Linked Data: from theory to implementation and application

Biswanath Dutta
bisu@drtc.isibang.ac.in

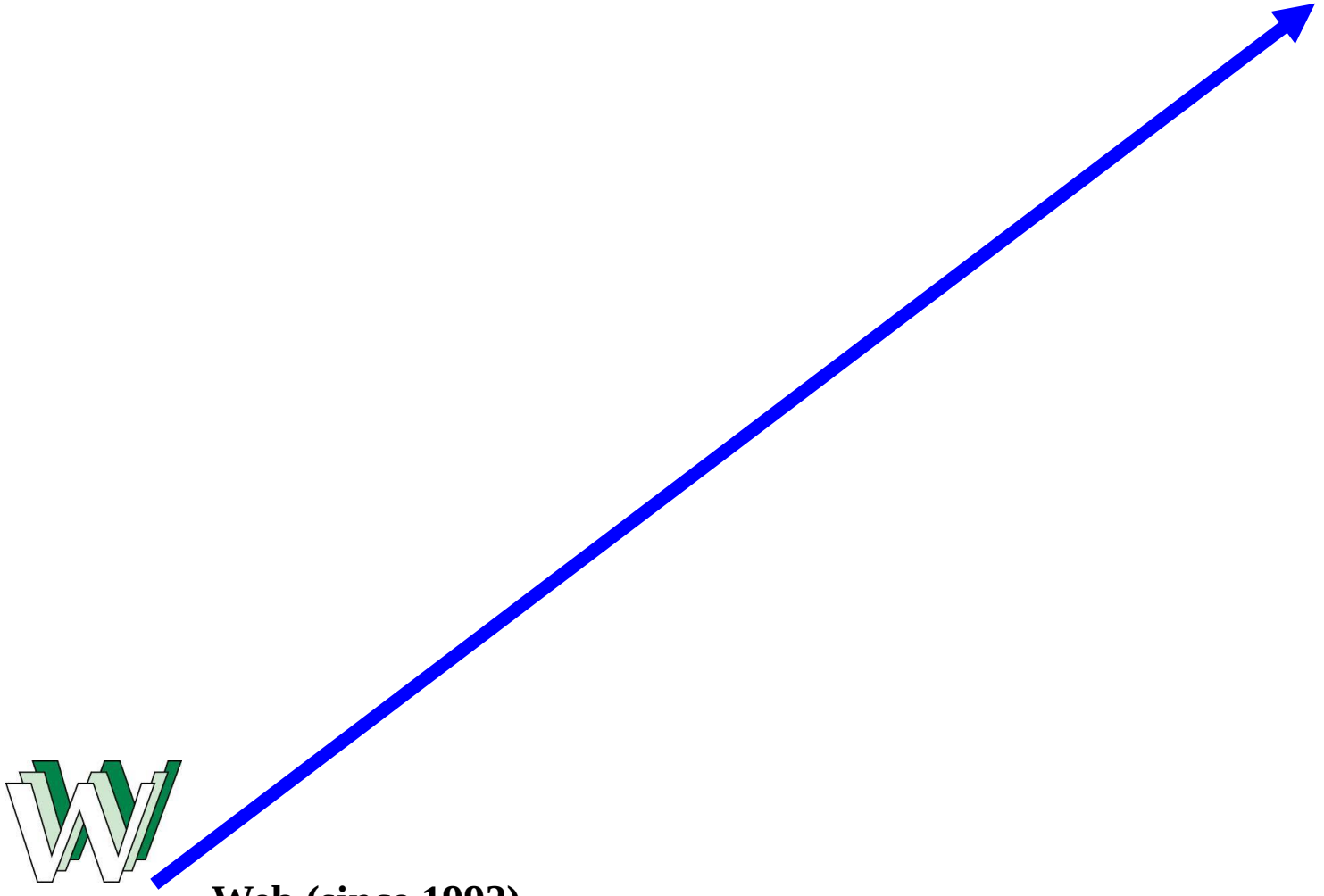
Documentation Research and Training Centre
Indian Statistical Institute
Bangalore, India

Overview

- Transition from Document Web to Data Web
- What is Data?
- Linked Data Fundamentals
- How to Create and Publish Linked Data?
- Linked Data Benefits
- Applications and Use Scenarios
- Some Open Issues
- Conclusion

Document Web to Data Web

Transition from **Web of Document** to **Web of Data**



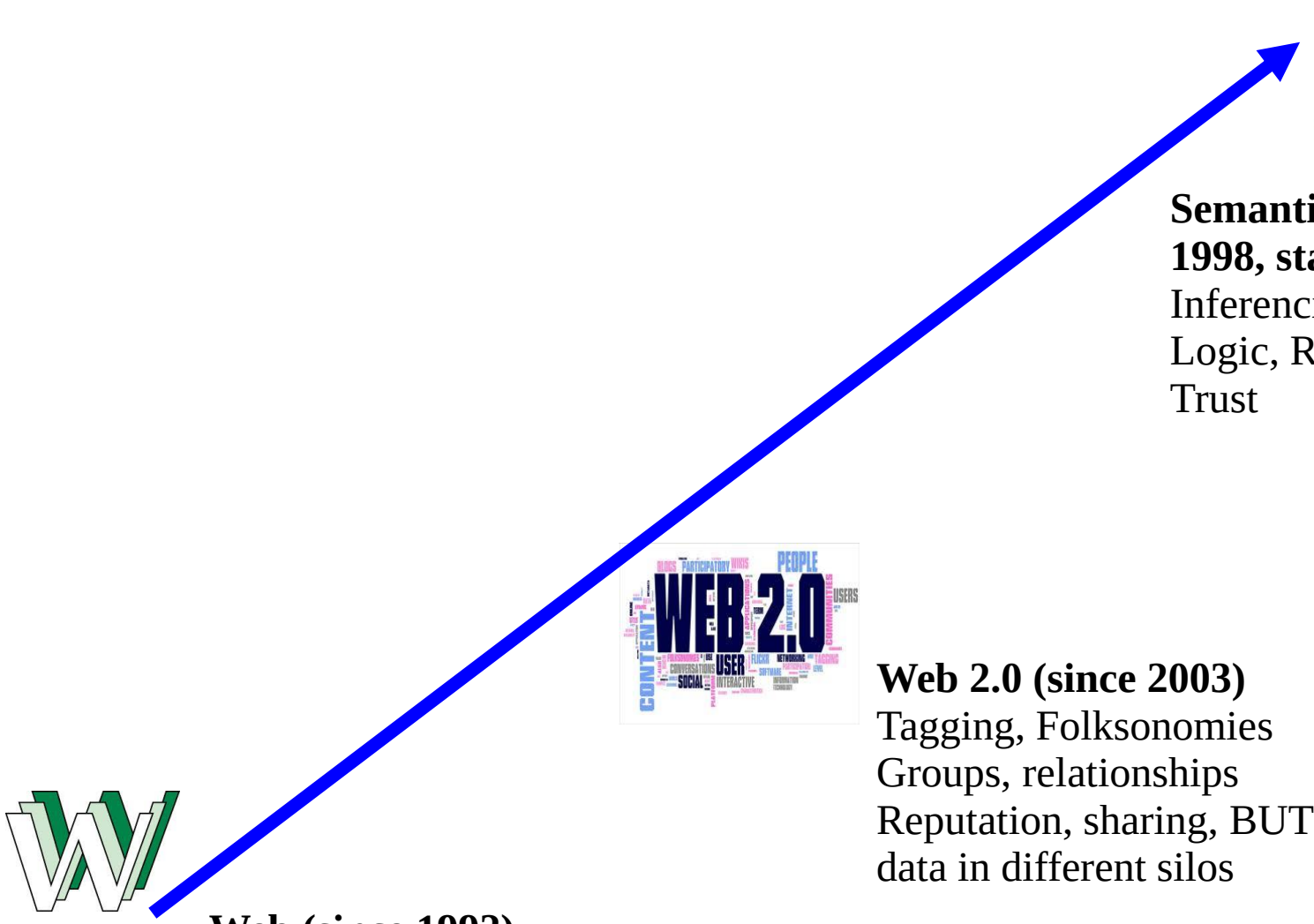
Web (since 1992)
HTTP, URL(as global unique Ids)
HTML (shared format, not structured, no expressive queries can be asked), CSS, JavaScript

Source: [2]

Can we Ask These Questions?

- **Search answers** for the following queries in the current search engines.
  - *Researchers actively working on semantic technology related topics in India.*
  - *Books on ontology related topics that are written by an Italian author stays in Trento.*
  - *Apartments near Bengali dominated area in Bangalore.*
  - *Guided tour providers with offices in Jaipur, Delhi and Bangalore.*
  - *Name of people employed by Government agencies in the year 2007.*
- The required information to answer the above queries is available on the Web, but the current Web search engines are not yet smart enough to understand and answer the queries.

Transition from **Web of Document** to **Web of Data**



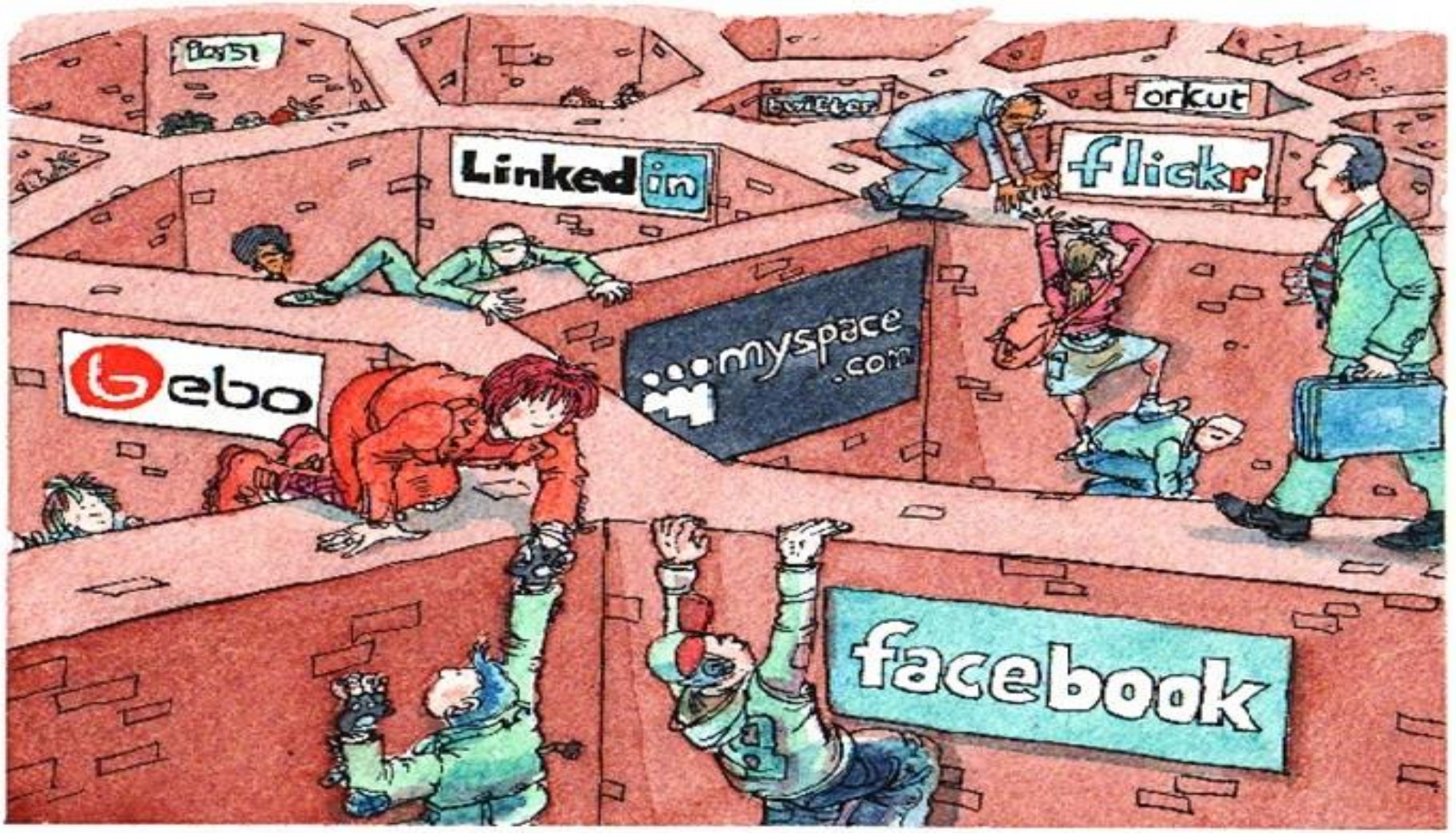
Semantic Web (vision 1998, started??)
Inferencing
Logic, Rules
Trust

Web 2.0 (since 2003)
Tagging, Folksonomies
Groups, relationships
Reputation, sharing, BUT
data in different silos

Web (since 1992)
HTTP, URL(as global unique Ids)
HTML (shared format, not structured, no expressive queries can be asked), CSS, JavaScript

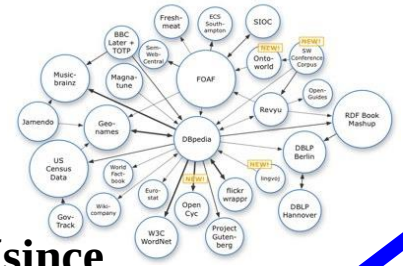
Source: [2]

Data in Silos (Web 2.0)



Source: Chris Bizer, Freie Universität Berlin, SRI International, Artificial Intelligence Center Menlo Park, USA, 24 July 2009

Transition from **Web of Document** to **Web of Data**



Semantic Data Web (since 2006)
De-referencability
RDF serializations

Semantic Web (vision 1998, started??)
Inferencing
Logic, Rules
Trust

Web 2.0 (since 2003)
Tagging, Folksonomies
Groups, relationships
Reputation, sharing, BUT
data in different silos

Web (since 1992)
HTTP, URL(as global unique Ids)
HTML (shared format, not structured, no expressive queries can be asked), CSS, JavaScript

Source: [2]

Semantic Data Web

- **Semantic Data Web** is a Web which is to complement text on Web pages with structured data and to intelligently **combine** and **integrate** such structured information from different sources.

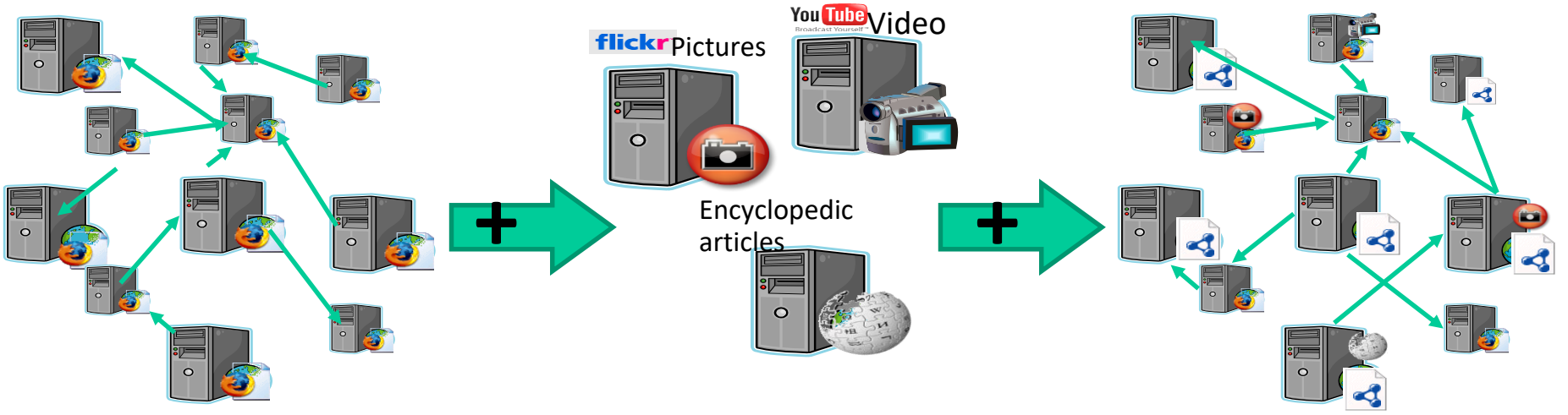
- Allow applications to operate on top of an unbounded set of data sources, via standardised access mechanisms.

Transition of Web

Web 1.0
Many Web sites consisting of unstructured, textual content

Web 2.0
Few large Web sites specialized with specific content types

Web 3.0
Many Web sites containing and semantically syndicating arbitrary structured content

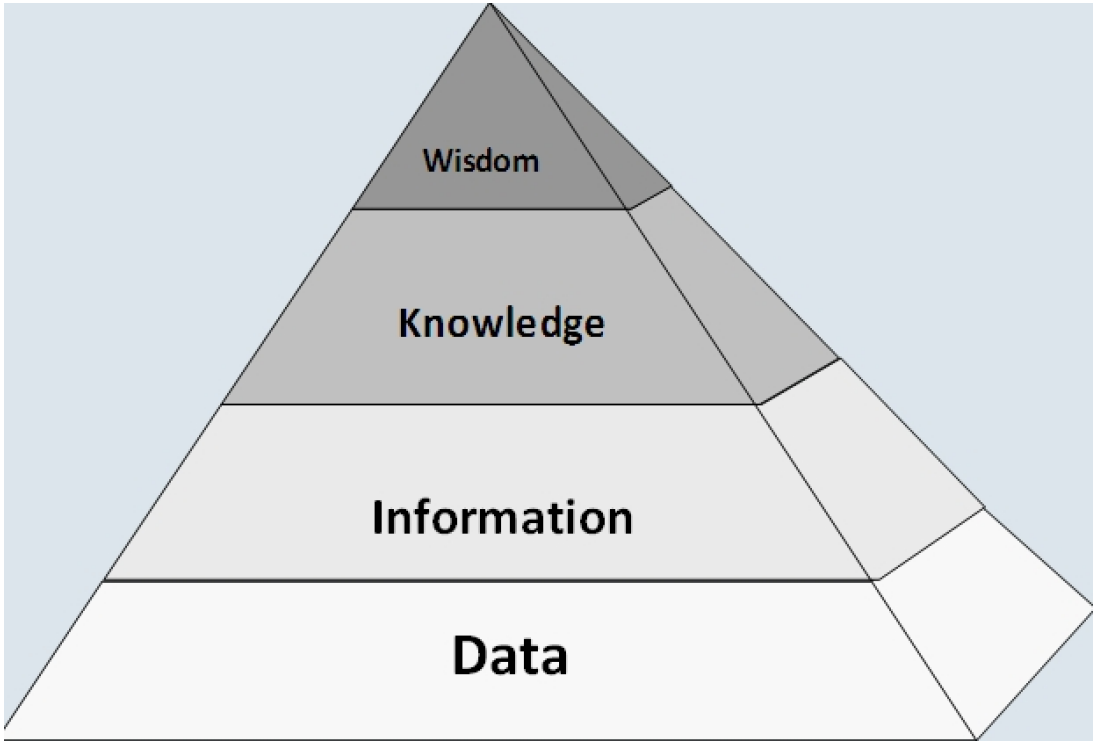


Source: [2]

What is Data?

Why we talk about Data?

Data is the basis for Information, Knowledge, and Wisdom, and what else??



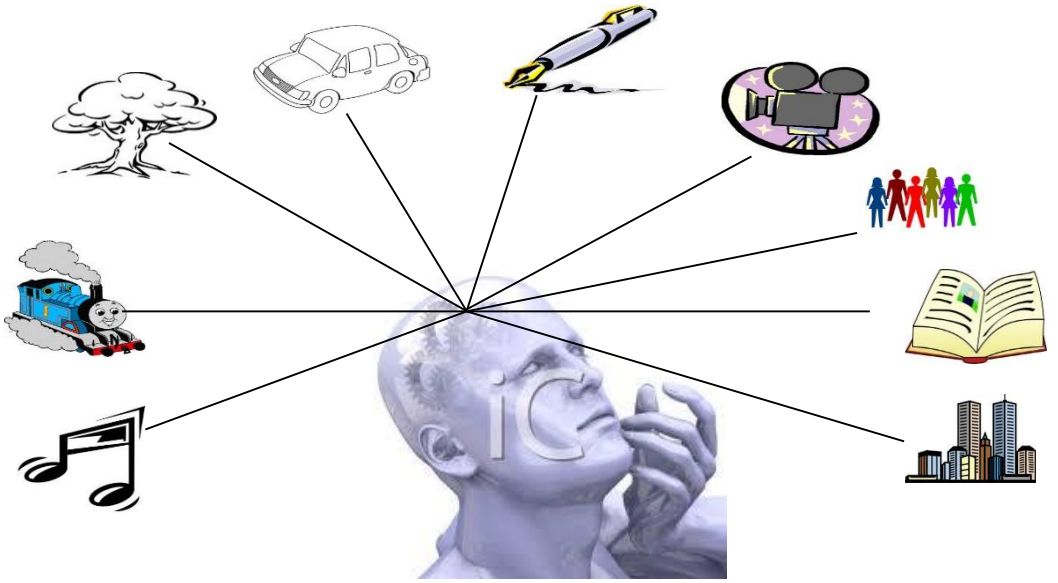
Source:

So, What is Data?

How we express observation in reusable form.

What is Observation?

- It is an entity relationship perception
- (entities are the distinctly identifiable things)



- Entity Relationships are expressed using statements constrained by language (signs, syntax, and relation semantics).
- Statements are represented using a variety of notations; persisted to paper or digital documents; and transmissible using a variety of serialization formats.

How Entity Relationships are Expressed?

- Entity relationships are expressed as statements constrained by language (signs, syntax, and relation semantics).
    - Entity Relationship Model (Network / Graph) Diagrams
    - CSV Tables (Spreadsheets & SQL Relational Databases)
    - RDF-Turtle, JSON-LD, RDF/XML, HTML+Microdata, HTML+RDFa Statements.
- Statements are represented using a variety of notations associated with a language.
- Statements are transmissible using a variety of serialization formats.

How Entities are Related?

John

Alex

Associated through some relationships
(e.g., teacher, student)

Entity **Attribute** **Value**

Entity Roles in Relationships

All entities play some role in a relationship

John

Entity (Subject)

teacher of

Attribute (Predicate)

Alex

Value (Object)

Types of Values

- Untyped Literals (Strings)

- Typed Literals
  - Numbers
  - Date
  - DateTime
  - Booleans
  - Float
  - Etc...

- References (Links)

Relationship Role Types

- **Entity-Attribute Value (EAV) model:**
  - Entity (observation focal point)
  - Attribute (observation attribute name)
  - Value (observation attribute value)

- **W3C's Resource Description Framework (RDF):**
  - Subject (observation focal point)
  - Predicate (observation attribute name)
  - Object (observation attribute value)

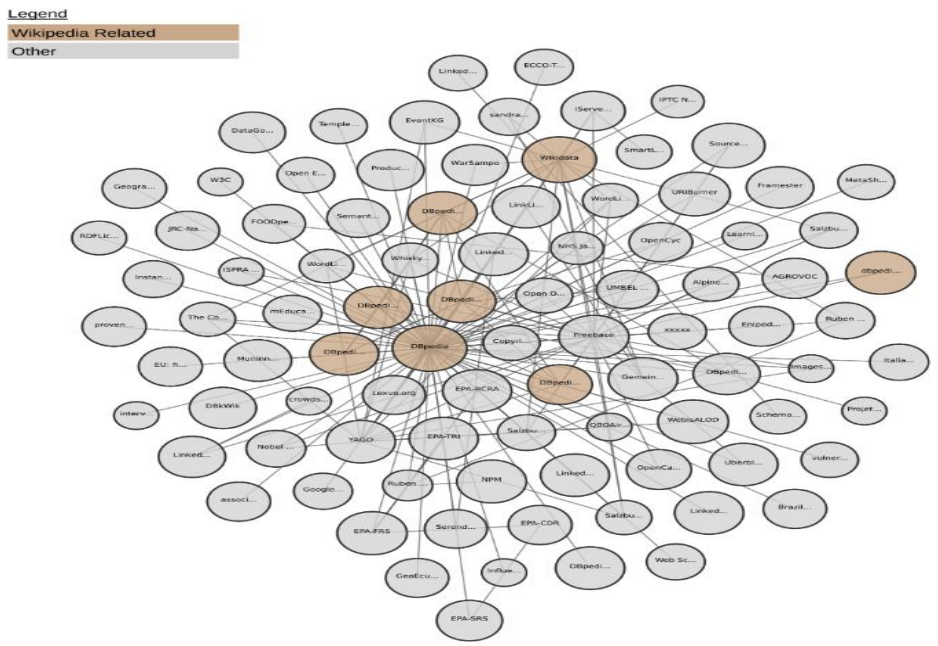
How an Entity is denoted?

- Entities are denoted through the **identifiers** (an identifier is a Sign (or Token) that Signifies (Names, Denotes or “Refers To”) an Entity.)
    - Absolute reference:
    - http://drtc.isibang.ac.in/people#biswanathDutta
    - Relative reference:
    - #biswanathDutta

Linked Data Fundamentals

What is Linked Data

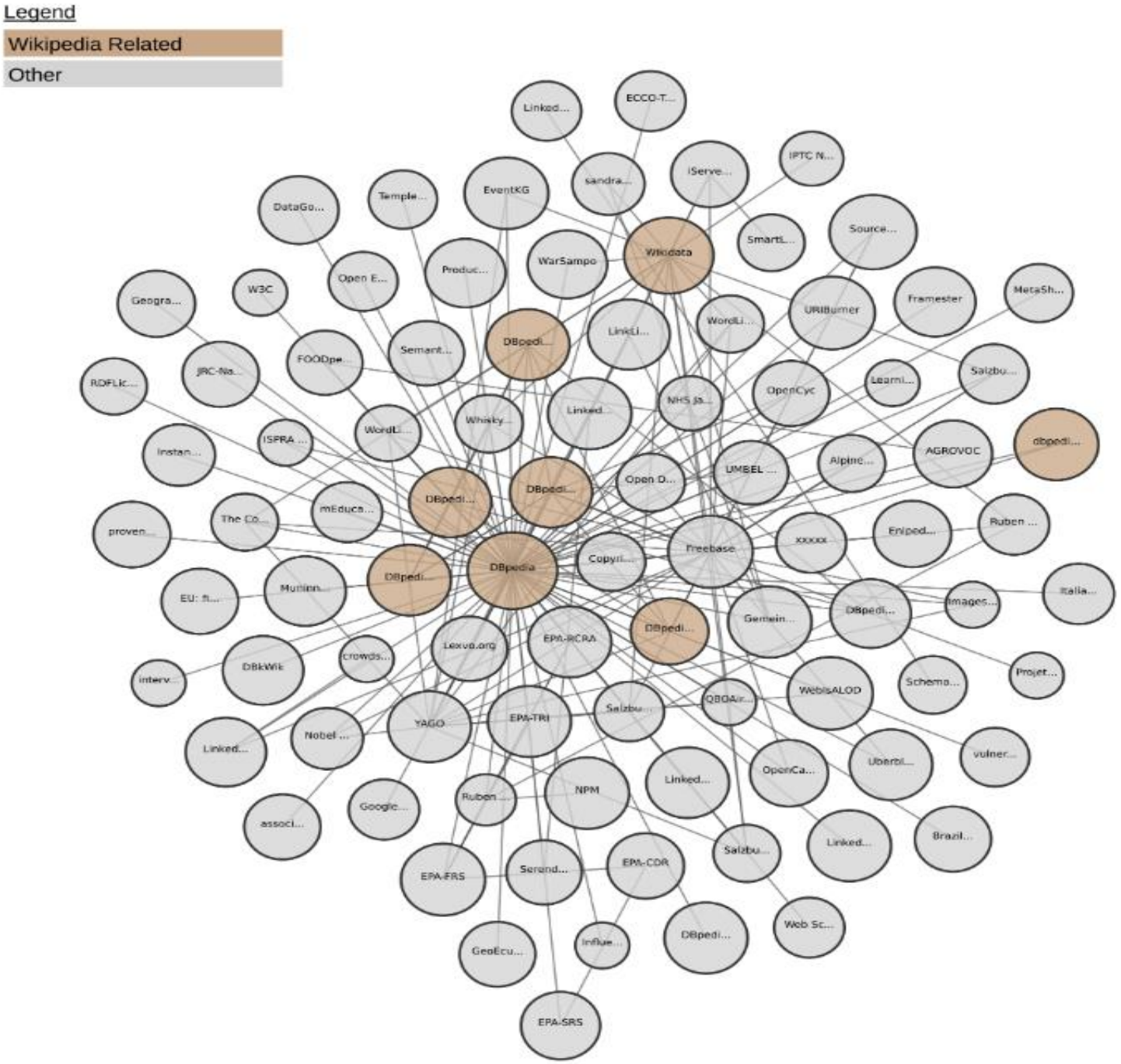
“A structured data interlinked with other data built upon the web standards, such as URI, HTTP URI, and RDF.”



Linked Data Cloud (https://lod-cloud.net/)

Open Linked Data

Is a Linked Data that is released under an open license and does not obstruct its reuse for free.

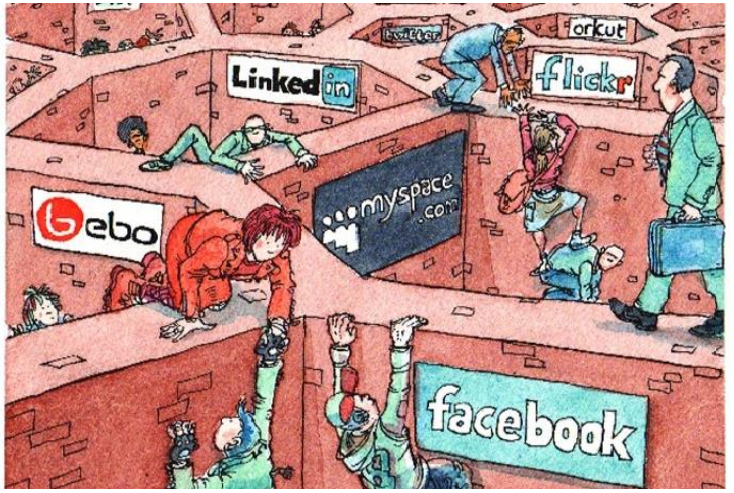


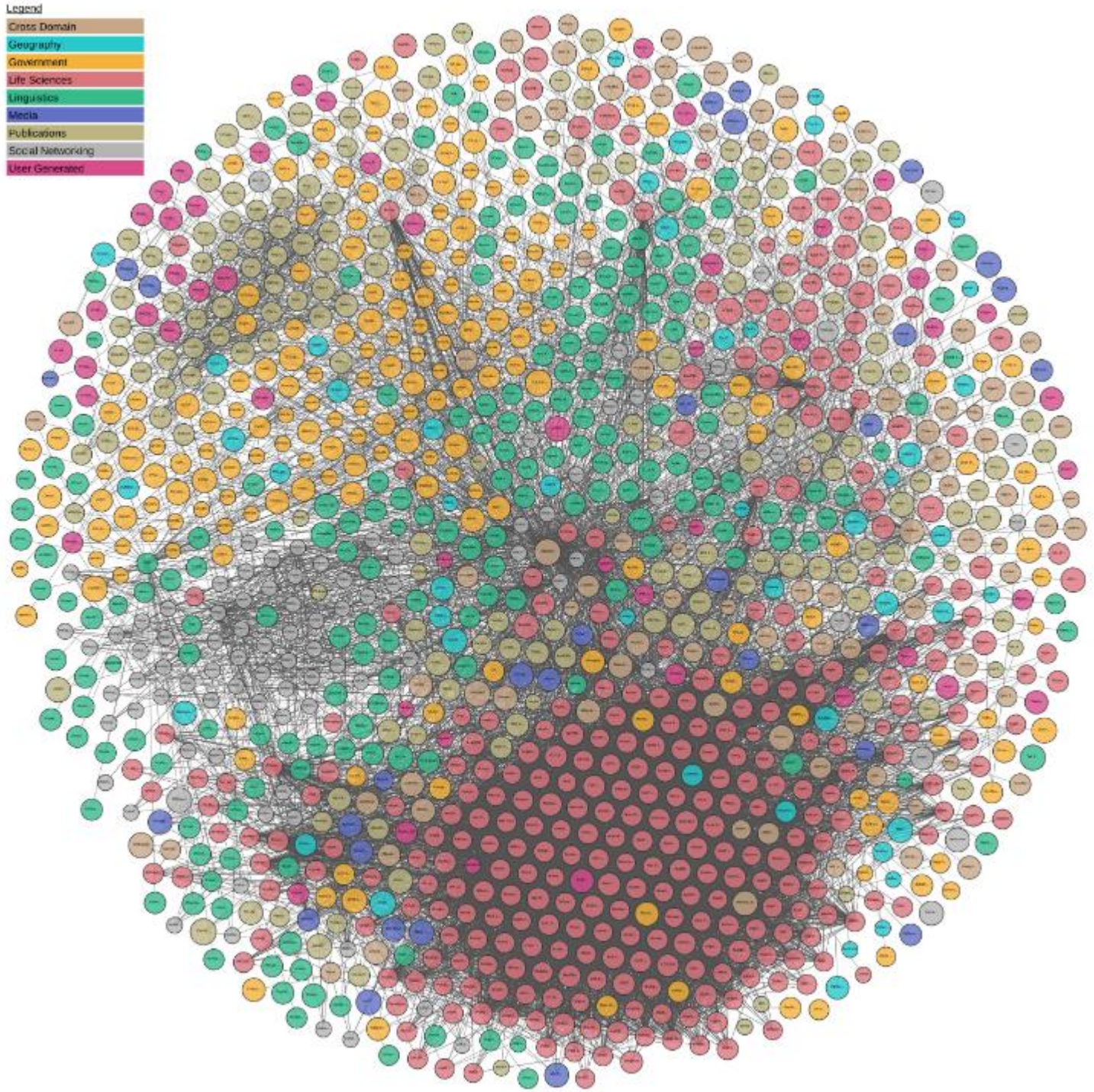
Subclouds by Domain (https://lod-cloud.net/)

Goal of Linked Data

The ultimate **goal** is to evolve the Web like a single global database (a global discovery space) by integrating the data across disparate data silos

so that both human and machine can explore and make optimum use of available data on the Web.



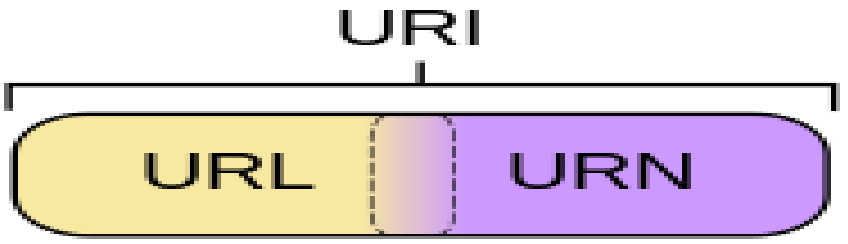


Linked Data Cloud (https://lod-cloud.net/)

Linked Data Technology Stack

- **URIs**: use URIs to name and denote entities unambiguously.
- **HTTP URIs**: use HTTP (data access protocol) URIs, so that people can look up those names. Also, the description of entities can be looked up using HTTP user agent.
- **RDF** (Resource Description Framework): use RDF to create human and machine readable statements describing the entities.
- **Vocabulary**: to describe data.

Uniform Resource Identifier (URI)



- "A Uniform Resource Identifier (**URI**) provides a simple and extensible means for identifying a resource." -- RFC 3986
    - Examples
        - http://drtc.isibang.ac.in/people/Biswanath_Dutta
        - http://dbpedia.org/page/Tajmahal
- The **URL** https://www.facebook.com/natgeo identifies the location from where a Web page can be retrieved.
- The **URN** urn:isbn:3-549-35469-6 identifies a book using its ISBN. Here ISBN system uniquely identifies books.

RDF

- RDF is an abstract data model
- It is allows structured data representation.
- It is a general method for conceptual description or modeling of information that is implemented in WWW, using a variety of serialization formats (e.g., XML, Notation 3 (N3), N-Triples, Turtle, etc.).
- It is based on the idea of identifying things using Web identifiers (i.e., *Uniform Resource Identifiers*, or *URIs*)
- RDF is to represent Entity Relationships and Relation Semantics using Statements.
- A RDF statement is tuple of <*Subject*, *Predicate*, *Object*>, where:
  - ***Subject*** is the resource, which is being described.
  - ***Predicate*** is a resource, which determines the type of the relationship.
  - ***Object*** is a resource , which represents the value of the attribute.

Entities in Relationships with Identifiers

http://drtc.isibang.ac.in/teacher/john

http://drtc.isibang.ac.in/student/alex

http://drtc.isibang.ac.in/education/element/teacherOf

Subject

Predicate

Object

Triple:
<http://drtc.isibang.ac.in/teacher/john, http://drtc.isibang.ac.in/education/element/teacherOf, http://drtc.isibang.ac.in/student/alex>

Vocabulary

- Vocabulary is required to describe your data.
- RDF is a framework, and does not have vocabulary to describe data.
- RDFS and OWL provide general (domain independent) vocabulary to describe data.
  - RDFS(RDF vocabulary Description Language, also known as RDF Schema) (e.g., rdfs:label, rdfs;resource, rdfs:comment, rdf:type)
  - OWL(The Web Ontology Language) (e.g., owl:sameAs, owl:equivalentProperty)
- Of course, we need more domain specific vocabulary to further describe our data.

Linked Data: Benefits, Applications and Use Scenarios

Linked Data Benefits

- Integrated Access to Data
- Data Enrichment
- Decentralization
- Organizational (!!) Visibility
- Shared Data (empowers the idea of re-usability!!)
- Data Maintenance, Data Currency
- Data Durability and Robustness - Linked Data retains its meaning across changes of format
- Federated Search Facility
- Promote Interdisciplinary Research
- Enhanced Peer Review
- Independency from Specific Data Format (!!)

Use Scenarios

Use Case I: Transparency: Some Government agencies are providing data on the Web, but in non-standard formats (Excel, PDF, etc). The lack of standardization of such data creates barriers for third parties to consume and analyze data, effectively reducing the transparency of government. Data related to **public spending, procurement** and **contracts** can be made available as Linked Open Data, facilitating the creation of applications for consumption and analysis of these data by third parties, increasing the effective transparency.

Use Case II: Government-Population Collaboration: Data of intelligence containing statistics on social assistance, education and violence can be made available as Linked Data, facilitating that others to consume and analyze data. This process maximizes the return on investment in collecting and curing of the data.

(Source: http://greco.ppgi.ufrj.br/gtlinkedbr/?q=en/usescenarios)

Use Scenarios

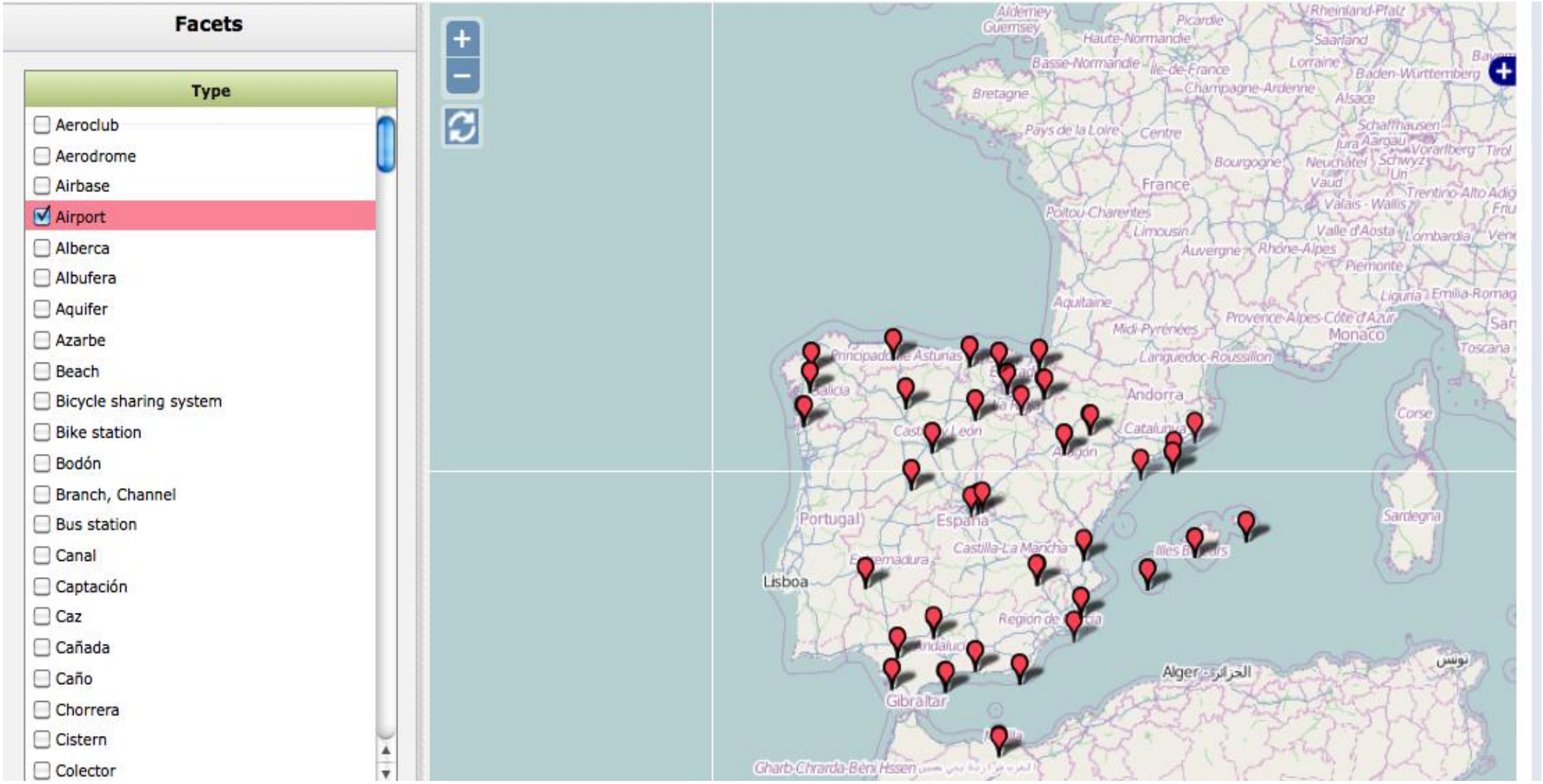
Use Case III: Visibility: Public utility services such as **trade information**, **transportation, events, traffic information**, etc. can be stored in Linked Data repositories for the development of various applications. For instance, a mobile application can be developed pointing the trade and public utilities for the tourists, major events happening around the city, and so forth.

Use Case IV: Inter-sectoral Integration: Databases of the government sectors can be integrated through Linked Data techiniques. This would allow different government sectors to work together in close proximity to the solution of common problems. Departments of social assistance, for example, can use data of education and health to make effective strategies to solve social problems.

(Source: http://greco.ppgi.ufrj.br/gtlinkedbr/?q=en/usescenarios)

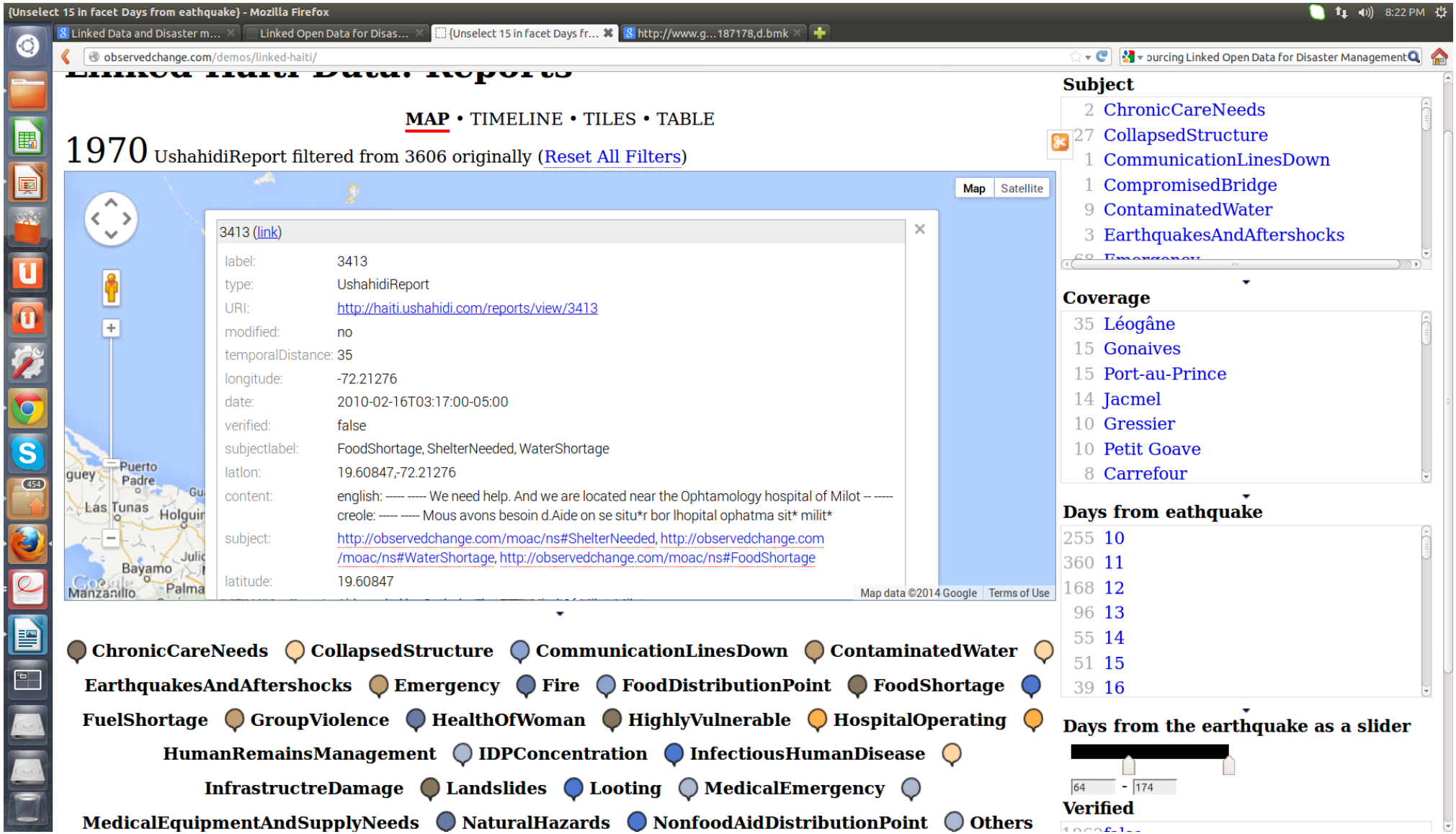
Applications

GeoLinked Data (http://geo.linkeddata.es/web/guest/visualizacion-beta)



Applications

Linked Haiti Data:Reports (http://observedchange.com/demos/linked-haiti/)



How to Create and Publish Linked Data?

What Data you have?

- Scientific, Experimental Data
- Statistical Data
- Business Data
- Library Data
- …

Where your Data Resides?

- In relational databases (MySql, PostgreSql)?
- In Spreadsheets?
- XML?
- Text files?
- …

Relational Data to RDF

DRTC:Book

id	title	author	year	price
1	Lord of Light	Roger Zelanzy	1967	INR 290
2	Creatures of Light and Darkness	Roger Zelanzy	1969	INR 310

```
@prefix dc: <http://purl.org/dc/elements/1.1/>
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix dcterms: <http://dublincore.org/documents/dcmi-terms/>
@prefix drtc:<http://drtc.isibang.ac.in/triples/book/>
@prefix shop:<http://www.example.com/doctype/element/>

drtc:book1   dc:title          drtc:Lord_of_Light
drtc:book1   rdf:type          shop:Book
drtc:book1   dc:creator        drtc:Roger_Zelanzy
drtc:book1   dcterms:created   “1969”^^xsd:date
drtc:book1   shop:price        “INR290”
```

DBPedia: Person

id	name	livesIn
1	Roger Zelanzy	USA
2	Chetan Bhagat	India

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix foaf: <http://xmlns.com/foaf/0.1/>
@prefix dbpedia: <http://dbpedia.org/resource/>

dbpedia:person1   foaf:name          “Roger Zelanzy”
dbpedia:person1   rdf:type           dbpedia:Person
dbpedia:person1   dbpedia:livesIn    “India”
```

RDF Reconciliation: Everyone Wins

- Your RDF Data can be reconciliated with other Linked Data bases using
  - Using SPARQL Endpoint
  - Through owl:sameAs, owl:equivalentClass, owl:equivalentProperty

- The advantages of reconciliation are
  - Enrichment of the data.
  - Reducing the individual efforts.
  - …

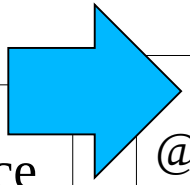
Data Reconciliation: Everyone Wins (2)

DRTC:Book

id	title	author	year	price
1	Lord of Light	Roger Zelanzy	1967	INR 290
2	Creatures of Light and Darkness	Roger Zelanzy	1969	INR 310

DBPedia: Person

id	name	livesIn
1	Roger Zelanzy	USA
2	Chetan Bhagat	India



```
@prefix dc: <http://purl.org/dc/elements/1.1/>
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix dcterms: <http://dublincore.org/documents/dcmi-terms/>
@prefix drtc:<http://drtc.isibang.ac.in/triples/book/>
@prefix shop:<http://www.example.com/doctype/element/>
@prefix foaf: <http://xmlns.com/foaf/0.1/>
@prefix dbpedia: <http://dbpedia.org/resource/>
@prefix owl: <<http://w3.org/2002/07/owl#>>

drtc:book1      dc:title          drtc:Lord_of_Light
drtc:book1      rdf:type          shop:Book
drtc:book1      dcterms:created   “1969”^^xsd:date
drtc:book1      shop:price        “INR290”

drtc:book1          dc:creator      drtc:Roger_Zelanzy
drtc:Roger_Zelanzy  owl:sameAs      dbPedia:person1

dbpedia:person1   rdf:type          dbpedia:Person
dbpedia:person1   foaf:name         “Rozer Zelanzy”
dbpedia:person1   dbpedia:livesIn   “USA”
```

Linked Data Tools

Conversion Tools

- Google Refine (http://code.google.com/p/google-refine/): a tool for working with messy data, cleaning it up, convert it from one format (say, spreadsheet) into RDF data. Also, allows linking converting data to databases like Freebase, DBPedia.

- **GraphDB OntoRefine** (https://graphdb.ontotext.com/): an upgraded version of the open-source OpenRefine data transformation tool. It allows the quick mapping of any structured data to a locally stored RDF schema in GraphDB.

- DB2RDF (http://sourceforge.net/projects/db2rdf/): convert data from relational data to RDF data. It also supports SPARQL endpoint for querying the converted data.
- Triplify (http://triplify.org/): converts relational data to RDF data.

Linked Data Tools

- Csv2rdf4lod-automation (https://github.com/timrdf/csv2rdf4lod-automation/wiki): a quick and easy way to produce an RDF encoding of data available in Comma-Separated-Values (CSV).

- GRDDL (http://www.w3.org/TR/grddl/): a W3C specification that defines a method for exposing XML as RDF via XSLT (a technology for mapping XML-to-XML).

- RDFTEF (http://rdftef.sourceforge.net/): converts XML documents consisting of a subset of TEI (Text Encoding Initiative) XML into RDF. TEI (http://www.tei-c.org/index.xml) is a standard used for epigraphic data. It is XML based.

Linked Data Tools

Triple Store

- Virtuoso (http://virtuoso.openlinksw.com/): a data server that supports various data representations including relational, XML and RDF. It provides an RDF triple-store and supports SPARQL endpoints. Its Sesame and Jena APIs allow it to be used with those products. (available under commercial license)
- Virtuoso Open Source Edition (http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSDownload)
- **Apache Jena Fuseki** (https://jena.apache.org/documentation/fuseki2/): a SPARQL server. It provides the SPARQL 1.1 protocols for query and update as well as the SPARQL Graph Store protocol.

Linked Data Tools (2)

Triple Store

- AllegroGraph (http://www.franz.com/agraph/allegrograph/): an RDF database with support for SPARQL queries and Prolog reasoning.
- Mulgara (http://www.mulgara.org/): a 100% Java RDF database which supports REST interfaces for SPARQL and also to insert, update or delete triples.
- Cliopatria (http://cliopatria.swi-prolog.org/home): an RDF database with web server, user management, SPARQL query support and Prolog reasoning.

Linked Data Publication

- Linked Data can be published as a downloadable RDF dump.
  - GitHub??
- As a complementary to the above, your data can also be exposed via SPARQL endpoint.
- Examples of SPARQL endpoints:
  - DBPedia endpoint: https://dbpedia.org/sparql
  - AGROVOC Thesaurus: https://agrovoc.uniroma2.it/sparql/
  - Gene Ontology endpoint: http://geneontology.org/sparql
  - FactForge: http://factforge.net/sparql
  - DBPedia: https://live.dbpedia.org/sparql
  - BBC Programmes and Music: http://lod.openlinksw.com/sparql/
  - European Environment Agency: http://semantic.eea.europa.eu/sparql

*(asterisk) mark the important SPARQL Endpoints.

Open Issues

- Quality of links and data
- Data fusion support: enhanced support for schema mapping
- Trust
- …

Check Whether Your Data is Qualified for 5 Star!

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

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

*** **Non-proprietary format** (e.g. CSV instead of excel)

**** All the above plus, use **open standards** (for instance, from W3C, RDF and SPARQL) to identify things, so that people can point at your stuff

***** All the above, plus: **Link your data** to other people's data to provide context

Conclusion

- Contribute to make your data more valuable
- Contribute to make your data more visible
- Contribute to make data more useful

Important Links on Linked Data

1. Idehen, Kingsley. Understanding Data.
2. Auer, S. From Document Web to a Web of Linked Data.
3. Auer, S., Dietzold, S., Lehmann, J., Hellmann, S. and Aumueller, D. Triplify - Linked Data ublication from Relational Databases.
4. Linked Open Data for Disaster Management: http://observedchange.com/projects/linked-open-data-for-disaster-management/
5. Linked Haiti Data Reports (Demo): http://observedchange.com/demos/linked-haiti/
6. Linked Data Tools. http://spqr.cerch.kcl.ac.uk/?page_id=94
7. Google Refine: https://code.google.com/p/google-refine/
8. RDF Refine: http://refine.deri.ie/qbExport
9. RDF Data Cube Vocabulary (for Statistical Data). http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html
10. SPARQL Endpoints. http://www.w3.org/wiki/SparqlEndpoints
11. Falcons (a Linked Data Browser): http://ws.nju.edu.cn/falcons/objectsearch/index.jsp

for your attention!

Question!

Email: bisu@drtc.isibang.ac.in