Web Data Management: A Short Introduction to Data Science

V. CHRISTOPHIDES

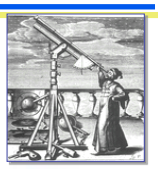
© Serge Abiteboul

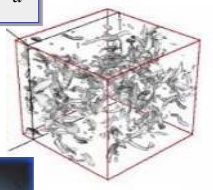
Department of Computer Science
University of Crete
ICS - FORTH, Heraklion, Crete

Shifting Paradigm in Sciences

- Thousand years ago: science was empirical
  - describing natural phenomena
- Last few hundred years: theoretical branch
  - using models, generalizations
- Last few decades: a computational branch
  - simulating complex phenomena
- The fourth paradigm today (eScience): data exploration unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/Knowledge stored in computer
  - Scientist analyzes data using data management services and statistics

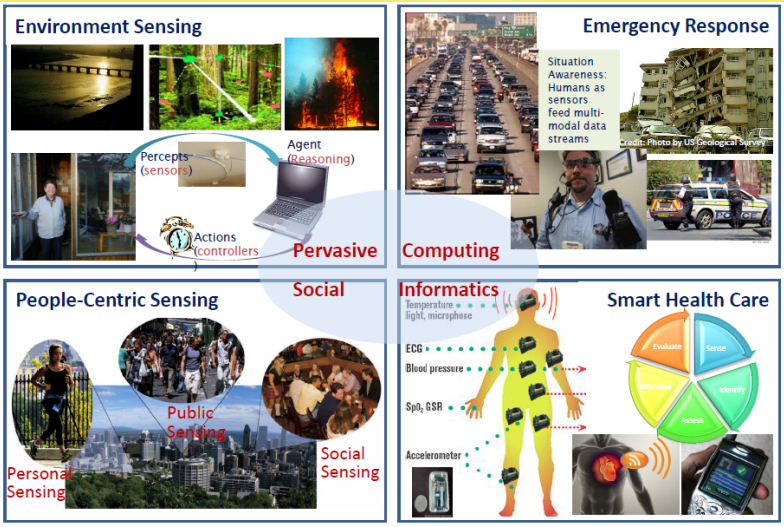
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$





© Jim Gray

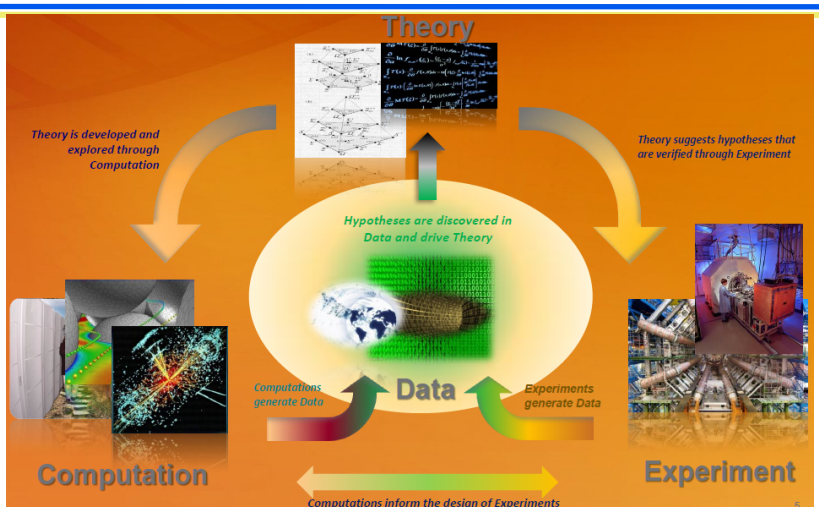
But also Ubiquitous Data Creation



© Sajal Das, Keith Marzullo

- Ubiquitous sensing & reasoning in physical, biological and cyber worlds to support data-driven decision making

Data-driven Discovery

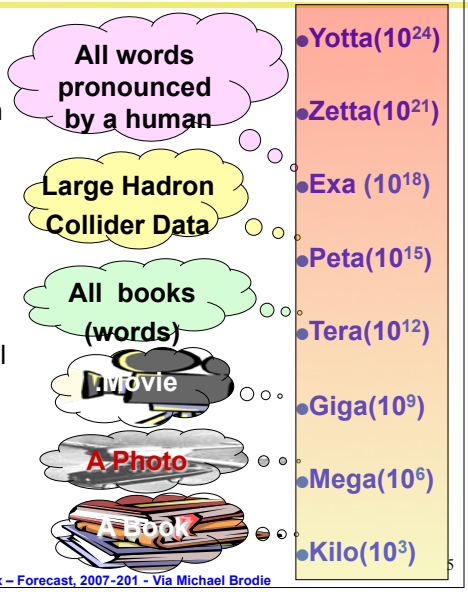


© JOHN R. JOHNSON

- Data-driven discovery is revolutionizing scientific exploration as well as engineering innovations

The Volume of Data Doubles Every 18 Months!

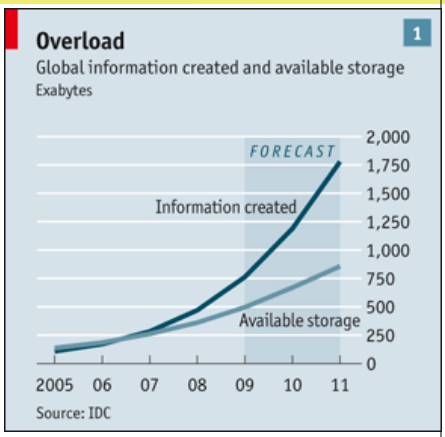
- **Mega** 10^6 : a big building in Heraklion
- **Giga** 10^9 : the entire human genome
- **Tera** 10^{12} : 200 T are all books ever written in any langue
- **Peta** 10^{15} : 100P is the amount of data produced in a single minute by the new particle collider at CERN
- **Exa** 10^{18}: 5 Exa is a transcript of all words ever spoken
- **Zetta** 10^{21} : 66 Zetta is the amount of visual information conveyed from the eyes to the brain
- **Yotta** 10^{24} : a holographic image of earth surface



Source: Cisco Visual Networking Index – Forecast, 2007-201 - Via Michael Brodie

Predictions of the "Industrial Revolution of Data"

- Data is the new "raw material of business" – Economist
- Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from
  - very large volumes (T -> Z)
  - of a wide variety of data (Structured -> Semi-structured -> Unstructured),
  - by enabling high velocity (Batch -> Streaming) capture, discovery, and/or analysis at a very high rate
- Data has complex interrelations
- Data has many free parameters
- Data is needed by many people



It is not possible to store all the data we produce!

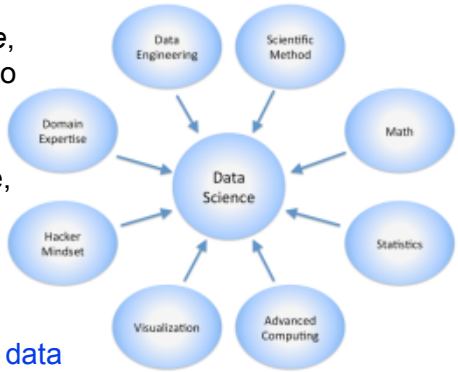
A single person / computer alone cannot do all the work!

All brand new Research Landscape

- More data is being collected than we can store
  - Analyse the data as it becomes available
  - Decide what to archive and what to discard
- Many data sets are too large to download
  - Analyse the data wherever it resides
- Many data sets are too poorly organized to be usable
  - Self-describing data to improve their quality
- Many data sets are heterogeneous in type, structure, semantics, organization, granularity, accessibility …
  - Integrate and customize access to federate data
- Utility of data limited by our ability to interpret and use it
  - Extract and visualize actionable knowledge using higher-lever data abstractions
- A large number of linked datasets may be exploited to identify real-word entities
  - built-in privacy preserving characteristics in management and analysis

Towards a Data Science

- Data Science aims to combine computer science, statistics and machine learning, visualization and human-computer interactions to *collect*, *clean*, *integrate*, *analyse*, *visualize*, *interact* with data to create data products
- Core Challenges:
  - Preparing Data (Noisy, Incomplete, Diverse, Streaming …)
  - Analyse Data (Scalable, Accurate, Real-time, Advanced Methods, Probabilities and Uncertainties ...)
  - Disseminate Analysis Results (i.e. data product) (Story-telling, Interactive, explainable)



Data Science Research Agenda

Acquisition, Storage, and Management of "Big Data"	Data Analytics	Data Sharing and Collaboration
Data representation, storage, and retrieval	Computational, mathematical, statistical, and algorithmic techniques for modeling high dimensional data	**Tools for distant data sharing**, real time visualization, and software reuse of complex data sets
New parallel data architectures, including clouds	Learning, inference, prediction, and knowledge discovery for large volumes of dynamic data sets	**Cross disciplinary model, information and knowledge sharing**
Data management policies, including privacy and secure access	Data mining to enable automated hypothesis generation, event correlation, and anomaly detection	Remote operation and real time access to distant data sources and instruments
Communication and storage devices with extreme capacities	Information infusion of multiple data sources	
Sustainable economic models for access and preservation		

Source Big Data R&D Initiative Howard Wactlar NIST Big Data Meeting June, 2012

9

What Changed with the Advent of the Web?

Source: flickr http://www.flickr.com/photos/docsearls/5500714140/

- Information was residing in isolated silos and manipulated with a variety of programming languages, management systems and infrastructures
- Thanks to open standards for exchanging information on the Web we have:
  - A uniform and universal access to information
  - Easy linking and re-used of information in different contexts
  - Network effects in adding value to information

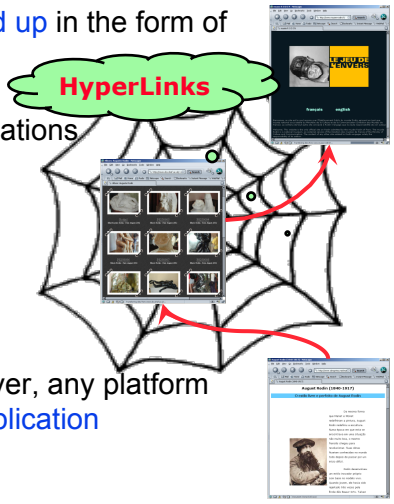
10

A Bit of Web History

- The Hypertext-centric View
  - Web seen as a collection of unstructured documents with hyperlinks
  - HTTP, HTML
- The Document-centric View
  - Web seen as a collection of structured documents
  - XML, XMLSchema, XPath, XQuery, XSL
- The Data-centric View
  - Web seen as a collection of linked datasets with optional semantic specifications
  - RDF, RDFS, SPARQL, OWL, …

The Secrets of HTML Success

- Information and its presentations are mixed up in the form of HTML pages
  - all intended for human consumption
  - many generated automatically by applications
- Everybody can write it:
  - HTML is simple
  - HTML is textual: it is human readable, you can use any editor, ...
- Everybody can read it:
  - HTML is portable on any platform
- Easy to fetch any Web page, from any server, any platform
  - The browser is the universal access application
- Everybody can search it:
  - Keyword-based Search Engines
- It connects pieces of information together through hypertext links



What's Wrong with HTML

- If written properly, normal HTML markup may reflect document presentation, but it cannot adequately represent the semantics & structure of data

Artist Name

Artifact Title

Date

Dimensions

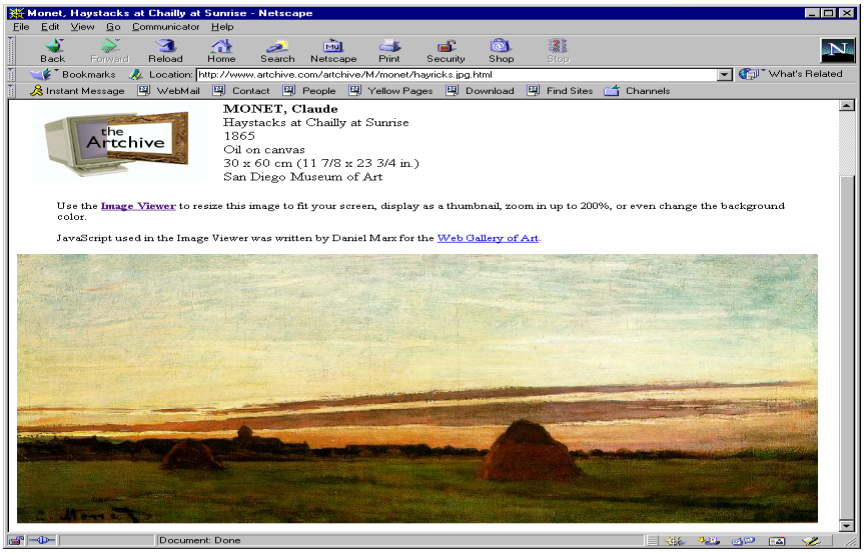
Material

Image Reference

Museum

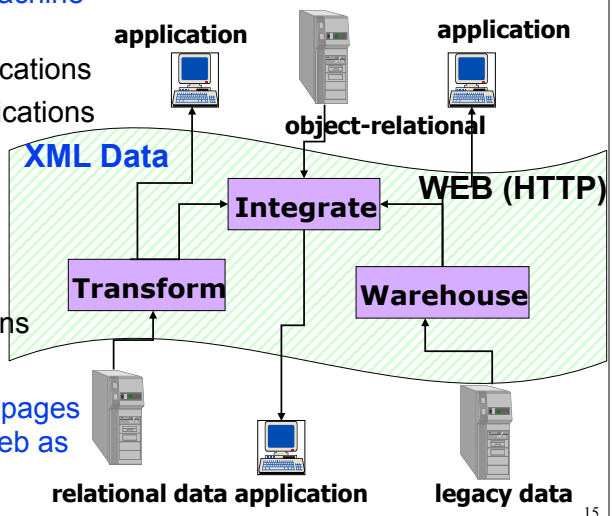
```
<B>MONET, Claude<B><BR>
Haystacks at Chailly at Sunrise<BR>
1865<BR>
Oil on canvas<BR>
30 x 60 cm (11 7/8 x 23 3/4 in.)<BR>
San Diego Museum of Art <BR>
<P>
<IMG SRC="http://192.41.13.240/artchive/
m/monet/hayricks.jpg">
```


HTML Document Presentation



The Need for Sharing Data on the Web

- Not only human but also machine processable information
  - XML generated by applications
  - XML consumed by applications
- Universal data exchange:
  - across platforms
  - across organizations
- More than Web browsers
  - Web-enabled Applications
- From a collection of HTML pages to data published on the Web as structured documents!



15

XML Data Representation: The Document View

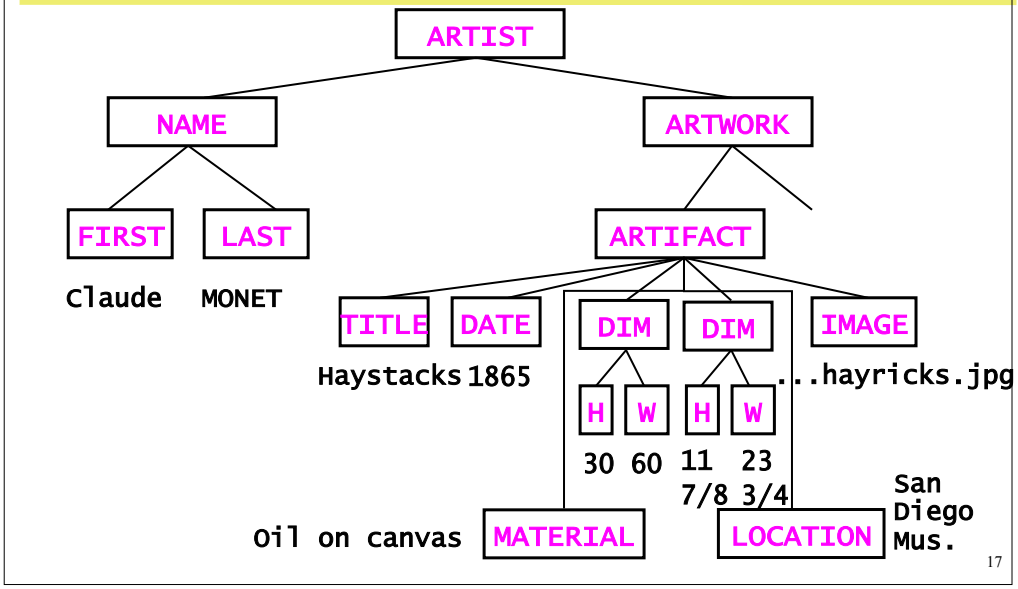
```
<ARTIST>
  <NAME>
    <FIRST>Claude</FIRST> <LAST>Monet</LAST>
  </NAME>
  <ARTWORK>
    <ARTIFACT>
      <TITLE>Haystacks at Chailly at Sunrise</TITLE>
      <DATE>1865</DATE>
      <MATERIAL>Oil on canvas</MATERIAL>
      <DIM Metric='cm'>
       <HEIGHT>30</HEIGHT><WIDTH>60</WIDTH></DIM>
      <DIM Metric='in'>
       <HEIGHT>11 7/8</HEIGHT><WIDTH>23 3/4</WIDTH></DIM>
      <LOCATION>San Diego Museum of Art</LOCATION>
      <IMAGE File='http://192.41.13.240/artchive/m/monet/hayricks.jpg'/>
    </ARTIFACT>
  </ARTWORK>
</ARTIST>
```

Element Name
Element Content
Attribute Name
Attribute Value
Empty Element

- User definable and domain specific markup

16

XML Data Representation: The Data View



The Secrets of XML Popularity

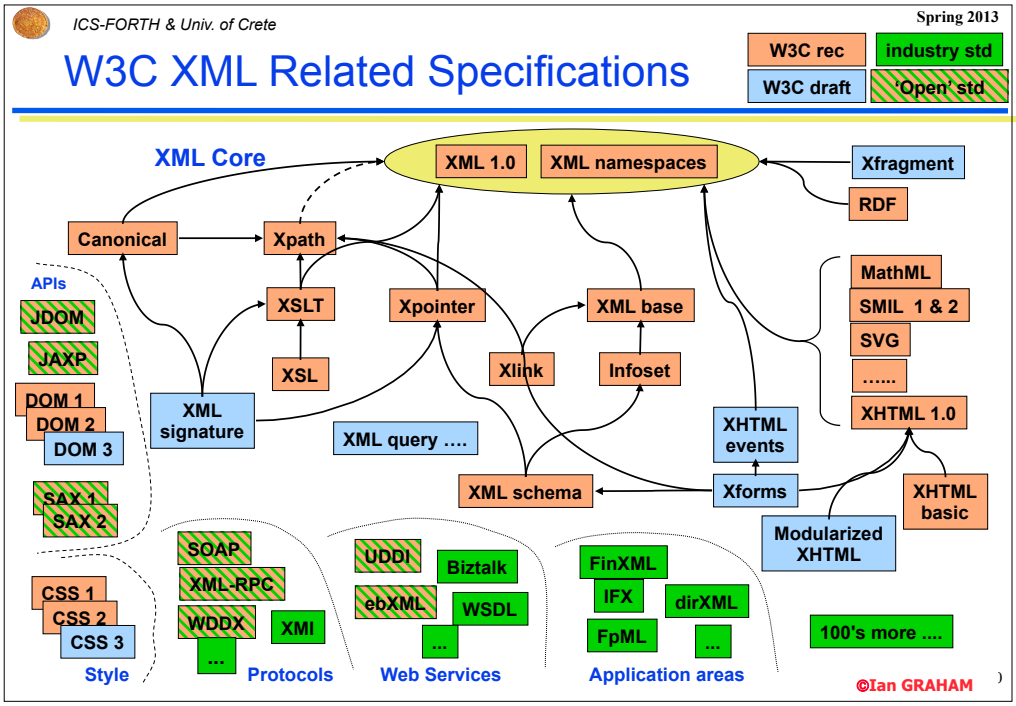
- It looks like HTML...
  - Simple, familiar, easy to learn, human-readable
  - Universal and portable
- …but it's more than HTML!
  - flexible: you can represent any information
  - extensible: you can represent it the way you want!
- Increasing precision in XML specifications
  - Well-Formed: already better than plain text
  - Valid: Structure conforms to a DTD or an XML Schema

XML Standardization Activities in W3C

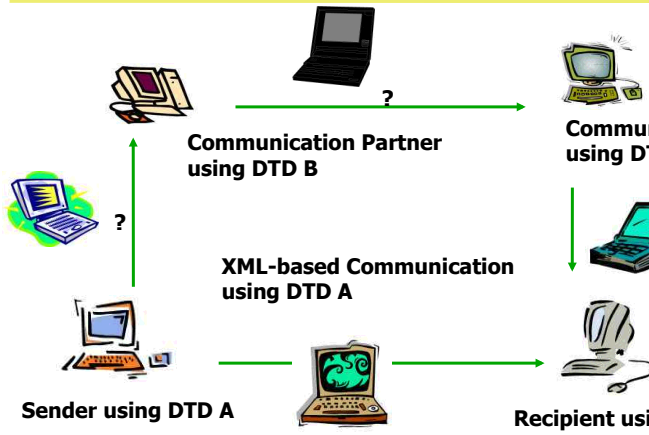
- Core XML WG
  - eXtensible Markup Language (XML 1.0), namespaces, Infoset
- XML Linking WG
  - XML Pointer Language (Xpointer), XML Linking language
- XML Schema WG
- XML Query WG
  - XML Data Model, Algebra and Query Language
- Document Object Model WG
- XSL WG
  - XPath (with XML Linking WG)
  - Transformation and stylesheet language (XSLT/XSL)

- Designing XML tools is a data management problem:
  - XML 1.0 to describe structured documents = Syntax for trees
  - XML data models to describe the information content = Data model for trees
  - XML schemas to describe the structure of information = Data definition language for trees
  - XML languages to describe information processing = Data manipulation language for trees

W3C XML Related Specifications



XML Limitations for Large Scale Data Sharing



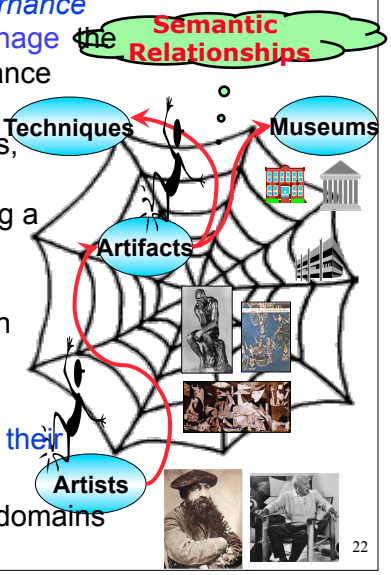
- Only feasible for closed collaboration
  - Partners in a small & stable community
  - Documents on a small & stable intranet

- XML makes no commitment on:
  1. Domain specific ontological vocabulary
  2. Ontological modelling primitives

=> requires pre-arranged agreement on 1 & 2

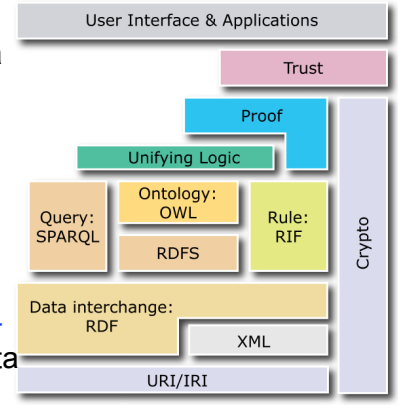
Beyond XML: Towards a Web of Meaning

- For the successful *sharing*, *access*, and *governance* of web data, the ability to understand and manage the semantics of the data is of paramount importance
  - The meaning of data can be extracted by other data such as public knowledge bases, structured/unstructured annotations, etc.
  - The meaning of data can be declared using a machine-processable and understandable knowledge representation language
- The "Semantic Web" was aiming to provide an information infrastructure for expressing the semantics of data & applications on the Web
  - Enable web resources to be accessible by their meaning rather than by keywords
  - Facilitate semantic interoperability across domains and independently-developed applications

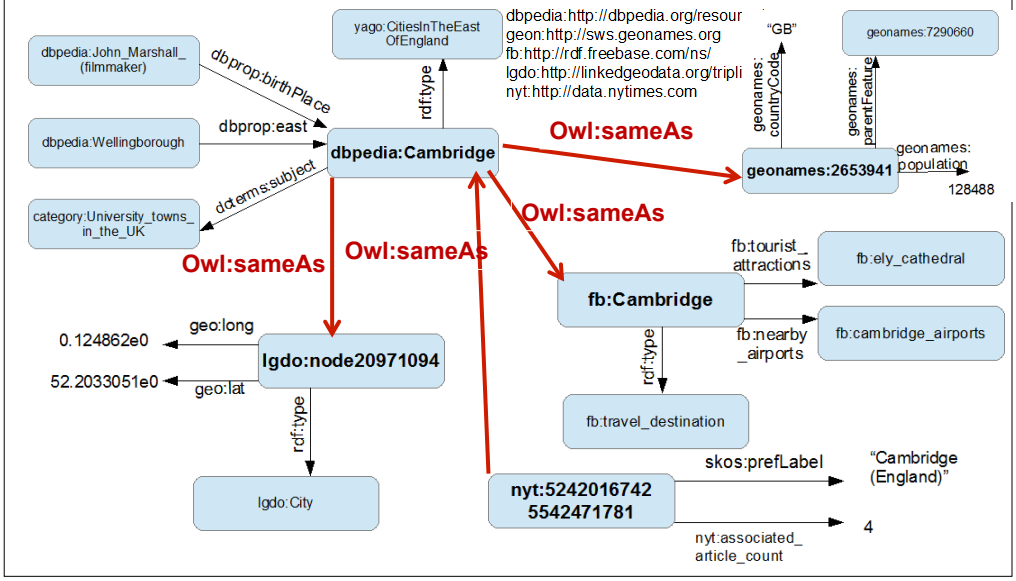


SW Standardization Activities in W3C

- Universal Resource Identifier (URI)-Internationalized Resource Identifier (IRI)
- Resource Description Framework (RDF) is a labeled directed graph model for web data
- RDF Schema (RDFS) allows to define the vocabulary of labels employed by web data
- Simple Protocol and RDF Query Language (SPARQL) provides graph pattern-matching support for web data
- Web Ontology Language (OWL) defines DL-based ontologies for reasoning over web data
- Rule Interchange Format (RIF) facilitates inference rule sharing and exchange from different languages and paradigms (first-order, logic programming/deductive databases, etc.)



An RDF Dataset Example

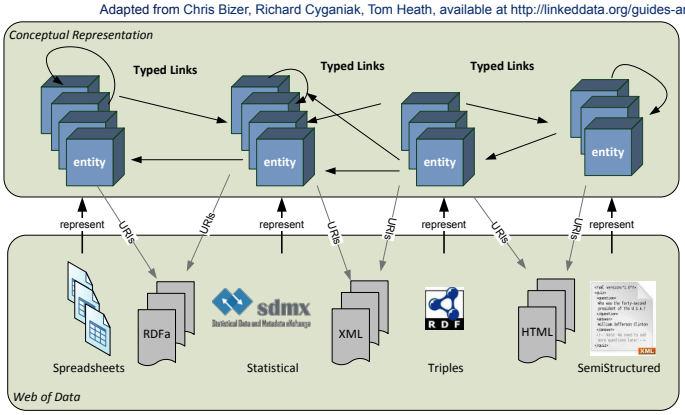


12

The Web of Data

A Web of things in the world, described by data on the Web

Adapted from Chris Bizer, Richard Cyganiak, Tom Heath, available at http://linkeddata.org/guides-and-tutorials



- Global data space connecting data from diverse domains and sources
  - Primary objects: "things" (or description of things)
  - Links between "things"
- Granularity of information: from entire data collections to atomic data

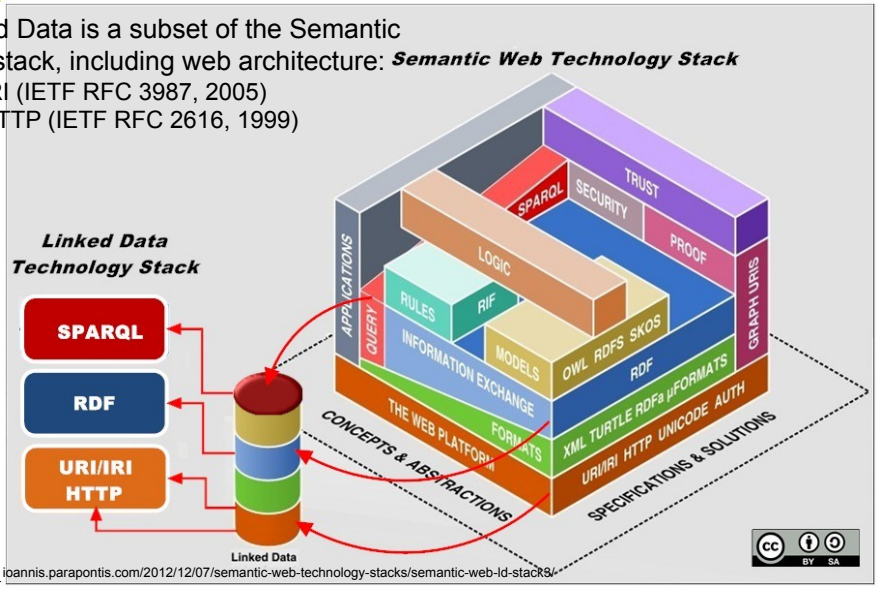
The Linked Data Principles

- Anyone can publish data on the Web by respecting a minimal set of syntactic conventions
  - Use URIs as names for things
  - Use HTTP URIs so that people can names
  - When someone looks up a URI, information
  - Include links to other URIs, so that discover more things

- Data is self-describing
  - Applications encountering data described by an unfamiliar vocabulary, they can resolve its URIs and understand the vocabulary terms by their RDFS or OWL definitions
- The Web of Data is open
  - Many common things are represented in multiple data sources
  - Discover new data sources at run-time by following links
  - Incremental data integration and provide mappings as you go

Linked Data Technology

Linked Data is a subset of the Semantic Web stack, including web architecture:
- IRI (IETF RFC 3987, 2005)
- HTTP (IETF RFC 2616, 1999)



ioannis.parapontis.com/2012/12/07/semantic-web-technology-stacks/semantic-web-ld-stacks/

Linked Data of Increasing Quality

Not all linked data is open and not all open data is linked!

★ Available on the web (whatever format) but with an open license, to be Open Data

★★Available as machine-readable structured data (e.g. excel vs. image scan of a table)

★★★ as (2) plus non-proprietary format (e.g. CSV instead of excel)

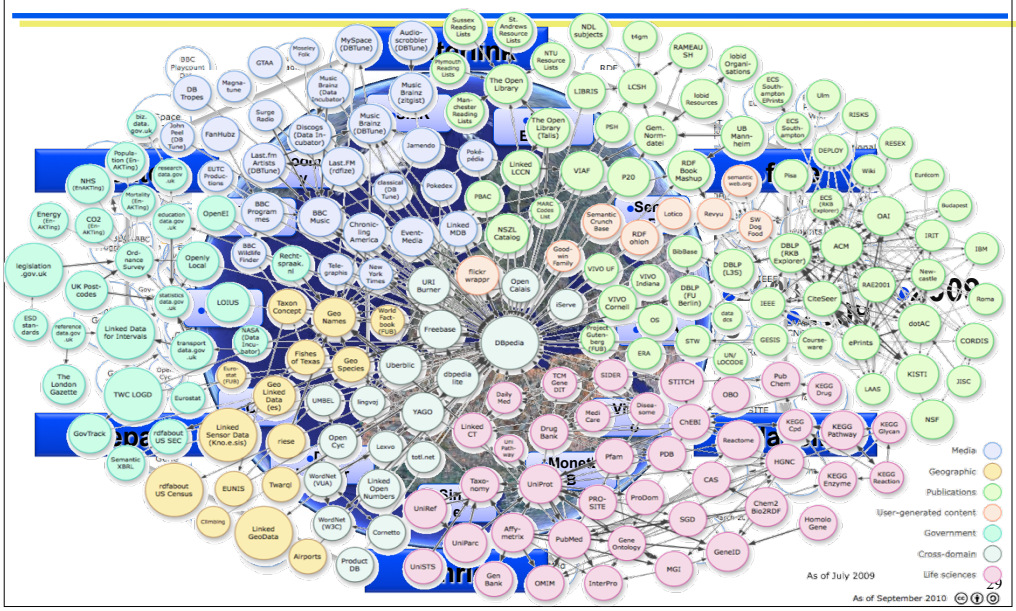
★★★★ as (3), plus using open standards from W3C (RDF and SPARQL) to identify things through dereferenceable HTTP URIs, to ensure effective access

★★★★★ as all the above plus establishing links between data of different sources

www.w3.org/DesignIssues/LinkedData.html

File format	Recommendations (on a scale of 0-5)
csv	★★★
xls	★
pdf	★
doc	★
xml	★★★★
rdf	★★★★★
shp	★★★★
ods	★★
tiff	★
jpeg	★
json	★★★
txt	★
html	★★

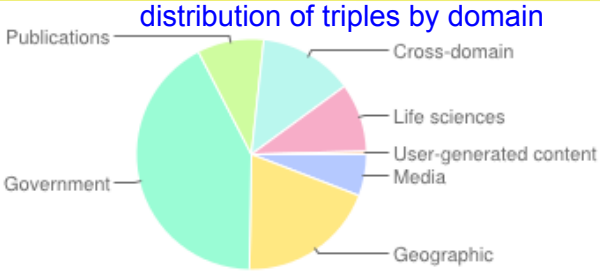
The Emerging Web of Data



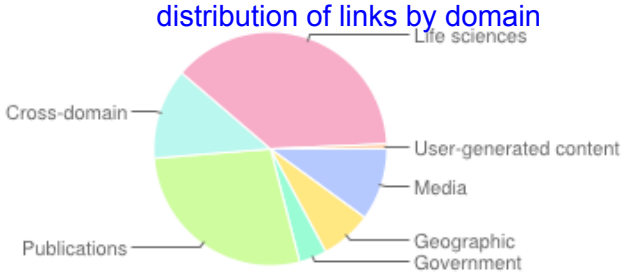
Types of Data in the Linking Open Data Cloud

- Over 60 % of all LOD sources use proprietary vocabularies
  - It's up to the data consumer to normalize the vocabularies
- Data sources that overlap in content use different identifiers for the same real-world entity
  - It's up to the data consumer to generate links

distribution of triples by domain



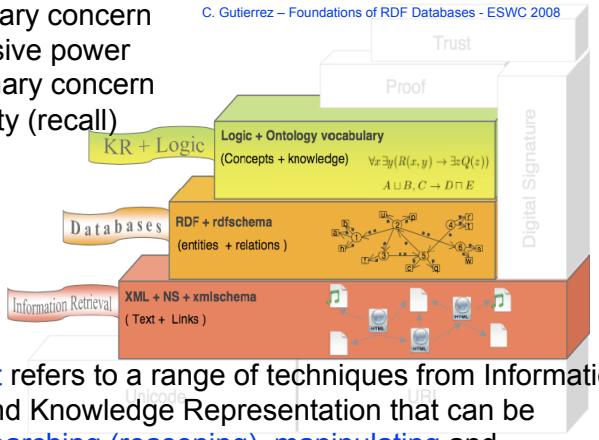
distribution of links by domain



http://www4.wiwiss.fu-berlin.de/lodcloud/state/ (Version 0.3, 09/19/2011)

Data Management at the Web Scale

As opposed to AI: DB primary concern is *scalability*. Then expressive power
As opposed to IR: DB primary concern is *precision*. Then scalability (recall)



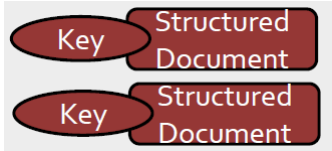
- Web Data Management refers to a range of techniques from Information Retrieval, Databases and Knowledge Representation that can be employed for storing, searching (reasoning), manipulating and aggregating data on the web based on its meaning
  - aims to support a more comprehensive usage of larger scale and more complex semantic datasets at lower cost

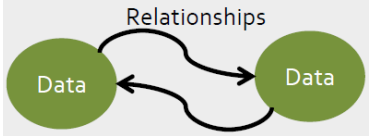
The Data Infrastructure Universe

- Key Value Stores
  - Schema-less system
- Column-oriented databases
  - Storage by column, not row

Name	Height	Eye Color
Bob	6'2"	Brown
Nancy	5'3"	Hazel

- Document Oriented Database
  - Stores documents that are semi-structured
  - Includes XML databases
- Graph Databases
  - Uses nodes and edges to represent data
  - Includes RDF stores
- Shaded RDBMS





The Data Integration "Sextant"

Data Integration	Database World	Web of Data	Web of Documents
virtual	Query-Mediators	Entity-centric Search	Meta-Search Engines
physical	Data Warehouses	Linked Data Portals	Text Search Engines

Execution Time: high / low

high | medium | low

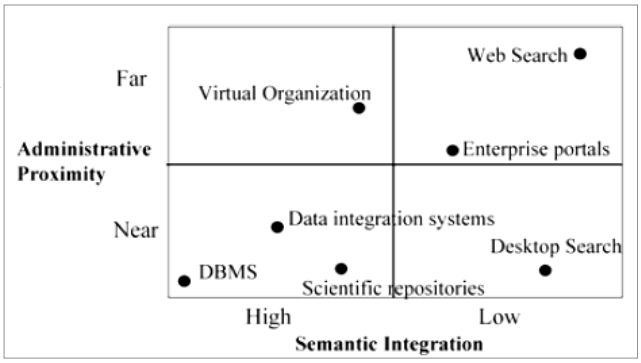
Schema/Structure Integration

Scalability: low → high

A Space of Existing Data Management Solutions

From Databases to Dataspaces: A New Abstraction for Information Management Michael Franklin, Alon Halevy and David Maier

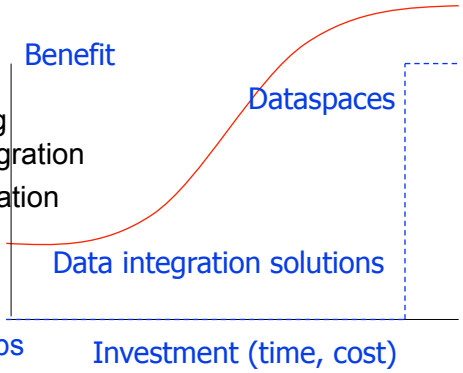
- More and More complex and heterogeneous environments
  - Many different types of systems
  - Many inter-related applications



- Administrative Proximity indicates how close the various data sources are in terms of administrative control
- Semantic Integration is a measure of how closely the schemas of the various data sources have been matched

DataSpaces: Collaborative Data Sharing Systems

- Data management from the enterprise to the masses:
  - Need support for collaboration
  - Help people structure their data
  - Pay-as-you go data management
- Data co-existence approach improving scalability w.r.t. the "schema first" integration
- A dataspace contains all of the information relevant to a particular organization regardless of its format and location
- We model a dataspace as a set of participants (datasets) and relationships (correspondences or mappings)
  - Some participants may support expressive query languages, while others are opaque and offer only limited interfaces for posing queries



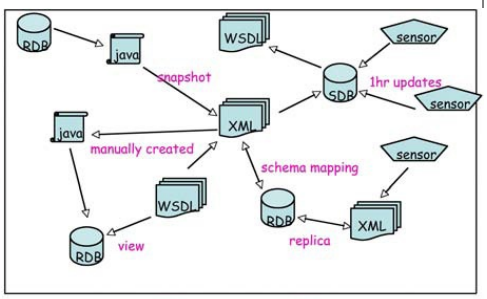
Dataspace Systems Architecture and Services

- Catalog and browse
  - Collection of data sources (schema, rate of change, accuracy…)
- Search and query
  - Query everything
  - Structured queries
  - Metadata queries
  - Monitoring
- Local store and index
  - Store associations between objects, increase availability,…
- Discovery
  - Locate new databases
- Source extension
  - Add query functionality,…

Catalog

Local Store & Index

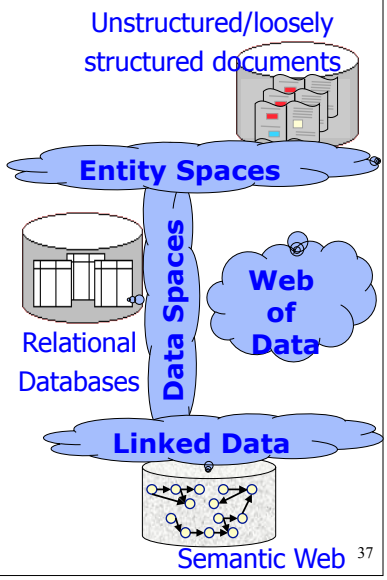
Search & query



Administration | Discovery | Enhancement

The Web of Data = Dataspaces + Web Semantics

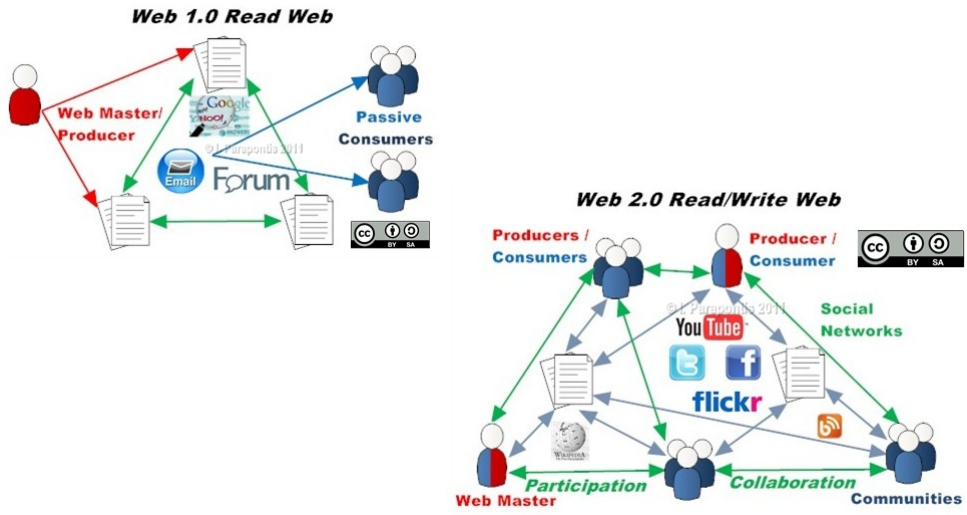
- Global information space where a variety of data co-exist along with a minimal structuring and naming conventions
- Participants
  - data published according to LOD guidelines and represented as fine-grained RDF triples
  - at best only partial knowledge of data semantics by ontologies/schemes
- Relationships
  - links constructed by humans (same as etc.) or by machines ("is view", "schema mapping", "created independently", ...)
- Pay-as-you go data management
  - The overall data integration effort is split between the data publisher, the data consumer and third parties



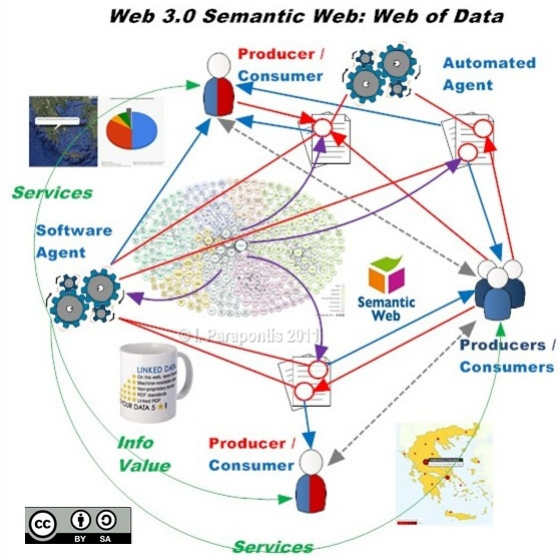
Few Discrepancies

- **Linked Data**
  - Method for decentralised data publishing and interlinking
  - Ecosystem (incl. people)
  - m:n mappings
  - Many small sources
  - Decentralised interlinking
  - No central catalog
- **Dataspaces**
  - Comprehensive architecture for data integration
  - Platform
  - 1:m mappings
  - Few large sources
  - Links in the local index
  - Central catalog

From Web 1.0 and 2.0



To Web 3.0

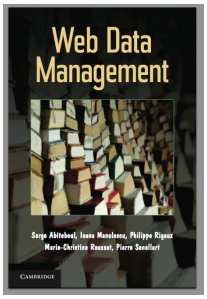


Tentative Course Schedule

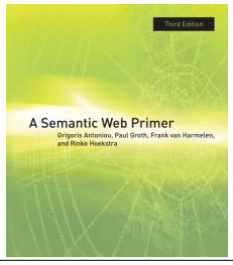
- Week 1 (Feb 26-28): Intro on the Web and Big Data
- Week 2 (March 5–7): XML, XML Schema
- Week 3 (March 12-14): XPath, XQuery
- Week 4 (March 19-21):
- Week 5 (March 26-28): RDF, RDFS, Linked Data
- Week 6 (April 2-4): SPARQL
- Week 7 (April 9-11): SPARQL
- Week 8 (April 16-18): Data Integration Architectures
- Week 9 (April 23-25): Query Mediation and Federation
- Week 10 (Mai 14-16): Data Mapping and Entity Resolution
- Week 11 (Mai 21-23): Students presentations
- Week 12 (Mai 28-30): Students presentations

Course Text Books

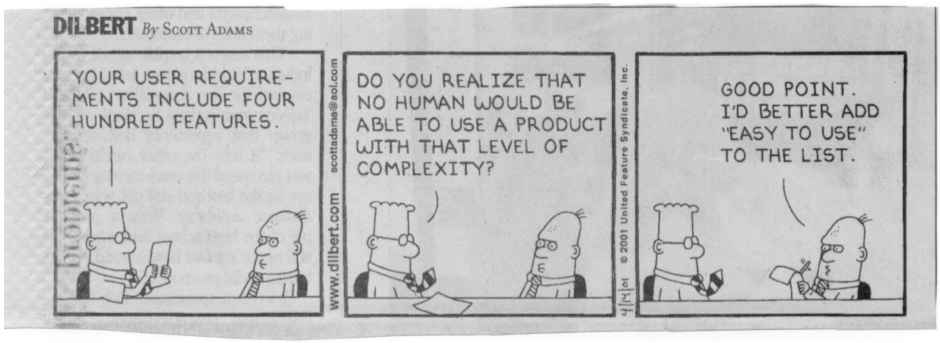
- *Web Data Management*, By Serge Abiteboul, Ioana Manolescu, Philippe Rigaux, Marie-Christine Rousset, Pierre Senellart Cambridge University Press, 2011
  - http://www.cambridge.org/gb/knowledge/isbn/item6564193/?site_locale=en_GB
  - Free download of a previous version webdam.inria.fr/Jorge/files/wdm.pdf



- *A Semantic Web Primer, Third Edition* By Grigoris Antoniou, Paul Groth, Frank van Harmelen and Rinke Hoekstra, MIT Press 2012
  - http://mitpress.mit.edu/books/semantic-web-primer-0
  - Free download of a previous version http://www.csd.uoc.gr/~hy566/SWbook.pdf



Course Organization



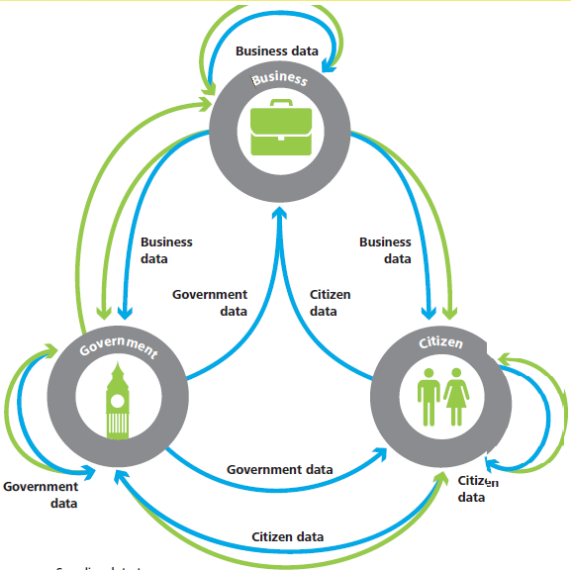
- Two Programming Exercises (30%):
  - XML and XPath
  - RDF and SPARQL
- A Research Project (30%):
  - Presentation of 2 research papers and written report
- Final Examination (40%):

Research Project Topics

- Adding Legacy Data to the Web of Data
  - Publish relational DB to the Web of Data: (e.g., R2RML: fixed schemas)
  - Extract data from Web documents (e.g., DBPedia: extraction from Wikipedia)
  - Convert to RDF data from other formats (e.g., RDFizers for Excel, JSon)
- Data Integration in the Web of Data
  - Crawl Linked DataSets (e.g., Sindice, Swget, LDSpider)
  - SPARQL Federation and Query Mediation (e.g., FedX, DARQ)
  - Data Mapping and Entity Resolution (e.g. Silk, Ldif)
  - Data Quality and Provenance (e.g., abstract SPARQL prov models)
- Web Data Storage and Access
  - SPARQL Engines and Optimization (e.g. native, rowstores and columnstores)
  - Continuous Query Processing (e.g., C-SPARQL)
  - Multi-Query Processing
  - Large Scale Data Management (e.g., Map-Reduce and Cloud)
- Benchmarking

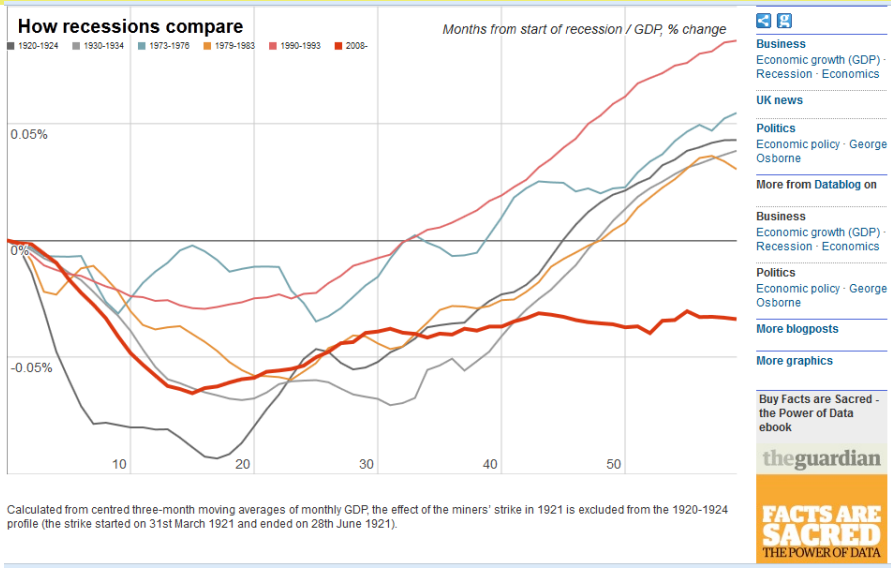
Killer Applications: Data Journalism

The Open Data Ecosystem



Source Deloite LLP

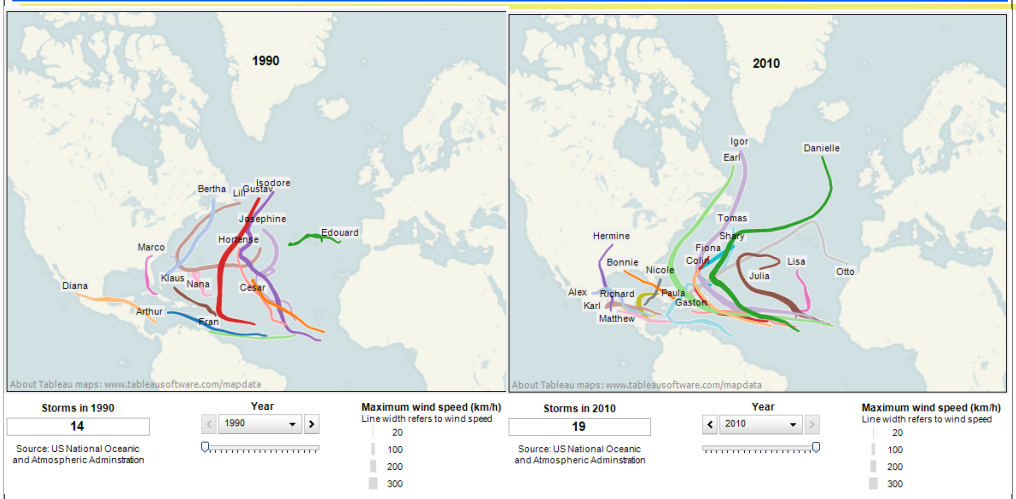
Recessions Compared: how does Britain's GDP Compare to Every Recession Since 1930?



Calculated from centred three-month moving averages of monthly GDP, the effect of the miners' strike in 1921 is excluded from the 1920-1924 profile (the strike started on 31st March 1921 and ended on 28th June 1921).

Friday 25 January 2013 www.guardian.co.uk/news/datablog/interactive/2012/sep/27/recessions-compared-1930-2008

Storm Warning: Prepare for a Busy Hurricane Season



12:42 31 May 2011 www.newscientist.com/blogs/shortsharpscience/2011/05/prepare-for-hurricane-season.html

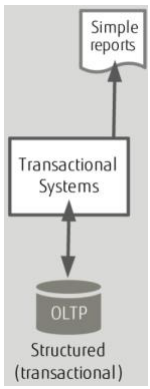
Που πάνε οι Φόροι μου;

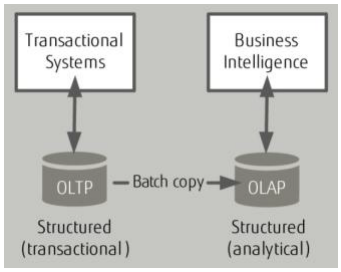
Στο διπλανό γράφημα εμφανίζεται η διαχρονική εξέλιξη των δημόσιων δαπανών που καταχωρίζονται στη ΔΙΑΥΓΕΙΑ από την έναρξη της μέχρι σήμερα. Τα ποσά έχουν ομαδοποιηθεί ανά εβδομάδα.

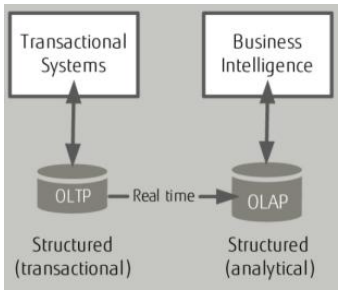
<< Γραφήματα

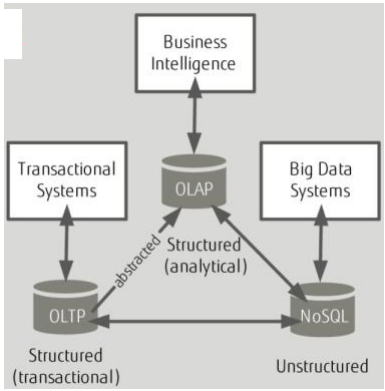
08/02/2013 http://publicspending.medialab.ntua.gr/tplotpaymentsweek.php

A Parallel with DBMS Evolution







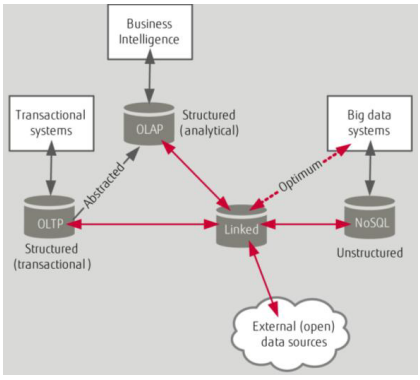


Fujitsu White Paper Linked Data: March 2012

51

A Parallel with DBMS Evolution

Data Model	Attributes	Flexibility	Age of data	Quality of data
Transactional	Known	Fixed schema	Short lived/current	High
Analytical	Aggregated	Fixed schema	Longer life/historical	(Typically) High
Unstructured	Unknown	Implied	Random	Low
External	Variable	Variable	Variable	Variable



52