Semantic Web Activity @ W3C

Ivan Herman

W3C Workshop on Multilingual Web

Luxembourg, March 15, 2012

Before going into details…

# What does the term "Semantic Web" mean to people?

衆盲
探象之圖

For some people, Semantic (Web) is…

- ▸ An intelligent system manipulating and analyzing knowledge bases
  - ▪ e.g., via big ontologies, vocabularies
- ▸ A means to manage large amount of data
- ▸ Improve search by adding structure to embedded data
- ▸ A means to *integrate* many different pieces of data
- ▸ And a mixture of all these…

Example: OWL as a knowledge base

Courtesy of Songmao Zhang, Academy of Mathematics and System Sciences, China
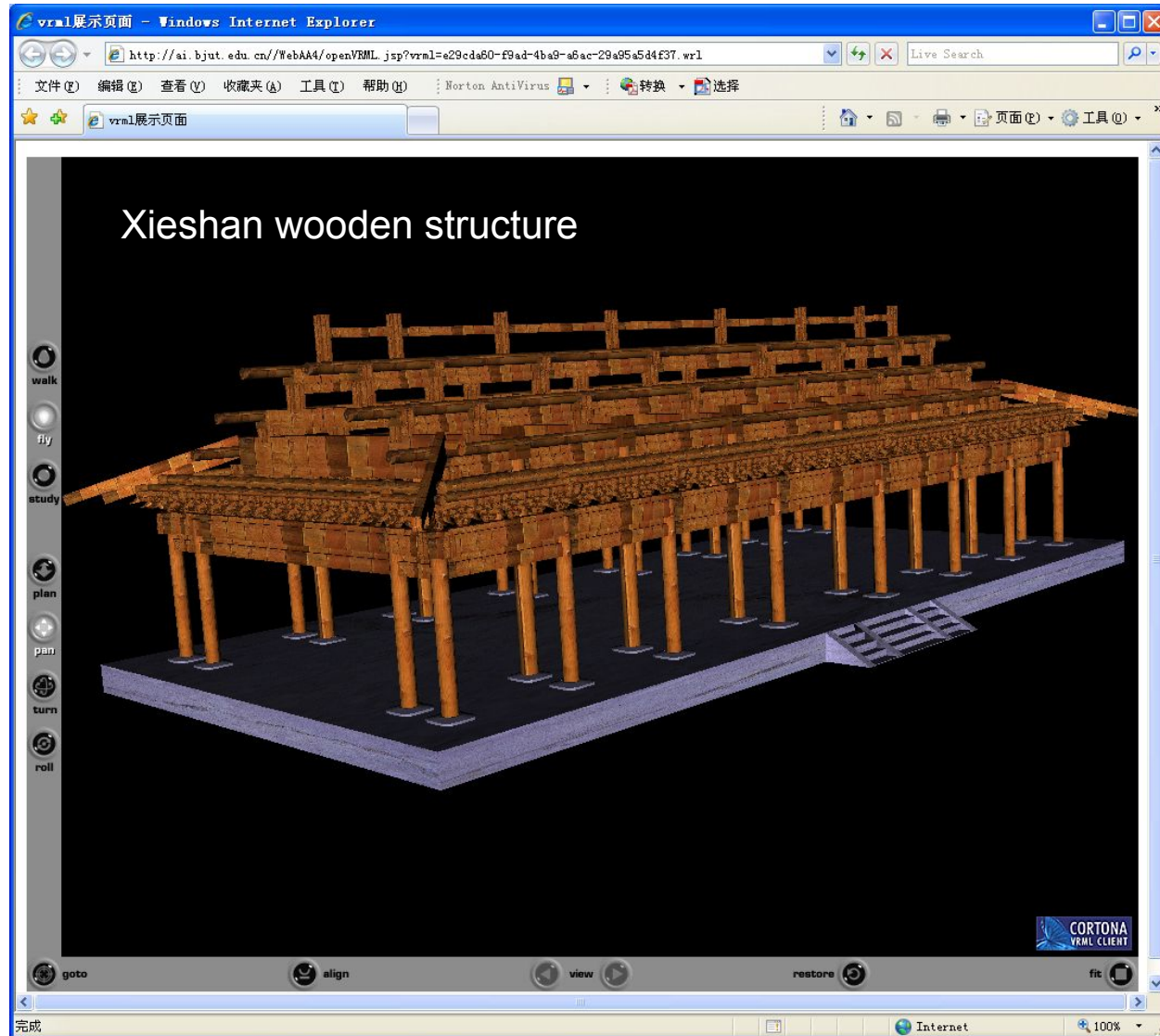
Example: making use of major ontologies

▸ Help in finding the best drug regimen for a specific patient

Courtesy of Erick Von Schweber, PharmaSURVEYOR Inc., (SWEO_Use Case)

Example: making use of linked data

Courtesy of the BBC

Example: making use of linked data

Courtesy of the BBC

Example: making use of structured data and search engine facilities

Example: making use of structured data and search engine facilities

Example: making use of structured data and search engine facilities

And that is all right!

- ▸ We have to acknowledge that the field has grown and has become multi-faceted
- ▸ All different "views" have their success stories
- ▸ There are also no clear and water-proof boundaries between the different views
- ▸ The question is: where is the *emphasis*?

Data on the Web

▸ There are more and more data on the Web

 ▪ government data, health related data, general knowledge, company information, flight information, restaurants,…

▸ More and more applications rely on the availability of that data

But: we do not want that!

Imagine…

‣ A "Web" where
  - documents are available for download on the Internet
  - but there would be no hyperlinks among them

CoCoDat - Collation of Cortical Data - Mozilla Firefox

File   Edit   View   History   Bookmarks   Tools   Help

http://www.coco

CoCoMac   DATABASES   ORT   EXAMPLES

CoCoDat: Collation of Cortical [s microcircuitry] Data

CoCoDat is a microcircuitry database that
published experimental reports. The data
and cellular compartment), as well as the

- Morphology
- Firing properties
- Ionic currents
- Ionic conductances
- Synaptic currents
- Connectivity

The database is available for download u
data tables but also a Search Board with p
manual or automatic relaxation of the sea

- Brain region
- Layer
- Neuron type

http://www.cocomac.org/cocodat/catalyzer/ind

Cell Centered Database - Mozilla Firefox

File   Edit   View   History   Bookmarks   Tools   Help

http://ccdb.ucsd.edu/sand/main?event=gallery&action=show&dpl=y   red database

Cell Centered Database™
National Center for Microscopy and Imaging Research
Gallery

Data | Search | Gallery | Dictionary | Publications | MyCCDB | Data Download | Contact us | Help

2D image   Reconstruction   Segmentation   Animation

py 23
D:
ge 071202tjd
e:
: P1188
ct Brain Maps
: Tom Deerinck
nt BioRad RTS 2000MP
d: Multiphoton
n: montage of cerebellar
cortex triple labeled for
Hoescht stain (blue), IP3
receptor (green) and
GFAP (red)

Z ● ◐ ⊕ PIP logged out

NeuronDB = Thalamic relay neuron - Overview (A) () - Mozilla Firefox

File   Edit   View   History   Bookmarks   Tools   Help

http://senselab.med.yale.edu   senselab

NeuronDB

Thalamic relay neuron

Back

Mode:	Overview Data/Search plus Connectivity plus Classical References/Notes Models
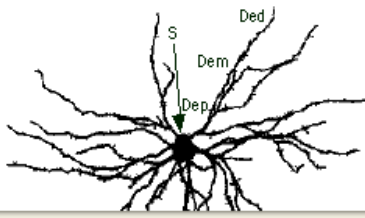
Region:	Distal equivalent dendrite Middle equivalent dendrite Proximal equivalent dendrite Soma Axon hillock Axon fiber Axon terminal All Compartments
Properties:	Receptors Channels Transmitters All Properties
Interoperation:	Gene and Chromosome Experimental Data (neurodatabase.org) Microscopy Data (CCDB)
Neuron type:	principal
Organism:	Vertebrates

Ded
S
Dem
Dep

1. Equivalent dendrite   Show other
2. Distal equivalent dendrite   Show other
3. Middle equivalent dendrite   Show other
4. Proximal equivalent dendrite   Show other
5. Soma   Show other

Done   Z ● ◐ ⊕ PIP logged out

Data on the Web is not enough…

▸ We need a proper infrastructure for a real *Web of Data*

- data is available on the Web
  - accessible via standard Web technologies
- data are *interlinked over the W*eb
  - the terms used for linkage are well defined

▸ I.e.: data can be *integrated* over the Web

This is what we want!

Photo credit "kxlly", Flickr

Semantic Web technologies should be at the service of such a Web of Data

Web of Data Applications

Stand Alone Applications

Browser Applications

Query

Inferencing

Common Format & Common Vocabularies

"Bridges"

Data on the Web
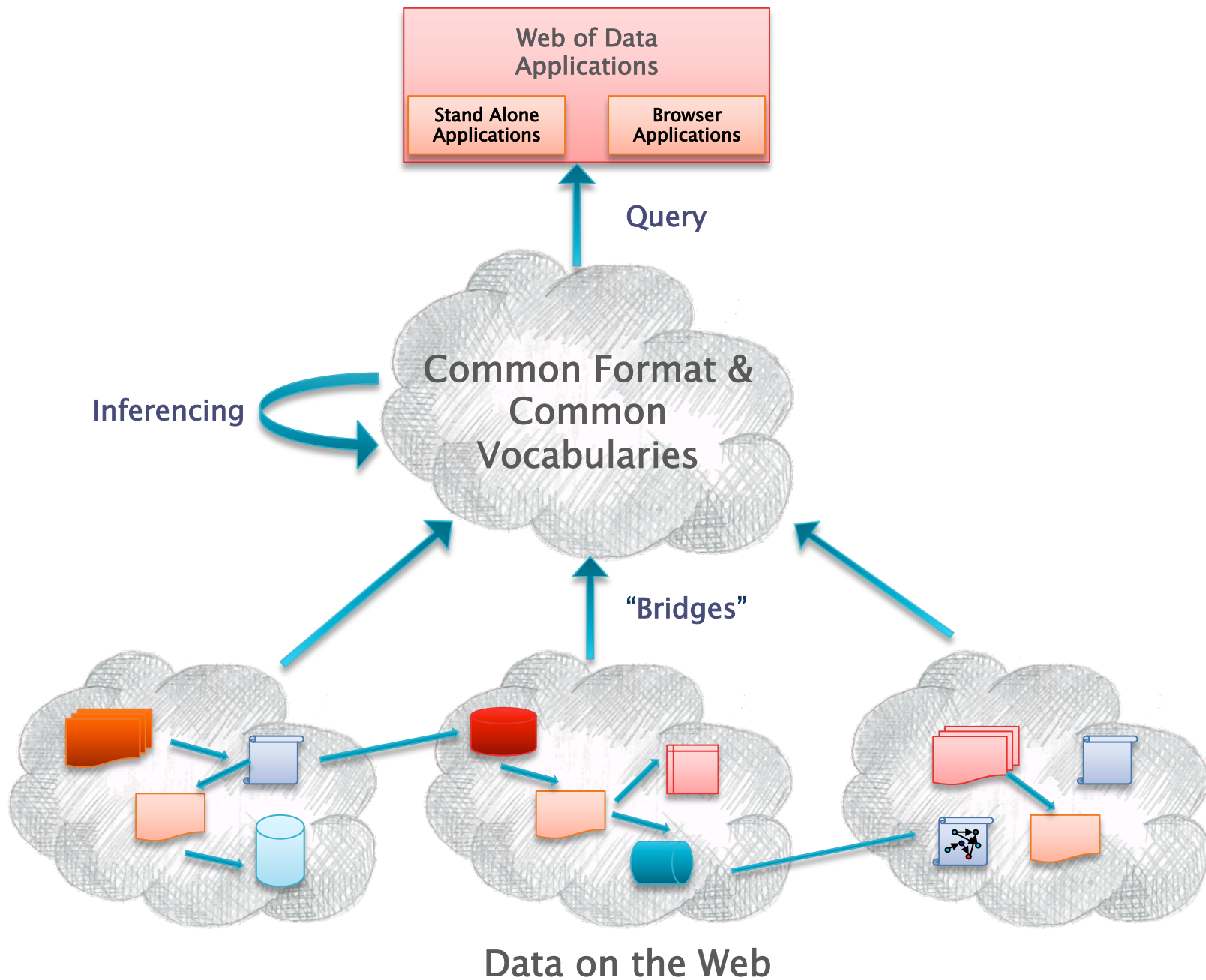
On a longer term…

- ▸ Data should be easily manipulated from an application
- ▸ Some sort of a "globally linked database"
- ▸ But this is still further down the road…

Web of Data Applications

Stand Alone Applications

Browser Applications

Query and Update

Inferencing

Common Format & Common Vocabularies

"Bridges"

Data on the Web

http://www.w3.org/2001/sw/

2012 My Mercurial My Feedly TR LocalData Private ▼ Mailing lists ▼ Social ▼ SW ▼ Python ▼ RDFa it! ▼ Bookmarklets ▼ To... ▼ Wolfram

W3C
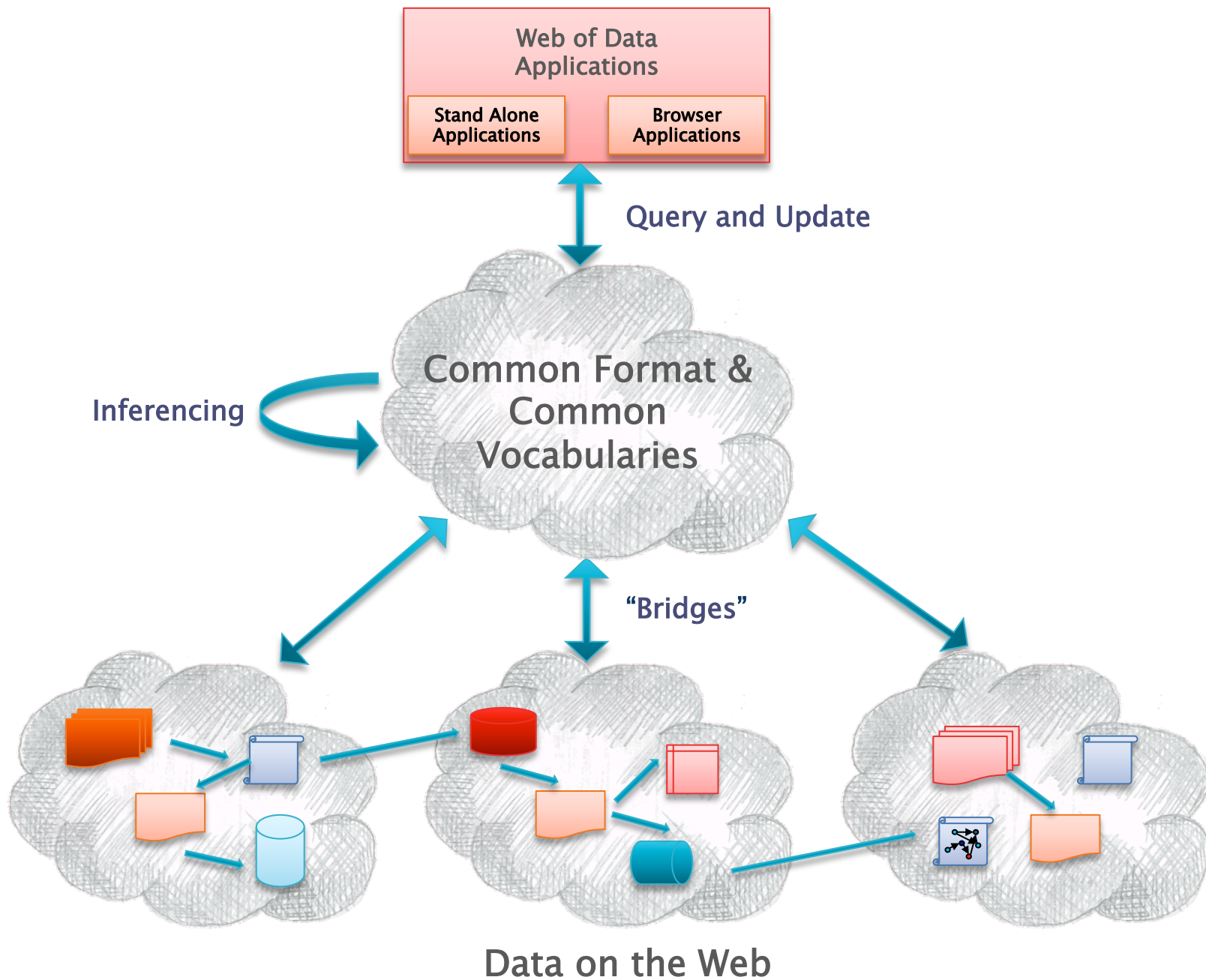
Views: desktop mobile print

STANDARDS PARTICIPATE MEMBERSHIP ABOUT W3C

W3C SEMANTIC WEB ACTIVITY

On this page → publications, interviews • presentations • active groups • completed groups • past groups

So… what is happening at W3C?

NEARBY

Activity news

Specifications

FAQ

Use Cases and Case Studies

List of Tools

On-line validators

List of Books

Semantic Web Logos and Buttons

SW Wiki

ACTIVITY RSS FEEDS

Activity news (RSS 1.0)

What is the Semantic Web?

The **Semantic Web** provides a common framework that allows **data** to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF). See also the separate FAQ for further information.

The Semantic Web is a web of data. There is lots of data we all use every day, and it is not part of the web. I can see my bank statements on the web, and my photographs, and I can see my appointments in a calendar. But can I see my photos in a calendar to see what I was doing when I took them? Can I see bank statement lines in a calendar?

Why not? Because we don't have a web of data. Because data is controlled by applications, and each application keeps it to itself.

The Semantic Web is about two things. It is about common formats for integration and combination of data drawn from diverse sources, where on the original Web mainly concentrated on the interchange of documents. It is also about language for recording how the data relates to real world objects. That allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing.

The (almost) past

▸ Some technologies are in the process of finalization
- SPARQL 1.1 (SPARQL Protocol and RDF Query Language)
- RDB2RDF (Relational Databases to RDF)
- RDFa 1.1 (RDF in attributes)

The present

▸ Some areas are subject of intensive work
  ▪ RDF update (Resource Description Framework)
  ▪ Provenance

The future

▸ We are discussing new works, new areas, e.g.,

- Linked Data Platform
- Access Control issues
- Constraint checking on Semantic Web data
- …

Link to specialized communities

▸ Various communities have different emphasis on which part of the Semantic Web they want to use

▸ W3C has contacts with some of those

- health care and life sciences (a separate IG is up and running)
- libraries, publishing
- financials
- the oil, gas, and chemicals community

Query RDF: SPARQL 1.1

Reminder…

▸ *SPARQL is a query language on RDF data*

▸ SPARQL is defined in terms of a protocol, to send query and results over the Web

▸ Is based on the idea of "graph pattern matching":

1. a graph pattern is described in the query, with real and unknown nodes ("variables")

2. if the pattern can match a portion of the graph, the unknown nodes are replaced by the "real" ones

3. resulting information is returned

▸ First version of SPARQL was published in 2008

SPARQL 1.1: adding missing features to SPARQL

▸ Nested queries (i.e., **SELECT** within a **WHERE** clause)

▸ Negation (**MINUS**, and a **NOT EXIST** filter)

▸ Aggregate function on search results (**SUM**, **MIN**,…)

▸ Property path expression (**?x foaf:knows+ ?y**)

▸ SPARQL UPDATE facilities (**INSERT**, **DELETE**, **CREATE**)

▸ Combination with entailment regimes

SPARQL 1.1 and RDFS/OWL/RIF
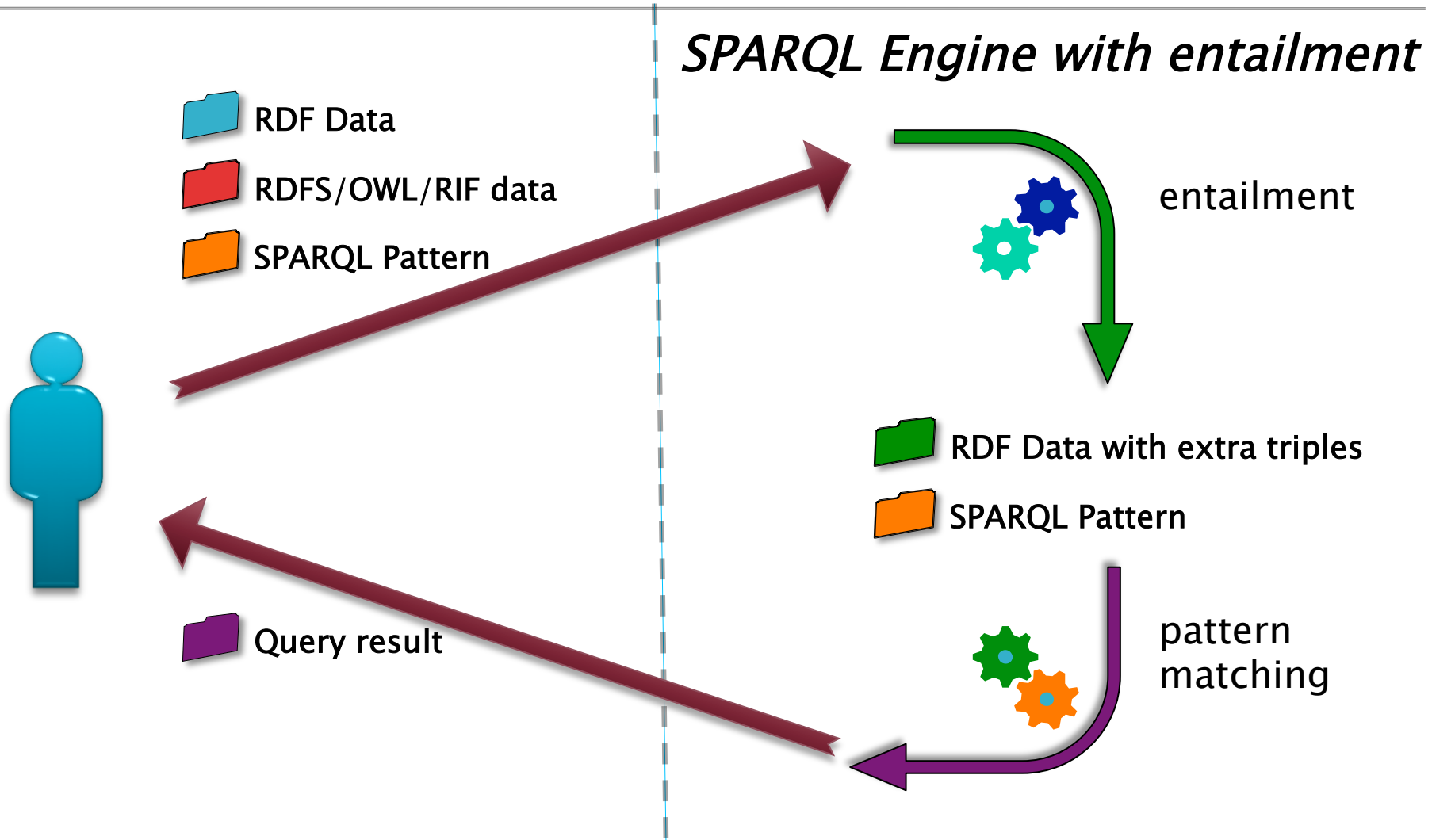
SPARQL as a unifying point
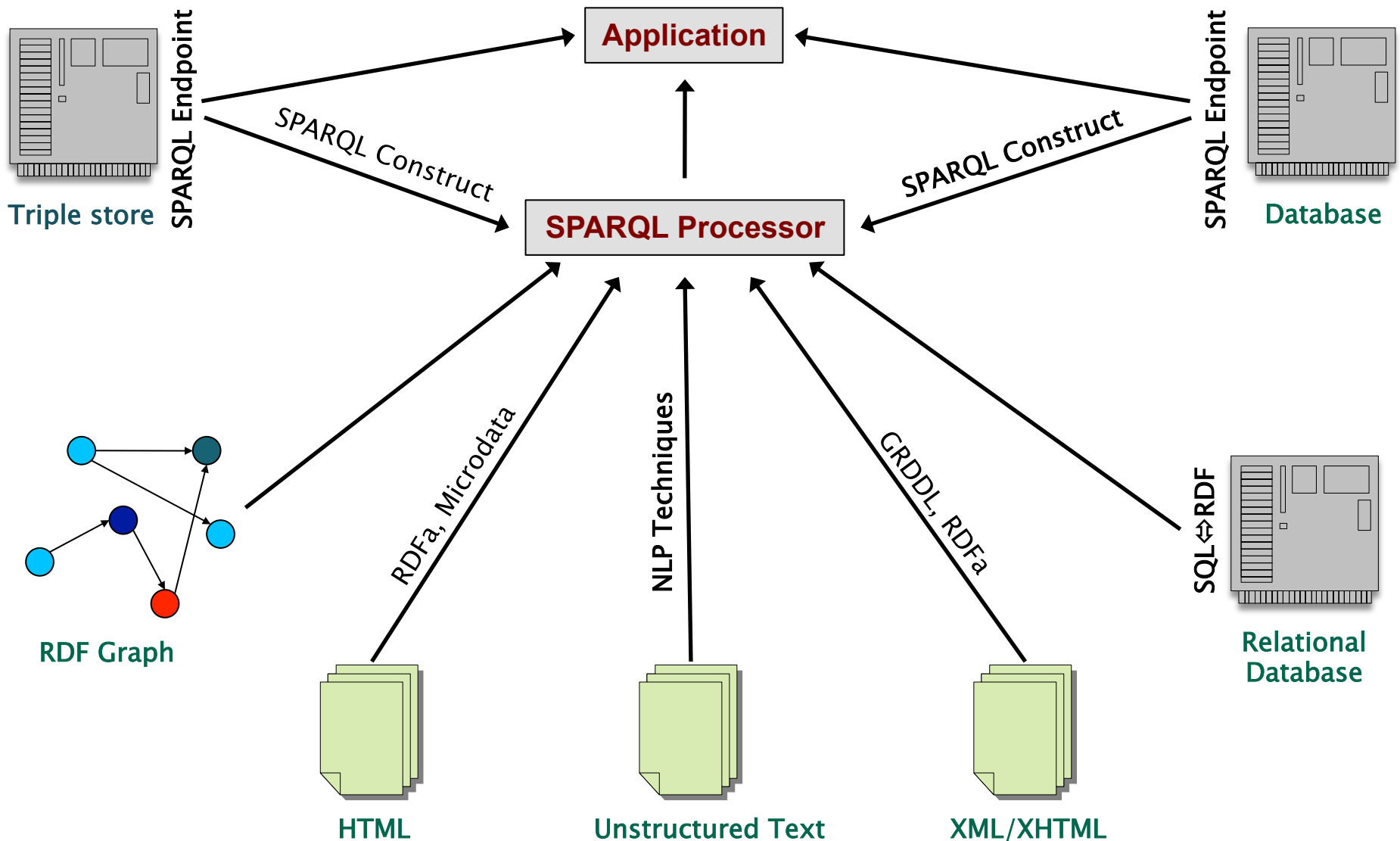
SPARQL 1.1 as a unifying point

Application

SPARQL Processor

SPARQL Endpoint

Triple store

Inferencing

SPARQL Construct
SPARQL Update

SPARQL Construct
SPARQL Update

SPARQL Endpoint

Database

RDF Graph

RDFa, Microdata

NLP Techniques

GRDDL, RDFa

SQL⇔RDF

HTML

Unstructured Text

XML/XHTML

Relational
Database

Inferencing

W3C Semantic Web
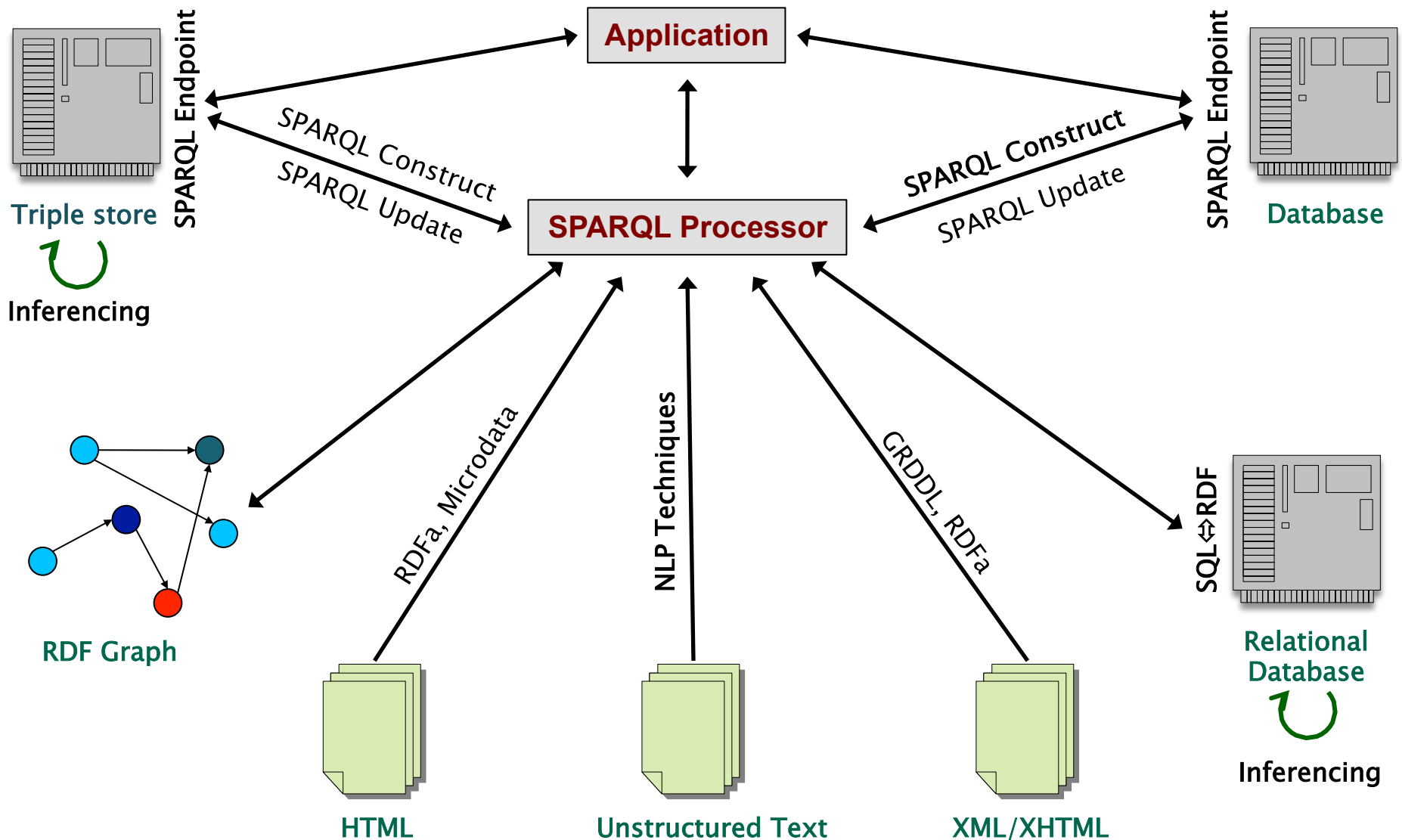
SPARQL 1.1 Status

▸ Technology has been finalized

▸ Goes to "candidate recommendation" soon

▸ Should be finished this summer

Access to Relational Databases

Relational Databases and RDF

▸ Most of the data on the Web is, in fact, in RDB-s

▸ Proven technology, huge systems, many vendors…

▸ Data integration on the Web must provide access to RDB-s

▸ *RDB2RDF provides means to "see" relational data as RDF*

RDF provides a common "view"

Common "view"
in
RDF

export
query

export
query
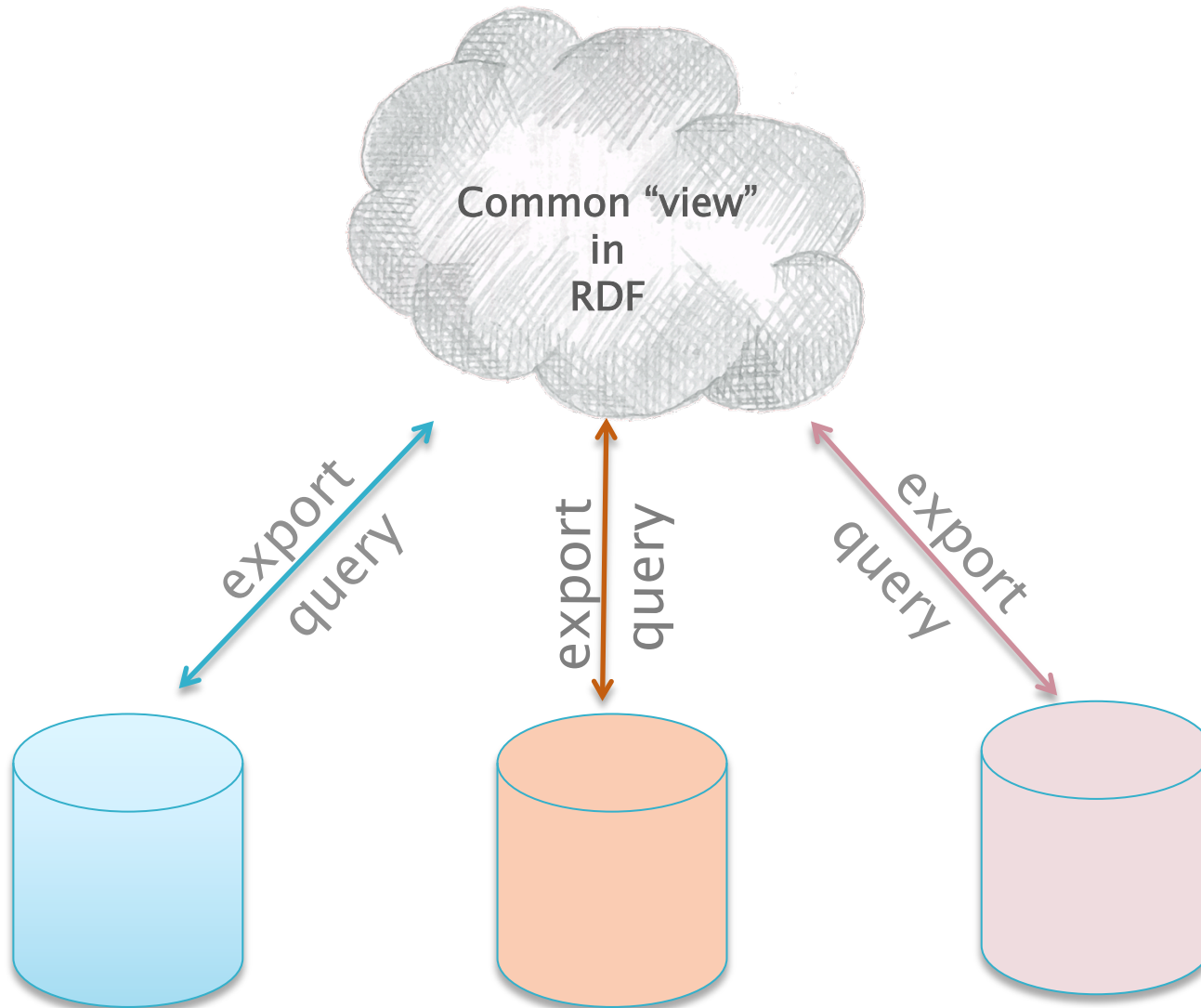
export
query

What is "export"?

- ▸ "Export" does not *necessarily* mean physical conversion

  - ▪ for very large databases a "duplication" would not be an option

  - ▪ systems may provide SPARQL⇔SQL "bridges" to make queries on the fly

- ▸ Result of export is a "logical" view of the RDB content

Simple export: Direct Mapping

‣ A canonical RDF "view" of RDB tables

‣ Only needs the information in the RDB Schema

Fundamental approach

Each column name provides a predicate

ISBN	Author	Title	Publisher	Year
0006511409X	id_xyz	The Glass Palace	id_qpr	2000
0007179871	id_xyz	The Hungry Tide	id_qpr	2004

Each row is a subject

Table references are URI objects

Cells are Literal objects

ID	Name	Homepage
id_xyz	Ghosh, Amitav	http://www.amitavghosh.com

Fundamental approach

ISBN	Author	Title	Publisher	Year
0006511409X	id_xyz	The Glass Palace	id_qpr	2000
0007179871	id_xyz	The Hungry Tide	id_qpr	2004

0006511409X ←⟨Book#ISBN⟩

⟨Book#Author⟩ → id_xyz

The Glass Palace ←⟨Book#Title⟩

2000 ←⟨Book#Year⟩

<Book/ISBN-0006511409>

<Book#ref-Author>

<Person/ID-id_xyz>

⟨Person#ID⟩ → id_xyz

ID	Name	Homepage
id_xyz	Ghosh, Amitav	http://www.amitavghosh.com

⟨Person#Name⟩

⟨Person#Homepage⟩

Ghosh, Amitav

http://www.amitavghosh.com

Tables

RDB
Schema
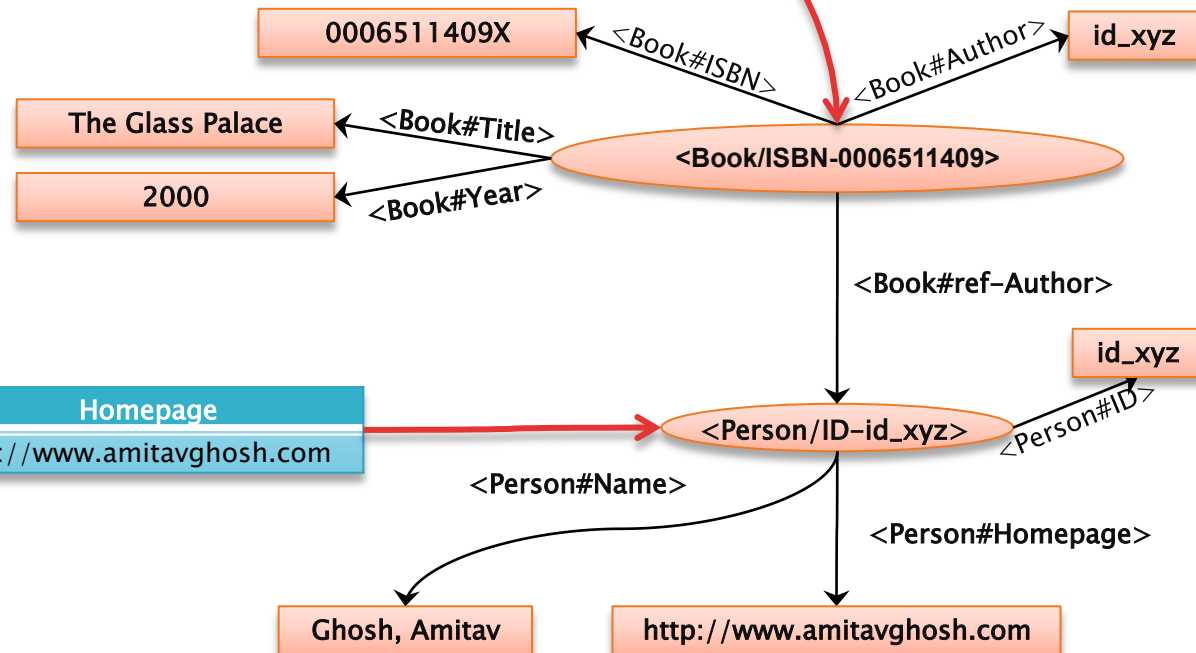
Direct
Mapping
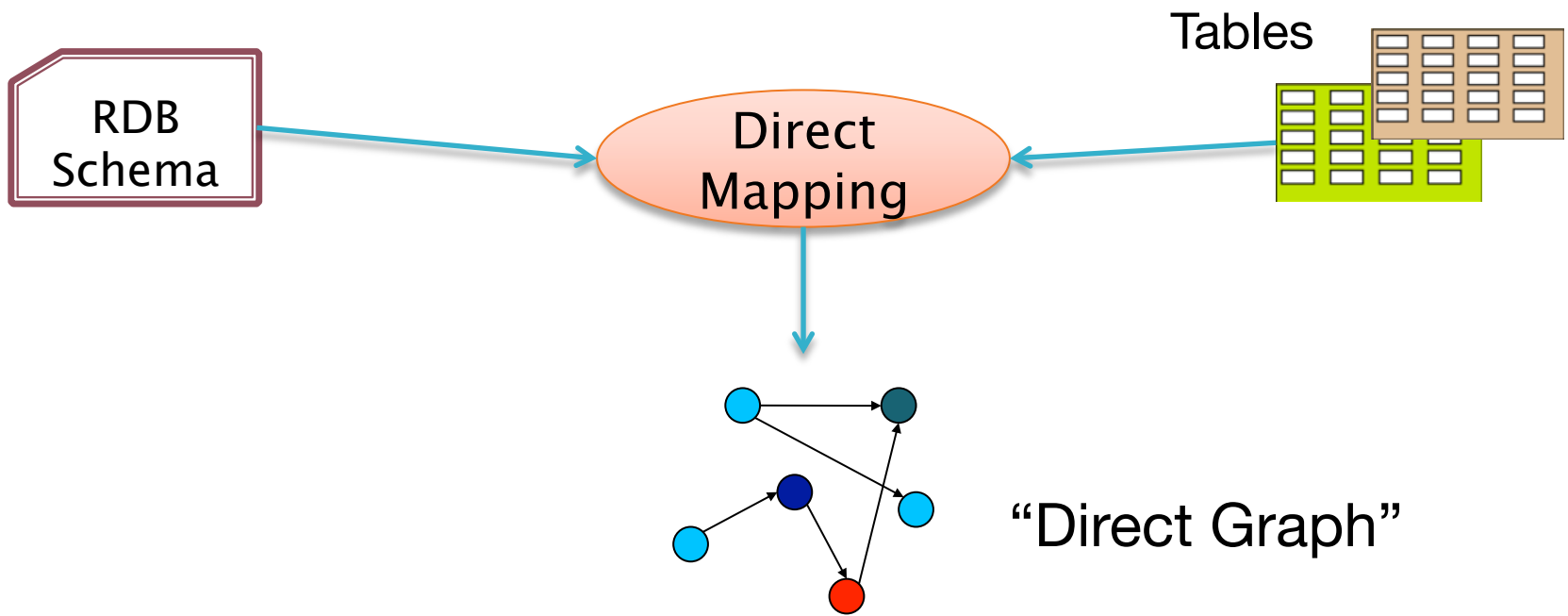
"Direct Graph"

Pros and cons of Direct Mapping

‣ Pros:

  ▪ Direct Mapping is simple, does not require any other concepts

  ▪ know the Schema ⇒ know the RDF graph structure

  ▪ know the RDF graph structure ⇒ good idea of the Schema(!)

‣ Cons:

  ▪ the resulting graph is not what the application really wants

RDB Schema

Tables

Direct Mapping

"Direct Graph"

Graph Processing
(Rules, SPARQL, …)

Final, Application Graph
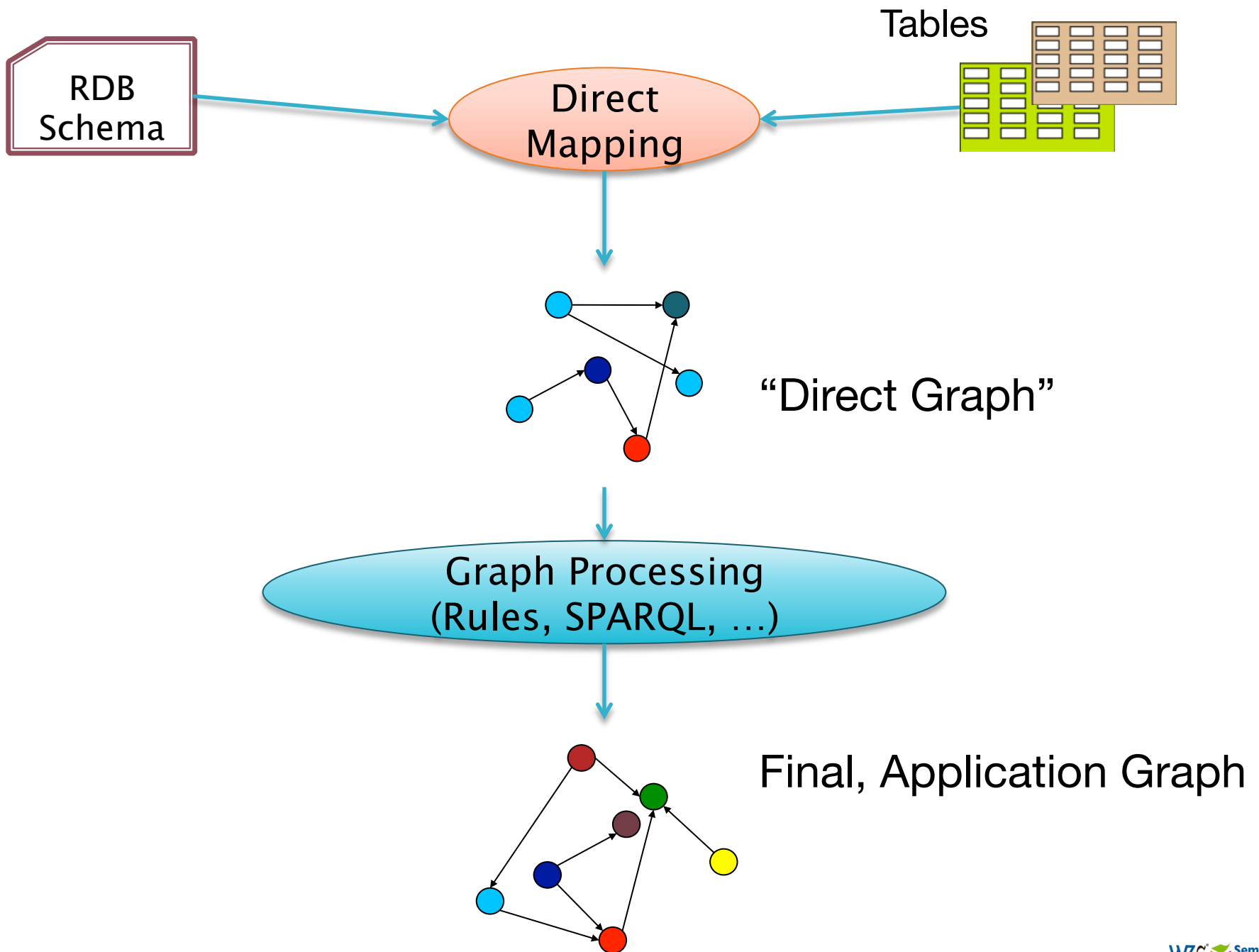
Beyond Direct Mapping: R2RML

‣ Separate vocabulary to control the details of the mapping, e.g.:

- finer control over the choice of the subject
- creation of URI references from cells
- predicates may be chosen from a vocabulary
- datatypes may be assigned
- etc.

‣ Gets to the final RDF graph with one processing step

RDB Schema

R2RML Instance

Tables

R2RML Mapping

Final, Application Graph
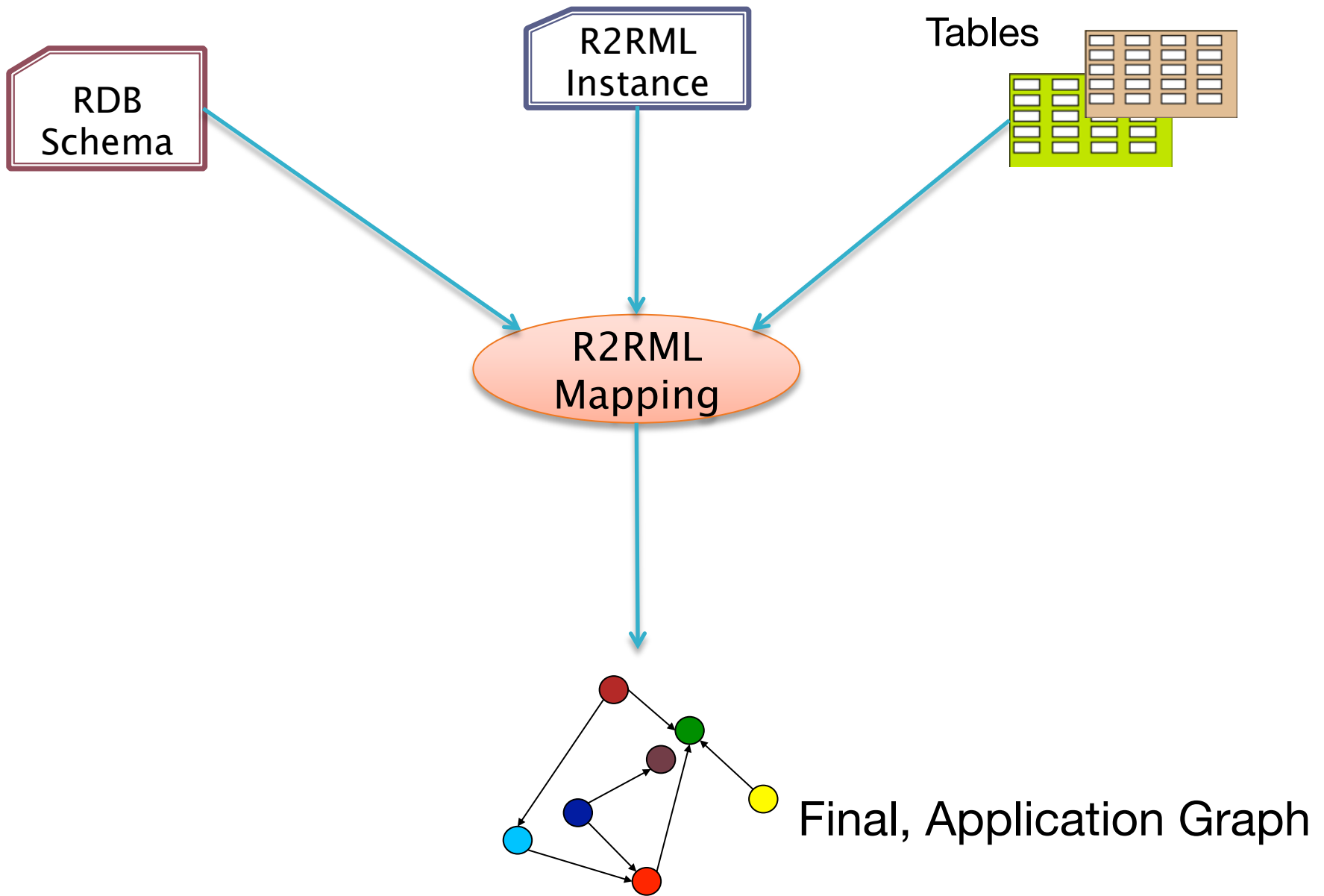
Relationships to the Direct Mapping

‣ Fundamentals are similar:

  ▪ each row is turned into a series of triples with a common subject

‣ Direct mapping is a "default" R2RML mapping

R2RML and Direct Mapping Status

▸ Technology has been finalized

▸ Both documents are in "Candidate Recommendation" phase

▸ Should be finished this summer

Structured data in HTML:
RDFa & microdata

HTML pages are a huge source of structured data

▸ Not necessarily large amount of data per page, but lots of them…

▸ Have become very valuable to search engines

- Google, Bing, Yahoo!, or Yandex (i.e., schema.org) all committed to use such data

▸ Two syntaxes have emerged at W3C:

- microdata with HTML5
- RDFa with HTML5, XHTML, and with XML languages in general

Example: making use of structured data and search engine facilities

RDFa and microdata: similarities

▸ Both have similar philosophies:
  ▪ the structured data is expressed *via attributes only* (no specialized elements)
  ▪ both define some special attributes
    • e.g., `itemscope` for microdata, `resource` for RDFa
  ▪ both reuse *some* HTML core attributes (e.g., `href`)
  ▪ both reuse the textual content of the HTML source, if needed

▸ RDF data can be extracted from both
  ▪ i.e., HTML+RDFa and HTML+microdata have become an additional source of Linked Data

RDFa and microdata: differences

▸ Microdata has been *underline optimized* for simpler use cases, concentrating on
- one vocabulary at a time
- tree shaped data
- no datatypes

▸ RDFa provides a full serialization of RDF in XML or HTML
- the price is an extra complexity compared to microdata

▸ RDFa 1.1 Lite is a simplified authoring profile of RDFa, very similar to microdata

RDFa 1.1 and microdata status

‣ **For RDFa 1.1**

- Technology has been finalized
- Is in "Candidate Recommendation"
- Should be finished this summer

‣ **For microdata**

- Technology has been finalized
- Is part of HTML5, hence its advancement depends on other technologies

Relevant W3C notes

‣ "HTML Data Guide"

- help publishers to choose what to use

- how to combine microdata and RDFa in one document

‣ Microdata to RDF

- defines a mapping of an HTML5+Microdata document to RDF

Cleaning up RDF

Reminder…

▸ *Resource Description Framework*: a graph-based model for (Web) data and its relationships

- has a simple `(subject,predicate,object)` model

- makes use of URI-s for the naming of terms

  • objects can also be Literals

- informally: defines named relationships (named links) among entities on the Web

- has different serialization formats

▸ Latest version was published in 2004

RDF cleanup (a.k.a. RDF1.1)

- ▸ Many issues have come up since 2004:
  - ▪ deployment issues
  - ▪ new functionalities are needed
  - ▪ underlying technology may have moved on (e.g., datatypes)
- ▸ The goal of the RDF Working Group is to refresh RDF
- ▸ NOT a complete reshaping of the standard!

Some new features

▸ Standardize Turtle as a serialization format

▸ Clean up some aspects of datatyping, e.g.:

- plain vs. typed literals
- details and role of `rdf:XMLLiteral`

▸ Proper definition for "named graphs"

- including concepts, semantics, syntax, …
  - obviously important for linked data access
  - but generates quite some discussions on the details

▸ etc.

Editorial improvements

▸ Cleanup the documents, make them more readable
- possibly rewrite all documents
- maybe a completely new primer
- new structure for the Semantics document

Status

- ‣ Work has begun a bit less than a year ago
- ‣ Turtle is almost finalized
- ‣ Agreement on most of the literal cleanup
- ‣ Lots of discussion currently on named graphs…

Provenance

The goal is simple…

▸ We should be able to express all sorts of "meta" information on the data

  ▪ creator: who played what role in creating the data (author, reviewer, etc.)

  ▪ view of the full revision chain of the data

  ▪ in case of a integrated data: which part comes from which original data and under what process

  ▪ what vocabularies/ontologies/rules were used to generate some portions of the data

  ▪ etc.

...the solution is more complicated

▸ Requires a complete model describing the various constituents (actors, revisions, etc.)

▸ The model should be usable with RDF

▸ Has to find a balance between

- simple ("scruffy") provenance: easily usable and editable
- complex ("complete") provenance: allows for a detailed reporting of origins, versions, etc.

▸ That is the role of the Provenance Working Group (started in 2011)

ex:aggregation   ex:chart

prov:used   prov:wasGeneratedBy

ex:illustrate

prov:wasGeneratedBy   prov:wasControlledBy

ex:aggregate   prov:wasControlledBy

DRAFT!!!

foaf:name

Derek

prov:used   prov:used

foaf:mbox

ex:regionList   ex:dataSet   mailto:derek@ex.com
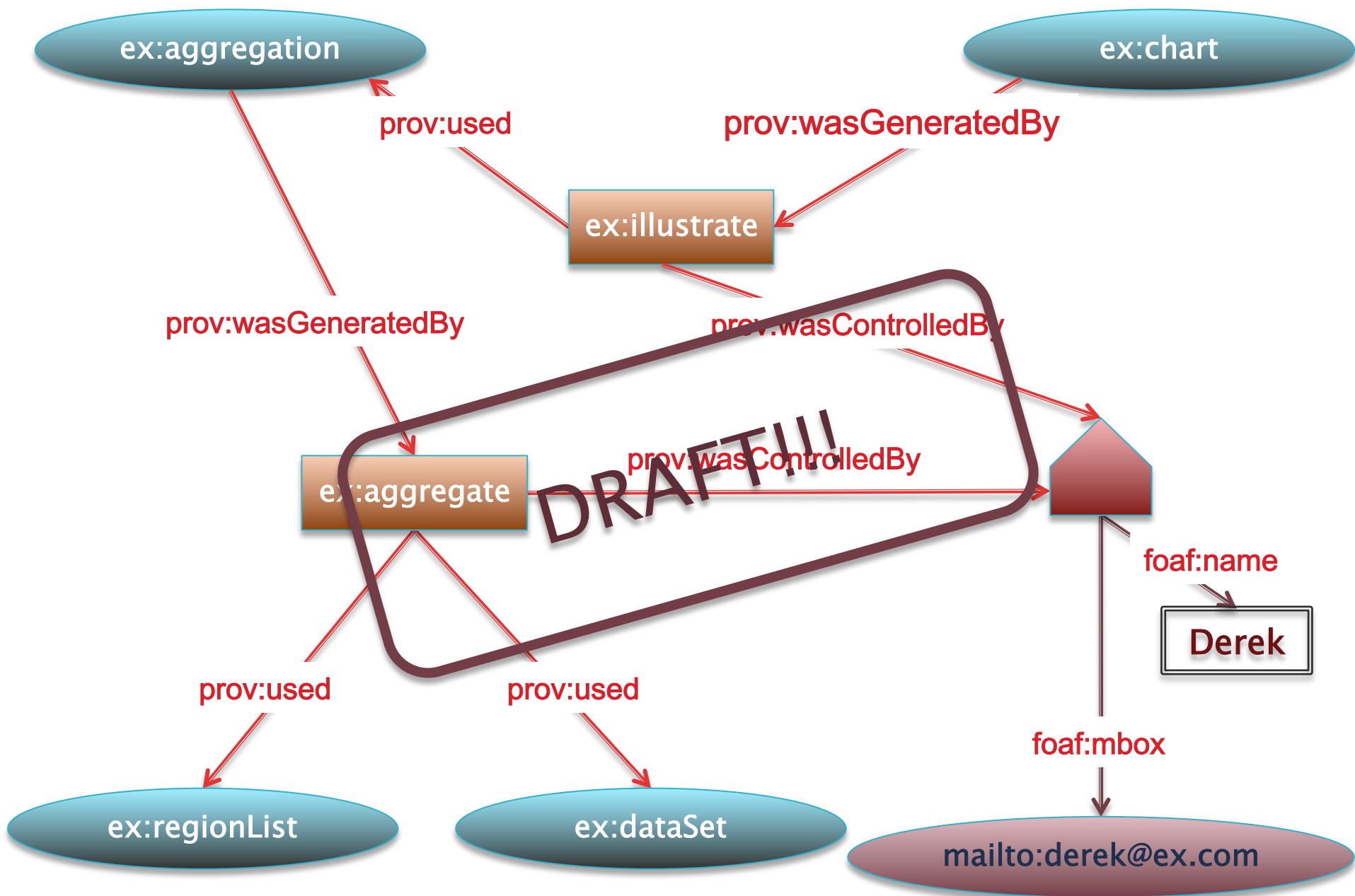
Status

- ‣ Drafts have been published
  - ▪ abstract data model, OWL version
  - ▪ primer
- ‣ The current effort is on simplifying the first versions

Linked Data Patterns

Linked Data: a seed for a Web of Data

- ▸ "Linked Data" is also a set of principles:
  - ▪ put things on the Web through URI-s
  - ▪ use HTTP URI-s so that things could be dereferenced
  - ▪ provide useful information (using standards) when a URI is dereferenced
  - ▪ *include links* to other URI-s
- ▸ RDF is an ideal vehicle to realize these principles

As of September 2011

Media
Geographic
Publications
User-generated content
Government
Cross-domain
Life sciences

(71) *Courtesy of Richard Cyganiak and Anja Jentzsch*
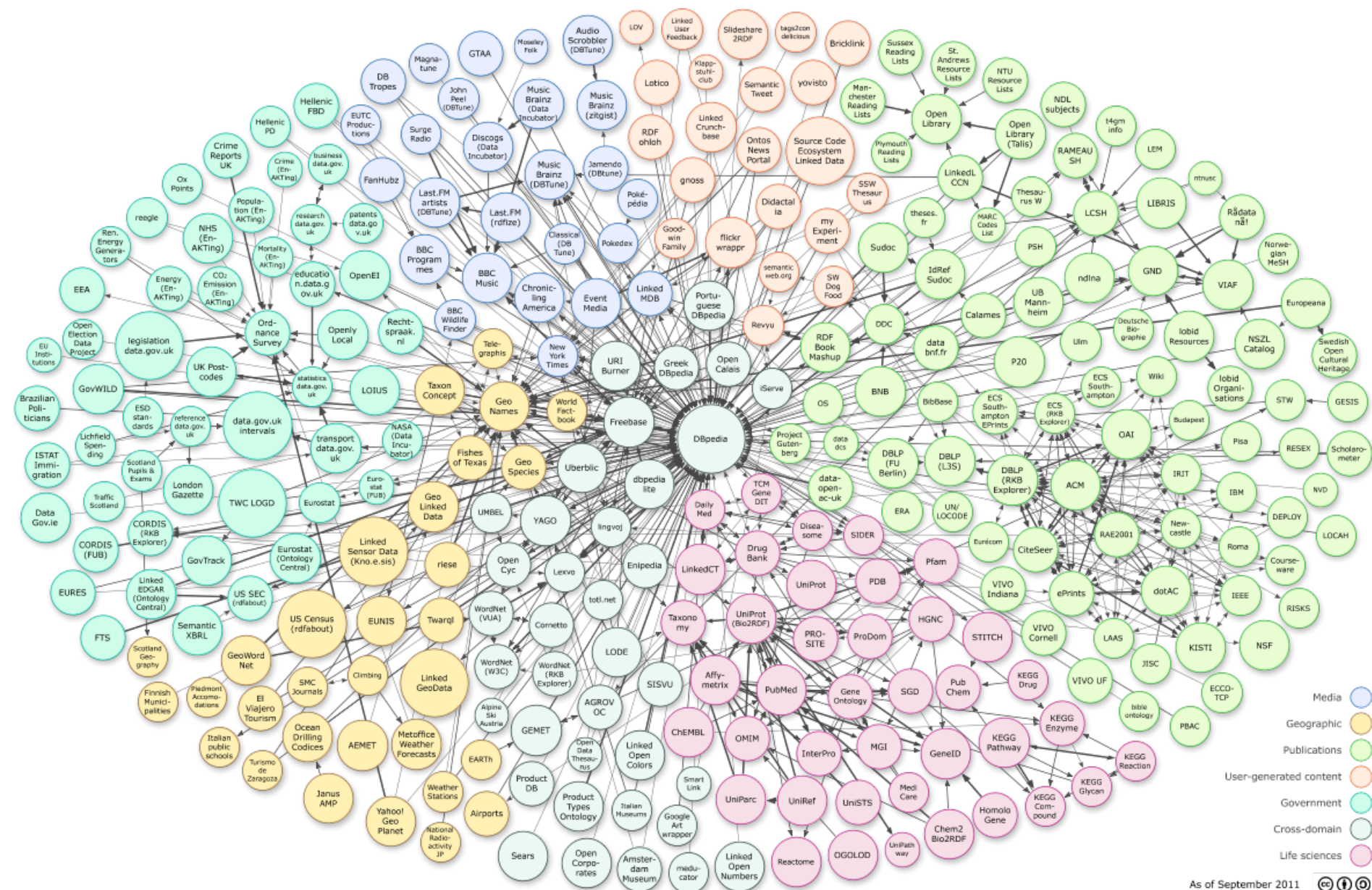
But: the number of links among datasets is still small

Courtesy of Frank van Harmelen, ISWC2011 keynote address
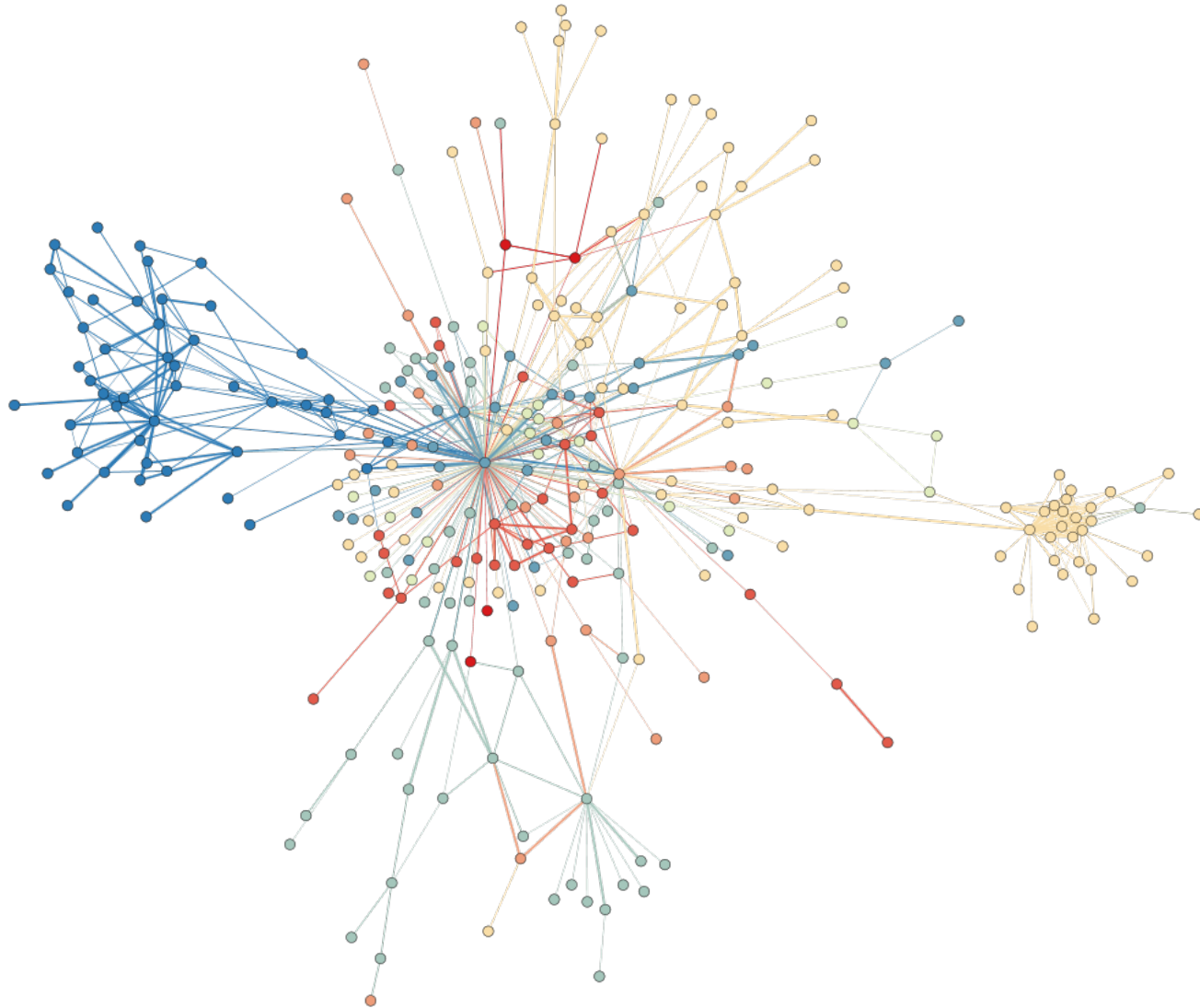
Linked Data offers major challenges for Semantic Web

▸ Scale: we are talking about billions of triples, increasing every day

▸ Highly distributed: data spread over the Web, connected via http links

▸ Very heterogeneous data of different origins

▸ Need not only to read but also to write the data (using the same concepts)

▸ Setting a SPARQL endpoint everywhere may not be realistic

W3C Semantic Web

Read only vs. Read/Write

▸ The current Linked Data work concentrates on *publishing* data for read-only usage

▸ The future requires an easy way to read *and* write data

   ▪ SPARQL UPDATE is a step in this direction

   ▪ but requires an additional SPARQL processor

Planned: Linked Data Platform WG

▸ General goal: to provide a "lower" level, HTTP based infrastructure to publish, read, write, or modify linked data

 ▪ typical usage: data intensive application in a browser, application integration using shared data…

▸ The infrastructure should be "low-cost", easy to implement and install

Planned: Linked Data Platform WG

▸ Two major work areas:

1. Define an HTTP protocol to
   - access and update RDF data through standard HTTP terms
   - define patch, paginated access
   - define a RESTful API

2. Linked Data Profiles: subsets of existing Semantic Web standards to be used for such HTTP based access
   - use only a subset of datatypes
   - use HTTP URI-s only
   - etc.

Planned: Linked Data Platform WG

▸ Two major work areas:

1. Define an HTTP protocol to
   - access and update RDF data through standard HTTP terms
   - (possibly) define patch, paginated access
   - define a RESTful API
2. Linked Data Profiles: subsets of existing Semantic Web standards to be used for such HTTP based access
   - use only a subset of datatypes
   - use HTTP URI-s only
   - etc.

Planned!!!

What else is on the horizon?

Further challenges raised by Linked Data

▸ Knowledge structures vs. data is very different: very shallow, simple vocabularies for huge sets of data

- The role of reasoning is different (vocabularies, OWL DL, etc., may not be feasible)

▸ Highly distributed SPARQL implementations are necessary

▸ etc.

For example: data vs. vocabularies

Courtesy of Frank van Harmelen, ISWC2011 keynote address
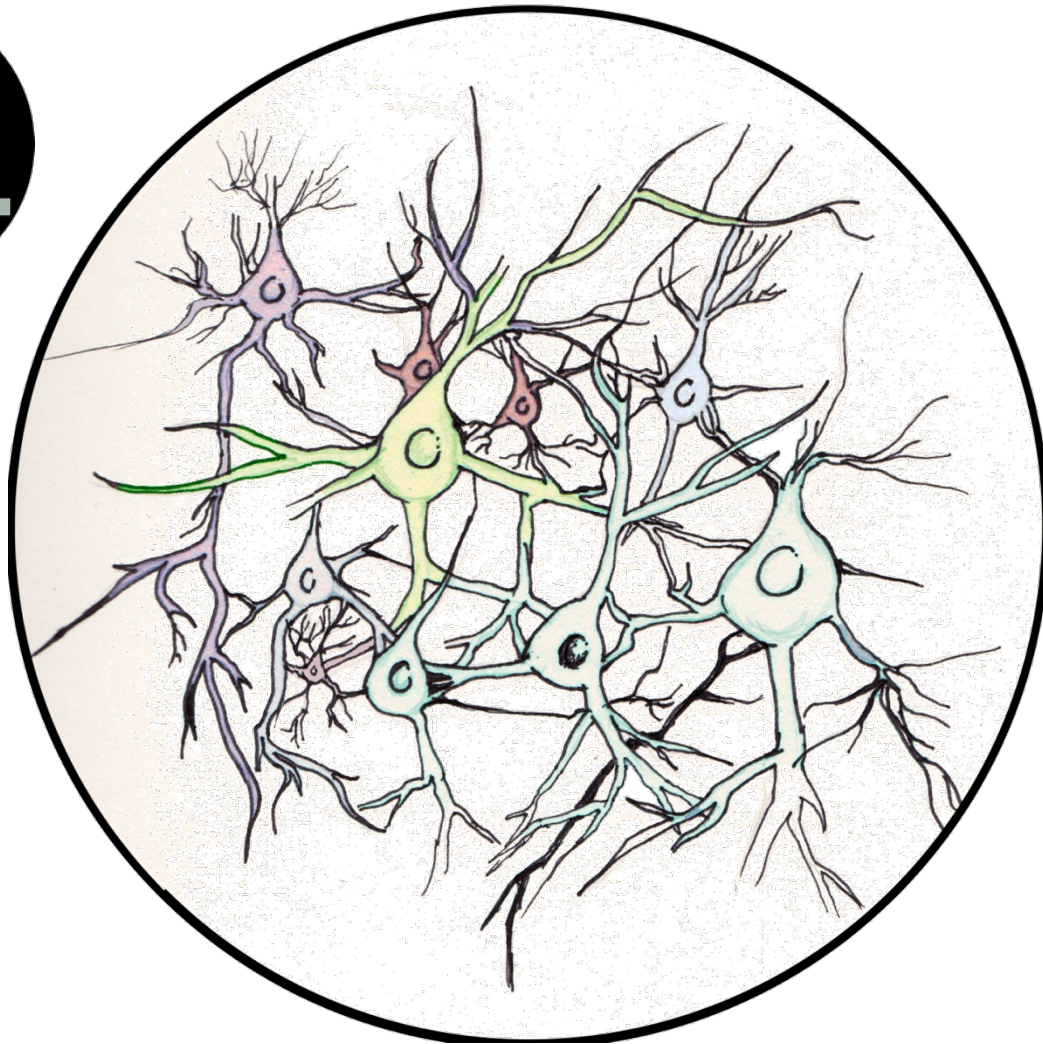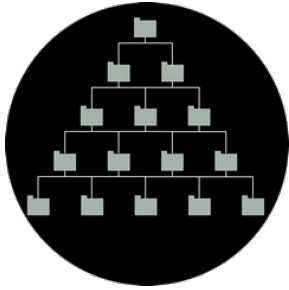
Other work areas in activity that are explored

- Profiles for the publication of Linked Data, e.g.,
  - further profiles of OWL
  - URI patterns
  - datatypes
  - usage of Bnodes
  - etc.
- JSON serialization of RDF

Other work areas in activity that are explored

▸ Standardized approaches for Access Control to data

▸ Reconsider rule languages for (e.g., for Linked Data applications)

▸ Constraint checking of Data

▸ API-s for client-side Web Application Developers

▸ …

To remember…

- *Data* on the Web is a major challenge
  - technologies are needed to use them, to interact with them, to integrate them
- Semantic Web technologies (RDF(S), vocabularies, SPARQL, etc.) can play a major role in publishing and using Data on the Web

Semantic and Multilingual Web?
(leaving my domain of expertise ☺)

‣ Relationships between Semantic Web and multilingualism is twofold:

- can SW help in achieving MLW?

- MLW challenges for SW?

SW helping MLW?

▸ Semantic Web has powerful technologies to *categorize* knowledge (e.g., SKOS and other vocabulary standards)
  - thesauri can be created with labels in different languages
  - some level of knowledge extraction and analysis could be done on those

▸ Via the Linked Data it is possible to interlink *information* different languages
  - DBpedia integrates the various Wikipedia instances

▸ It is possible to tag texts using the same terms (e.g., via stable URI-s)

▸ etc.

These may help in "binding", translating, etc., information in different languages

MLW challenges for SW?

SW technologies and practice has to consider the challenges of MLW

▸ RDF has a very simple way of representing literals (copied from XML): single language tag. Is it enough?

▸ Ontologies/vocabularies are typically monolingual, terms are mostly English…

▸ Practice of vocabulary design very often forgets about MLW issues (first name, last name…)

▸ IRI equivalence is a major headache in practice

▸ etc.

I am looking forward to the discussions!

Thank you for your attention

These slides are also available on the Web:

http://www.w3.org/2012/Talks/0315-Luxembourg-IH/