Linked Open Data
A Way to Extract Knowledge from Global Datastores

Bebo White
SLAC National Accelerator Laboratory

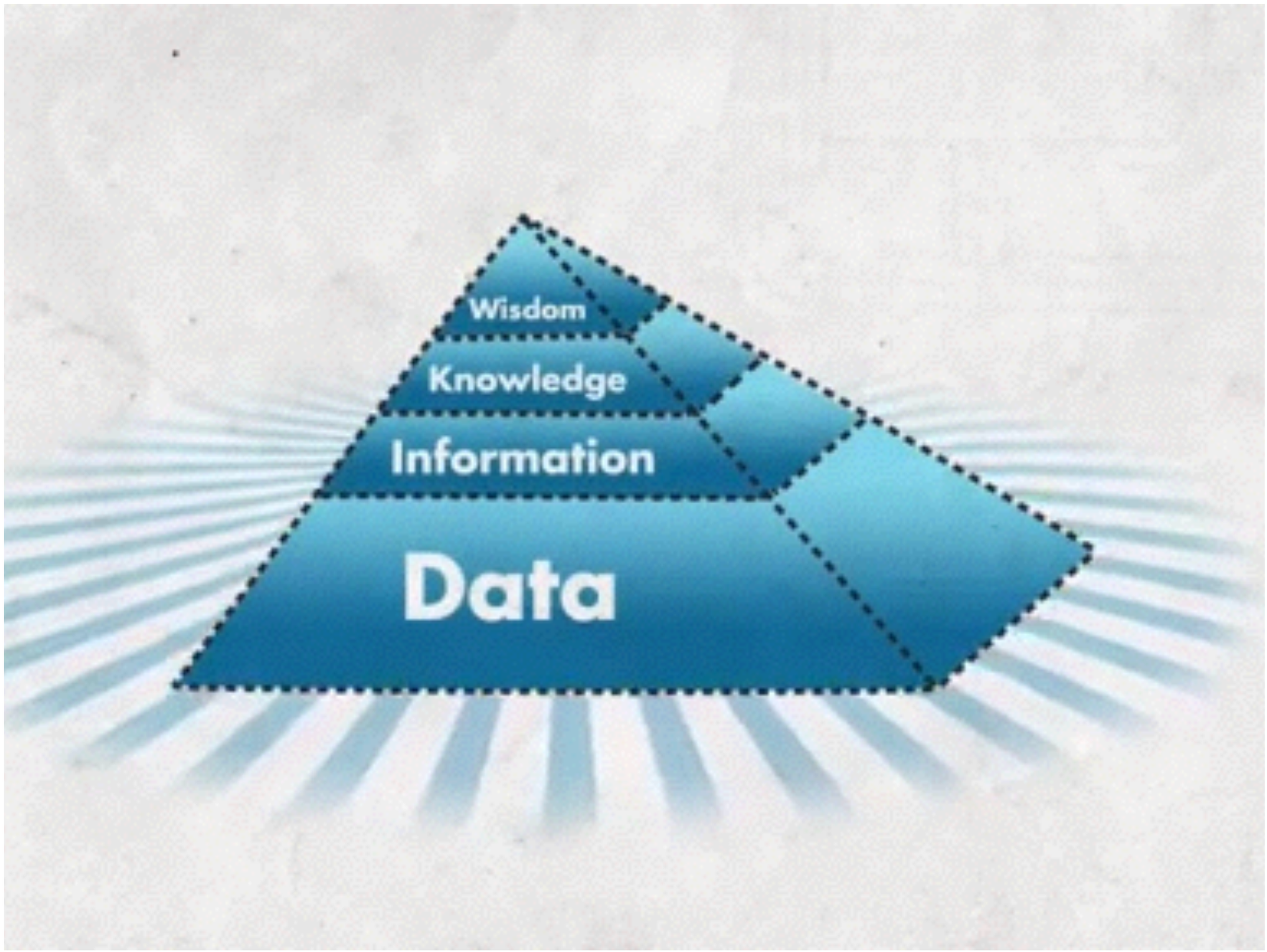
HKU Expert Address
18 September 2014

"Developments in science and information processing have changed the meaning of the verb, 'to know.' It used to mean 'having information stored in one's memory.' It now means the process of having access to information and knowing how to use it."

---Herbert Simon

How can we build *knowledge* from all the facts/data on the Web?

Knowledge means data with context


Wisdom
Knowledge
Information
Data

Perceptions of Web Content

- The Web is generally thought of being composed of pages, documents
- We have been able to insert some data
  - Images <img src="….">
  - Multimedia
- Web 2.0 mashups provided a new way of thinking about a "Web of Data" but it was awkward to obtain
  - APIs
  - "Screen-scraping"

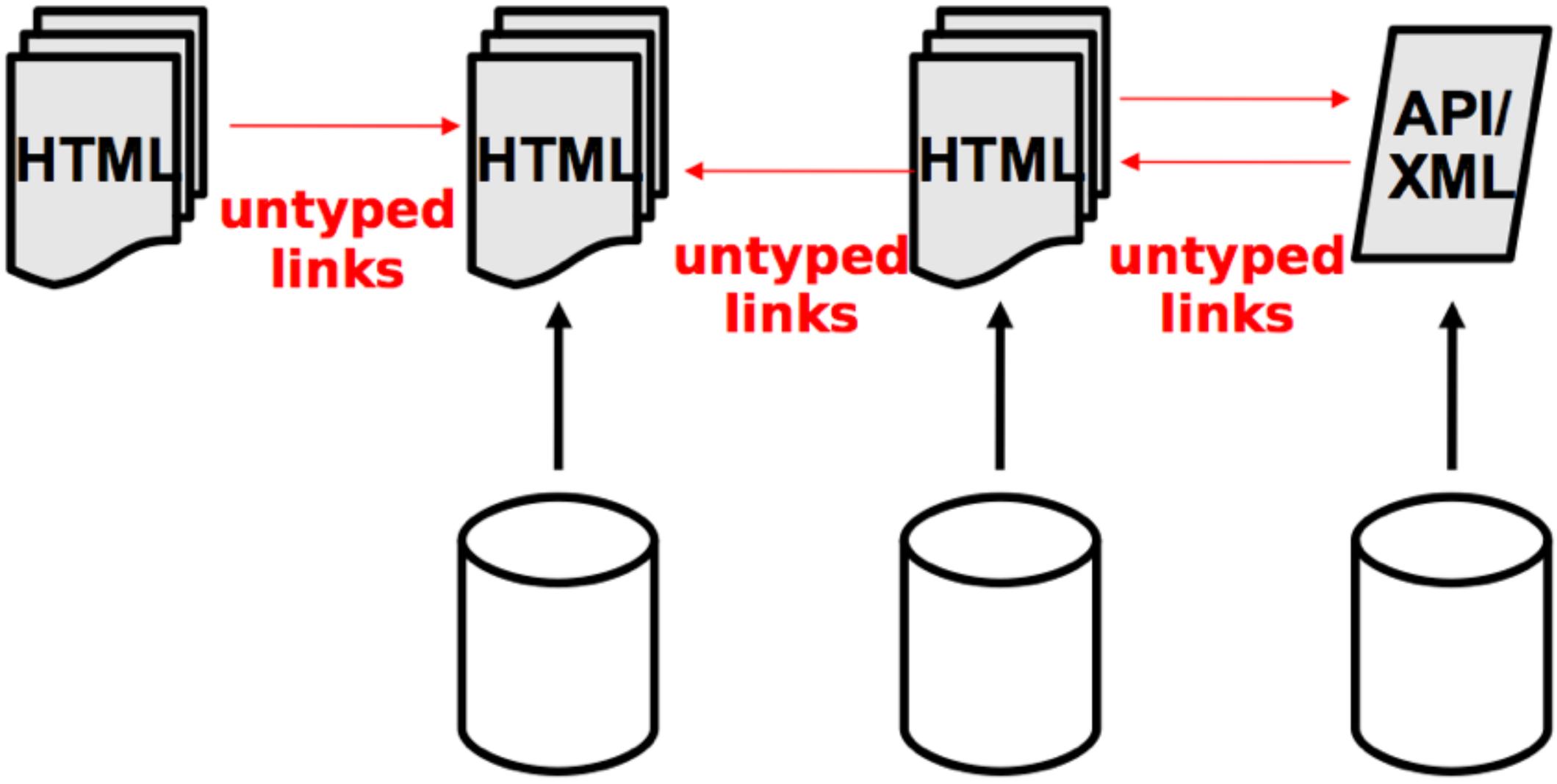
The Web of Documents

- Analogy
  - A global filesystem
- Designed for
  - Human consumption
- Primary objects
  - Documents (or sub-parts of)
- Links between
  - Documents (or sub-parts of)
- Degree of structure in objects
  - Fairly low
- Semantics of content and links
  - Implicit

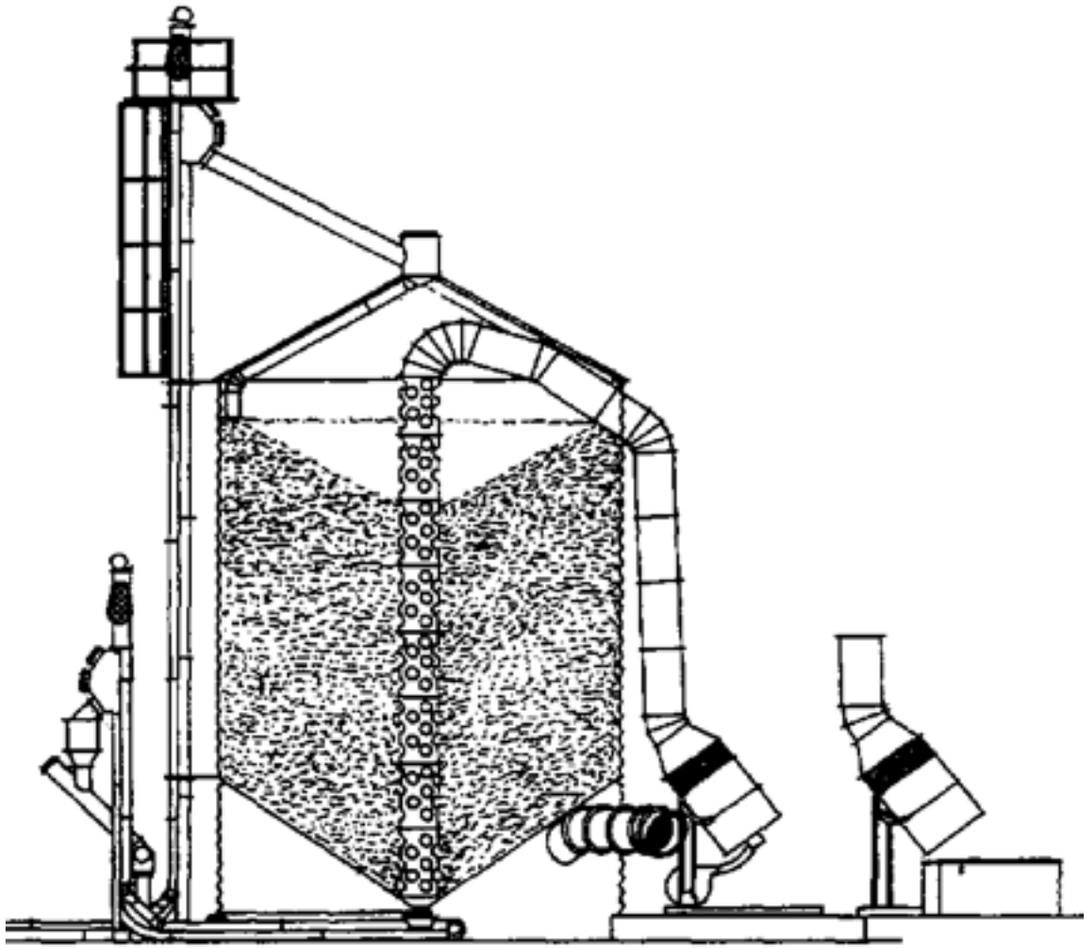
The Web of Documents: Issues

- Simplicity
  - Loosely structured data, untyped links, disconnected data
- Integration
  - Show me all the publications from HKU PhD students in Computer Science
- Querying
  - Which papers have I written with colleagues outside the US?

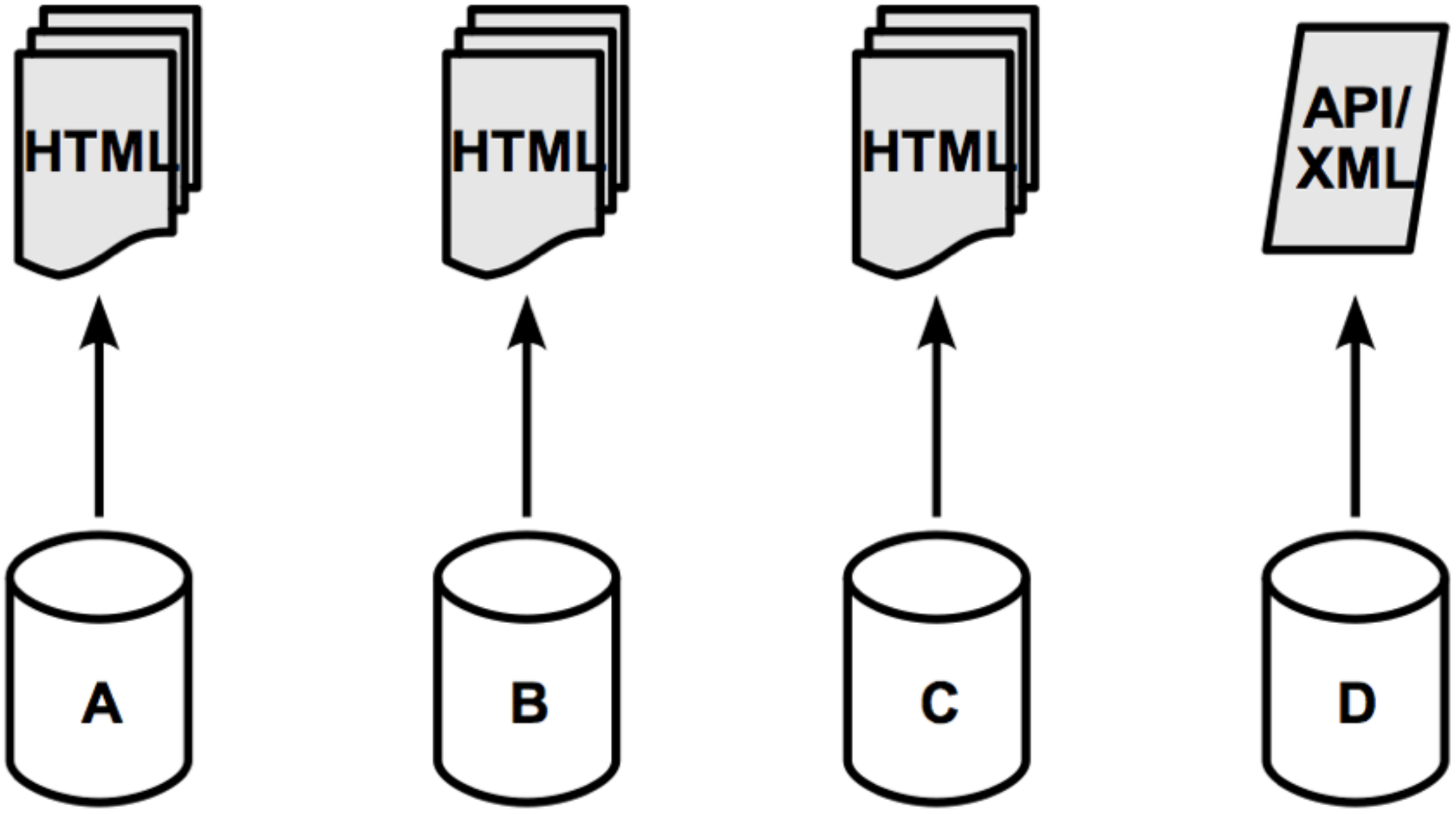
The Web of Linked Documents



“Data Silos” on the Web



“Data Silos” on the Web



How About Open Data? (1/2)

- **Interoperability** to ensure broad and easy use
- Human AND machine readable, i.e., data + metadata
- In common **open formats** using **open standards**
- Smooth and **cost efficient data integration**, i.e., reuse
- Can **generate effects** - local, regional, national, global

How About Open Data? (2/2)

- Anyone can publish it
    - Individuals
    - Companies/Institutions
    - Governments
- Who can use it?
    - Politicians and decision makers
    - Public administration and project developers
    - Citizens (citizen analysts)
    - Economy and Industry
    - (Data) journalists, media, and publishers
    - Academia and Science

DATA.GOV

DATA TOPICS IMPACT APPLICATIONS DEVELOPERS CONTACT

The home of the U.S. Government's open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

GET STARTED

SEARCH OVER 108,606 DATASETS

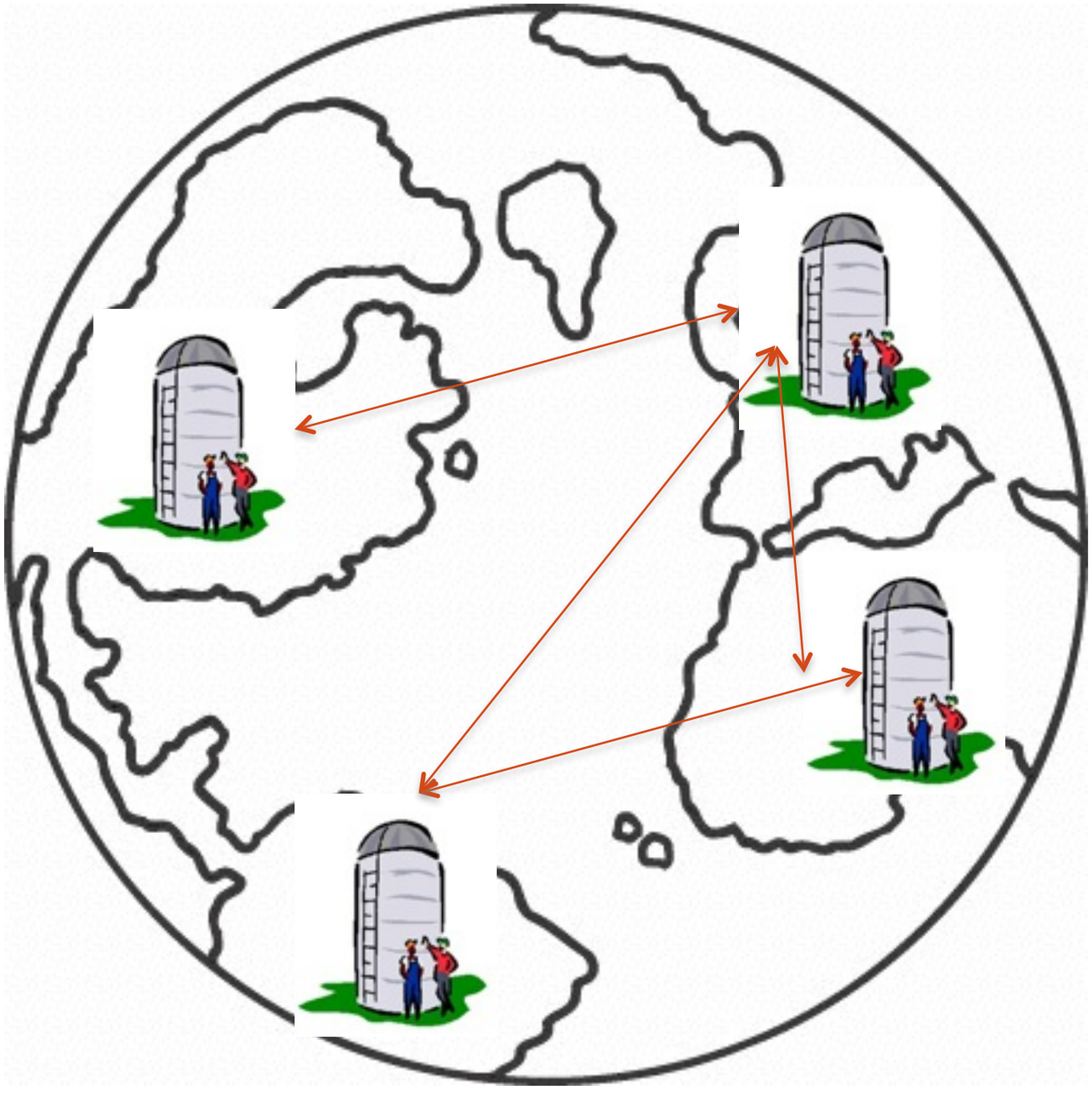
Monthly House Price Indexes

BROWSE TOPICS

Agriculture
Climate
Education
Energy
Finance
Geospatial
Global Development
Health
Jobs & Skills
Public Safety
Science & Research
Weather

MORE TOPICS

A World Wide Network of Data Silos



5 Stars for Open Data (Tim B-L)

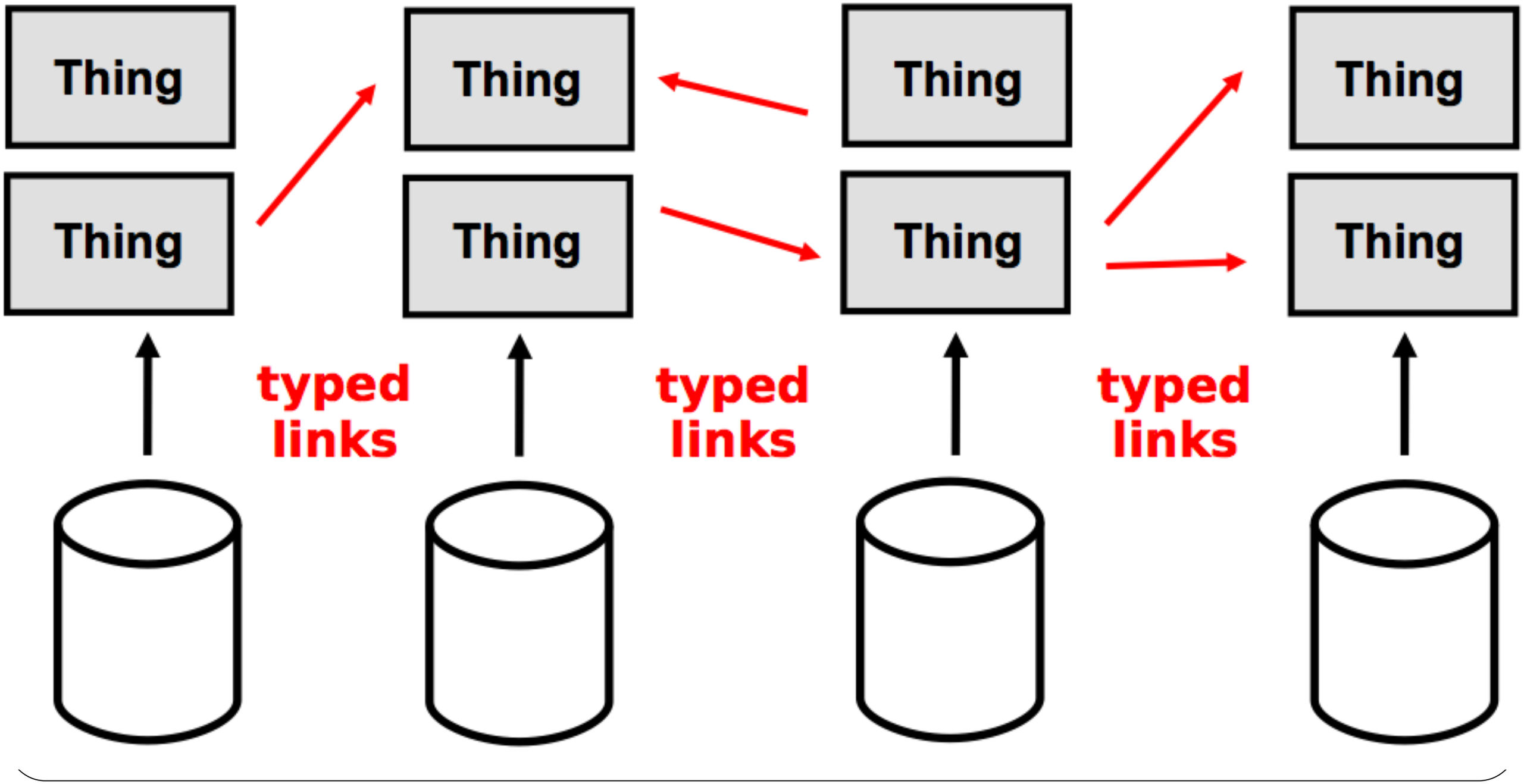
- Make content available on the Web (whatever format) - 1 ★
- Make content available as structured data - 2 ★
- Use a non-proprietary format - 3 ★
- Use URLs to identify things - 4 ★
- Link data to other data to provide context - 5 ★

The Web of Linked Data

- Analogy
  - A global database
- Designed for
  - Machines first, humans later
- Primary objects
  - Things (or descriptions of things)
- Links between
  - Things
- Degree of structure in (descriptions of) things
  - High
- Semantics of content and links
  - Explicit

The Web of Linked Data

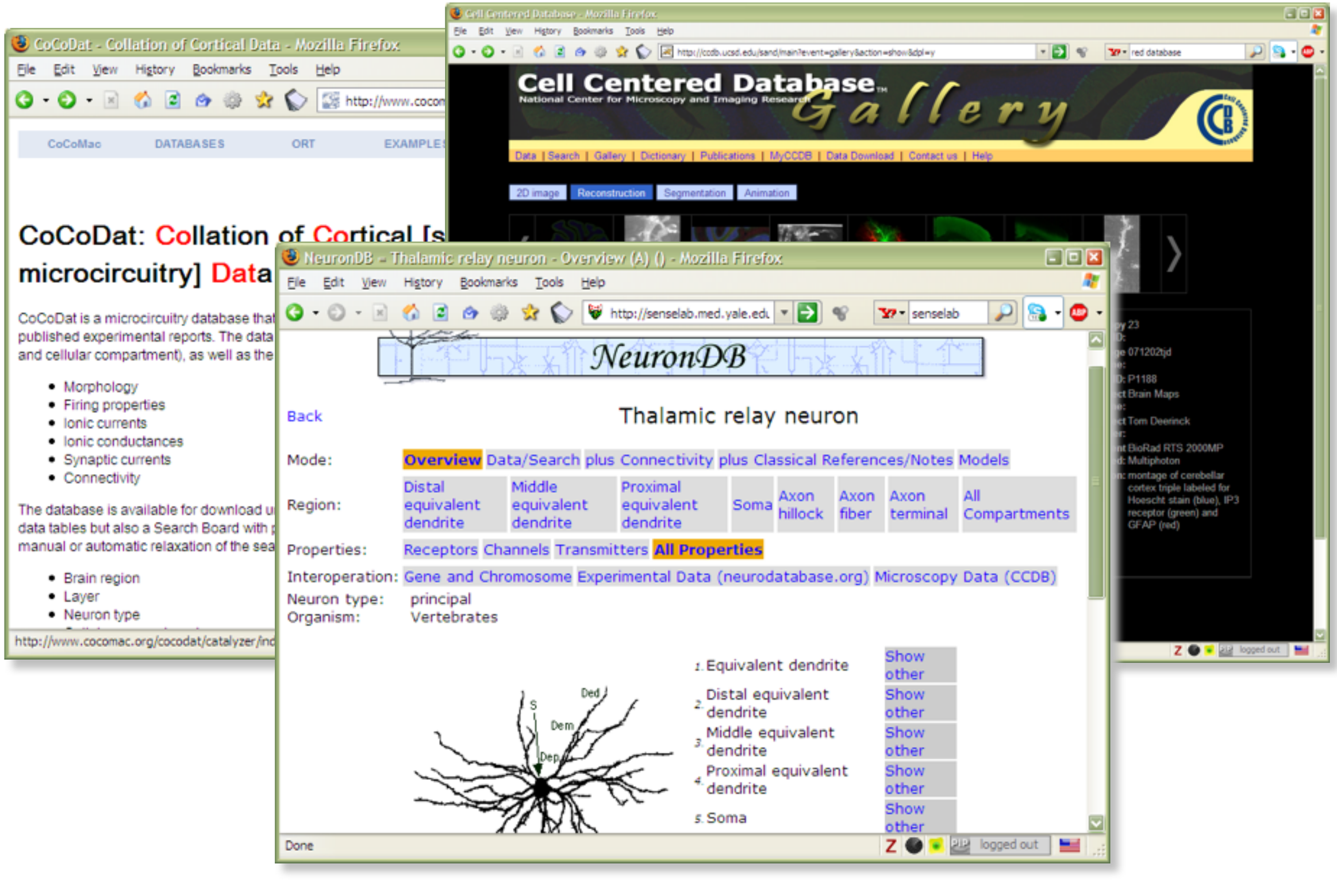
Don't just link the *documents*, link the *things*



Imagine…

- A “Web” where
  - Documents are available for download on the Internet
  - But there would be no hyperlinks among them

And the Problem is Real



Data on the Web (1 ★) is Not Enough

- 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 Web
  - ie, data can be integrated over the Web
- This is where Semantic Web technologies come in

Costs & Benefits of ★ Web Data

- As a consumer/publisher
  - You can read it
  - You can print it
  - You can store it locally
  - You can export it to another system
  - You can modify it
  - You can share it (with conditions)

Costs & Benefits of ★★Web Data

- Everything from one ★ plus:
  - You can directly process it with proprietary software to aggregate it, perform calculations, visualizations, etc.
  - You can export it into another structured format

Costs & Benefits of ★★★ Web Data

- Everything with 2 ★ plus:
- You can manipulate the data in any way without software restrictions

Costs & Benefits of ★★★★Web Data

- Everything with ★★★ plus:
  - You can link to it from anywhere
  - You can bookmark it
  - You can reuse parts of the data
  - You can combine the data with other data
  - You have complete control over the data items and can optimize their access
  - Other publishers can now link to your data making it

Costs & Benefits of ★★★★★ Web Data

- Everything with ★★★★ plus:
  - You can discover more (related) data while consuming the data
  - You can learn about the data schema
  - You can make your data discoverable
  - You increase the value of your data

Linked Data Principles

- Use URIs as names of things
  - Anything, not just documents
  - You are not your homepage
  - Information resources and non-information resources
- Use HTTP URIs
  - Globally unique names, distributed ownership
  - Allows people to look up those names
- Provide useful information in RDF
  - When someone looks up a URI
- Include RDF links to other URIs
  - To enable discovery of related information

RDF

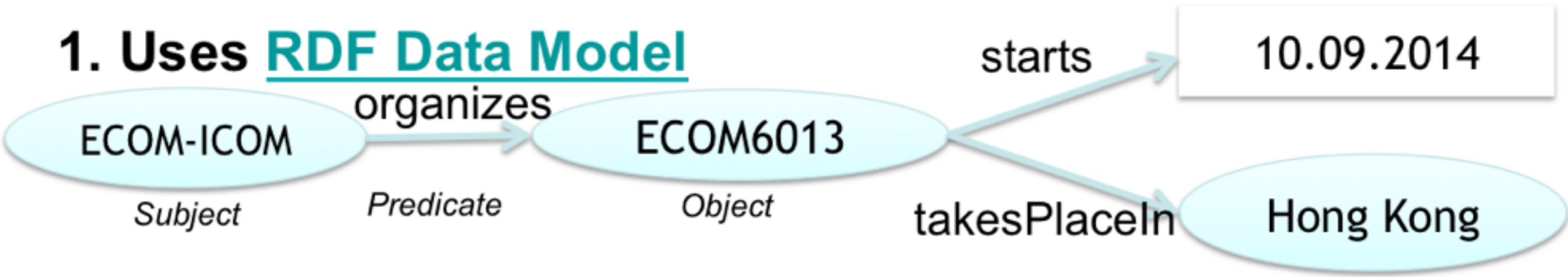
- A data format for describing things and their interrelationships
  - Standardized (XML)
  - Easily parsed by machines

Code	Title	Instructor
ECOM6013: exam	E-commerce technologies	Prof. Bebo White
ECOM6029	E-business transformation	Prof. Ali Farhoomand (HKU)
ECOM6030	Web 2.0 strategy and innovation	Prof. Amy Shuen (CEIBS)
ECOM6031: exam	Fundamentals of e-commerce security	Dr. KP Chow (HKU)
ECOM6035	Developing business models for digital media and online games	Mr. Peter Looms (University of Copenhagen)
ICOM6012: exam	Internet infrastructure technologies	Prof. Lawrence Yeung (HKU)
ICOM6043: exam	Information architecture	Prof. Renato Iannella (NETHA)

Session time: weekday: 6:45pm - 9:45pm; weekend: 9:30am - 12:30pm; 2:00pm - 5:00pm; 6:45pm - 9:45pm
Venues: Please refer to the module homepages

MON	TUE	WED	THU	FRI	SAT	SUN
1-Sep	2-Sep	3-Sep	4-Sep	5-Sep	6-Sep	7-Sep
	ICOM6012			ICOM6012		
8-Sep	9-Sep	10-Sep	11-Sep	12-Sep	13-Sep	14-Sep
	Mid-Autumn Festival	ECOM6013		ECOM6013	ICOM6043 - Lab (pm)	ECOM6013 (am+pm)
15-Sep	16-Sep	17-Sep	18-Sep	19-Sep	20-Sep	21-Sep
ECOM6013	ICOM6012	ECOM6013		ECOM6013	ICOM6043 -Lab (pm)	ECOM6013 (am+pm)
22-Sep	23-Sep	24-Sep	25-Sep	26-Sep	27-Sep	28-Sep

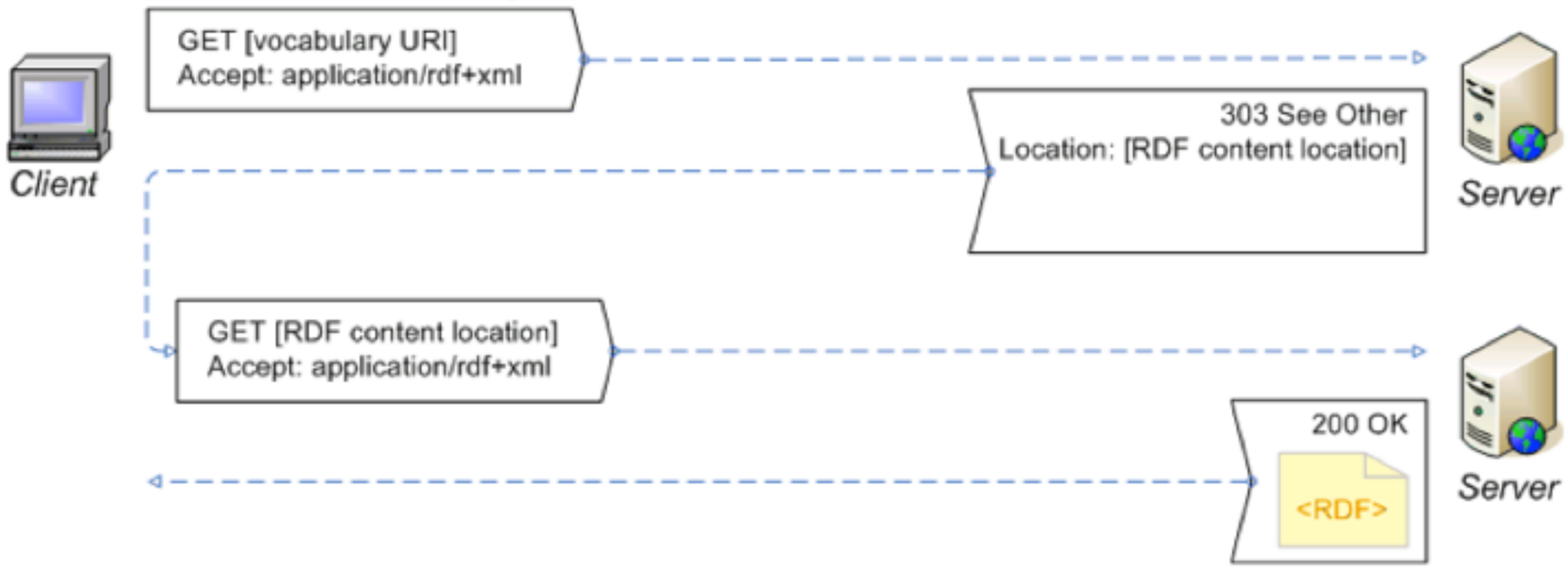
1. Uses RDF Data Model



2. Is serialized in triples:

```
ECOM-ICOM      organizes      ECOM6013 .
ECOM6013       starts         "20140910^^xsd:date .
ECOM6013       takesPlaceIn   Hong Kong.
```

3. Uses Content-negotiation



Why Publish Linked Data? (Why Be 5-Star?)

- Ease of discovery
- Ease of consumption
  - Standards-based data sharing
- Reduced redundancy
- Added value
  - Build ecosystems around your data/content

Publishing Linked Open Data

- Identify and analyze your data
- Clean your data (?)
- Model your data (URI schema, vocabularies)
- Select and specify license(s)
- Convert your data to RDF
- Link your data to other data
- Publish and promote your Linked Open Data
- Watch others use it (become 5-star!)

Consuming Linked Open Data

- Specify concrete use cases
- Evaluate relevant data sources and data sets
- License clearing (check respective licenses)
- Create data consumption patterns
- Manage alignment, caching, and update mechanisms
- Create mash-ups, GUIs, services, and applications on top of the data
- Establish sustainable new partnerships

URI SCHEME FOR US DATA.GOV

Submitted by jahendler on Fri, 2010-11-19 01:21 **URI Schemes**

1 0

On the data.gov Web site, George Thomas and I have proposed that we need an effort to help develop a URI solution for the resolving of linked data on data.gov. We explain:

Data.gov now hosts a set of Resource Description Framework (RDF) documents containing triples created by converting a number of the Data.gov datasets into this format, making over 6.4 billion triples of open government data available to the community ...

The URI scheme chosen is a very simple one for the time being, designed to allow users to easily explore and extend the data. A proposal is being developed with RPI, one of the Data.gov community leaders, for a new encoding of datasets converted from CSV (and other formats) to RDF. We're looking forward to a design discussion to determine the best scheme for persistent and dereferenceable government URI naming with the international community and the World Wide Web Consortium to promote international standards for persistent government data (and metadata) on the World Wide Web.

The RPI proposal that is mentioned is discussed at: **http://data-gov.tw.rpi.edu/wiki/A_Proposal_for_Governmental_Data_URIs**

Active Group topics

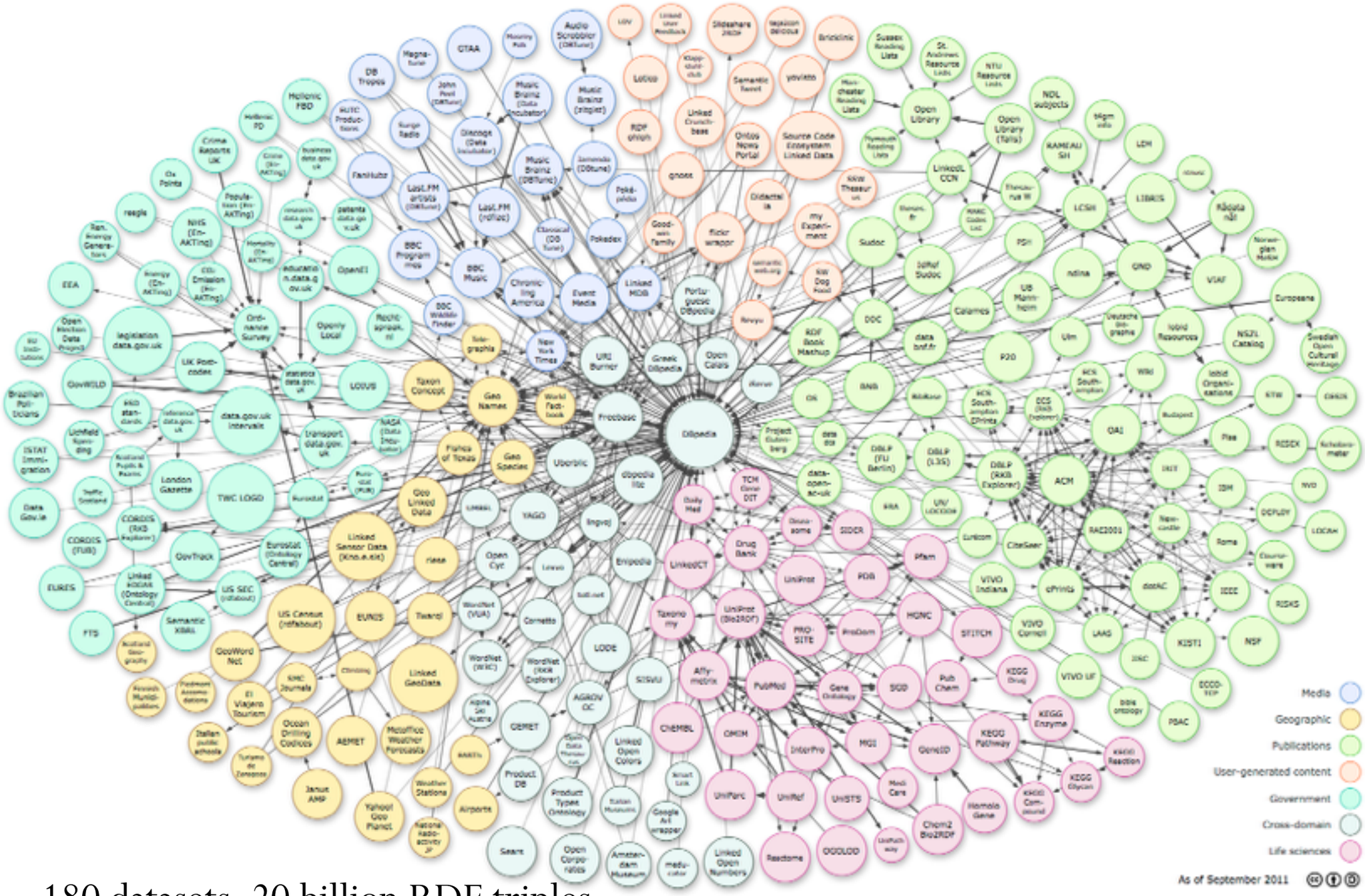
- **Status of NRC-Regulated Complex Materials Sites Undergoing Decommissioning** | Last comment 2 days ago
- **Annual 2008 Electric Power Industry Data** | Last comment 2 days ago
- **Alternative Transportation Fuels (ATF) and Alternative Fueled Vehicles (AFV) 2008** | Last comment 2 days ago
- **Annual 2008 Electric Generator Report (EIA-860)** | Last comment 2 days ago

The Linking Open Data Project

The Linking Open Data Project

- Community project with W3C support
- Take existing open data sets
- Make them available on the Web in RDF
- Interlink them with other data sets
- Began in early 2007

The Linked Open Data Cloud



180 datasets, 20 billion RDF triples

DBpedia Version 2014 released

September 9, 2014 - 10:58 am by ChrisBizer -

Hi all,

we are happy to announce the release of **DBpedia 2014**.

The most important improvements of the new release compared to DBpedia 3.9 are:

1. the new release is based on updated Wikipedia dumps dating from April / May 2014 (the 3.9 release was based on dumps from March / April 2013), leading to an overall increase of the number of things described in the English edition from 4.26 to 4.58 million things.

2. the DBpedia ontology is enlarged and the number of infobox to ontology mappings has risen, leading to richer and cleaner data.

The English version of the DBpedia knowledge base currently describes **4.58 million things**, out of which 4.22 million are classified in a consistent ontology (http://wiki.dbpedia.org/Ontology2014), including 1,445,000 persons, 735,000 places (including 478,000 populated places), 411,000 creative works (including 123,000 music albums, 87,000 films and 19,000 video games), 241,000 organizations (including 58,000 companies and 49,000 educational institutions), 251,000 species and 6,000 diseases.

DBpedia

Querying Wikipedia like a Database.

Recent Posts

DBpedia Version 2014 released
DBpedia Spotlight V0.7 released
Call for Ideas and Mentors for GSoC 2014 DBpedia + Spotlight joint proposal (please contribute within the next days)

Categories

DBpedia Blog | Get Involved | Get Help

About / News
Applications
Use Cases
Datasets

DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web. DBpedia allows you to ask sophisticated queries against Wikipedia, and to link the different data sets on the Web to Wikipedia data. We hope that this work will make it easier for the huge amount of information in Wikipedia to be used in some new interesting ways. Furthermore, it might inspire new mechanisms for navigating, linking, and improving the encyclopedia itself.

About: SLAC National Accelerator Laboratory

An Entity of Type : Feature, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

The SLAC National Accelerator Laboratory, originally named Stanford Linear Accelerator Center, is a United States Department of Energy National Laboratory operated by Stanford University under the programmatic direction of the U.S. Department of Energy Office of Science.

Property	Value
dbpedia-owl:abstract	▪ The SLAC National Accelerator Laboratory, originally named Stanford Linear Accelerator Center, is a United States Department of Energy National Laboratory operated by Stanford University under the programmatic direction of the U.S. Department of Energy Office of Science. The SLAC research program centers on experimental and theoretical research in elementary particle physics using electron beams and a broad program of research in atomic and solid-state physics, chemistry, biology, and medicine using synchrotron radiation. ▪ Das SLAC National Accelerator Laboratory (SLAC) ist eine Forschungseinrichtung des Department of Energy der USA. Es handelt sich um einen Linearbeschleuniger für Elektronen und Positronen. Datei:Stanford-linear-accelerator-usgs-ortho-kaminski-5900. jpg Luftbild des Linearbeschleunigers in Stanford ▪ SLAC on vuonna 1962 perustettu Stanfordin yliopiston yhteydessä toimiva hiukkaskiihdytinlaboratorio Kaliforniassa Yhdysvalloissa. SLAC keskittyy alkeishiukkasten kokeelliseen ja teoreettiseen tutkimukseen. ▪ Le Centre de l'accélérateur linéaire de Stanford (en anglais Stanford Linear Accelerator Center) est un laboratoire de physique dépendant du Département de l'Énergie des États-Unis et opéré par Université Stanford. Ses activités de recherche se concentrent sur la physique des particules théorique et expérimentale, et depuis quelques années s'ouvrent à la photonique au travers du projet LCLS. L'accélérateur de particules de 3,2 km de long situé sur le site est le plus long accélérateur linéaire au monde. Fichier:Stanford-linear-accelerator-usgs-ortho-kaminski-5900. jpg Vue aérienne de l'accélarateurlinéaire]]

Linked Open Data Publication Strategies: Application in Networking Performance Measurement Data

Renan F. Souza[1], Les Cottrell[2] , Bebo White[2], Maria L. Campos[1], Marta Mattoso[1]

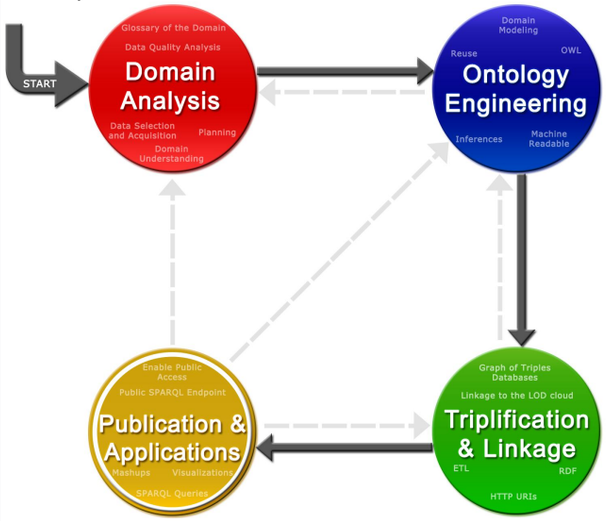
[1]Federal University of Rio de Janeiro, Brazil, [2]SLAC National Accelerator Laboratory

ABSTRACT

- Most of the data published on the web is unstructured or does not follow a standard.
- It makes the data harder to be retrieved and interchanged between different data sources
- Linked Open Data (LOD) technologies are applied in a scenario that deals with a large amount of computer network measurement data.
- As a result, we generated more structured data, hence easier to be retrieved, analyzed, and more interoperable.
- The challenges of processing large amount of data to: transform it into a standard format (RDF); link it to other data sources; and analyze and visualize the transformed data are discussed.
- An ontology that aims to minimize the number of triples is proposed and a discussion on how ontologies may impact query performance is presented.
- We emphasize the advantages of having the data in RDF format and show use cases on the scenario of the project.

RESEARCH DESIGN AND METHODOLOGY

We proposed the following methodology for Linked Open Data publication:



In the end, we want to have structured, retrievable, and publicly accessible PingER data directly linked to DBpedia [8], Geonames [9], and Freebase [10]. Also, indirectly, to any other data source on the LOD cloud [12].

www.pingerlod.slac.stanford.edu

SYSTEM MODELING AND RESULTS

Domain Analysis

PingER Ping end-to-end reporting

Understanding PingER Project's domain

- It envolves data about network performance measurement
- 80 monitoring nodes
- 800 monitored nodes
- 8000 pairs of nodes (monitor-monitored)
- 160 countries, several cities within each country
- 16 network metrics (e.g. TCP throughput, packet loss, average RTT)
- Hourly data, since 1998
- Data can be applied to many different situations such as economical, geographical, and seasonal events.

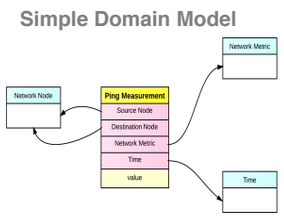
Data stored in multiple flat CSV files

Problem and Strategies

- Hard to query the CSV files to retrieve specific data, comparing to traditional DBMS
- Hard to produce informative graphs, reports, and dashboards
- Data not interoperable with other data sources

- Data could be published in an open standard format to enable wider consumption
- **Semantic Web** and **Linked Open Data** strategies can be applied to publish PingER structured data in an open standard web format, enabling complex queries to the data and interoperability with other external data sources.

Simple Domain Model



How a ping measurement can be modeled

Ontology Engineering

Ontology Reuse

- An Ontology is needed to model the domain following W3C recommendations.
- Reusing existing ontologies supports the idea of standardization and interoperability within LOD community.

Reuse Evaluation

- Semantic expressivity
- Completeness in relation to the domain
- Impacts on query performance

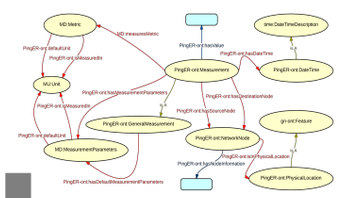
Ontologies being reused

- Geonames [3]
- W3C Time Ontology [4]
- MOMENT [5][6]

MOMENT Ontology

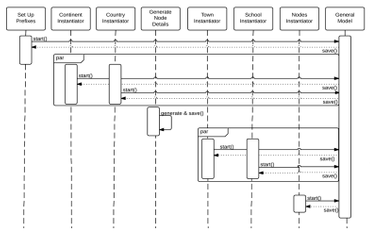
Geonames Ontology

PingER LOD Ontology



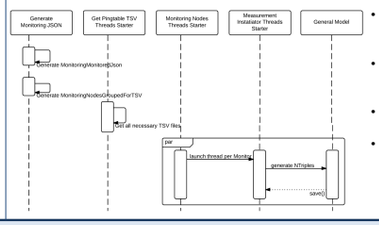
Triplification & Linkage

ETL Process for General Data



- Parallel and Distributed approach to triplify multiple CSV files
- ETL - Extract data from the CSV files, Transform it into Triples format, and Load it into the RDF DBMS
- While the data is being transformed into triples, it is also being linked to external data sources in the LOD cloud.
- Each process is independent, hence can be simultaneously executed in different machines.

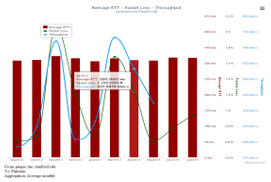
ETL Process for Measurement Data



- Each ETL process for Measurement data is responsible for a single network metric and a single time aggregation
- 11 network metrics (throughput, packet loss, etc) and 3 time aggregations (daily, monthly, and yearly)
- 33 processes that can run in distributed machines
- Each process is further parallelized

Publication & Applications

Mutiple Network Metrics



- It takes advantages of well-structured data with a schema, in a very expressive format (RDF).
- It explores complex SPARQL queries to capture precisely what is being searched.
- Any possible combination of paramaters is able to be retrieved.

Network Metrics vs. University Metrics

- Illustration of a mashup of PingER data with Dbpedia [9] data about universities (information about number of students, endowment, etc).
- Using this graph, one could visually verify that well-funded universities have better network connectivity.

Network Metrics vs. % of GDP Invested in Research and Development

- PingER data mashed up with World Bank [11] Data.
- It is possible to verify how the countries have invested in Research and Development throughout the years
- And how it has affected network connectivity.

CONCLUSIONS

This work followed the methodology proposed to publish Linked Open Data applied in a real scenario that deals with big datasets about internet measurement. This methodology is based on:

- Domain analysis: understanding the domain and selecting which should be triplified.
- Ontology engineering: reuse evaluation and number of triples minimization
- Triplification project based on a parallel and distributed approach, linking to other data sources in the LOD cloud
- Publication: Enabling public access to both the data and the ontology in a standard, open, structured, and interoperable format, utilizing Semantic Web and LOD technologies.

Results: SPARQL Endpoint is available to query and to interoperate the data; RDF dump of the database is available; and the Ontology is public in OWL format.

FUTURE WORK

- Utilizing complex SPARQL queries (those that are common in database with OLAP characteristics) on the PingER LOD database is still taking undesirable amount of time.
- Thus, in terms of query performance, more research is needed to provide an efficient way of querying very large Triple Stores with OLAP characteristics.

REFERENCES

1. PingER Project. (2014) PingER - Ping end-to-end reporting. [online]. http://www-iepm.slac.stanford.edu/pinger/
2. World Wide Web Consortium. (2013) Semantic Web. [Online]. http://www.w3.org/standards/semanticweb/
3. Geonames. (2013) GeoNames Ontology. [Online]. http://www.geonames.org/ontology/documentation.html
4. World Wide Web Consortium. (2006) Time Ontology. [Online]. http://www.w3.org/TR/owl-time/
5. Sathya Rao, "Monitoring and measurement in the next generation technologies," 2010.
6. European Telecommunications Standards Institute, "Measurement Ontology for IP traffic (MOI); Requirements for IP traffic measurement ontologies development," 2012.
7. Giancarlo Guizzard, "Uma abordagem metodológica de desenvolvimento para e com reúso, baseada em ontologias formais de domínio," 2000.
8. DBpedia. (2014) DBpedia. [Online]. http://dbpedia.org/About
9. Geonames. (2014) About GeoNames. [Online]. http://www.geonames.org/about.html
10. Freebase. (2014) Freebase. [Online]. http://www.freebase.com/
11. World Bank. (2014) The World Bank Data. [online]. http://data.worldbank.org/
12. Richard Cyganiak and Anja Jentzsch. (2012) Linking Open Data Cloud Diagram. [Online]. http://lod-cloud.net/

ACKNOWLEDGEMENT

- This work was supported in part by the Department of Energy contract DE-AC02-76SF0051.
- This work was supported in part by the Rio de Janeiro State Science Foundation (FAPERJ).

PingER Reuse Possibilities?

- What other parties might be interested in published PingER linked open data?
  - Network/Internet analysis researchers
  - Telecoms
  - IT managers at participating sites
  - “Digital Divide”/Education - “quality of service”
  - Emergency services
  - Etc.,etc.
- What other parties might be interested in merging PingER linked open data?

The Power of Linked Open Data

- Enables Web-scale data publishing and discovery
- Everything is a potential resource
- Everything can be annotated
- Easy to extend and add new properties
- Easy to merge new RDF graphs and data ontologies
- Easy use and re-use on top of common schemas AND schema mapping
- Allows complex querying of multiple distributed data sources and systems

Why is Linked Open Data Important?

- Because in many cases it's our data!
- Efficiency, reducing redundancy
- Promotes a digital society
- Opens the door to data innovation and discovery
- Holds the promise of creating from data
  - Knowledge
  - Wisdom
  - Benefit for all

Thanks for Your Attention!

Questions? Comments?

bebo@slac.stanford.edu

bebo@baiyongxue.com