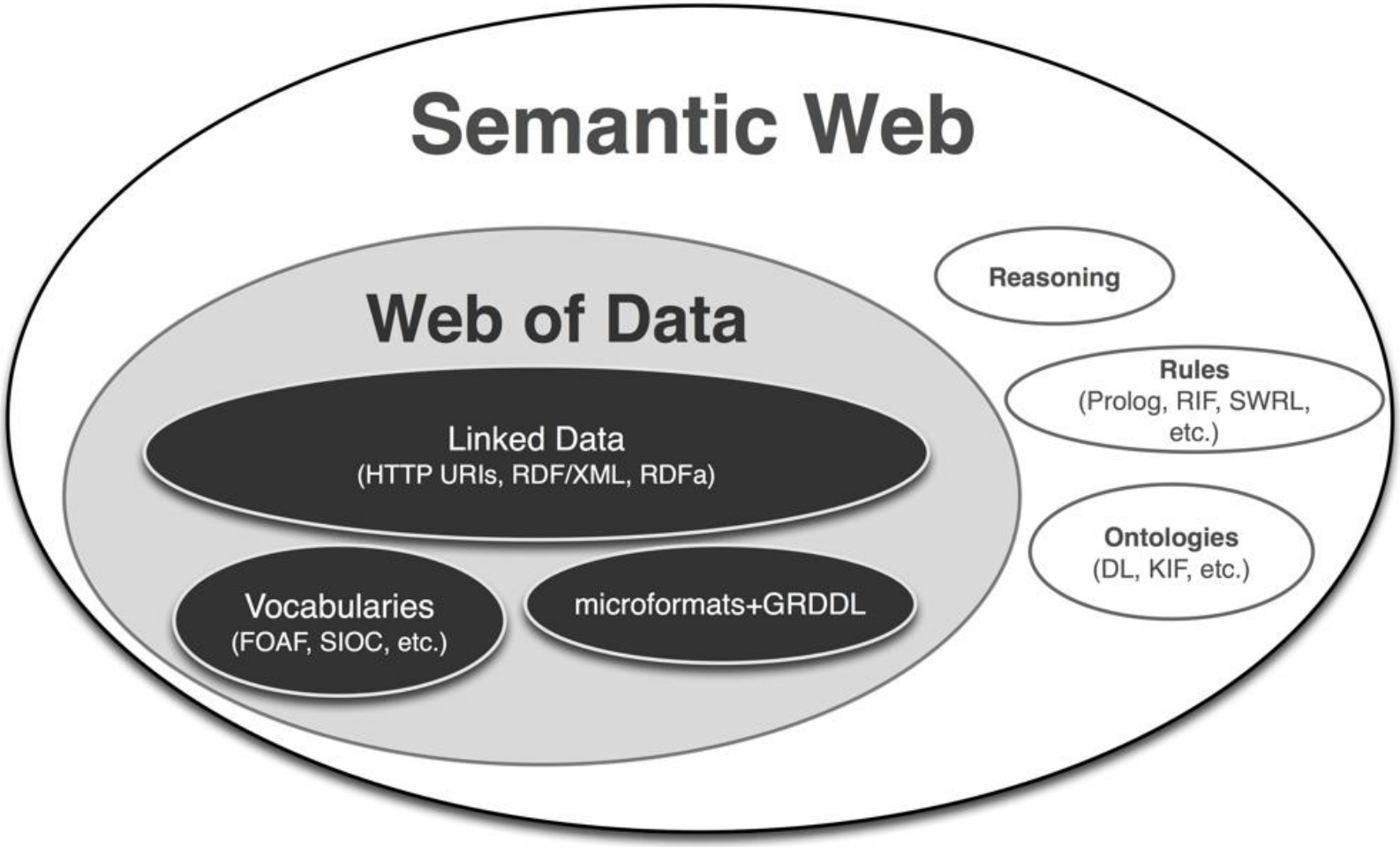
Web of Data

- Problems
  - Data available on the Web may be opaque to search engines
  - Deep Web databases are hard to discover, integrate and access
- Solution
  - Complement Web pages with structured linked data
  - Integrate structured linked data from different sources



Key concepts and technologies

Web of Data



Key concepts and technologies

Web of Data

- URI – Uniform Resource Identifier
  - a compact sequence of characters
    that identifies an abstract or physical resource
- Examples

```
http://lattes.cnpq.br/0400232298849115
http://purl.org/dc/elements/1.1/creator
http://sw.opencyc.org/concept/Mx4rwDC_HpwpEbGdrcN5Y29ycA
http://zitgist.com/music/artist/76c9a186-75bd-436a-85c0-823e3efddb7f
```

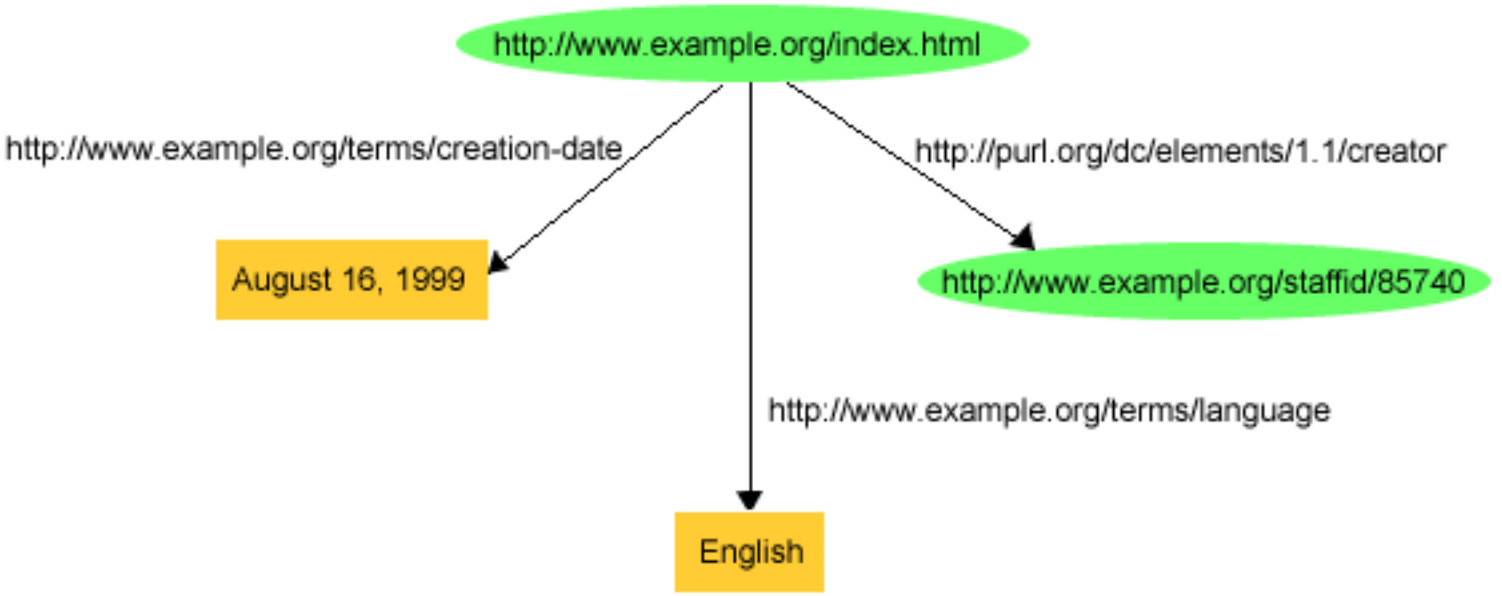
Key concepts and technologies

Web of Data

- RDF – Resource Description Framework

```
ex:index.html  exterms:creation-date  "August 16, 1999" .
ex:index.html  exterms:language       "English" .
ex:index.html  dc:creator  http://www.example.org/staffid/85740 .
```

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dc="http://purl.org/dc/elements/1.1/"
         xmlns:exterms="http://www.example.org/terms/">
  <rdf:Description rdf:about="http://www.example.org/index.html">
    <exterms:creation-date>August 16, 1999</exterms:creation-date>
    <exterms:language>English</exterms:language>
    <dc:creator rdf:resource="http://www.example.org/staffid/85740"/>
  </rdf:Description>
</rdf:RDF>
```




```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dc="http://purl.org/dc/elements/1.1/"
         xmlns:exterms="http://www.example.org/terms/">
  <rdf:Description rdf:about="http://www.example.org/index.html">
    <exterms:creation-date>August 16, 1999</exterms:creation-date>
    <exterms:language>English</exterms:language>
    <dc:creator rdf:resource="http://www.example.org/staffid/85740"/>
  </rdf:Description>
</rdf:RDF>
```

Key concepts and technologies

Web of Data

- Ontologies
  - Schema layer of RDF
  - Define classes and properties, and their constraints
  - Typically written in RDFS or OWL
- Examples
  - DCMI Metadata Terms (“Dublin Core”)
  - FOAF – The Friend of a Friend Vocabulary
  - FRBR – Functional Requirements for Bibliographic Records
  - VoID – Vocabulary of Interlinked Datasets

Key concepts and technologies

Web of Data

- Remarks about ontology design
  - “Same-old-conceptual-design”
  - “Lavoisier Principle” applies
    - Reuse known ontologies as much as possible
  - Ontology = Vocabulary + Axioms
    - Axioms (or constraints) capture the semantics of the terms
- Example
  - The Music ontology uses the FOAF, FRBR, Event and Timeline ontologies

Topics

- INCT for Web Science
- Web of Data
  - Motivation
  - Key concepts and technologies
  - Linked Data
  - Open Government Data
- Web of Data at the INCT for Web Science
- Conclusions

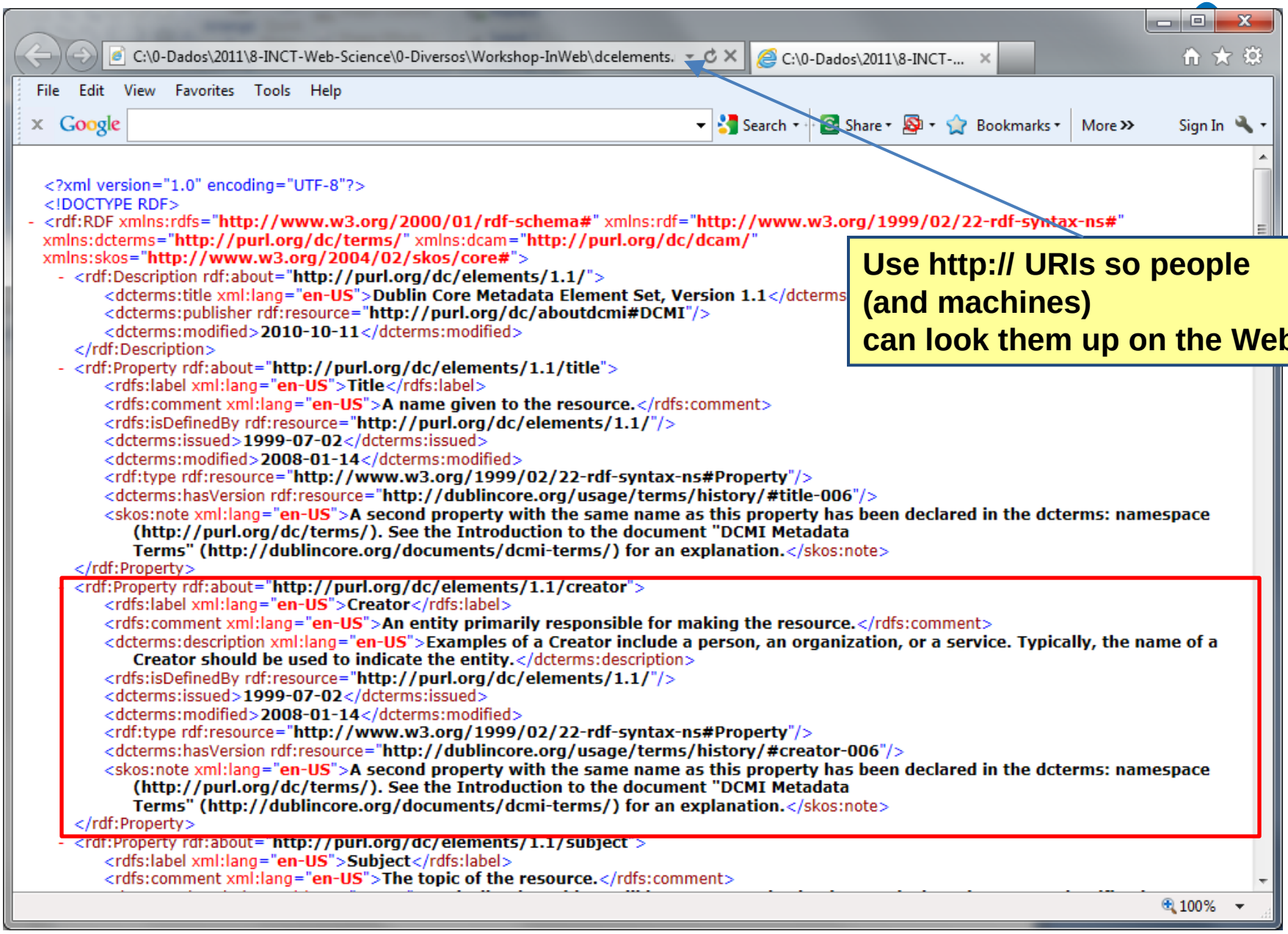
Linked Data

Web of Data

- ‘Linked Data Principles’ (in plain terms)
  1. Use URIs to identify the “things” in your data
  2. Use http:// URIs so people (and machines) can look them up on the Web
  3. When a URI is looked up, return a description of the “thing” (in RDF format)
  4. Include links to related “things”

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


wiki.dbpedia.org : Online A
About: Janis Joplin
Concept: "Janis Joplin" (Mx
sw.opencyc.org/concept/Mx4rwDC_HpwpEbGdrcN5Y29ycA
bdri.dcc.ufam.edu.b...
Strikes in Greece
About: Janis Joplin
OpenCyc (Current): [http://sw.opencyc.org/concept/Mx4rwDC_HpwpEbGdrcN5Y29ycA]
OpenCyc (Versioned): [http://sw.opencyc.org/2009/04/07/concept/Mx4rwDC_HpwpEbGdrcN5Y29ycA]
Search
Use http:// URIs so people (and machines) can look them up on the Web
cYcorp
OpenCyc Individual: Janis Joplin
Unique ID: [Mx4rwDC_HpwpEbGdrcN5Y29ycA]
English ID: [JanisJoplin]
English Aliases: ["Joplin"]
Instance of: female person, musician
Wikipedia:
http://en.wikipedia.org/wiki/Janis_Joplin
Same as:
http://dbpedia.org/resource/Janis_Joplin
Related to (broader):
Related to (narrower):
Copyright © 2001-2009 Cycorp, Inc.


C:\0-Dados\2011\8-INCT-Web-Science\0-Diversos\Workshop-InWeb\dcelements.
C:\0-Dados\2011\8-INCT-...
File Edit View Favorites Tools Help
Google
Search
Share
Bookmarks
More
Sign In
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE RDF>
<rdf:RDF xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dcterms="http://purl.org/dc/terms/" xmlns:dcam="http://purl.org/dc/dcam/"
xmlns:skos="http://www.w3.org/2004/02/skos/core#">
<rdf:Description rdf:about="http://purl.org/dc/elements/1.1/">
<dcterms:title xml:lang="en-US">Dublin Core Metadata Element Set, Version 1.1</dcterms
<dcterms:publisher rdf:resource="http://purl.org/dc/aboutdcmi#DCMI"/>
<dcterms:modified>2010-10-11</dcterms:modified>
</rdf:Description>
<rdf:Property rdf:about="http://purl.org/dc/elements/1.1/title">
<rdfs:label xml:lang="en-US">Title</rdfs:label>
<rdfs:comment xml:lang="en-US">A name given to the resource.</rdfs:comment>
<rdfs:isDefinedBy rdf:resource="http://purl.org/dc/elements/1.1/"/>
<dcterms:issued>1999-07-02</dcterms:issued>
<dcterms:modified>2008-01-14</dcterms:modified>
<rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
<dcterms:hasVersion rdf:resource="http://dublincore.org/usage/terms/history/#title-006"/>
<skos:note xml:lang="en-US">A second property with the same name as this property has been declared in the dcterms: namespace (http://purl.org/dc/terms/). See the Introduction to the document "DCMI Metadata Terms" (http://dublincore.org/documents/dcmi-terms/) for an explanation.</skos:note>
</rdf:Property>
<rdf:Property rdf:about="http://purl.org/dc/elements/1.1/creator">
<rdfs:label xml:lang="en-US">Creator</rdfs:label>
<rdfs:comment xml:lang="en-US">An entity primarily responsible for making the resource.</rdfs:comment>
<dcterms:description xml:lang="en-US">Examples of a Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.</dcterms:description>
<rdfs:isDefinedBy rdf:resource="http://purl.org/dc/elements/1.1/"/>
<dcterms:issued>1999-07-02</dcterms:issued>
<dcterms:modified>2008-01-14</dcterms:modified>
<rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
<dcterms:hasVersion rdf:resource="http://dublincore.org/usage/terms/history/#creator-006"/>
<skos:note xml:lang="en-US">A second property with the same name as this property has been declared in the dcterms: namespace (http://purl.org/dc/terms/). See the Introduction to the document "DCMI Metadata Terms" (http://dublincore.org/documents/dcmi-terms/) for an explanation.</skos:note>
</rdf:Property>
<rdf:Property rdf:about="http://purl.org/dc/elements/1.1/subject">
<rdfs:label xml:lang="en-US">Subject</rdfs:label>
<rdfs:comment xml:lang="en-US">The topic of the resource.</rdfs:comment>
100%
Use http:// URIs so people (and machines) can look them up on the Web

Linked Data

Web of Data

- ‘Linked Data Principles’
  1. Data is strictly separated from formatting and presentational aspects
  2. Data is self-describing
     - If an application consuming Linked Data encounters data described with an unfamiliar vocabulary, the application can dereference the URIs that identify vocabulary terms in order ***to find their definition***

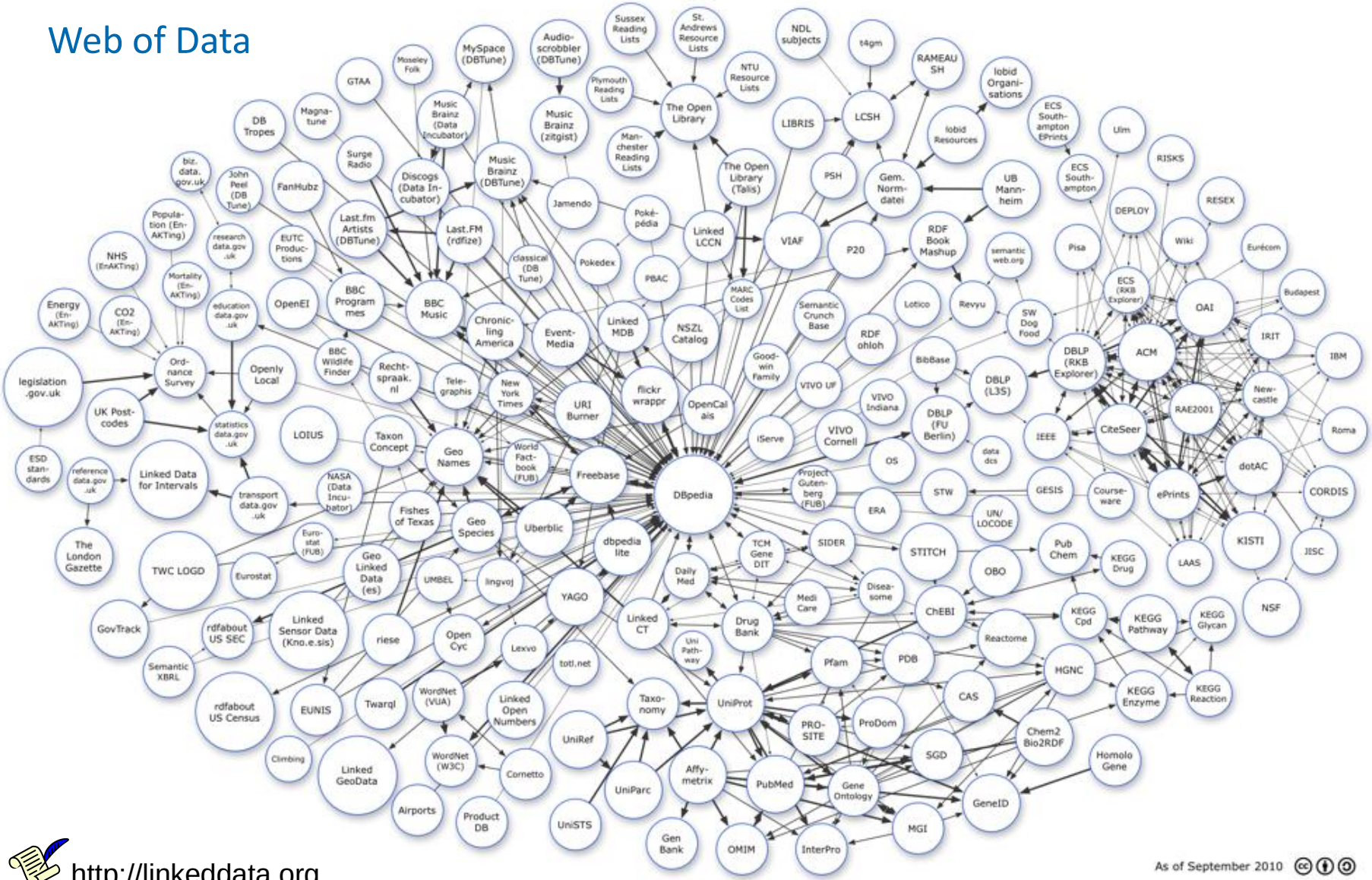
Linked Data

Web of Data

- ‘Linked Data Principles’ (cont.)
  3. Uses HTTP as a standardized data access mechanism and RDF as a standardized data model
     - simplifies data access when compared to Web APIs, which rely on heterogeneous data models and access interfaces
  4. The Web of Data is open
     - applications do not have to be implemented against a fixed set of data sources, but they can discover new data sources at run-time by following RDF links

Linked Data

Web of Data

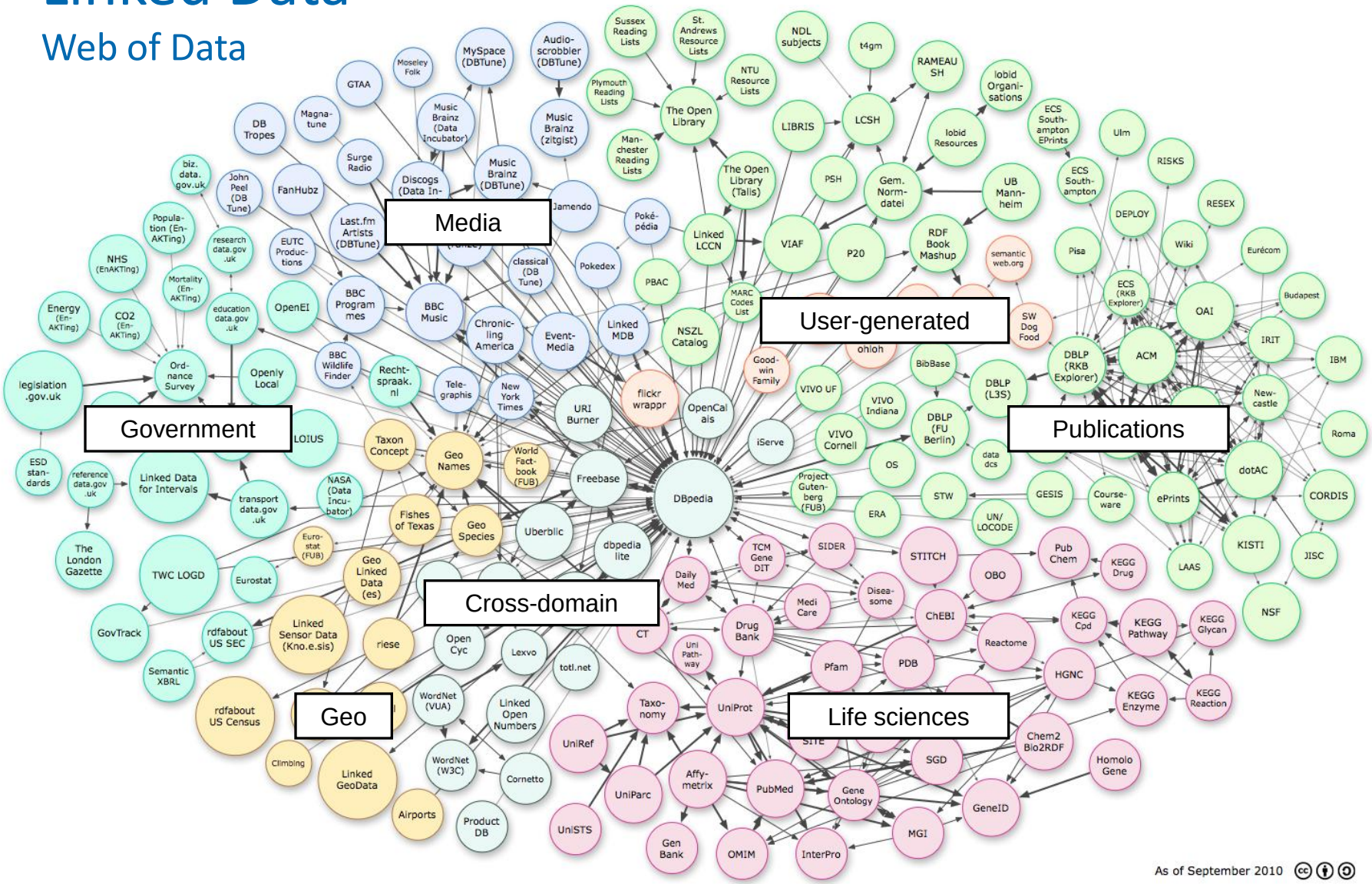


http://linkeddata.org

As of September 2010

Linked Data

Web of Data



Linked Data

Web of Data

- LOD (Nov. 2010)
  - ~200 datasets, ~30 billion triples, ~400 million links
  - Distribution of triples by domain (in millions)

Domain	Jun 2009	Nov 2010	% Growth
Geographic	3.097	5.904	91%
Libraries	212	2.237	955%
Media	698	2.453	252%
Life Sciences	2.429	2.664	10%
Cross Domain	214	1.999	834%
User Generated	76	57	-25%
Government	0	11.613	-

Topics

- INCT for Web Science
- Web of Data
  - Motivation
  - Key concepts and technologies
  - Linked Data
  - Open Government Data
- Web of Data at the INCT for Web Science
- Conclusions

Open Government Data

Web of Data

- Open government data
  - public government information
    – such as government records –
    that is shared with the public digitally,
    over the Internet, in open raw formats,
    and
    ways that make it accessible and readily available to all
    to promote analysis and allow reuse –
    such as the creation of data mashups

Open Government Data

Web of Data

(Where is…)?

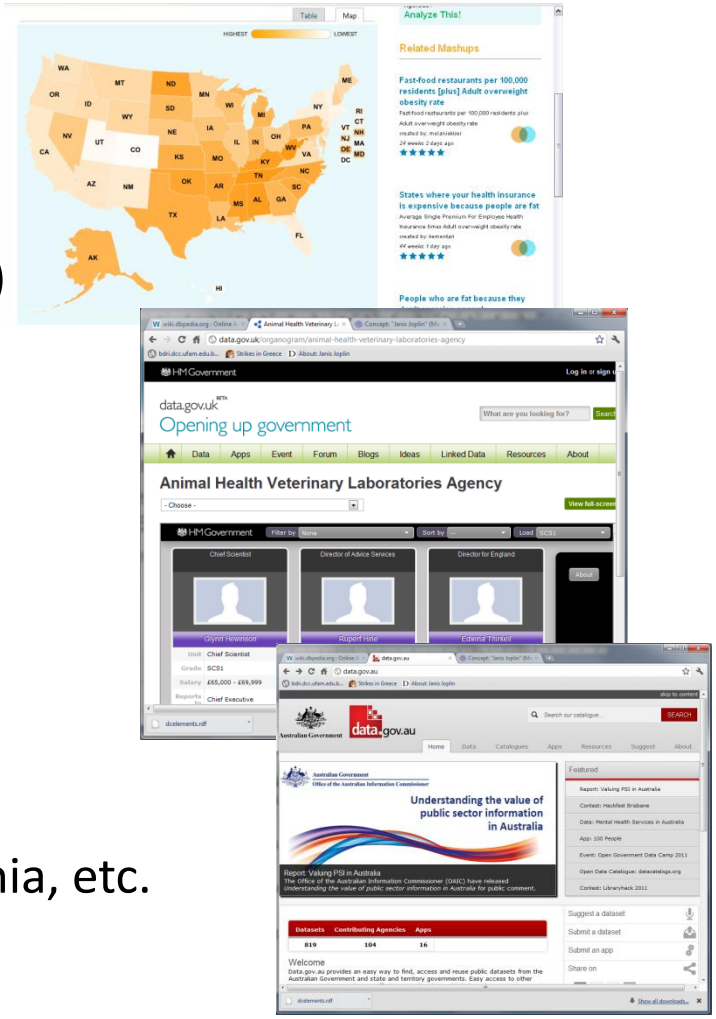
Open Government Data

Web of Data

Open Government Data

Web of Data

- OGD around the world
  - United States:
    - data.gov
      (rel. May/2009; redesign. May/2010)
  - United Kingdom:
    - data.gov.uk
      (released in January/2010)
  - Australia:
    - data.australia.gov.au
      (released in October/2009)
  - Other:
    - Canada, New Zealand, Norway, Estonia, etc.



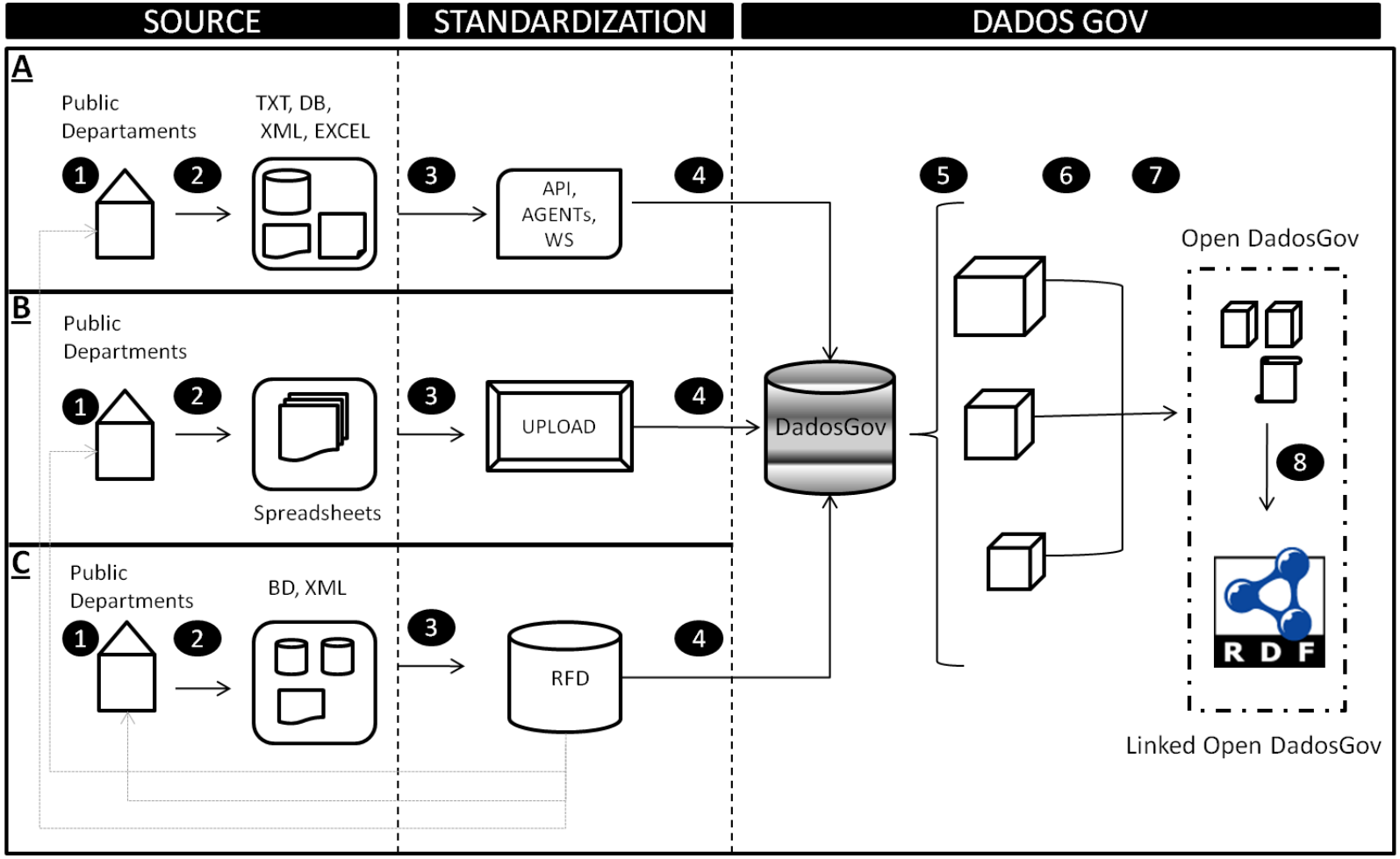
Open Government Data

Web of Data

- OGD in Brazil
  - Federal Law 131 - 27/05/2009
    - Establishes norms for publishing, in real time, detailed information about budget and financial resources
  - Action Plan - 15/09/2011
    - Stimulates the use of new technologies in the management and provision of public services and access to public information
    - Related to the participation of Brazil as member of the Open Government Partnership

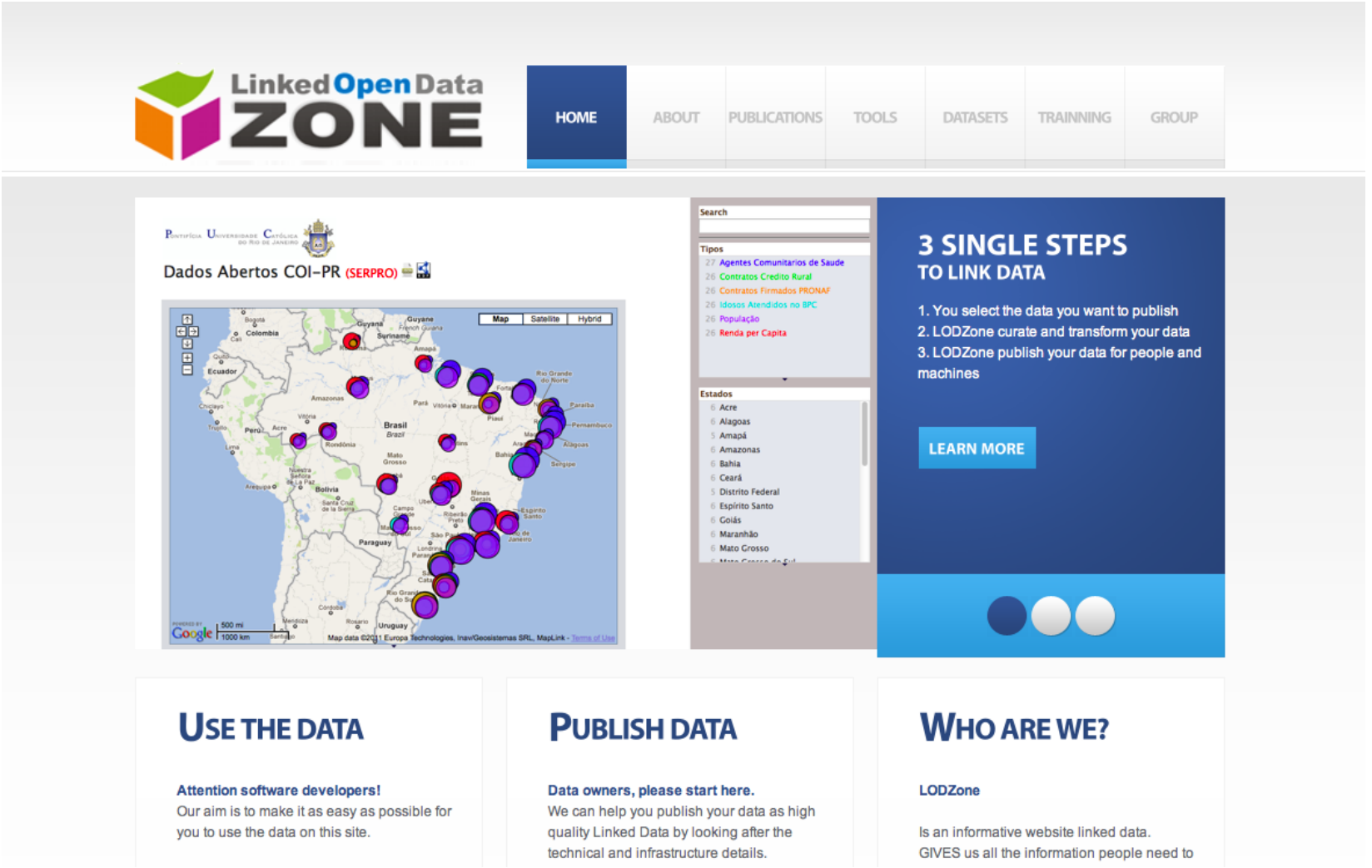
Open Government Data

Web of Data



Open Government Data

Web of Data



Topics

- INCT for Web Science
- Web of Data
- Web of Data at the INCT for Web Science
  - Design of Linked Data Sources
  - Publication of Linked Data
  - Consumption of Linked Data
- Conclusions

Design of Linked Data Sources

Web of Data at the INCT for Web Science

- Research Goals
  - Investigate correctness criteria for application ontologies
  - Develop methods and tools to
    - recommend ontologies
    - specify ontology mappings
    - find identical resources
    - retain design rational
  - Develop large scale experiments

CASANOVA, M. A., BREITMAN, K.K., FURTADO, A. L., VIDAL, V.M.P., MACEDO, J.A.F., GOMES, R. V., SALAS, P.E. The Role of Constraints in Linked Data In: ODBASE 2011, Hersonissos, Greece.

Correctness Criteria for Application Ontologies

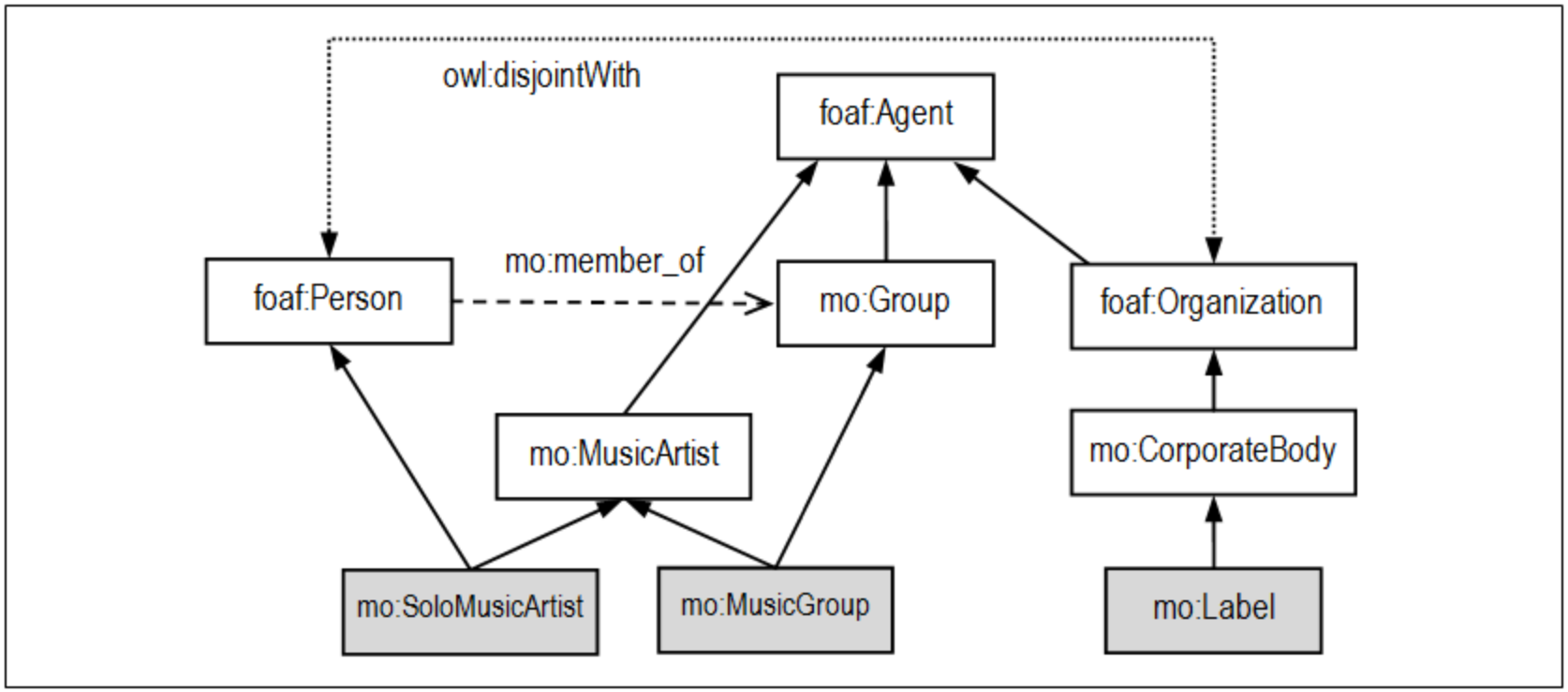
Design of Linked Data Sources

- Typical Design Process
  1. Select one or more *domain ontologies*
  2. Design the *application ontology* based on the domain ontologies
- Example
  - Domain Ontology = music ontology
  - Application Ontology = JanisJoplinFanClub

Correctness Criteria for Application Ontologies

Design of Linked Data Sources

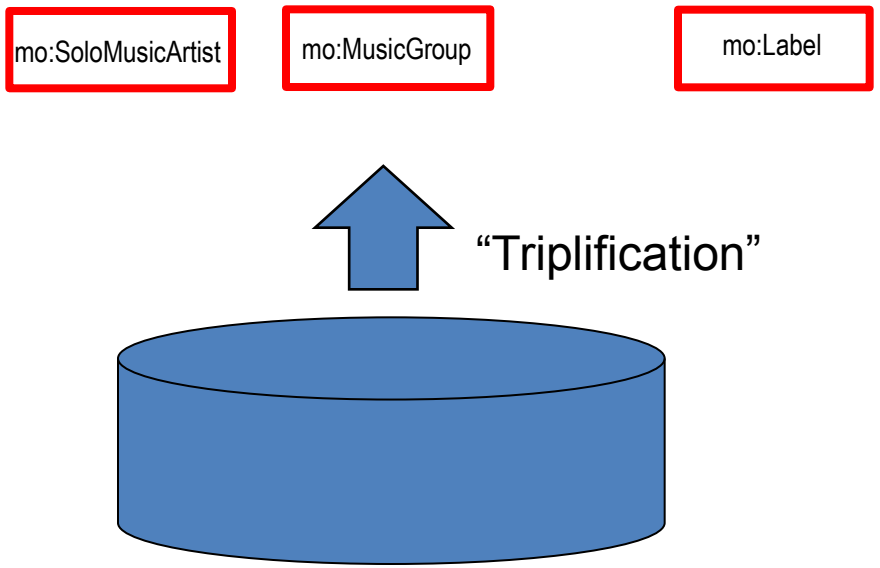
- Typical Design Process



Correctness Criteria for Application Ontologies

Design of Linked Data Sources

- Typical Design Process



Correctness Criteria for Application Ontologies

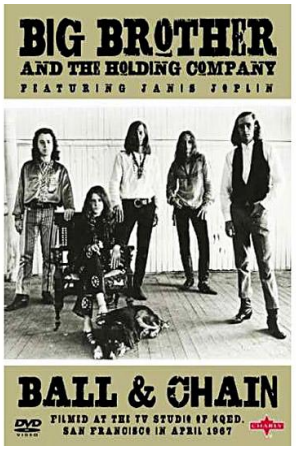
Design of Linked Data Sources

mo:SoloMusicArtist

mo:MusicGroup

mo:Label

(uri1, mo:SoloMusicArtist, “Janis Joplin”)
(uri2, mo:MusicGroup, “Big Brother and the Holding Company”)
(uri1, mo:Label, “Columbia”)



Correctness Criteria for Application Ontologies

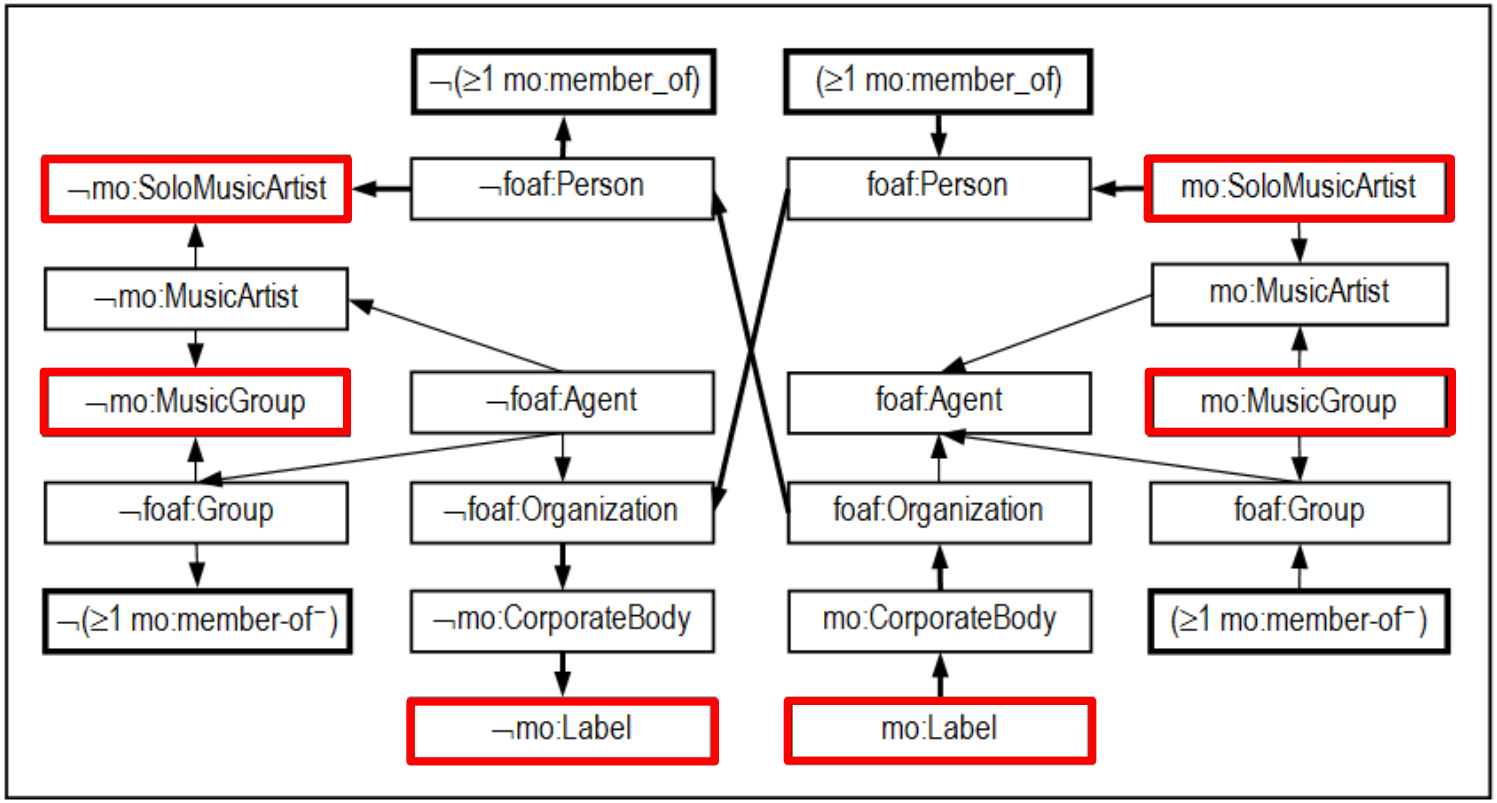
Design of Linked Data Sources

- Questions:
  - Are your constraints right?
  - Have you forgotten any constraint?

Correctness Criteria for Application Ontologies

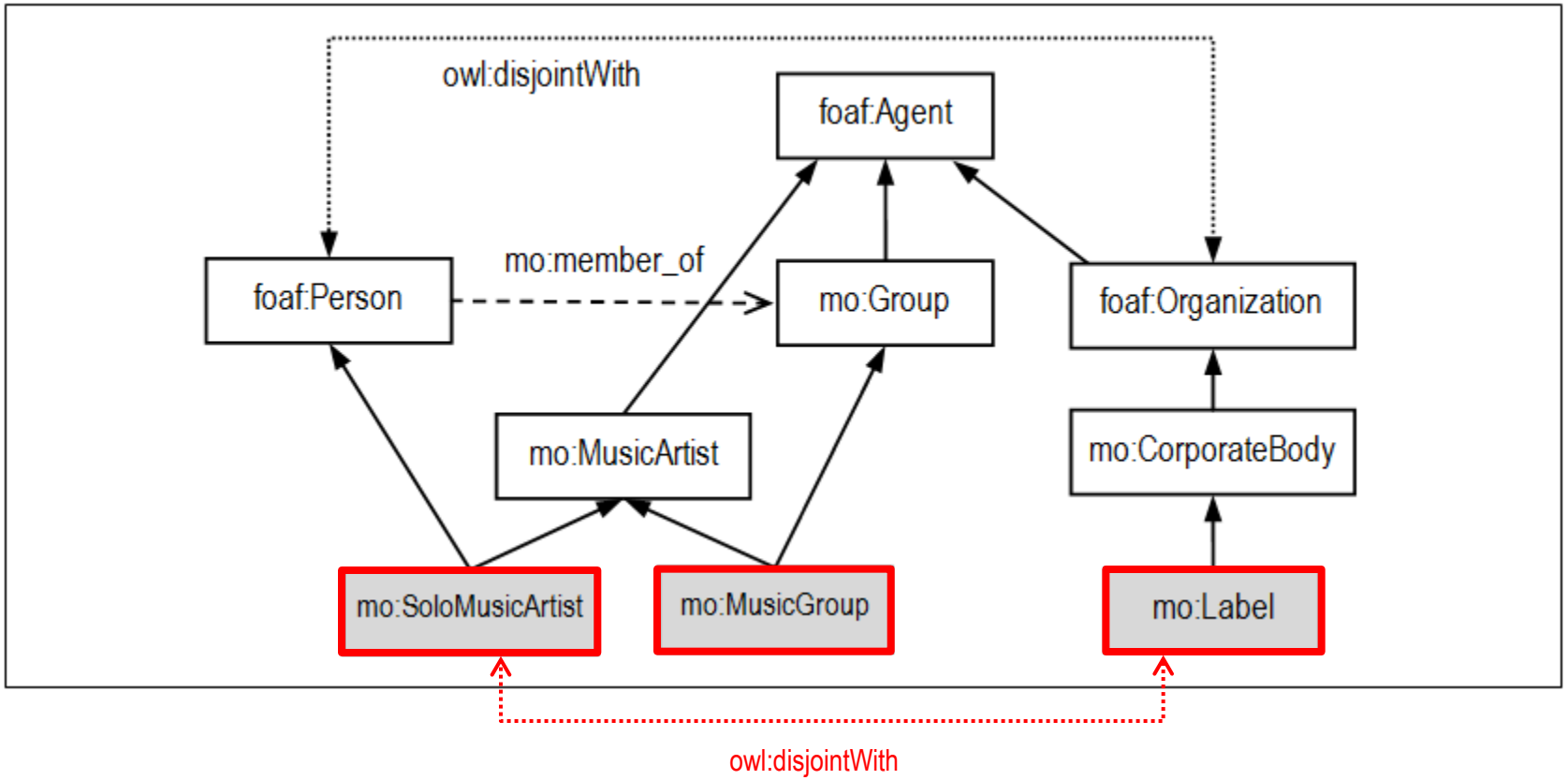
Design of Linked Data Sources

$A \cap B = \emptyset$
iff
$A \subseteq \neg B$
iff
$B \subseteq \neg A$



Correctness Criteria for Application Ontologies

Design of Linked Data Sources



Correctness Criteria for Application Ontologies

Design of Linked Data Sources

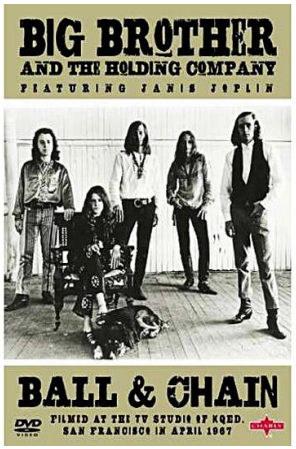
Inconsistent data

mo:SoloMusicArtist

mo:MusicGroup

mo:Label

(uri1, mo:SoloMusicArtist, “Janis Joplin”)
(uri2, mo:MusicGroup, “Big Brother and the Holding Company”)
(uri1, mo:Label, “Columbia”)



Correctness Criteria for Application Ontologies

Design of Linked Data Sources

- Ontology design revised
  - "Lavoisier Principle" applies
    - Reuse known ontologies as much as possible
  - Ontology = Vocabulary + Axioms
    - Axioms (or constraints) capture the semantics of the terms

  - Application ontology constraints must capture the semantics of the (application ontology) terms and they must be derived from the domain ontology constraints

Correctness Criteria for Application Ontologies

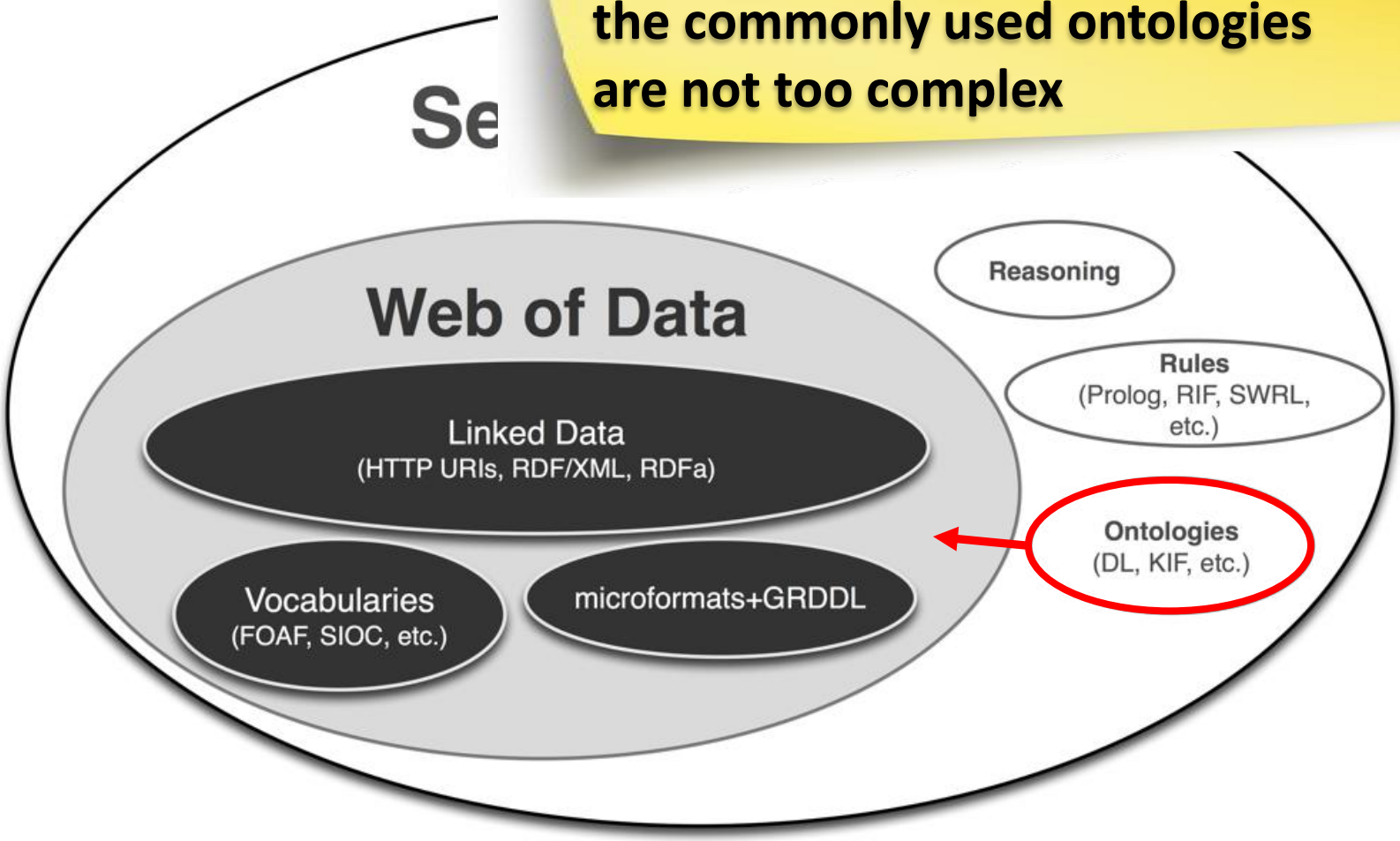
Design of Linked Data Sources

- Research Goals
  - Investigate correctness criteria for application ontologies
- Result
  - Correctness criteria for application ontologies = open (or closed) fragments of the domain ontology
  - Algorithms to *construct* the application ontology constraints (for lightweight ontologies based on DL Lite core)

Key concepts and

Web of Data

- **The deductive services must be of a different nature**
- **There is empirical evidence that the commonly used ontologies are not too complex**



Topics

- INCT for Web Science
- Web of Data
- Web of Data at the INCT for Web Science
  - Design of Linked Data Sources
  - Publication of Linked Data
  - Consumption of Linked Data
- Conclusions

Publication of Linked Data

Web of Data at the INCT for Web Science

- Research Goals
  - Develop methods and tools to publish linked datasets
    - Triplification of relational databases
    - Triplification of data cubes
    - Non-standard publication of opaque data
  - Develop methods and tools to
    - rematerialize linked datasets
    - store and index linked datasets in the Cloud
  - Develop large scale experiments

Publication of Linked Data

Web of Data at the INCT for Web Science

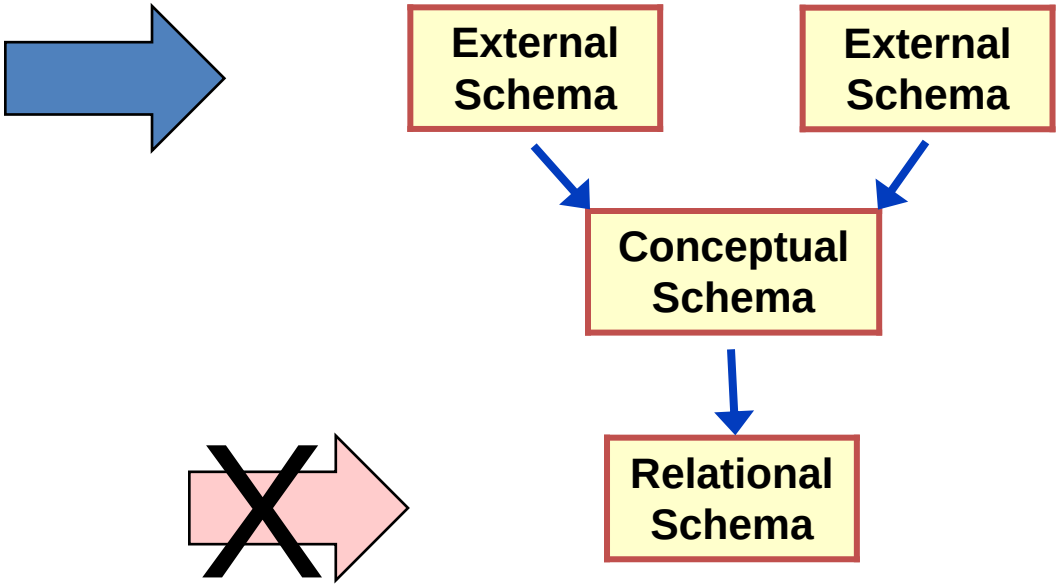
- Triplification of relational databases
  - Design guidelines
  - Std-Trip Tool

SALAS, P.E., BREITMAN, K.K., VITERBO, J., CASANOVA, M. A. Interoperability by Design Using the Std-Trip Tool: an a priori approach In: I-SEMANTICS 2010, 2010, Graz, Austria.

Triplification of Relational Databases

Publication of Linked Data

- Basic design guideline
  - published data should be meaningful to the external world

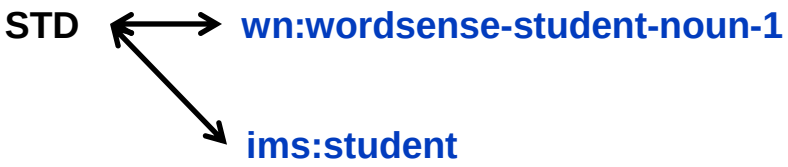


Triplification of Relational Databases

Publication of Linked Data

- Design guideline #1
  - Use a vocabulary meaningful to the external world
  - Example

STD	S-PK	S-ID	S-SSN	S-NM	S-LV
	1	102.01532	103957-99	John	3



Triplification of Relational Databases

Publication of Linked Data

- Design guideline #2
  - Avoid publishing internal keys, internal domain values, etc.

STD	S-PK	S-ID	S-SSN	S-NM	S-LV
	1	102.01532	103957-99	John	3

UNIV	U-PK	U-NM	U-Web
	32	PUC-Rio	www.puc-rio.br

LV	L-PK	L-NM
	1	BA
	2	MSc
	3	DSc
	4	PhD

ENRL	S-PK	U-PK
	1	32

Triplification of Relational Databases

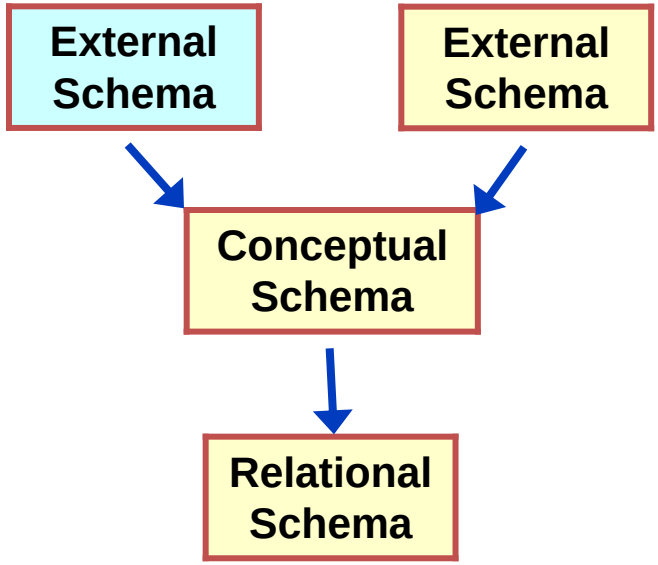
Publication of Linked Data

- Design guideline #3
  - De-normalize tables

STUDENT	SSN	Name	Level
	103957-99	John	DSc

UNIVERSITY	Name	Website
	PUC-Rio	www.puc-rio.br

ENROLL	Student	University
	103957-99	www.puc-rio.br



Triplification of Relational Databases

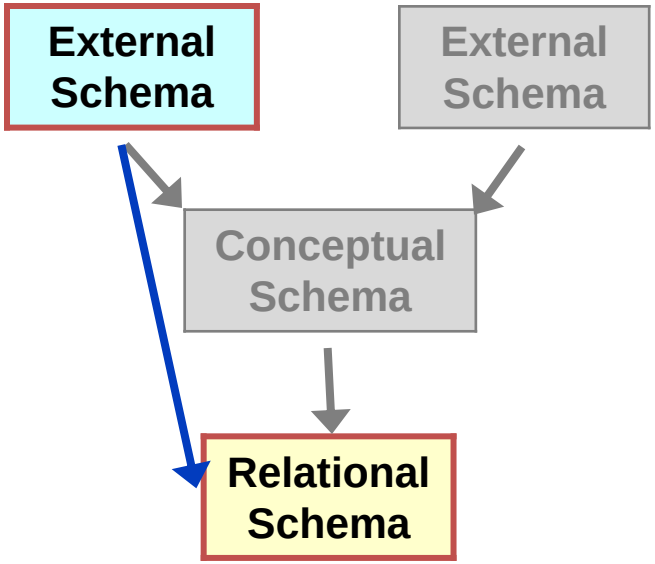
Publication of Linked Data

- Std-Trip Tool - External Schema Design
  - helps users specify an external schema, using the entity-relationship model

    Entity(Student)

    Entity(University)

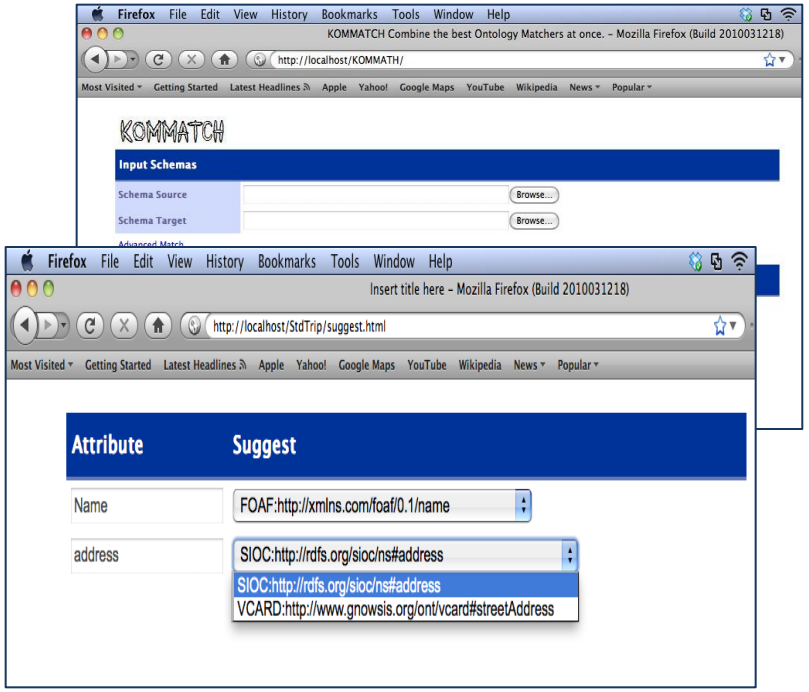
    Relationship(Student, University)
  - helps users specify how to map the external schema into the database internal schema



Triplification of Relational Databases

Publication of Linked Data

- Std-Trip Tool – Vocabulary Selection
  - helps users select a vocabulary for the external schema
    - locate published vocabularies
    - match distinct vocabularies



STD ⟷ wn:wordsense-student-noun-1

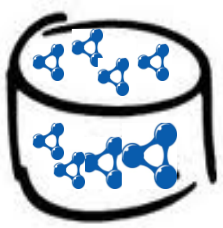
ims:student

Triplification of Relational Databases

Publication of Linked Data

- Std-Trip Tool – Triplification
  - based on the previous steps, the tool…
    - materializes the external schema
    - triplifies the materialized data

TRIPLES SET

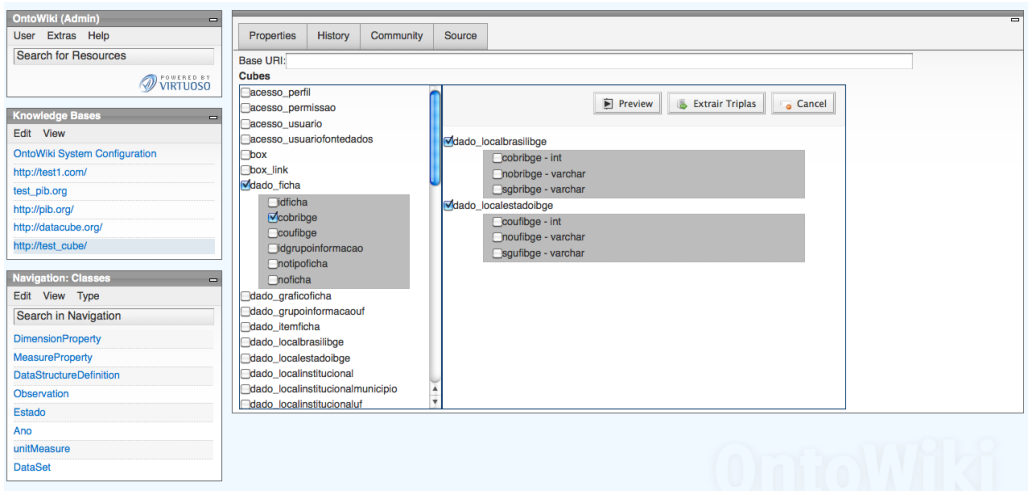


```
<http://example/address/1> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.gnowsis.org/ont/vcard#Address> .
<http://example/address/1> <http://www.gnowsis.org/ont/vcard#streetAddress> "47 MySakila Drive - Alberta" .
<http://example/address/1> <http://www.gnowsis.org/ont/vcard#city> "Lethbridge" .
<http://example/address/2> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.gnowsis.org/ont/vcard#Address> .
<http://example/address/2> <http://www.gnowsis.org/ont/vcard#streetAddress> "28 MySQL Boulevard - QLD" .
<http://example/address/2> <http://www.gnowsis.org/ont/vcard#city> "Woodridge" .
<http://example/address/3> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.gnowsis.org/ont/vcard#Address> .
<http://example/address/3> <http://www.gnowsis.org/ont/vcard#streetAddress> "23 Workhaven Lane - Alberta" .
<http://example/address/3> <http://www.gnowsis.org/ont/vcard#city> "Lethbridge" .
<http://example/address/4> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.gnowsis.org/ont/vcard#Address> .
<http://example/address/4> <http://www.gnowsis.org/ont/vcard#streetAddress> "1411 Lillydale Drive - QLD" .
<http://example/address/4> <http://www.gnowsis.org/ont/vcard#city> "Woodridge" .
<http://example/address/5> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.gnowsis.org/ont/vcard#Address> .
<http://example/address/5> <http://www.gnowsis.org/ont/vcard#streetAddress> "1913 Hanoi Way - Nagasaki" .
<http://example/address/5> <http://www.gnowsis.org/ont/vcard#postalcode> "35200" .
<http://example/address/5> <http://www.gnowsis.org/ont/vcard#city> "Sasebo" .
```

Triplification of Linked Data

Web of Data at the INCT for Web Science

- Triplification of data cubes
  - Basic design guideline
  - OlapImport Prototype



SALAS, P.E., BREITMAN, K.K., SORER, A., CASANOVA, M.A. Publishing Statistic Data on the Web. (to be submitted to ESWC 2012).

Triplification of Data Cubes

Publication of Linked Data

- Basic design guideline
  - published data should include
    a description of the dimensions of the cube
    and of their domains

Cube	Id	D1	D2	…	Dn	M
	k	d1	d2	…	dn	v

S	P	O
k	D1	d1
k	D2	d2
	…	
k	Dn	dn
k	M	v
D1	T	t1
	…	
d1	U	t2
	…	

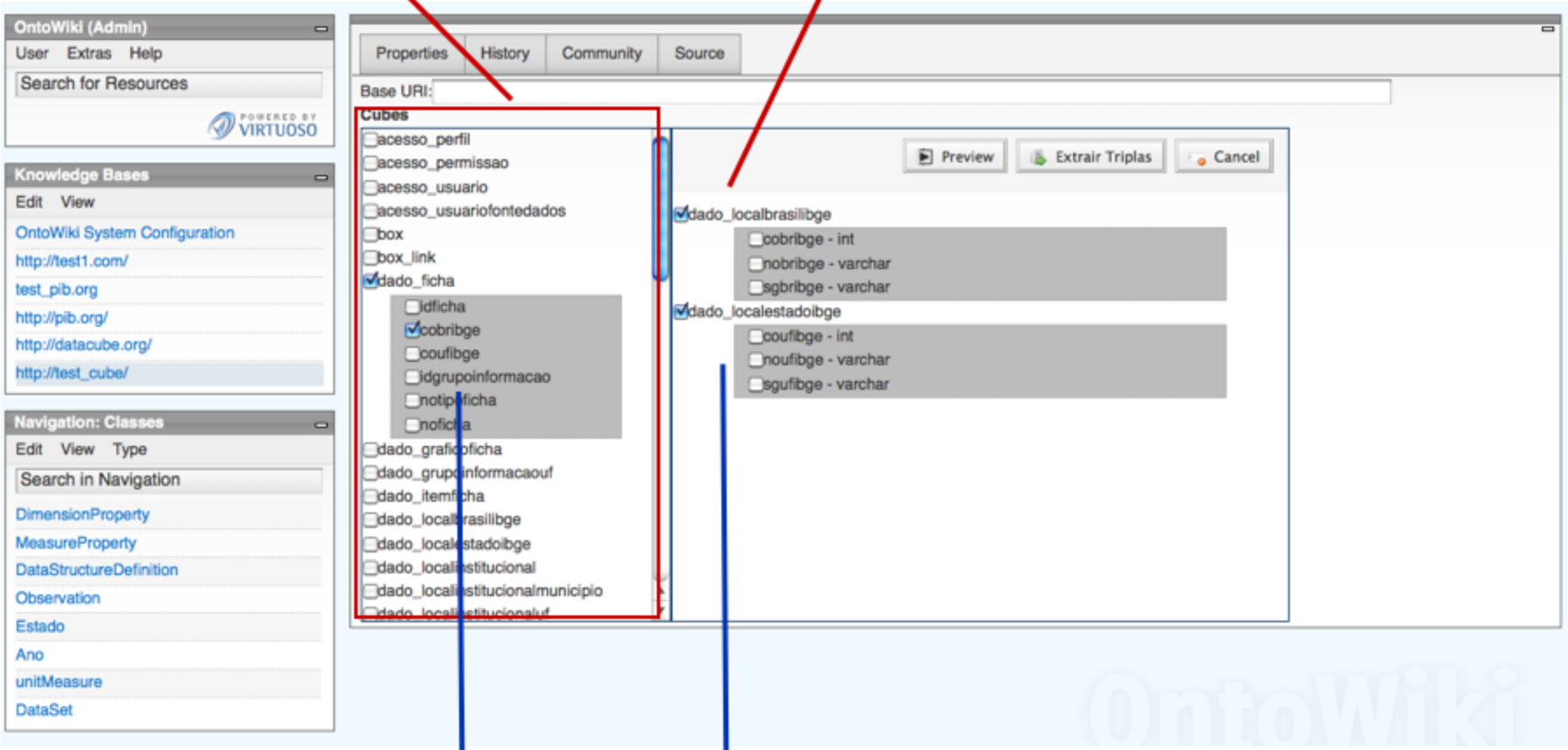
Triplification of Data Cubes

Publication of Linked Data

- OlapImport Prototype
  - Implemented as a plug-in of the OntoWIKI framework
  - (interface)
  - (sample data)

Tables (Facts)

Suggest dimensions available for the Fact selected



Select: The column which will be used for "value" (e.g lifeExpectancy)

Select: The column which will be used for "label"

A fact identified by a URI

Dimensions identified by URIs

Dimension values ident. by URIs

callret-0	p	o
http://test_cube/c1649ef1a4ac3ced22d1fcfb7982485c /item/5aee62babb5d6ded8671ef36b97ab909	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://purl.org/linked-data/cube#Observation
http://test_cube/c1649ef1a4ac3ced22d1fcfb7982485c /item/5aee62babb5d6ded8671ef36b97ab909	http://test_cube/c1649ef1a4ac3ced22d1fcfb7982485c /dimension/Estado	http://test_cube/c1649ef1a4ac3ced22d1fcfb7982485c /dimension/Estado/Pernambuco
http://test_cube/c1649ef1a4ac3ced22d1fcfb7982485c /item/5aee62babb5d6ded8671ef36b97ab909	http://test_cube/c1649ef1a4ac3ced22d1fcfb7982485c /dimension/Ano	http://test_cube/c1649ef1a4ac3ced22d1fcfb7982485c /dimension/Ano/1985
http://test_cube/c1649ef1a4ac3ced22d1fcfb7982485c /item/5aee62babb5d6ded8671ef36b97ab909	http://purl.org/linked-data/sdmx /2009/attribute#unitMeasure	http://test_cube/c1649ef1a4ac3ced22d1fcfb7982485c /attribute/Year
http://test_cube/c1649ef1a4ac3ced22d1fcfb7982485c /item/5aee62babb5d6ded8671ef36b97ab909	http://test_cube/c1649ef1a4ac3ced22d1fcfb7982485c /measure/lifeExpectancy	711548

Value

“Variable name” identified by a URI

callret-0	p	o
http://test_cube/c1649ef1a4ac3ced22d1fcfb7982485c/dimension/Estado	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://purl.org/linked-data/cube#DimensionProperty
http://test_cube/c1649ef1a4ac3ced22d1fcfb7982485c/dimension/Estado	http://www.w3.org/2000/01/rdf-schema#label	Estado
http://test_cube/c1649ef1a4ac3ced22d1fcfb7982485c/dimension/Estado	http://www.w3.org/2000/01/rdf-schema#subPropertyOf	http://localhost/ontowiki/sdmx-dimension/subUF
http://test_cube/c1649ef1a4ac3ced22d1fcfb7982485c/dimension/Estado	http://purl.org/linked-data/cube#concept	http://localhost/ontowiki/sdmx-concept/conceptUF

callret-0	p	o
http://test_cube/c1649ef1a4ac3ced22d1fcfb7982485c/dimension/Estado /Pernambuco	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://test_cube/c1649ef1a4ac3ced22d1fcfb7982485c/dimension /Estado
http://test_cube/c1649ef1a4ac3ced22d1fcfb7982485c/dimension/Estado /Pernambuco	http://www.w3.org/2000/01/rdf-schema#label	Pernambuco

Publication of Linked Data

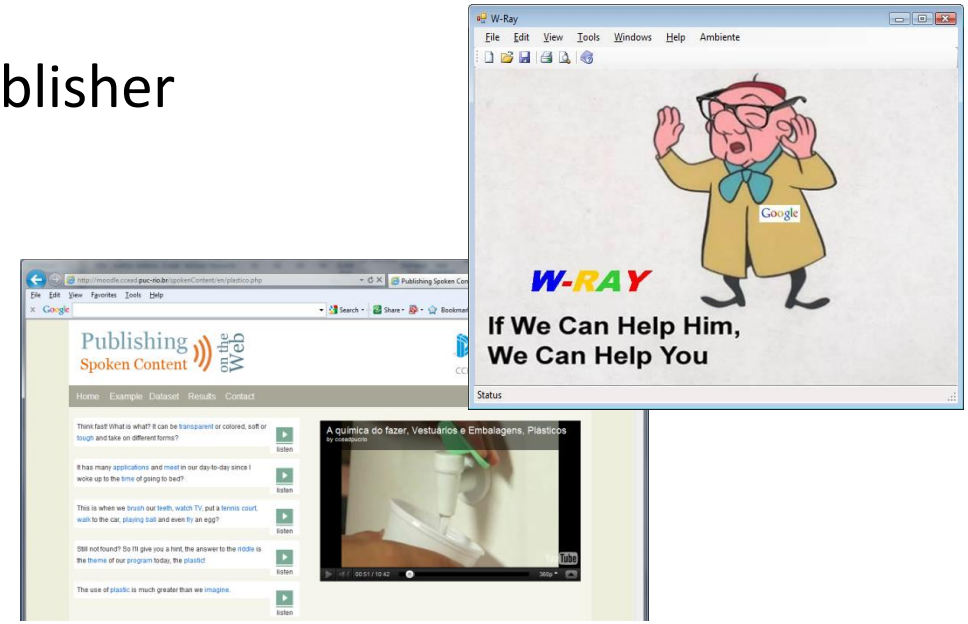
Web of Data at the INCT for Web Science

- Non-standard publication of opaque data
  - traditional search engines cannot discover “opaque data” by following hyperlinks
  - “opaque data”
    - data stored in databases and accessed through query interfaces
    - multimedia data

Publication of Linked Data

Web of Data at the INCT for Web Science

- Non-standard publication of opaque data
  - Basic strategy
  - W-Ray Tool
  - Spoken Content Publisher



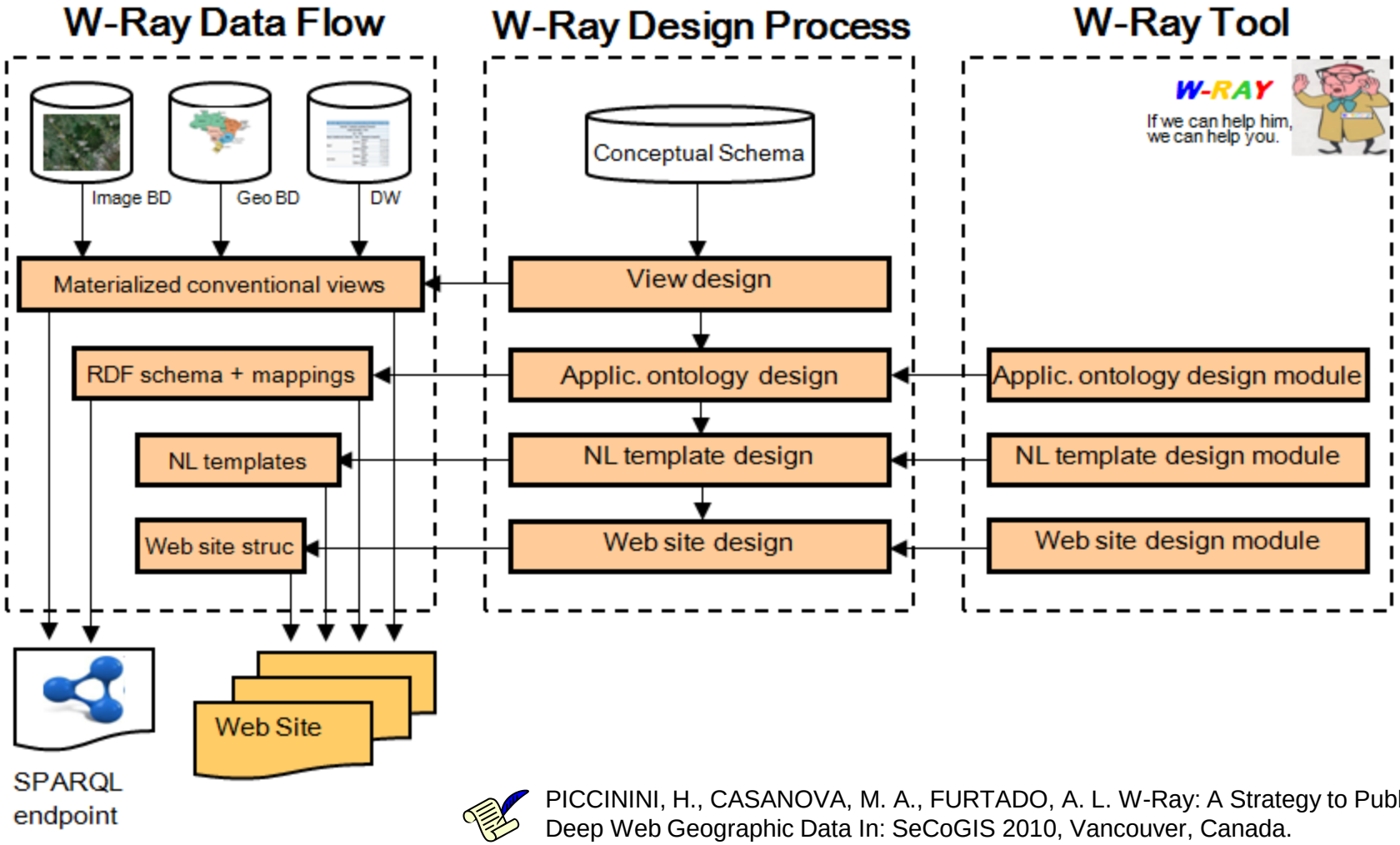
Non-standard Publication of Opaque Data

Publication of Linked Data

- Basic strategy
  - create natural language (NL) sentences
    that describe opaque data,
    and publish the sentences as static Web pages
  - use RDF triples instead on natural language sentences
    - a triple (S,P,O) is equivalent to “S has relationship P with O”
    - a triple (S,P,v) is equivalent to “S has property P with value v”
  - (a combination of the two, including RDFa)

Non-standard Publication of Opaque Data

Publication of Linked Data



PICCININI, H., CASANOVA, M. A., FURTADO, A. L. W-Ray: A Strategy to Publish Deep Web Geographic Data In: SeCoGIS 2010, Vancouver, Canada.

Non-standard Publication of Opaque Data

Publication of Linked Data

- Example – “Satellite Images”
  - extract image parameters
  - query a GIS for ‘hydrographic feature’
    - Feature(“Rodrigo de Freitas, Lagoa - Brazil”, lakes, contains)
    - Feature(“Comprido, Rio – Brazil”, streams, contains)
    - Feature(“Maracana, Rio – Brazil, streams, contains)

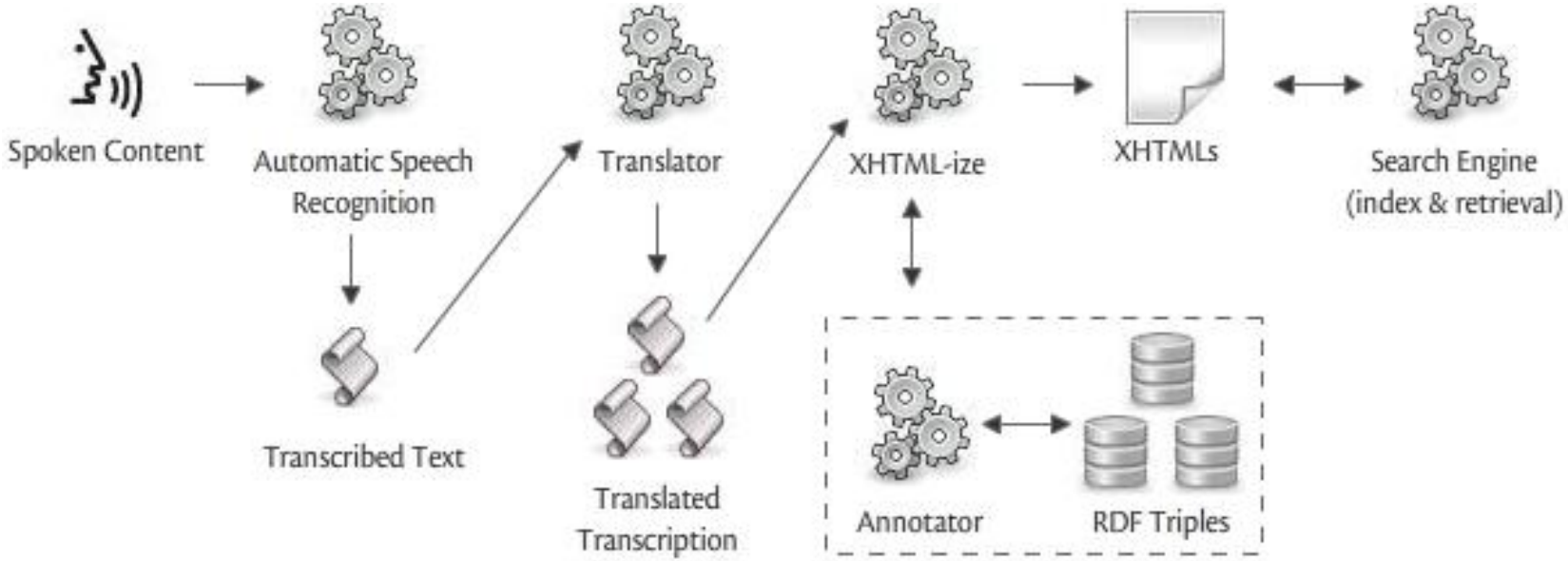
The image of Rio de Janeiro, Brazil, *contains the lake* “Rodrigo de Freitas” *and the streams* “Comprido” and “Maracanã”.

Image fragment of the City of Rio de Janeiro from the Web site “Brazil seen from Space”

Non-standard Publication of Opaque Data

Publication of Linked Data

- Spoken Content Publisher



NUNES, B.P., CASANOVA, M. A. A tool to publish descriptions of digital audio and video objects on the Web. (to be submitted to ICME 2012).


http://moodle.ccead.puc-rio.br/spokenContent/en/plastico.php
Publishing Spoken Content...
File Edit View Favorites Tools Help
Google
Search
Share
Bookmarks
AutoFill
marcoa...
Publishing Spoken Content on the Web
Departamento de Informática
CCEAD PUC RIO
PUC RIO
Home Example Dataset Results Contact
Think fast! What is what? It can be transparent or colored, soft or tough and take on different forms?
listen
It has many applications and meet in our day-to-day since I woke up to the time of going to bed?
listen
This is when we brush our teeth, watch TV, put a tennis court, walk to the car, playing ball and even fry an egg?
listen
Still not found? So I'll give you a hint, the answer to the riddle is the theme of our program today, the plastic!
listen
The use of plastic is much greater than we imagine.
listen
A química do fazer, Vestuários e Embalagens, Plásticos
by cceadpucrio
00:51 / 10:42
360p

Publication of Linked Data

Web of Data at the INCT for Web Science

- Research Goals
  - Develop methods and tools to publish linked datasets
    - Triplification of relational databases
    - Triplification of data cubes
    - Non-standard publication of opaque data
- Results
  - Std-Trip Tool
  - OlapImport Prototype
  - W-Ray Tool
  - Spoken Content Publisher

Topics

- INCT for Web Science
- Web of Data
- Web of Data at the INCT for Web Science
  - Design of Linked Data Sources
  - Publication of Linked Data
  - Consumption of Linked Data
- Conclusions

Consumption of Linked Data

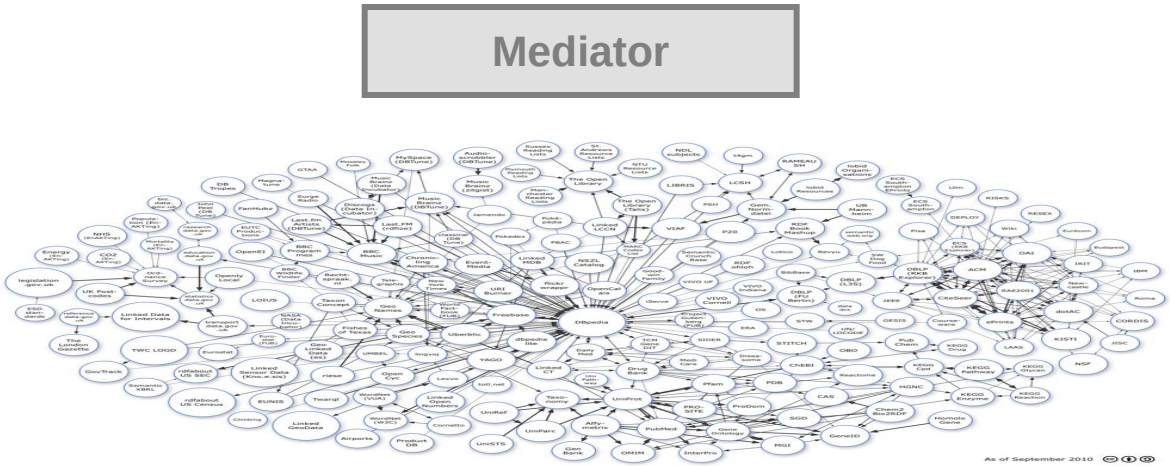
Web of Data at the INCT for Web Science

- Research Goals
  - Develop SPARQL query mediators, including
    - Runtime optimization
    - Semantic optimization, using ontology constraints
    - Post-processing optimization, including data de-duplication and isolation of data inconsistencies
  - Develop methods and tools to
    - optimize exploratory SPARQL queries
    - optimize SPARQL queries over linked data stored in the Cloud
    - design linked data mashups
  - Develop large scale experiments

Consumption of Linked Data

Web of Data at the INCT for Web Science

- SPARQL query mediator
  - Challenges
  - Mediator – major features

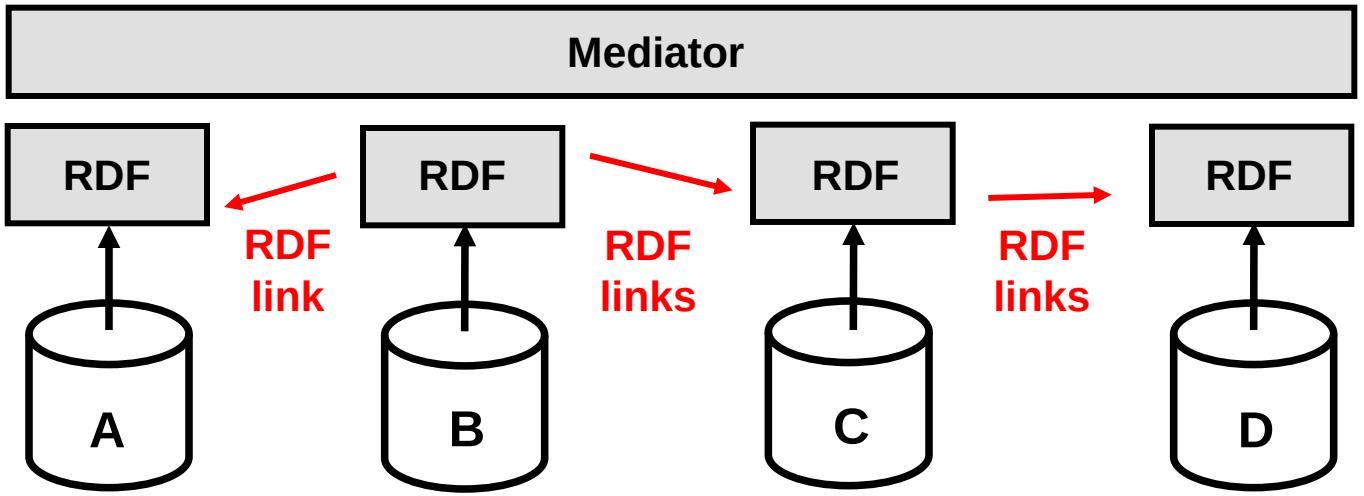


VIDAL, V.M.P., MACEDO, J.A.F., PINHEIRO, J.C., CASANOVA, M.A., PORTO, F.A.M. Query Processing in a Mediator Based Framework for Linked Data Integration. *International Journal of Business Data Communications and Networking* (IJBDCN), v.7, p.29 - 47, 2011.

SPARQL Query Mediator

Consumption of Linked Data

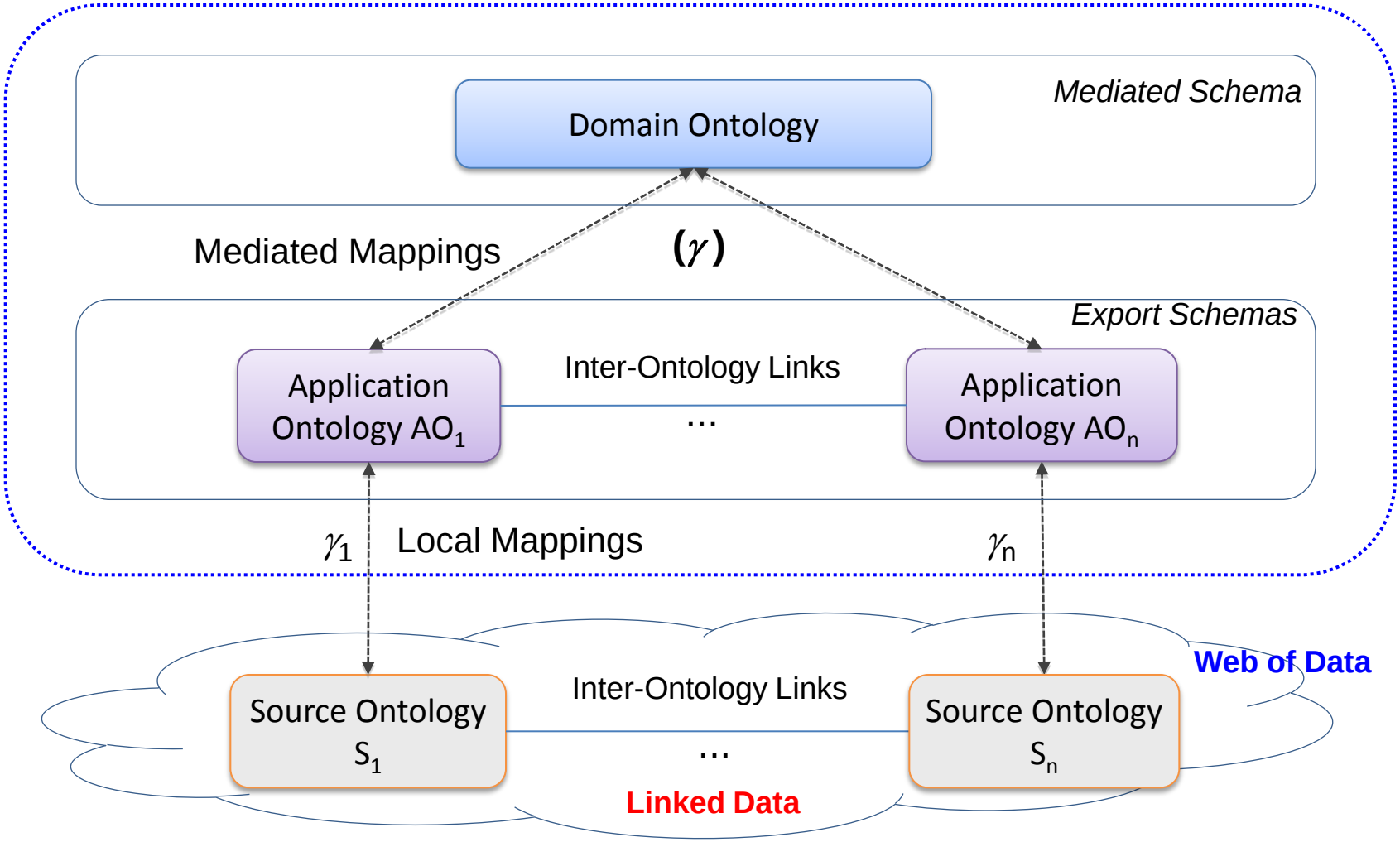
- Challenges
  - How to “complete” data to improve query results
  - How to process large volumes of linked data
    - Distributed, unreliable sources
    - Heterogeneous vocabularies

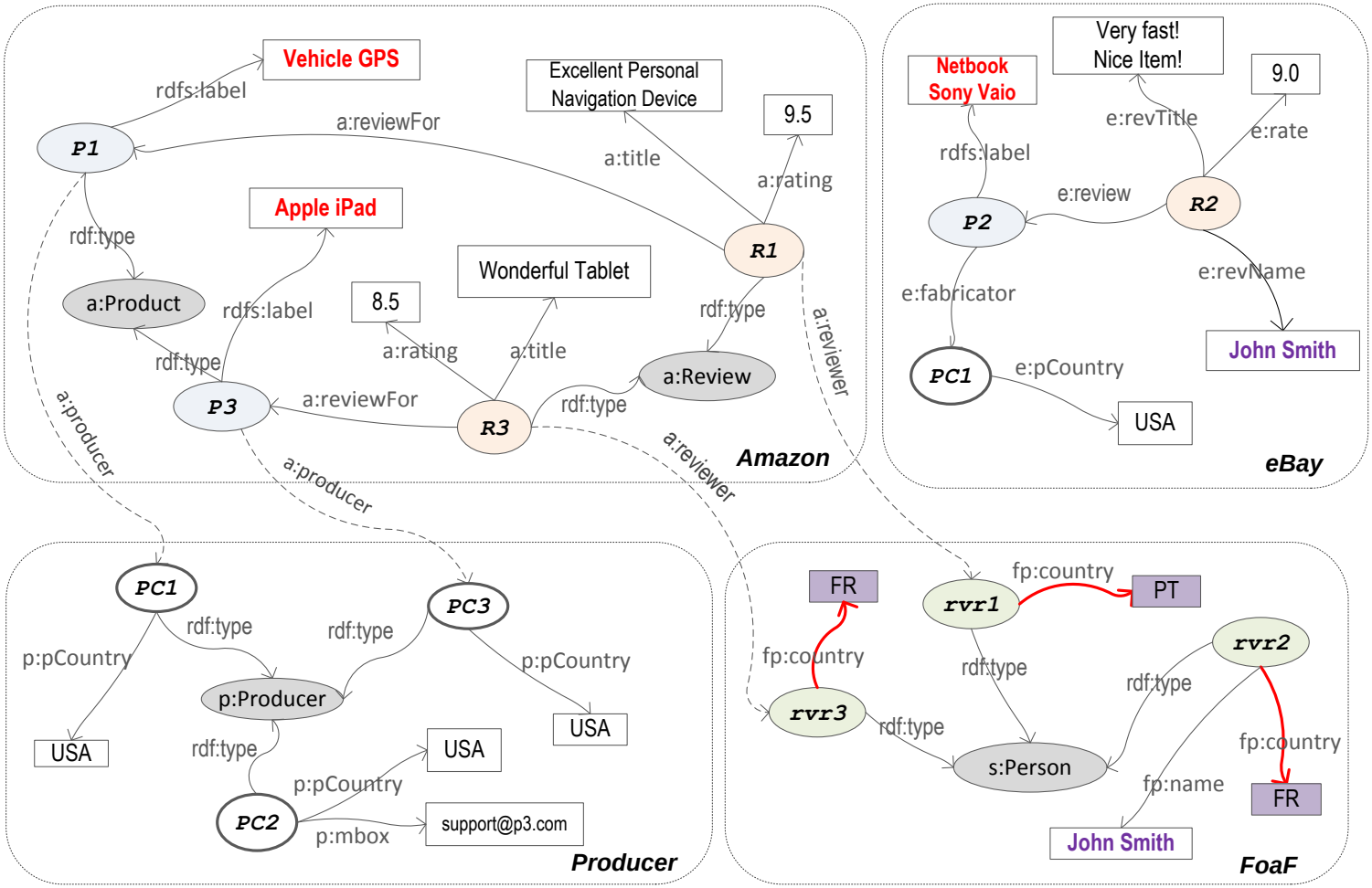


SPARQL Query Mediator

Consumption of Linked Data

- Mediator - major features
  - Query reformulation algorithm
    - Uses inter-ontology links at compile-time
  - Join algorithm
    - Accounts for SPARQL endpoint variability
    - Set-Bind-Join for Linked Data + Pipelining Hash-Join
    - Adaptive strategy

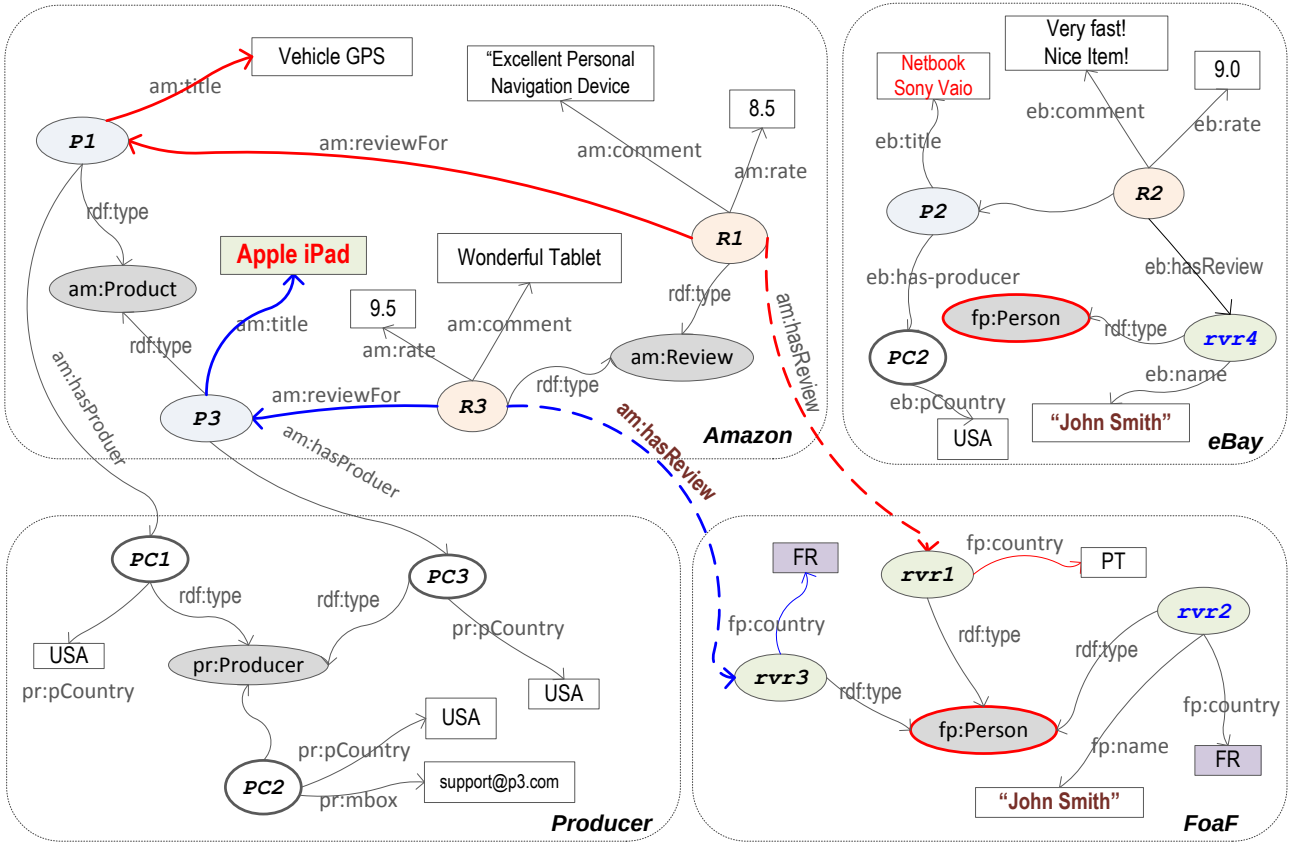

Mediated Schema
Domain Ontology
Mediated Mappings
(γ)
Export Schemas
Application Ontology AO1
Inter-Ontology Links
…
Application Ontology AOn
γ1
Local Mappings
γn
Web of Data
Source Ontology S1
Inter-Ontology Links
…
Source Ontology Sn
Linked Data


Vehicle GPS
rdfs:label
Excellent Personal Navigation Device
9.5
a:reviewFor
P1
a:title
a:rating
Apple iPad
rdf:type
R1
Wonderful Tablet
a:Product
rdfs:label
8.5
rdf:type
a:rating
a:title
rdf:type
a:Review
P3
a:reviewFor
R3
rdf:type
a:reviewer
a:producer
a:producer
a:reviewer
Amazon
Very fast!
Nice Item!
Netbook
Sony Vaio
9.0
e:revTitle
e:rate
rdfs:label
e:review
P2
R2
e:fabricator
e:revName
John Smith
PC1
e:pCountry
USA
eBay
PC1
rdf:type
rdf:type
PC3
p:pCountry
p:pCountry
p:Producer
USA
USA
rdf:type
USA
p:pCountry
PC2
p:mbox
support@p3.com
Producer
FR
fp:country
rvr1
PT
fp:country
rvr2
rdf:type
rdf:type
rvr3
rdf:type
fp:country
s:Person
fp:name
FR
John Smith
FoaF

```
SELECT ?t
WHERE  {
   ?r rdf:type s:Review .
   ?r s:hasReview ?rvr .
   ?rvr s:country ?ct .
   FILTER(?ct, 'FR')
   ?r s:reviewFor ?p .
   ?p s:title ?t . }
```

Query returns only

"**Apple iPad**"



Query returns only

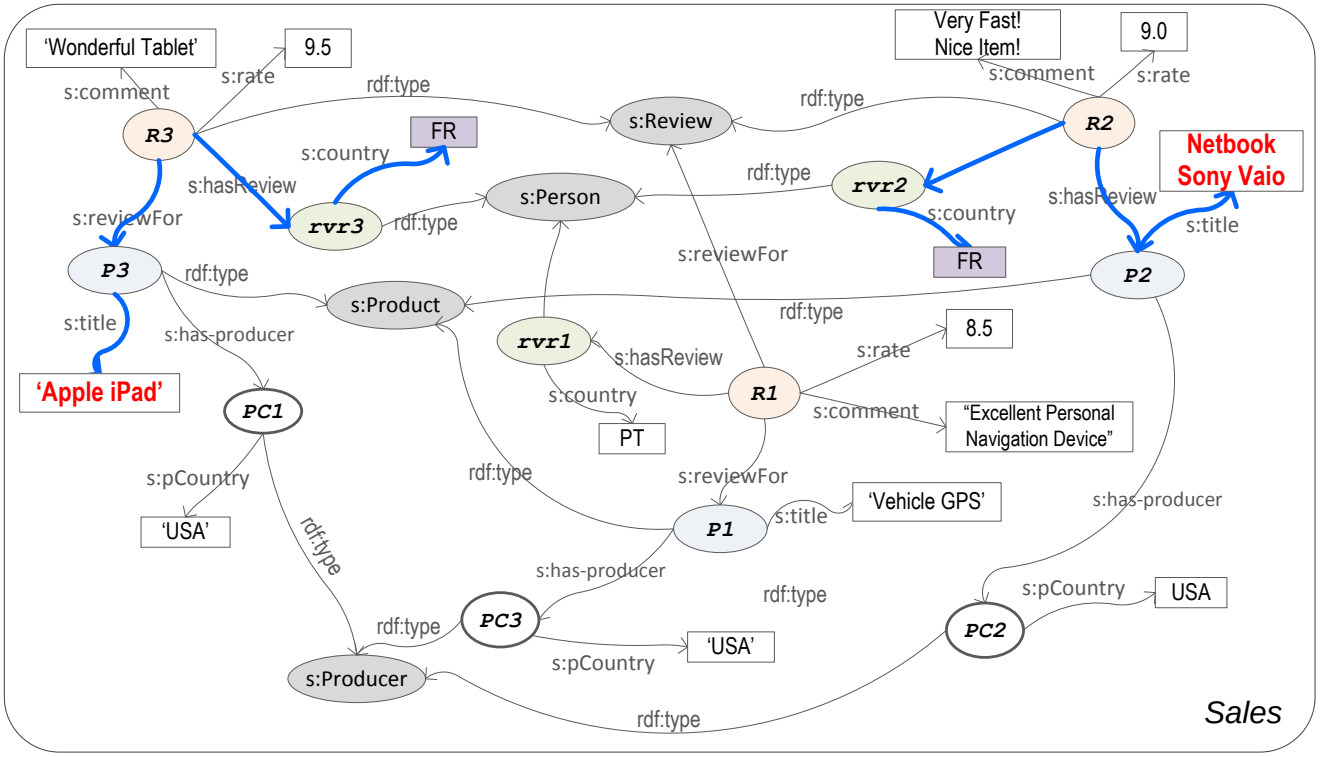
"**Apple iPad**"

same-as rule:
allows to infer
*same-as(**rvr4**, **rvr2**)*

same-as(rvr4, rvr2) ⇐
eb:Person(rvr4), eb:name(rvr4, "John Smith"),
fp:Person(rvr2), fp:name(rvr2, "John Smith")

Query now returns
“*Apple iPad*”
“*Netbook Sony Vaio*”

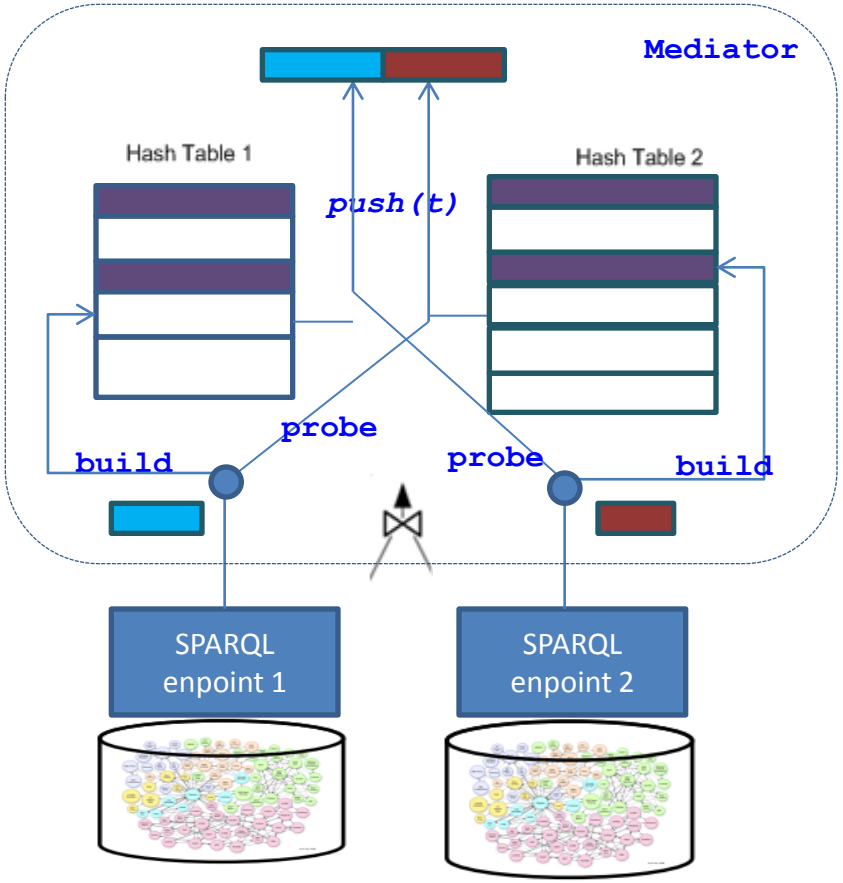
same-as rules:
allow the query rewriting process to complete the data



SPARQL Query Mediator

Consumption of Linked Data

- Pipelining Hash Join (PHJ)
  - Uses pipelining to achieve a high degree of parallelism
  - Hash tables constructed in parallel from tuples obtained from both sources
- Relevant feature
  - Non-blocking algorithm



SPARQL Query Mediator

Consumption of Linked Data

- Set-bind-join
  - Variation of bind-join
  - Uses semi-joins to reduce the data volume transferred from SPARQL endpoints to the mediator
  - (Example)

Step 1. Scan

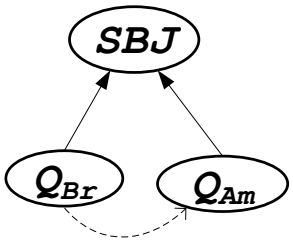
Sales (s)

```
SELECT ?t,  ?r
WHERE {
   ?b rdf:type s:Book .
   ?b s:title ?t .
   ?b br:rate ?r  . FILTER (?r >= 7.0)  }
```

Retrieve the title and rating of books with rating greater than 7

BookReview (br)

```
SELECT ?isbn , ?r
 WHERE {
     ?b br:isbn ?isbn .
     ?b br:rate ?r  FILTER (?r >= 7.0) . }
```



Amazon (am)

```
SELECT ?t,  ?isbn
WHERE {
  ?b rdf:type am:Book .
   ?b am:title ?t .
   ?b am:isbn ?isbn .  }
```

$isbn_i$	r_i
11	9.0
14	8.0
15	7.5
17	8.5

Set

Endpoint BookReview

?isbn	?r
11	9.0
13	5.0
14	8.0
15	7.5
16	5.0
17	8.5

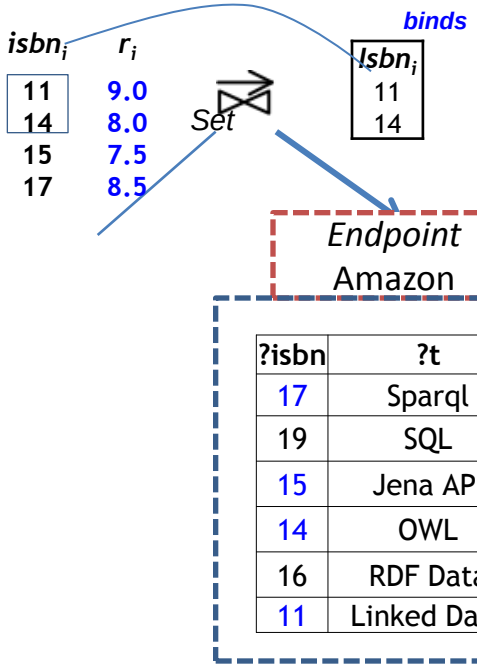
Endpoint Amazon

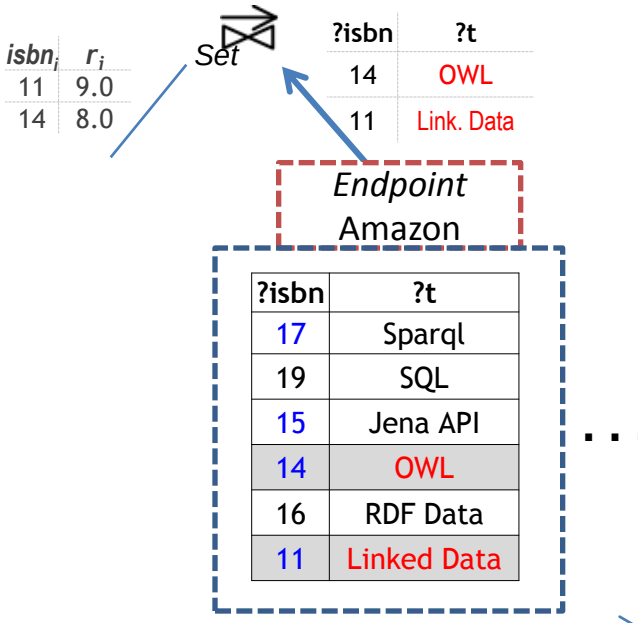
?isbn	?t
17	Sparql
19	SQL
15	Jena API
14	OWL
16	RDF Data
11	Linked Data

Step 2. Set-binds

Amazon (am)

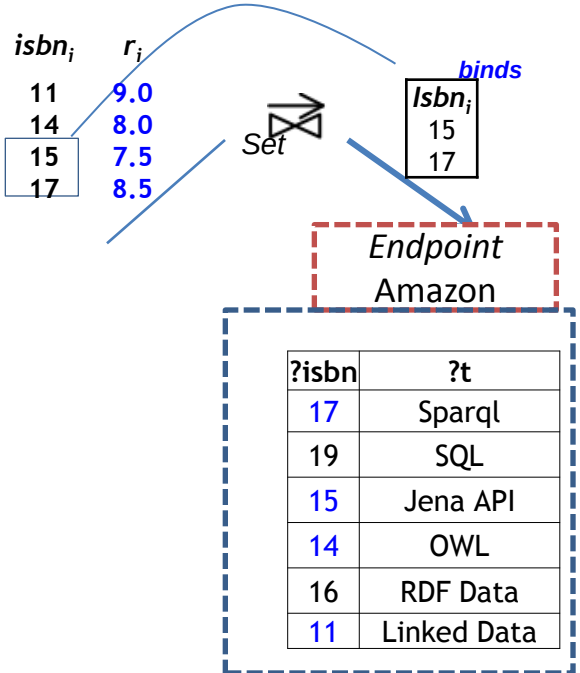
```
SELECT ?t, ?isbn
FROM http://amazon
WHERE {
  ?b rdf:type am:Book .
  ?b am:title ?t .
  ?b am:isbn ?isbn . }
BINDINGS ?isbn
{ ('11' e '14') }
```

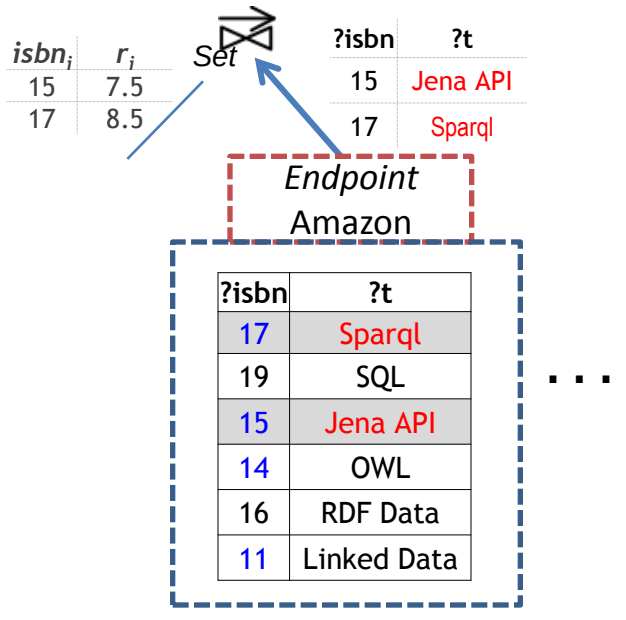




Amazon (am)

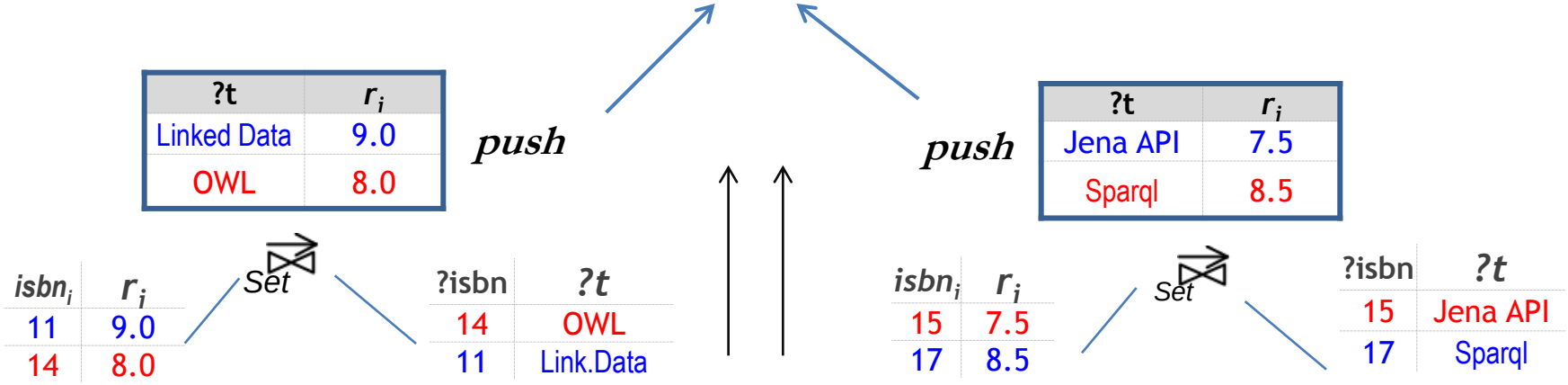
```
SELECT ?t, ?isbn
FROM http://amazon
WHERE {
  ?b rdf:type am:Book .
  ?b am:title ?t .
  ?b am:isbn ?isbn . }
BINDINGS ?isbn
{ ('15' e '17') }
```





Step 3. Join at the mediator

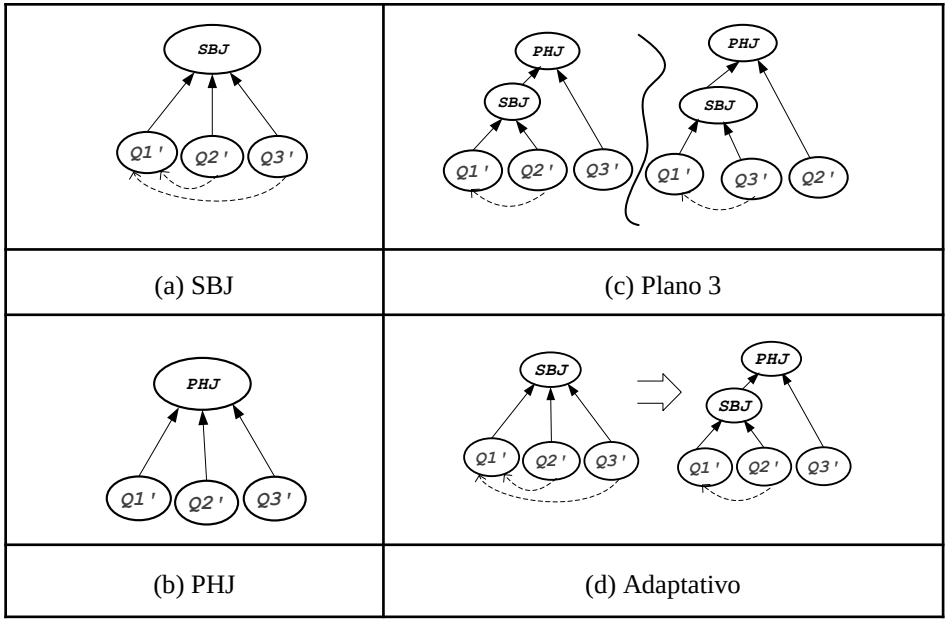
Required only if the subquery that retrieves the *binds* contributes with other values to the query answer

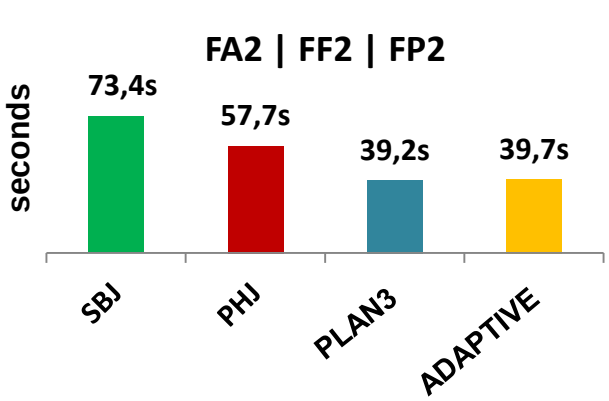


SPARQL Query Mediator

Consumption of Linked Data

- Adaptive strategy
  - Alternates the join strategy, at runtime, between SBJ and PHJ
  - Early results suggest to improve the adaptive strategy to deal with fault tolerance problems





SPARQL Query Mediator

Consumption of Linked Data

- Lessons learned:
  - The use of “sameAs” may explode query rewriting

    ?b prop ?d
    +
    ?a sameAs ?b
    ?d sameAs ?c

    (?b prop ?d or ?a prop ?d or
    ?b prop ?c or ?a prop ?c)

  - Linked Data sources are highly unreliable!

Consumption of Linked Data

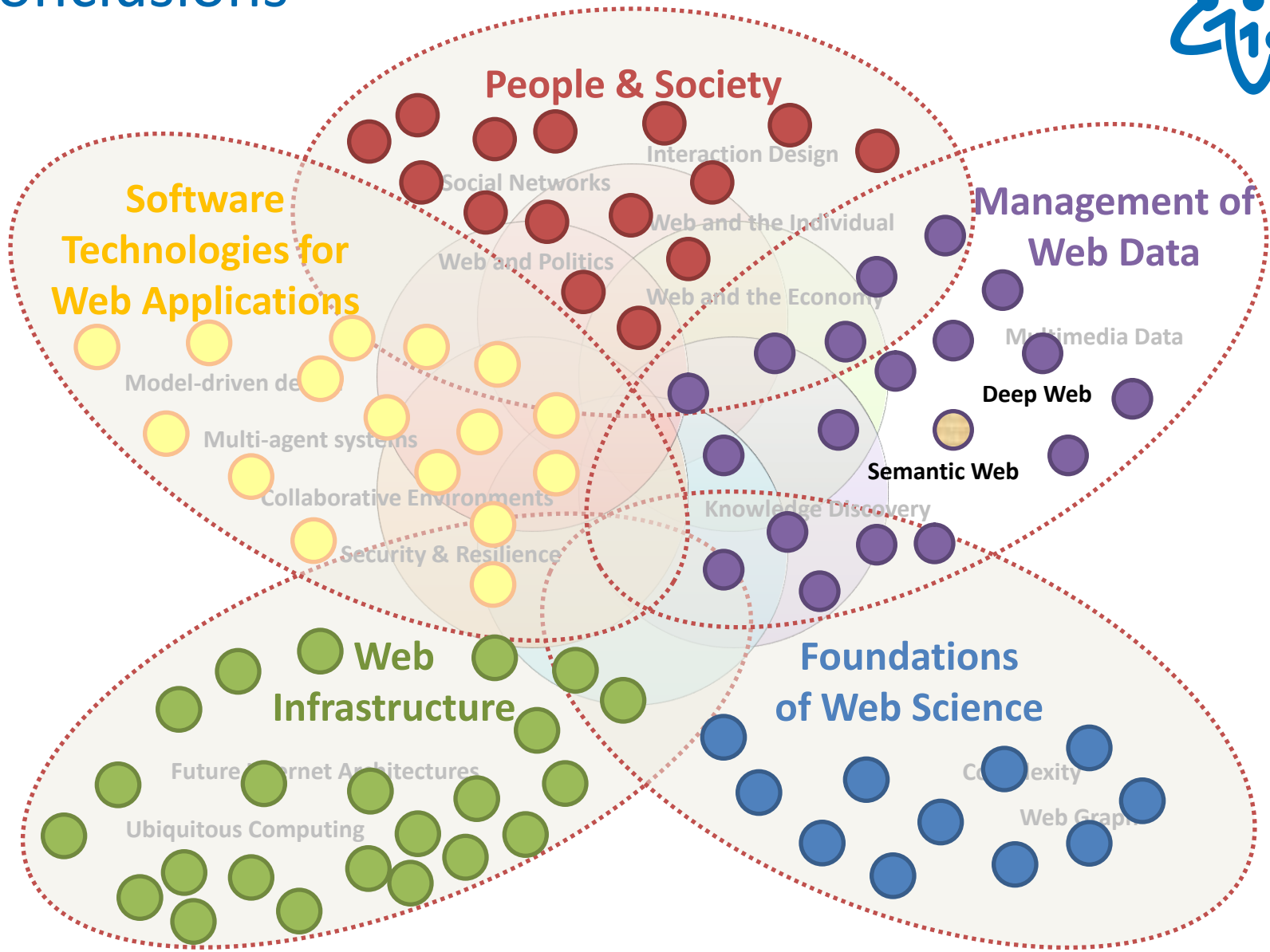
Web of Data at the INCT for Web Science

- Research Goals
  - Develop SPARQL query mediators, including
    - Runtime optimization
    - Semantic optimization, using ontology constraints
    - Post-processing optimization, including data de-duplication and isolation of data inconsistencies
- Results
  - Query mediator
    - Uses inter-ontology links at compilation time
    - Features runtime optimization

Topics

- INCT for Web Science
- Web of Data
- Web of Data at the INCT for Web Science
- Conclusions

Conclusions



WEB SCIENCE BRASIL

Brazilian Institute for Web Science Research

Contacts

Web of Data Group

A.L. Furtado, M.A. Casanova, K. Breitman

V.M.P. Vidal, J.A.F. Macedo

J. Viterbo F., L.A.P.P. Leme

INCT Web Science

webscience@inf.puc-rio.br

www.webscience.org.br

www.webscience.org.br/wiki

CNPq

Conselho Nacional de Desenvolvimento Científico e Tecnológico

FAPERJ

Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro