Web Mining

CSE 537 Artificial Intelligence, Spring 2016

Group #: 3

Author: Feiqiao Wang

Student ID: 104965863

Professor: Anita Wasilewska

Topics Covered Today

Motivation to choose the topic

What is web mining and why need web mining?

How to collect data from web?

Web mining methods summary

Web mining use cases review

Controversial issue of web mining

Semantic web and semantic web mining

Motivation

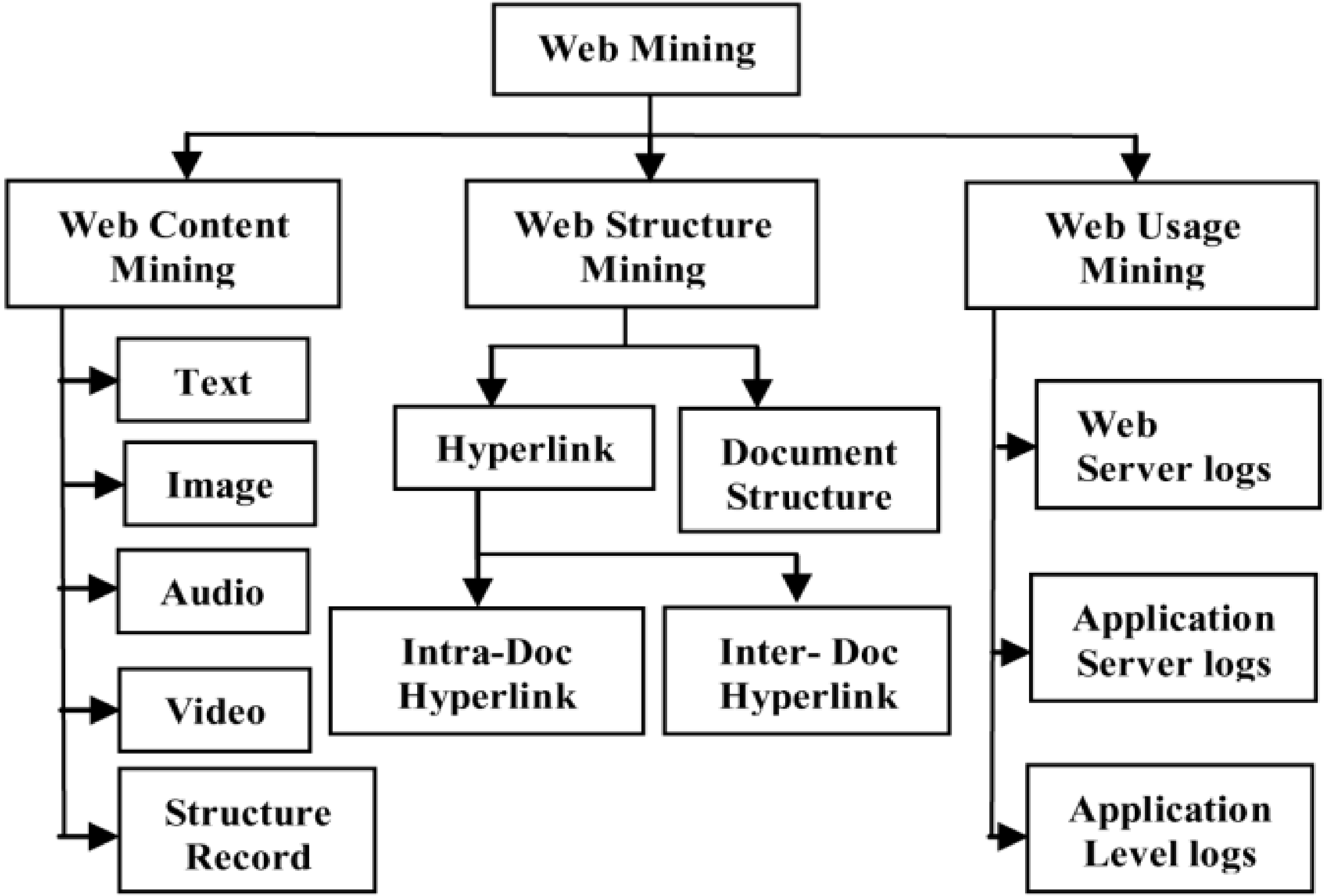
Extend the topics teaching in class;

Share the knowledge and learn from each other;

What is Web Mining?

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services. Web mining is subset of data mining.

Web Mining Categories



(Source: K.Dharmarajan-Scholar, “CURRENT LITERATURE REVIEW - WEB MINING “, Elysium Journal, September 2014, Volume-1, Special Issue-1)

Why need web mining?

- Better search result.
- Business intelligence
- Competitive intelligence
- Pricing analysis
- Events
- Product data
- Popularity
- Reputation (credit card score calculation etc..)
- Other …

Web Mining vs Data Mining

Scale – Huge dataset for web mining, small to large dataset for the traditional data mining;

Access – For web mining, Data is extracted explicitly or in most case inexplicitly (hidden) with web crawler. For traditional data mining, we access data explicitly from local database or from web.

Structure – A traditional data mining task gets information from a database, which provides some level of explicit structure. A typical web mining task is processing unstructured or semi-structured data from web pages.

Web Mining – History

- Term first used in [E1996], defined in a ‘task oriented’ manner
- Alternate ‘data oriented’ definition given in [CMS1997]
- 1st panel discussion at ICTAI 1997 [SM1997]
- Continuing forum
  - WebKDD workshops with ACM SIGKDD, 1999, 2000, 2001, 2002, … ; 60 – 90 attendees
  - SIAM Web analytics workshop 2001, 2002, …
- Special issues of DMKD journal, SIGKDD Explorations
- Papers in various data mining conferences & journals
- Surveys [MBNL 1999, BL 1999, KB2000]

How to collect data from web?

- Human copy-and-paste
- Text grepping and regular expression matching
- HTTP programming
- HTML parsers
- DOM parsing
- Web-scraping software
- Vertical aggregation platforms
- Semantic annotation recognizing
- Computer vision web-page analyzers

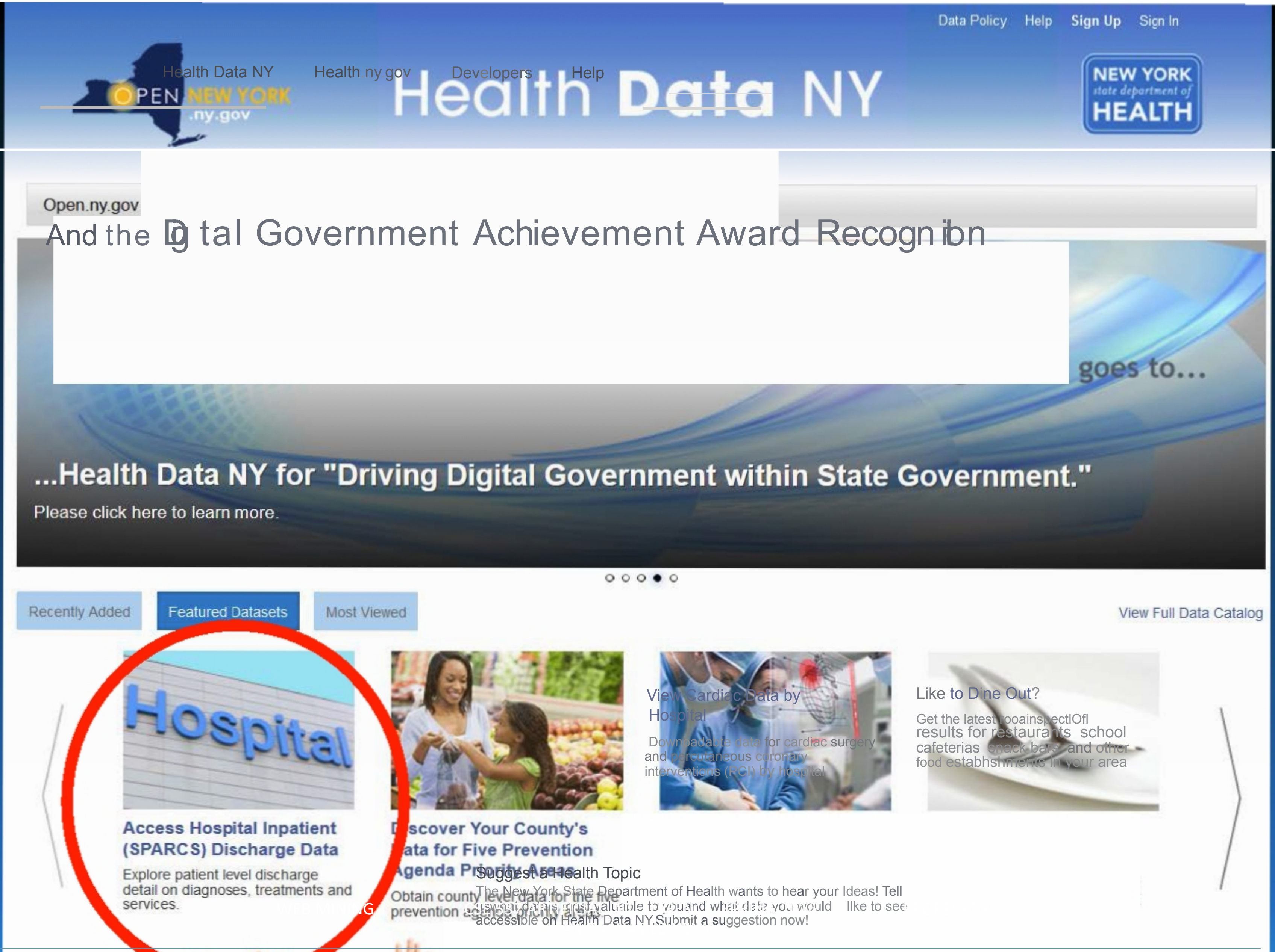
Example of Web data collection:

Clickstream is the recording of the parts of the screen a computer user clicks on while web browsing or using another software application.

As the user clicks anywhere in the webpage or application, the action is logged on a client or inside the web server, as well as possibly the web browser, router, proxy server or ad server.

Clickstream analysis is useful for web activity analysis, software testing, market research, and for analyzing employee productivity.

Another Example of Open Data Source From Web



Data Policy Help **Sign Up** Sign In

Health Data NY Health ny gov Developers Help

Health Data NY

Open.ny.gov

And the Digital Government Achievement Award Recognition

goes to...

...Health Data NY for "Driving Digital Government within State Government."

Please click here to learn more.

Recently Added | Featured Datasets | Most Viewed

View Full Data Catalog

Access Hospital Inpatient (SPARCS) Discharge Data

Explore patient level discharge detail on diagnoses, treatments and services.

Discover Your County's Data for Five Prevention Agenda Priority Areas

Obtain county level data for the five prevention agenda priority areas.

View Cardiac Data by Hospital

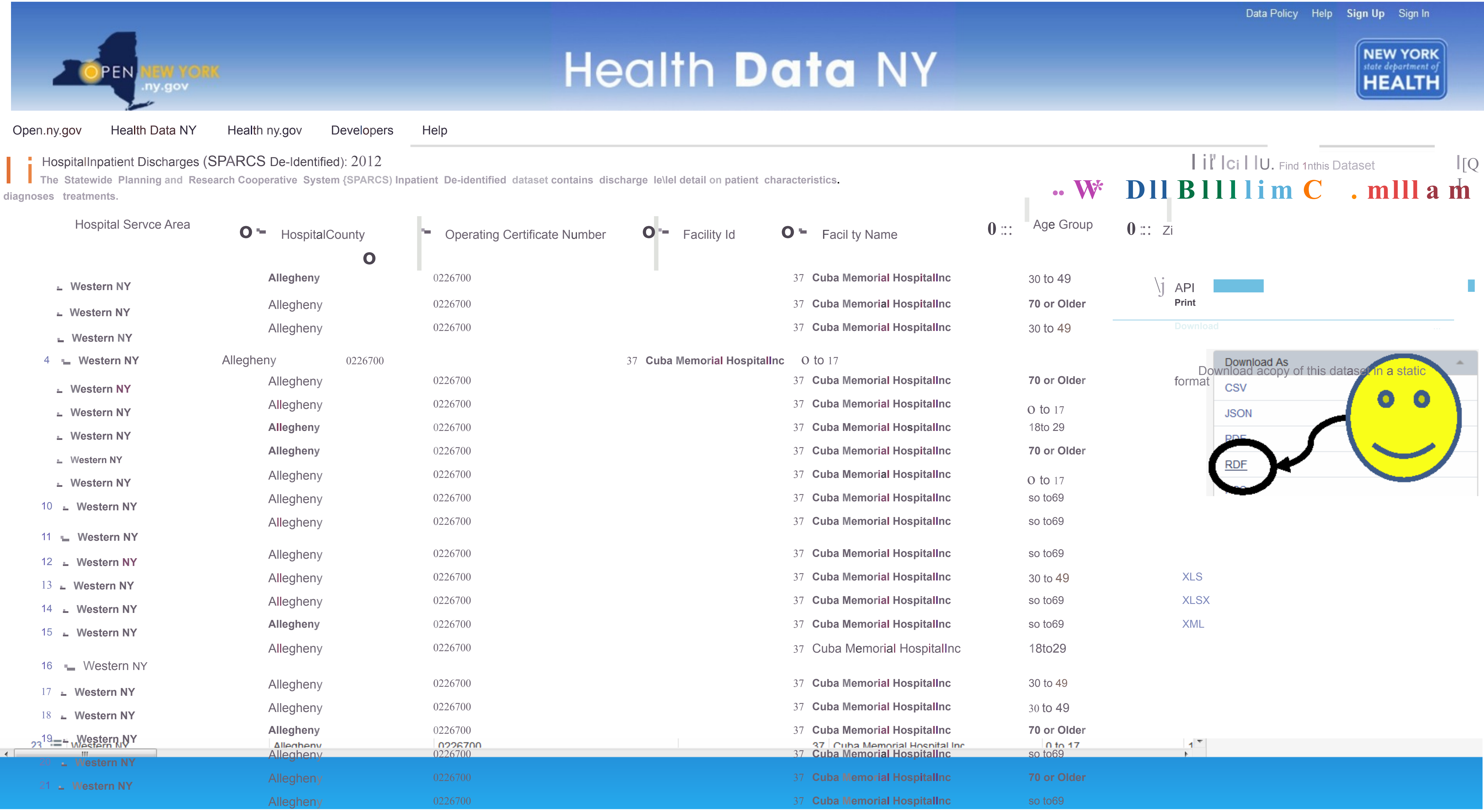
Downloadable data for cardiac surgery and percutaneous coronary interventions (PCI) by hospital

Like to Dine Out?

Get the latest food inspection results for restaurants school cafeterias snack bars and other food establishments in your area

Suggest a Health Topic

The New York State Department of Health wants to hear your Ideas! Tell us what data is most valuable to you and what data you would like to see accessible on Health Data NY. Submit a suggestion now!



Data Policy Help Sign Up Sign In

OPEN NEW YORK .ny.gov

Health Data NY

NEW YORK state department of HEALTH

Open.ny.gov Health Data NY Health ny.gov Developers Help

HospitalInpatient Discharges (SPARCS De-Identified): 2012

The Statewide Planning and Research Cooperative System (SPARCS) Inpatient De-identified dataset contains discharge le\lel detail on patient characteristics. diagnoses treatments.

Find 1nthis Dataset

	Hospital Servce Area	HospitalCounty	Operating Certificate Number	Facility Id	Facil ty Name	Age Group
	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	30 to 49
	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	70 or Older
	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	30 to 49
4	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	O to 17
	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	70 or Older
	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	O to 17
	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	18to 29
	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	70 or Older
	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	O to 17
10	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	so to69
11	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	so to69
12	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	so to69
13	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	30 to 49
14	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	so to69
15	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	so to69
16	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	18to29
17	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	30 to 49
18	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	30 to 49
19	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	70 or Older
23	Western NY	Allegheny	0226700	37	Cuba Memorial Hospital Inc	0 to 17
20	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	so to69
21	Western NY	Allegheny	0226700	37	Cuba Memorial HospitalInc	70 or Older
		Allegheny	0226700	37	Cuba Memorial HospitalInc	so to69

Download As

Download acopy of this dataset in a static format

- CSV
- JSON
- PDF
- RDF
- XLS
- XLSX
- XML

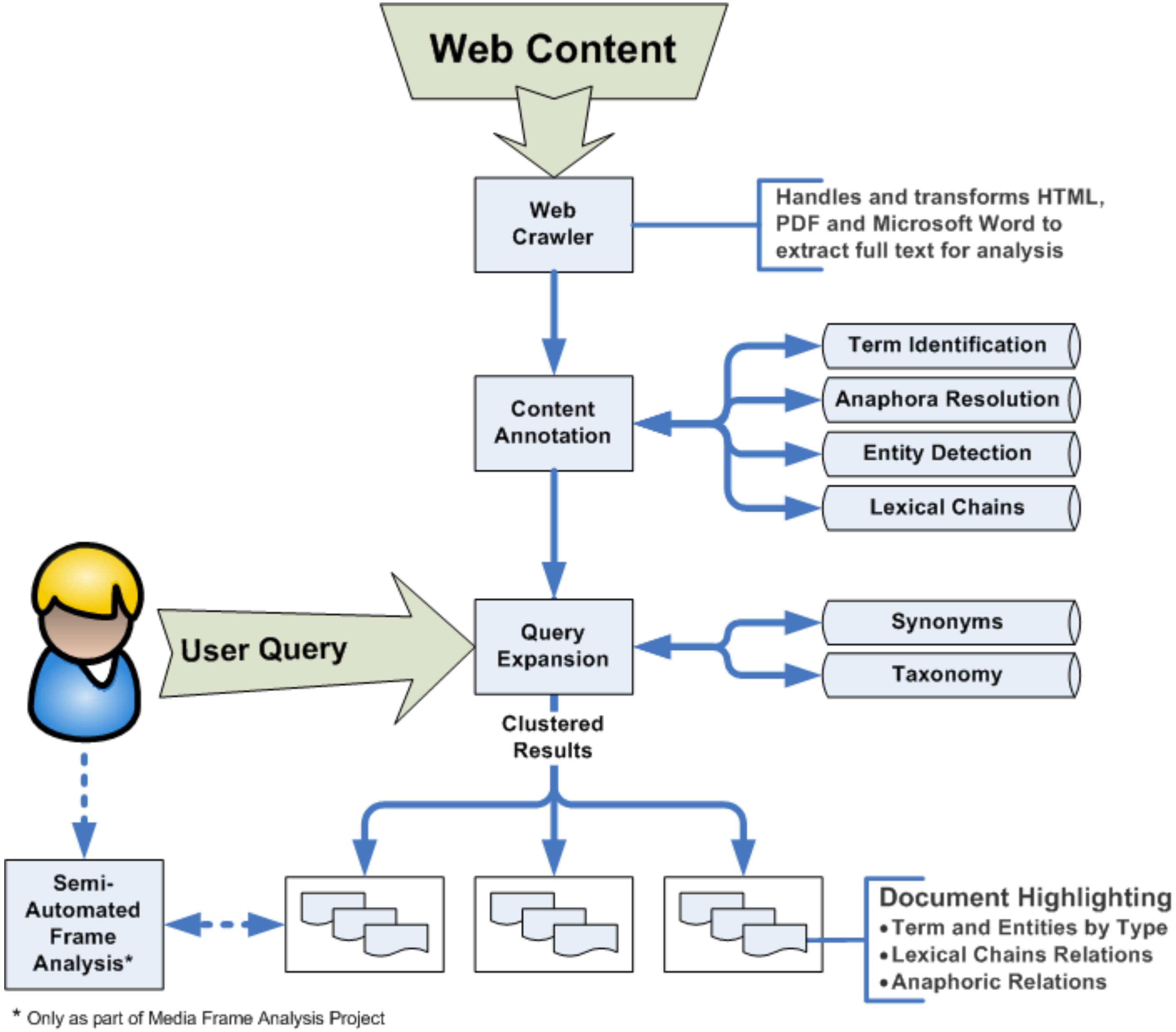
Web mining methods summary

What is web content mining ?

It describes the discovery of useful information from Web documents. Basically, Web content consists of several types of data such as text, image, audio, video, metadata as well as hyperlinks. Research in mining multiple types of data is now termed multimedia-data mining. We could consider multimedia-data mining as an instance of Web-content mining. The Web content data consist of unstructured data such as free text, semi-structured data such as HTML documents, and a more structured data such as tables and database- generated HTML pages. The goal of Web-content mining is mainly to assist or to improve information-finding or filtering the information. Building a new model of data on the Web, more sophisticated queries other than the keywords-based search could be asked.

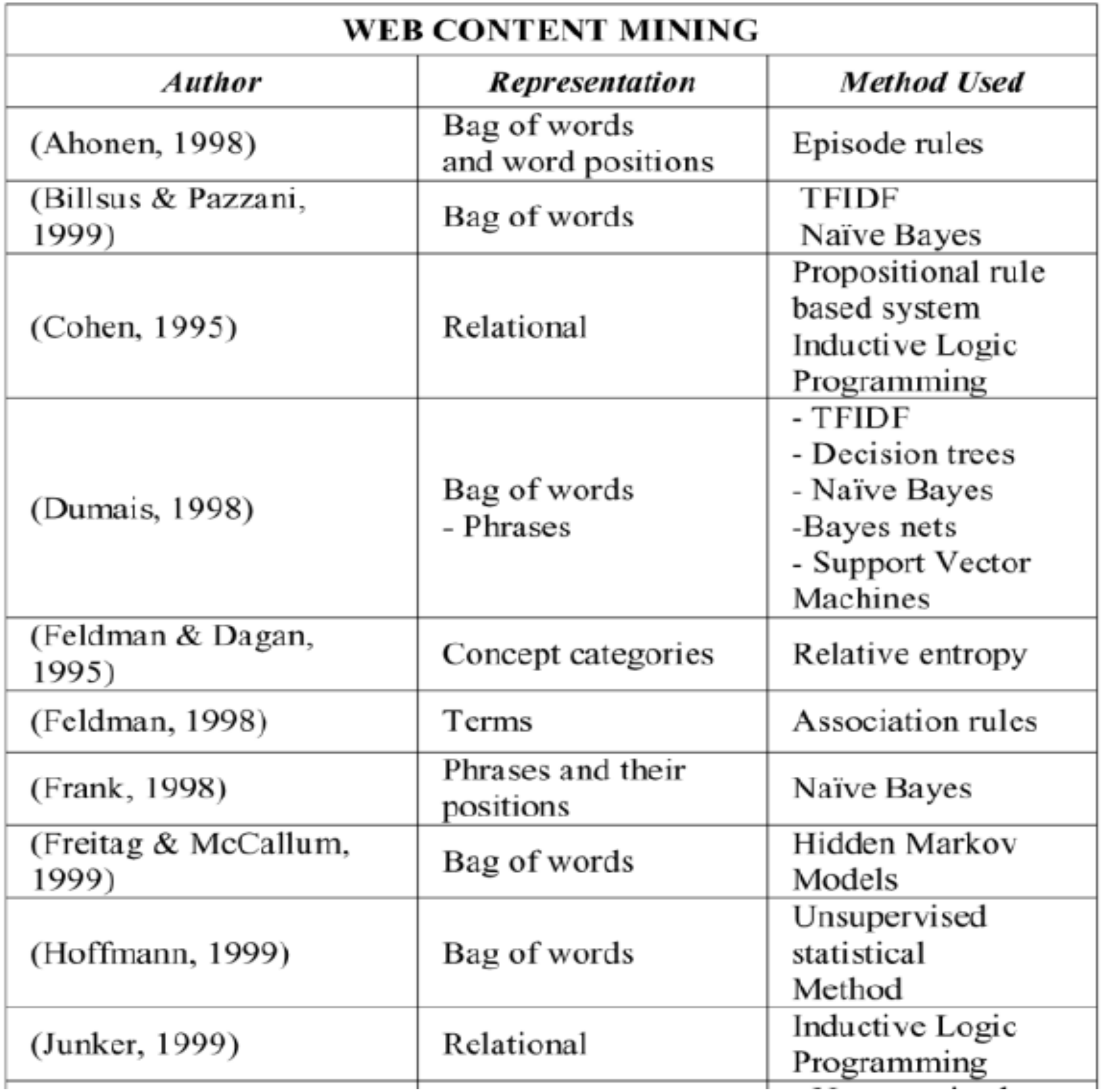
4 steps of Web Content Mining

- Collect – fetch the content from the Web
- Parse – extract usable data from formatted data (HTML, PDF, etc)
- Analyze – tokenize, rate, classify, cluster, filter, sort, etc.
- Produce – turn the results of analysis into something useful (report, search index, etc)


Web Content
Web Crawler
Handles and transforms HTML, PDF and Microsoft Word to extract full text for analysis
Content Annotation
Term Identification
Anaphora Resolution
Entity Detection
Lexical Chains
User Query
Query Expansion
Synonyms
Taxonomy
Clustered Results
Semi-Automated Frame Analysis*
Document Highlighting
•Term and Entities by Type
•Lexical Chains Relations
•Anaphoric Relations
* Only as part of Media Frame Analysis Project

(Source: Google Images)

WEB CONTENT MINING USING DIFFERENT ALGORITHMS



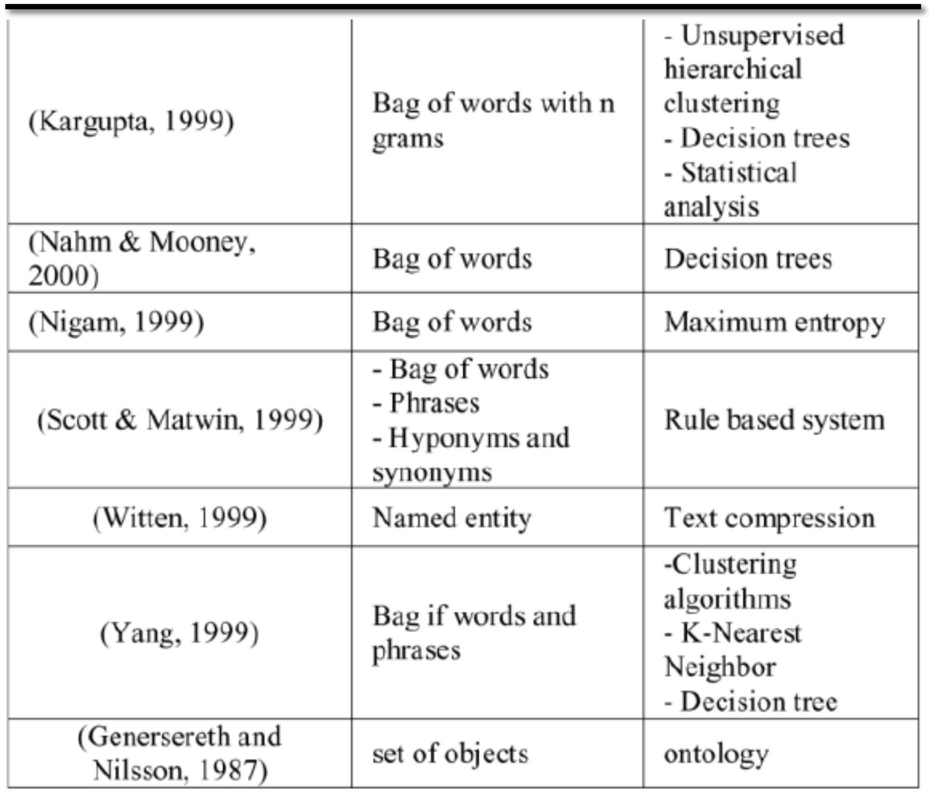
WEB CONTENT MINING		
Author	***Representation***	***Method Used***
(Ahonen, 1998)	Bag of words and word positions	Episode rules
(Billsus & Pazzani, 1999)	Bag of words	TFIDF Naïve Bayes
(Cohen, 1995)	Relational	Propositional rule based system Inductive Logic Programming
(Dumais, 1998)	Bag of words - Phrases	- TFIDF - Decision trees - Naïve Bayes -Bayes nets - Support Vector Machines
(Feldman & Dagan, 1995)	Concept categories	Relative entropy
(Feldman, 1998)	Terms	Association rules
(Frank, 1998)	Phrases and their positions	Naïve Bayes
(Freitag & McCallum, 1999)	Bag of words	Hidden Markov Models
(Hoffmann, 1999)	Bag of words	Unsupervised statistical Method
(Junker, 1999)	Relational	Inductive Logic Programming

(Source: K.Dharmarajan-Scholar, “CURRENT LITERATURE REVIEW - WEB MINING “, Elysium Journal, September 2014, Volume-1, Special Issue-1)

Continue …

WEB CONTENT MINING USING DIFFERENT ALGORITHMS

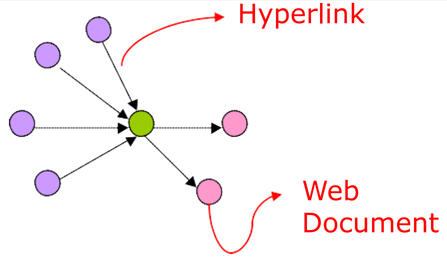
(Source: K.Dharmarajan-Scholar, “CURRENT LITERATURE REVIEW - WEB MINING “, Elysium Journal, September 2014, Volume-1, Special Issue-1)



(Kargupta, 1999)	Bag of words with n grams	- Unsupervised hierarchical clustering - Decision trees - Statistical analysis
(Nahm & Mooney, 2000)	Bag of words	Decision trees
(Nigam, 1999)	Bag of words	Maximum entropy
(Scott & Matwin, 1999)	- Bag of words - Phrases - Hyponyms and synonyms	Rule based system
(Witten, 1999)	Named entity	Text compression
(Yang, 1999)	Bag if words and phrases	-Clustering algorithms - K-Nearest Neighbor - Decision tree
(Genersereth and Nilsson, 1987)	set of objects	ontology

What is Web Structure Mining?

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages



Web Graph Structure

Web Structure Mining can be the process of discovering structure information from the Web

- This type of mining can be performed either at the (intra-page) document level or at the (inter-page) hyperlink level
- The research at the hyperlink level is also called *Hyperlink Analysis*

Motivation to study Hyperlink Structure

- Hyperlinks serve two main purposes.
  - ✓ Pure Navigation.
  - ✓ Point to pages with authority* on the same topic of the page containing the link.
- This can be used to retrieve useful information from the web.

*** - a set of ideas or statements supporting a topic**

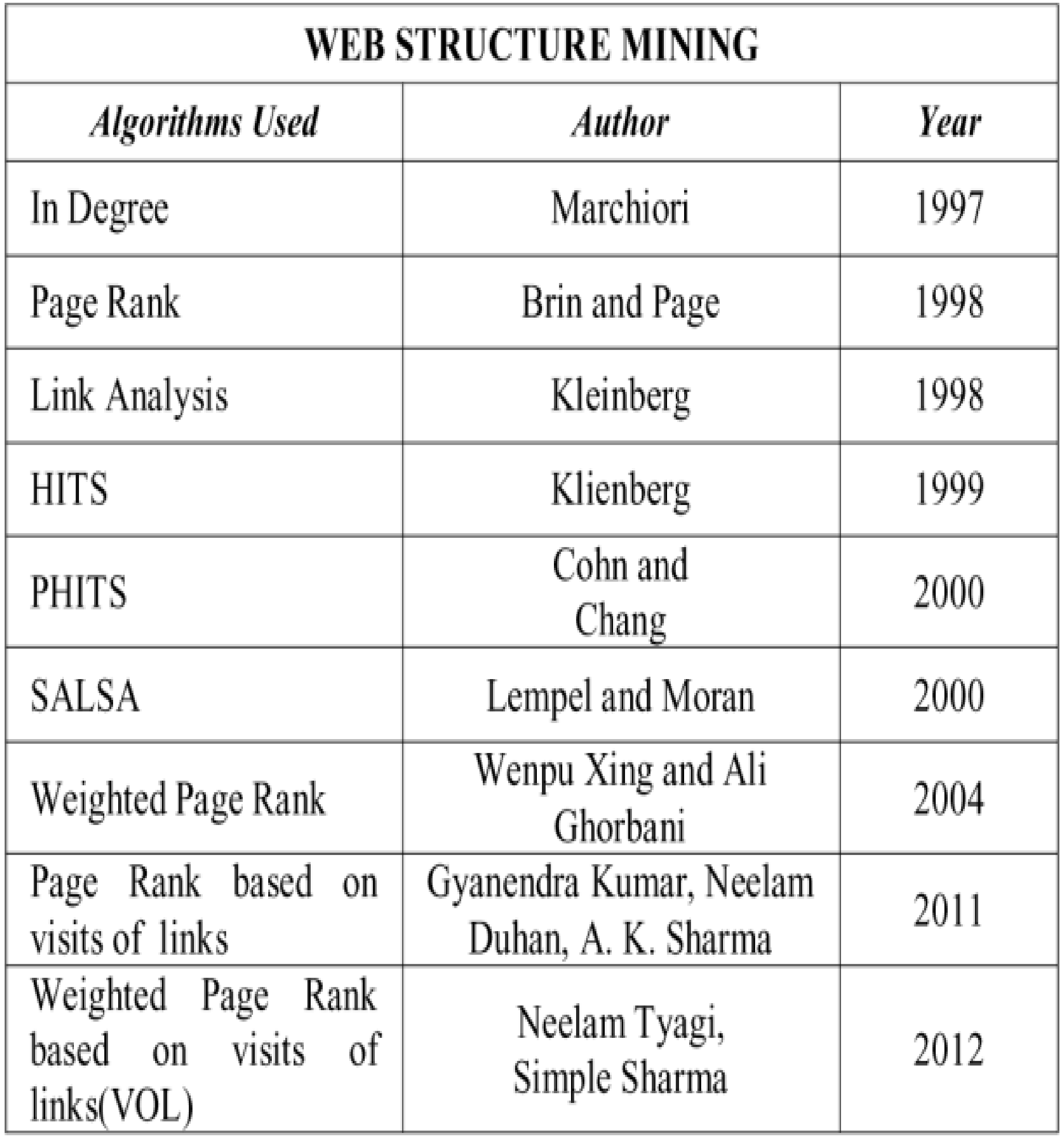
Web Structure Terminology(1)

- ***Web-graph:*** A directed graph that represents the Web.
- ***Node:*** Each Web page is a node of the Web-graph.
- ***Link:*** Each hyperlink on the Web is a directed edge of the Web-graph.
- ***In-degree*****:** The in-degree of a node, *p,* is the number of distinct links that point to *p.*
- ***Out-degree*****:** The out-degree of a node, *p,* is the number of distinct links originating at *p* that point to other nodes.

Web Structure Terminology(2)

- ***Directed Path***: A sequence of links, starting from *p* that can be followed to reach *q.*
- ***Shortest Path:*** Of all the paths between nodes *p* and *q,* which has the shortest length, i.e. number of links on it.
- ***Diameter***: The maximum of all the shortest paths between a pair of nodes *p* and *q,* for all pairs of nodes *p* and *q* in the Web-graph.

WEB STRUCTURE MINING USING DIFFERENT ALGORITHMS

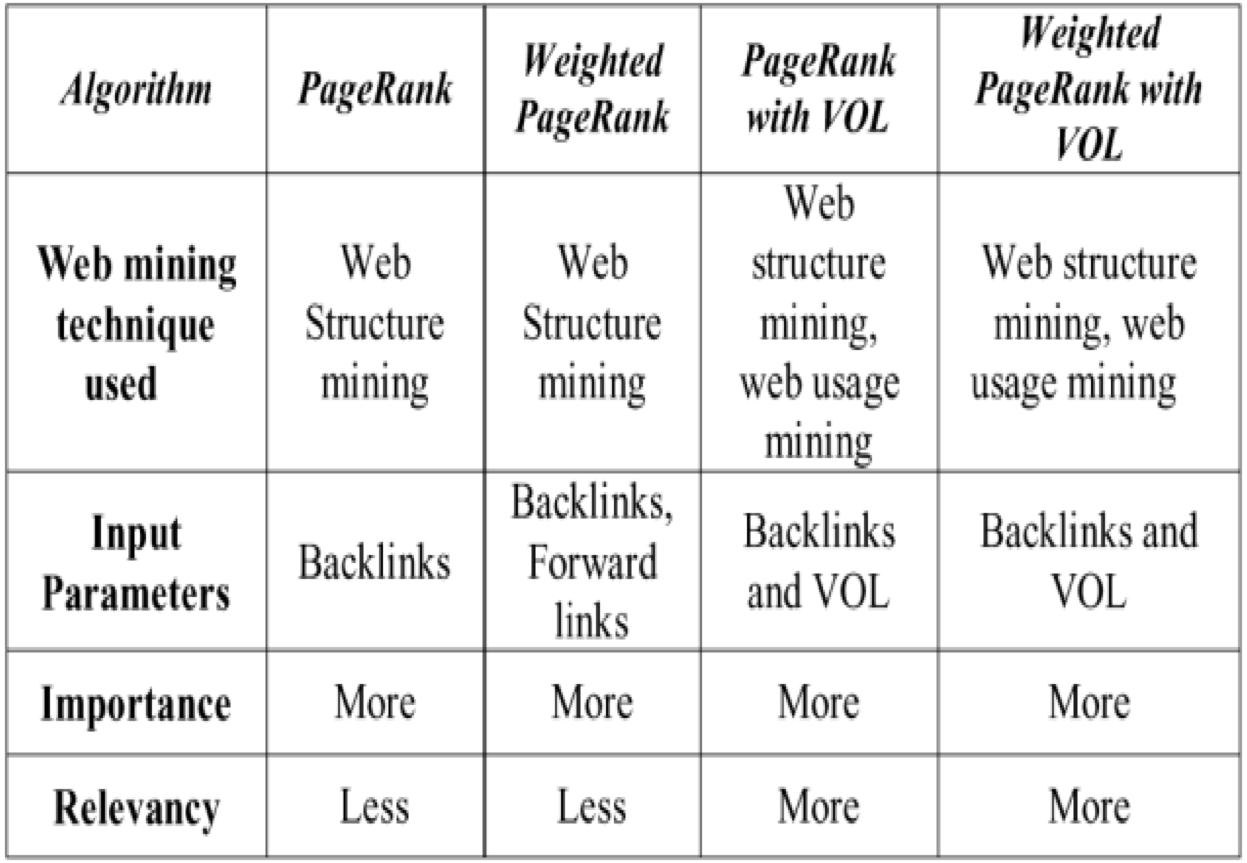


WEB STRUCTURE MINING

Algorithms Used	*Author*	*Year*
In Degree	Marchiori	1997
Page Rank	Brin and Page	1998
Link Analysis	Kleinberg	1998
HITS	Klienberg	1999
PHITS	Cohn and Chang	2000
SALSA	Lempel and Moran	2000
Weighted Page Rank	Wenpu Xing and Ali Ghorbani	2004
Page Rank based on visits of links	Gyanendra Kumar, Neelam Duhan, A. K. Sharma	2011
Weighted Page Rank based on visits of links(VOL)	Neelam Tyagi, Simple Sharma	2012

(Source: K.Dharmarajan-Scholar, “CURRENT LITERATURE REVIEW - WEB MINING “, Elysium Journal, September 2014, Volume-1, Special Issue-1)

COMPARISON OF DIFFERENT WEB STRUCTURE ALGORITHMS



Algorithm	*PageRank*	*Weighted PageRank*	*PageRank with VOL*	*Weighted PageRank with VOL*
Web mining technique used	Web Structure mining	Web Structure mining	Web structure mining, web usage mining	Web structure mining, web usage mining
Input Parameters	Backlinks	Backlinks, Forward links	Backlinks and VOL	Backlinks and VOL
Importance	More	More	More	More
Relevancy	Less	Less	More	More

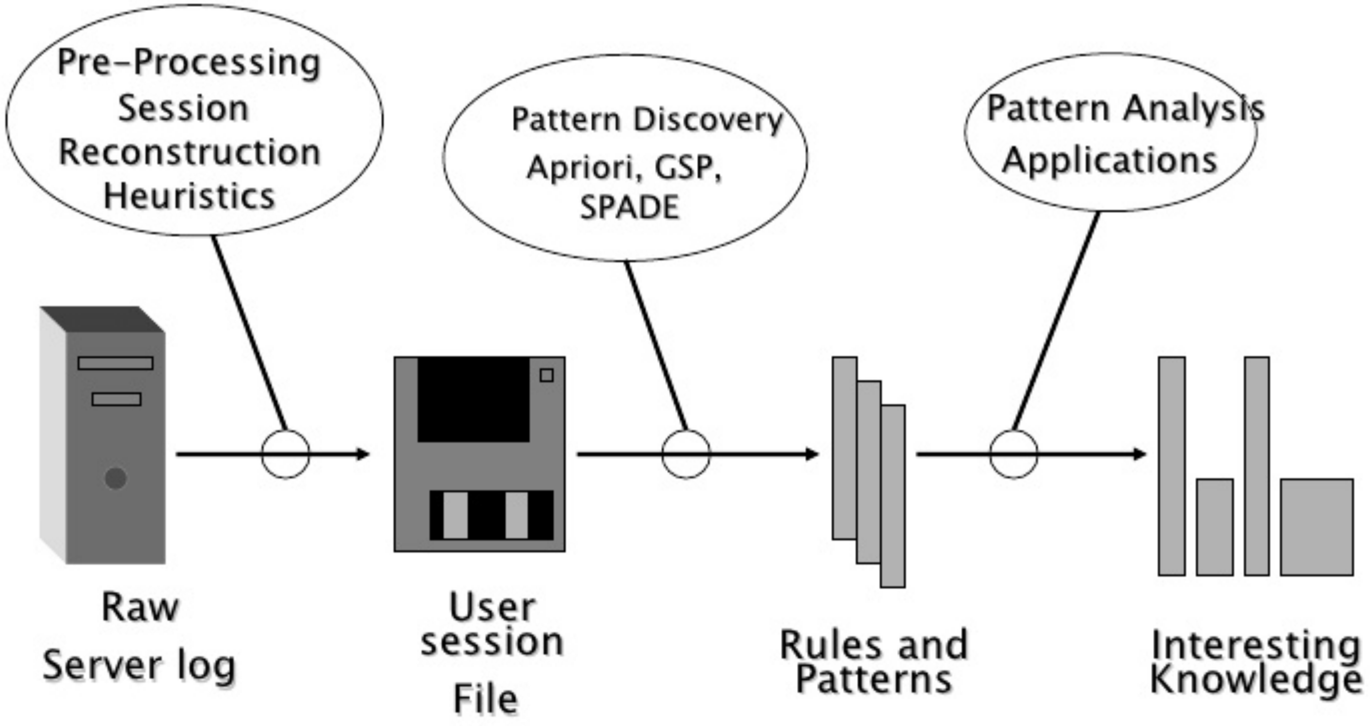
(Source: K.Dharmarajan-Scholar, "CURRENT LITERATURE REVIEW - WEB MINING ", Elysium Journal, September 2014, Volume-1, Special Issue-1)

What is Web Usage Mining?

- A *Web* is a collection of inter-related files on one or more *Web servers*
- *Web Usage Mining*
  - Discovery of meaningful patterns from data generated by client-server transactions on one or more Web localities
- Typical Sources of Data
  - automatically generated data stored in server *access* logs, *referrer* logs, *agent* logs, and client-side *cookies*
  - user profiles
  - meta data: page attributes, content attributes, usage data

Web Mining

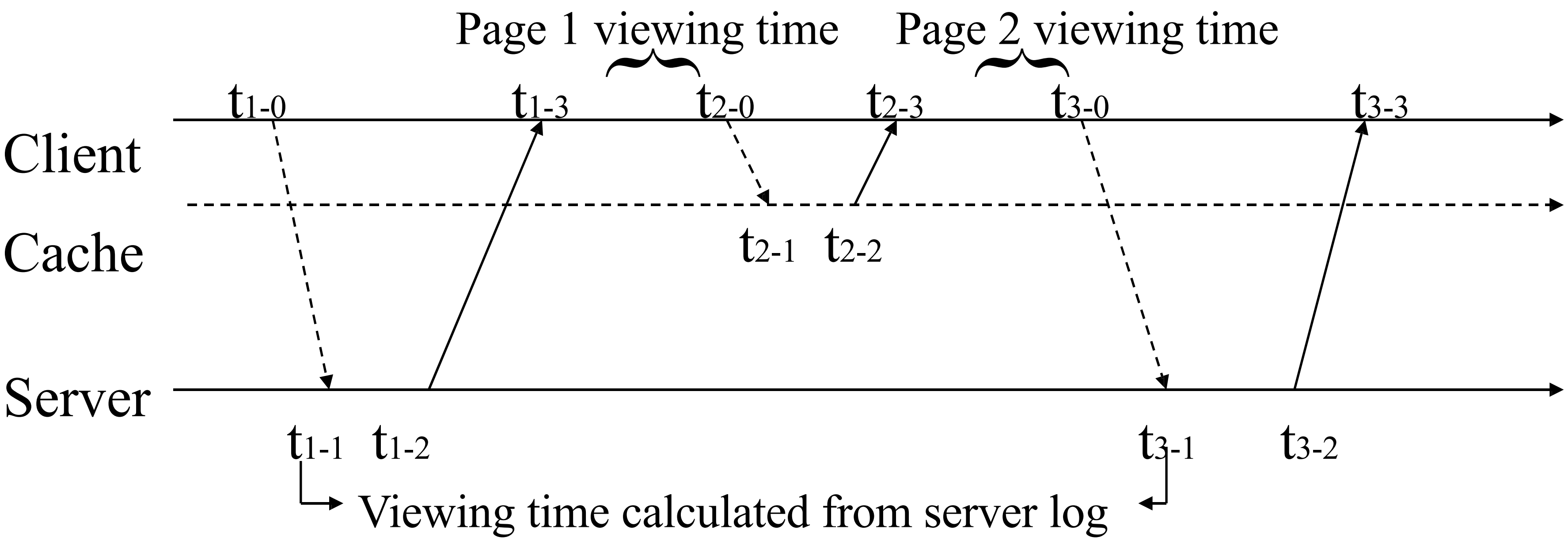
Phases of Web Usage Mining



(Source: Google Images)

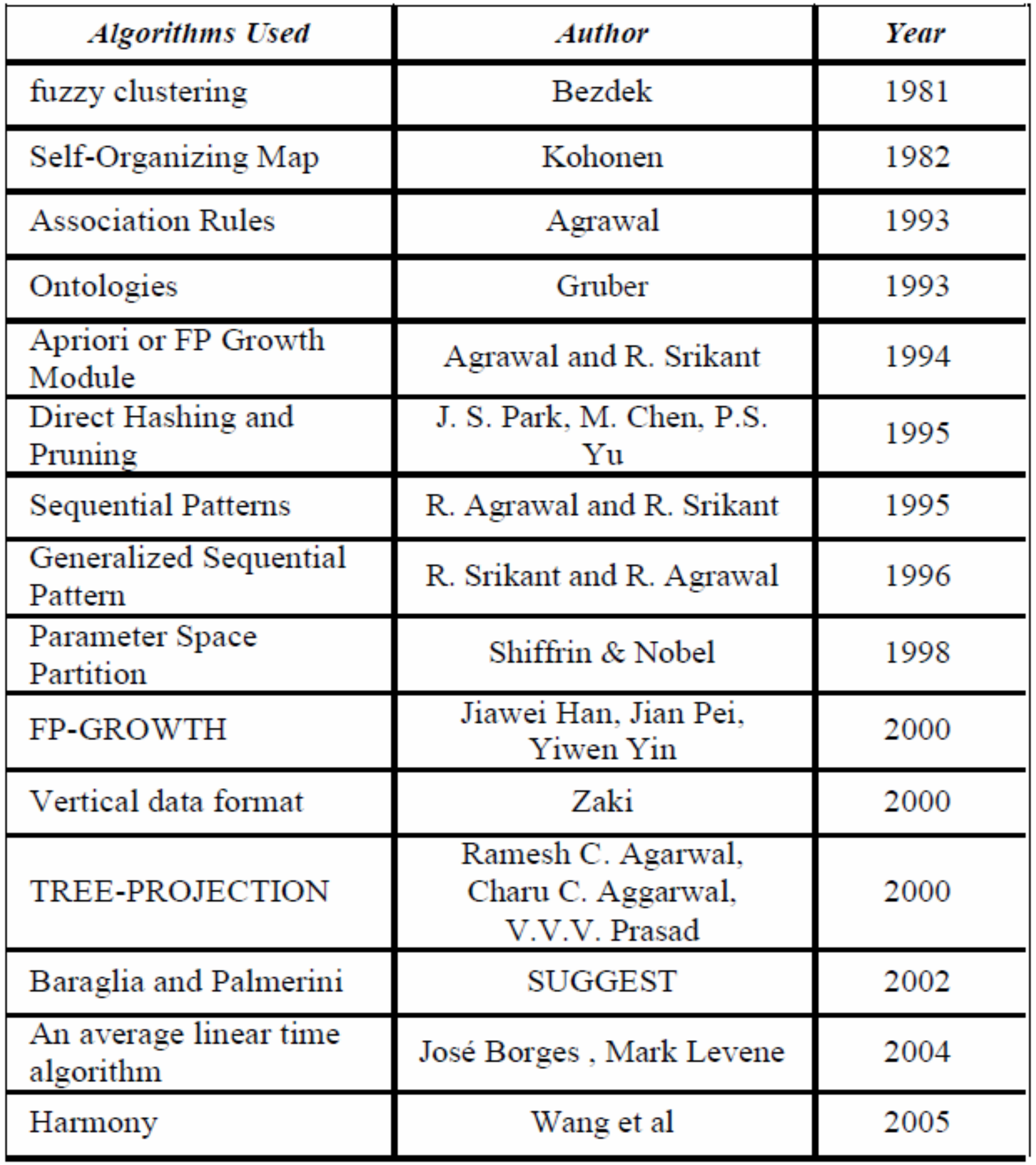
Missed Page Views at Server

- Viewing time for cached pages



(Source PDF file: Web Mining : Accomplishments & Future Directions , Jaideep Srivastava, University of Minnesota, USA)

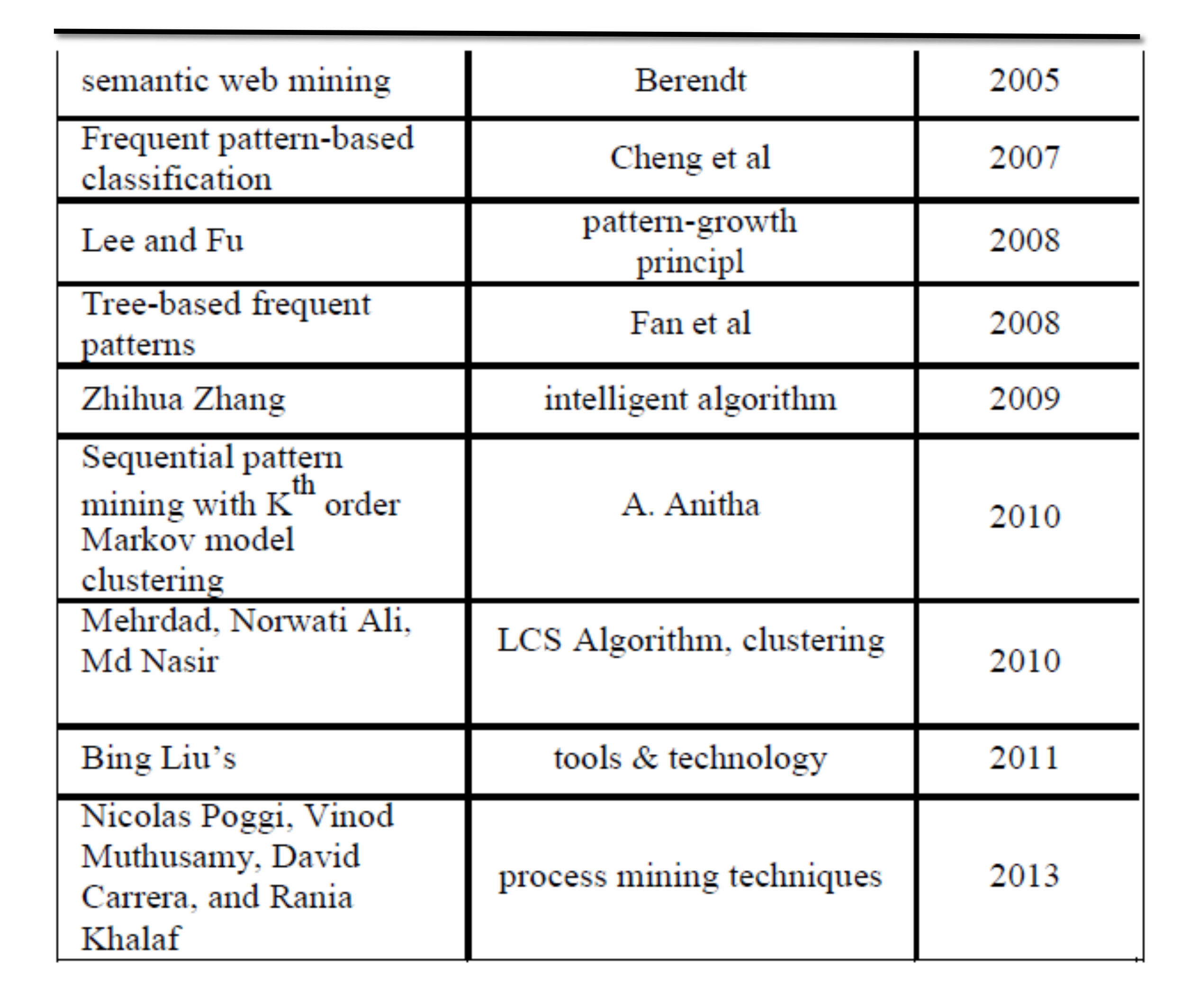
WEB USAGE MINING USING DIFFERENT ALGORITHMS



Algorithms Used	*Author*	*Year*
fuzzy clustering	Bezdek	1981
Self-Organizing Map	Kohonen	1982
Association Rules	Agrawal	1993
Ontologies	Gruber	1993
Apriori or FP Growth Module	Agrawal and R. Srikant	1994
Direct Hashing and Pruning	J. S. Park, M. Chen, P.S. Yu	1995
Sequential Patterns	R. Agrawal and R. Srikant	1995
Generalized Sequential Pattern	R. Srikant and R. Agrawal	1996
Parameter Space Partition	Shiffrin & Nobel	1998
FP-GROWTH	Jiawei Han, Jian Pei, Yiwen Yin	2000
Vertical data format	Zaki	2000
TREE-PROJECTION	Ramesh C. Agarwal, Charu C. Aggarwal, V.V.V. Prasad	2000
Baraglia and Palmerini	SUGGEST	2002
An average linear time algorithm	José Borges , Mark Levene	2004
Harmony	Wang et al	2005

(Source: K.Dharmarajan-Scholar, “CURRENT LITERATURE REVIEW - WEB MINING “, Elysium Journal, September 2014, Volume-1, Special Issue-1)

Continue …



semantic web mining	Berendt	2005
Frequent pattern-based classification	Cheng et al	2007
Lee and Fu	pattern-growth principl	2008
Tree-based frequent patterns	Fan et al	2008
Zhihua Zhang	intelligent algorithm	2009
Sequential pattern mining with Kth order Markov model clustering	A. Anitha	2010
Mehrdad, Norwati Ali, Md Nasir	LCS Algorithm, clustering	2010
Bing Liu's	tools & technology	2011
Nicolas Poggi, Vinod Muthusamy, David Carrera, and Rania Khalaf	process mining techniques	2013

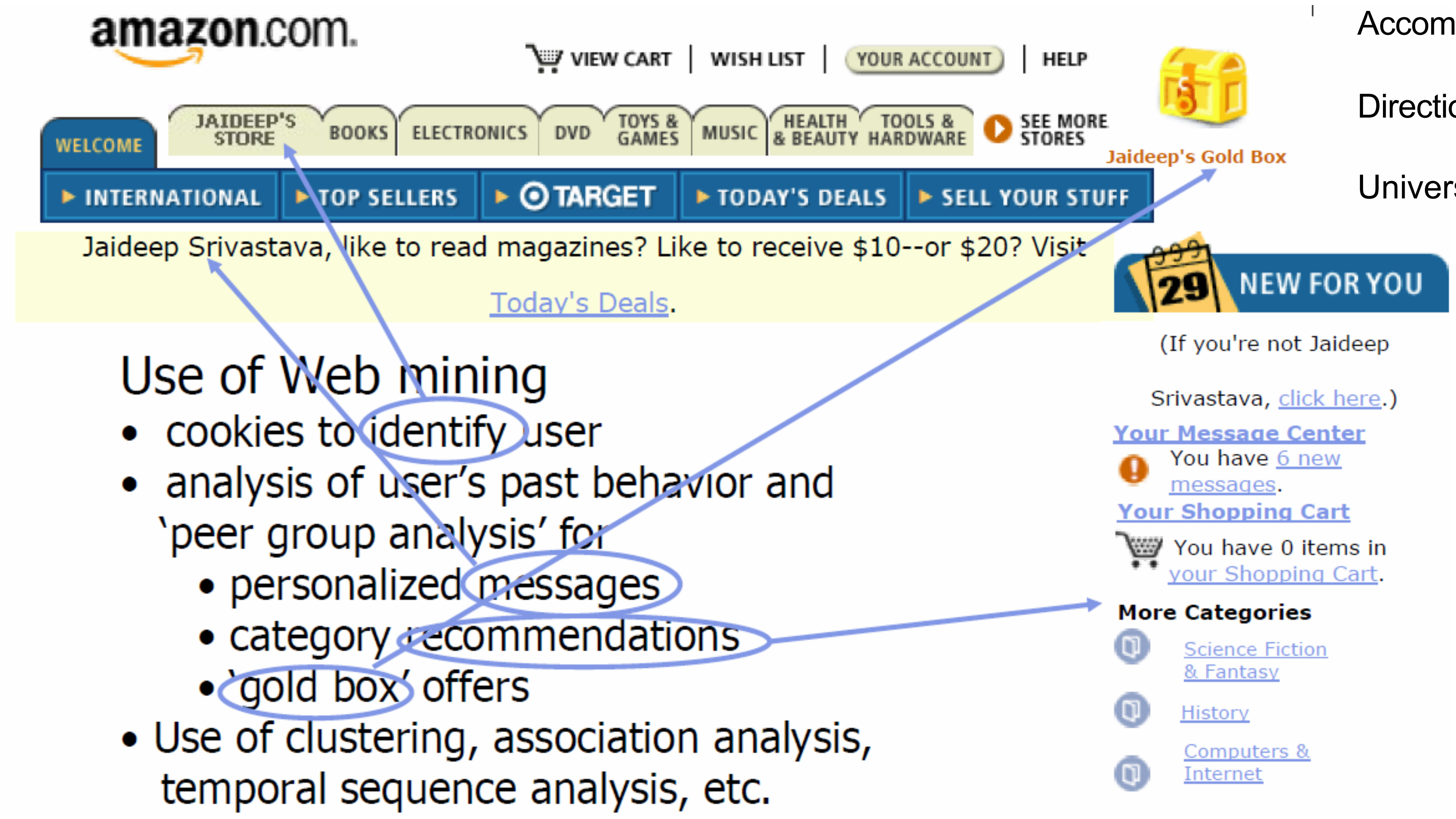
WEB USAGE MINING USING DIFFERENT ALGORITHMS

(Source: K.Dharmarajan-Scholar, “CURRENT LITERATURE REVIEW - WEB MINING “, Elysium Journal, September 2014, Volume-1, Special Issue-1)

Web Mining Use Cases Review

Use Case 1: Recommendation System

(Source PDF file: Web Mining : Accomplishments & Future Directions , Jaideep Srivastava, University of Minnesota, USA)



Jaideep Srivastava, like to read magazines? Like to receive $10--or $20? Visit Today's Deals.

Use of Web mining

- cookies to identify user
- analysis of user's past behavior and 'peer group analysis' for
  - personalized messages
  - category recommendations
  - 'gold box' offers
- Use of clustering, association analysis, temporal sequence analysis, etc.

(If you're not Jaideep Srivastava, click here.)

Your Message Center

You have 6 new messages.

Your Shopping Cart

You have 0 items in your Shopping Cart.

More Categories

- Science Fiction & Fantasy
- History
- Computers & Internet

Use Case 2: Google Search Page Ranking

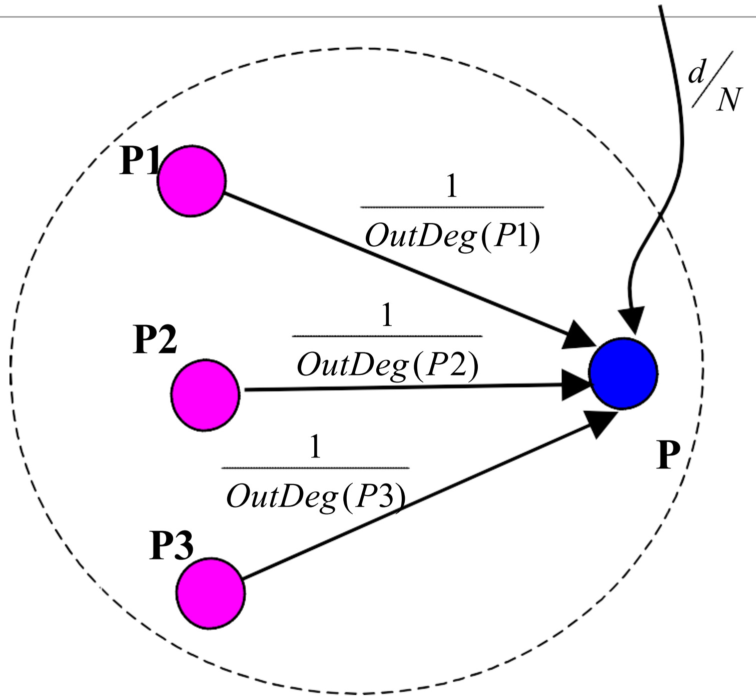
PageRank Formula:

$$PR(p) = d/n + (1-d) \sum_{(q,p)\in G} \left(\frac{PR(q)}{Outdegree(q)}\right)$$

Here, *n* is the number of nodes in the graph and OutDegree(q) is the number of hyperlinks on page q. Intuitively, the approach can be viewed as a stochastic analysis of a random walk on the web graph. The first term in the right hand side of the equation is the probability that a random web surfer arrives at a page p by typing the URL or from a bookmark; or may have a particular page as his/her homepage. Here d is the probability that the surfer chooses a URL directly, rather than traversing a link$_5$ and 1−d is the probability that a person arrives at a page by traversing a link. The second term in the right hand side of the equation is the probability of arriving at a page by traversing a link.

Continue…

Use Case 2: Google Search Page Ranking



Key idea

Rank of a web page depends

on the rank of the web pages pointing to it

(Source PDF file: Web Mining : Accomplishments & Future Directions , Jaideep Srivastava, University of Minnesota, USA)

Use Case 3: Advertisement serving ;

To offer what customers need and disseminate the promotion to the target community to keep their customers. Company like DoubleClick does this type of business for their clients.

Use Case 4: Social Media Network Data Mining

Collect data from social media network, such as Facebook, Twitter etc. to answer some question, for example, “who will win the presidential election”, ”How the disease spread out globally”.

Use Case 5: Fight against terrorism

Government agencies are using this technology to classify threats and fight against terrorism. The predicting capability of mining applications can benefit society by identifying criminal activities

- .
- .
- .

Use Case N …

Controversial Issue of Web Mining

- The web usage mining may cause the ***invasion of privacy***.
- The companies collecting the data for a specific purpose might use the data for a totally different purpose, and this essentially ***violates the user's interests***. no law preventing them from selling or trading the data.
- Some mining algorithms might ***use controversial attributes like sex, race, religion, or sexual orientation to categorize individuals***. These practices might be against the anti-discrimination legislation. The applications make it hard to identify the use of such controversial attributes, and there is no strong rule against the usage of such algorithms with such attributes.

What is Semantic Web?

A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities

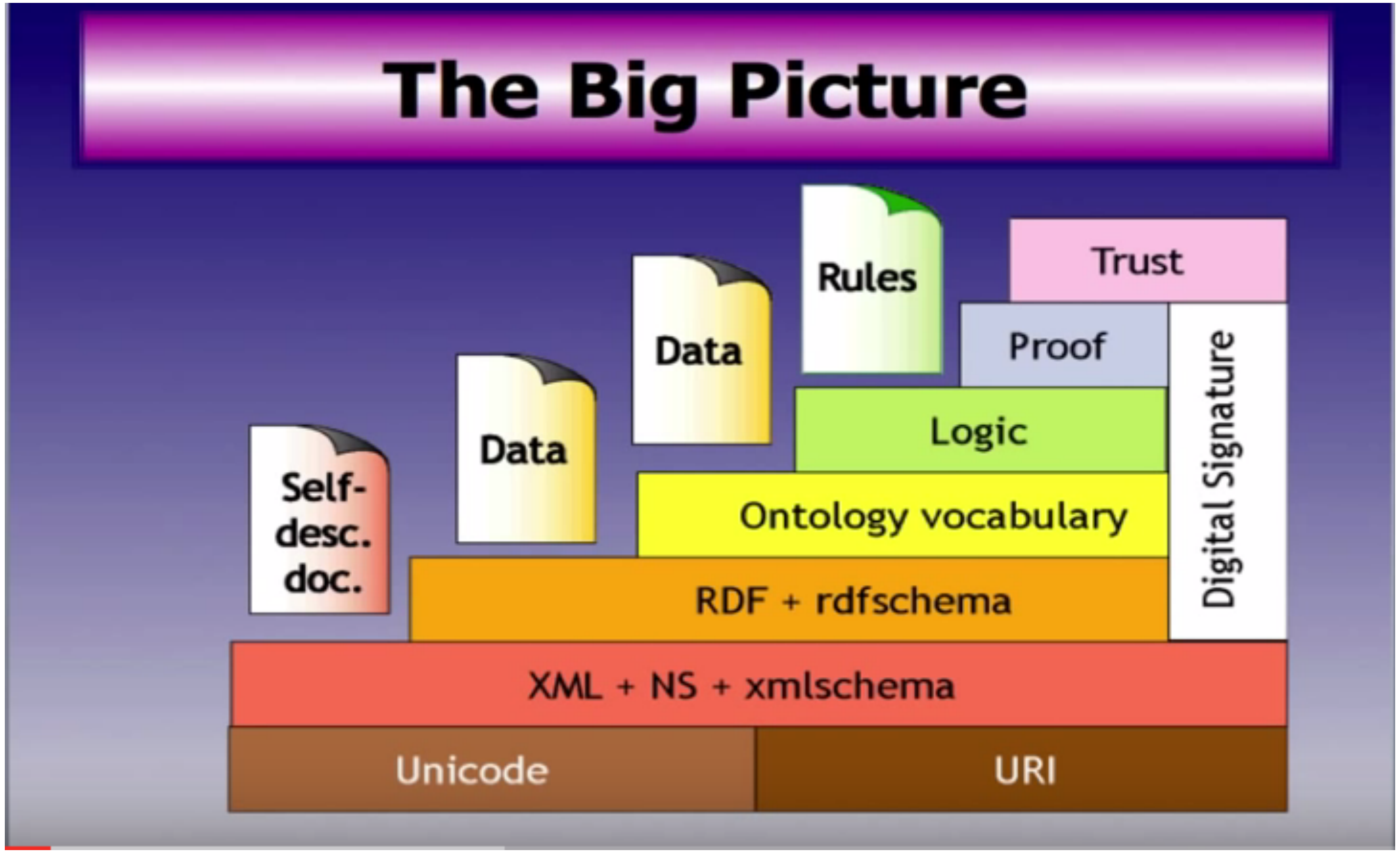
By Tim Berners-Lee, James Hendlerand OraLassila

May 17, 2001

Why need Semantic Web ?

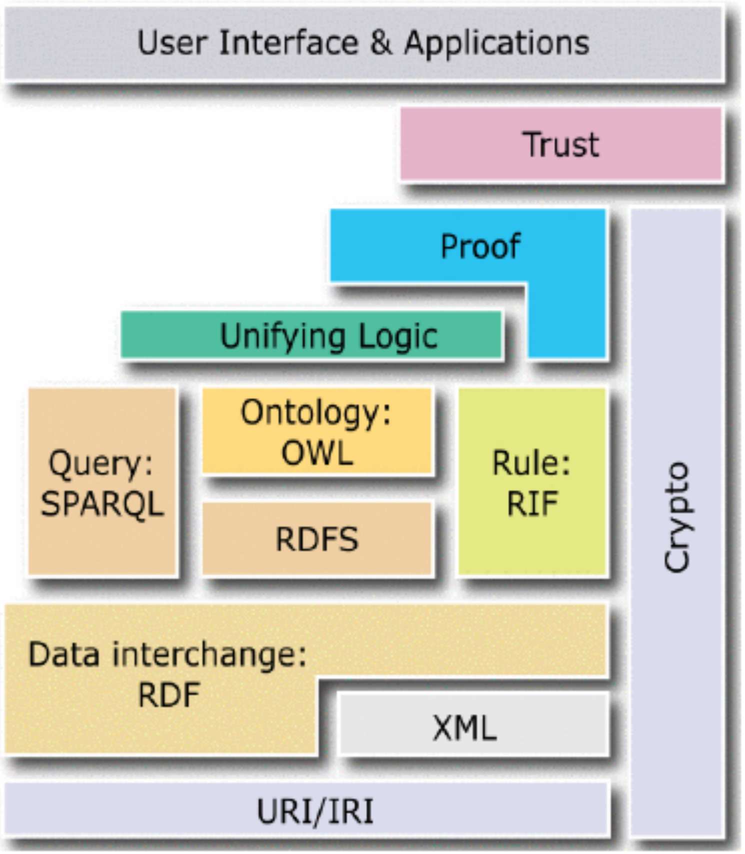
- Huge amount of data is interpretable by humans only; machine support is limited.
- Berners-Lee suggested to enrich the Web by machine processable information which supports the user in his tasks
- To reach this goal the Semantic Web will be built up in different levels, the one we care about is ontologies.
- Make data sharing feasible in an automatically manner.
- Refine Data mining algorithms and enhance quality of web mining result (attributes reduction and rules pruning in classification).

About Semantic Web



(Source: https://www.youtube.com/watch?v=rhgUDGtT2EM)

Semantic Technology Stack



- **Basic Technologies**
  - *URI*
    - Uniform Resource Identifier
  - *RDF*
    - Resource Description Framework
  - *RDFS*
    - RDF Schema
  - *OWL*
    - Web ontology language
  - SPARQL
    - Protocol and Query Language

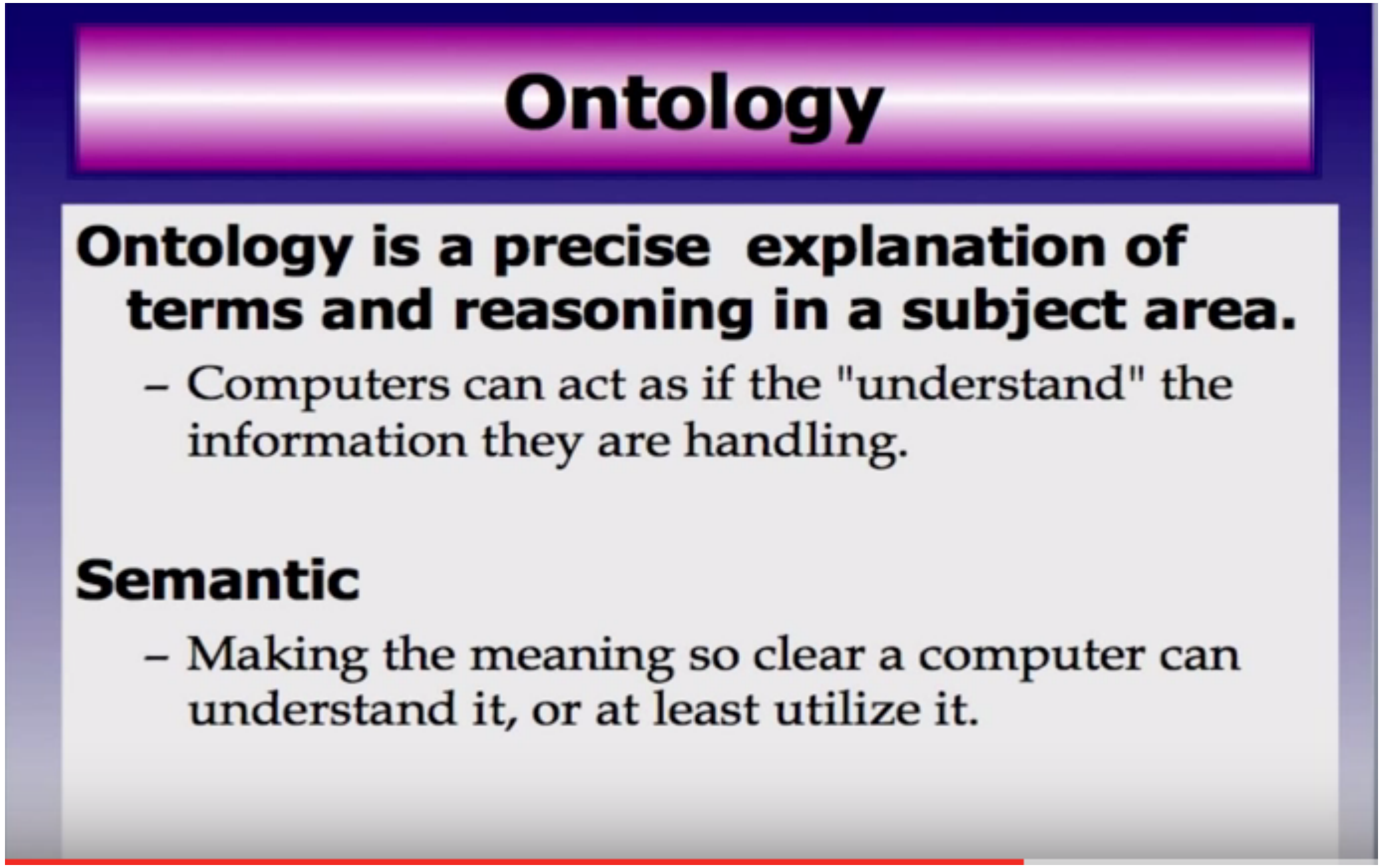
(Source: https://www.youtube.com/watch?v=rhgUDGtT2EM)

Key Features of Semantic Web

- ONTOLOGY ------ OWL
- RDF ---- SPARQL
- LINKED DATA

What is an Ontology ?

- An **ontology** is a formal explicit description of concepts in a domain of discourse ,properties of each concept describing various features and instances of the concept
- An ontology together with a set of individual **instances** of classes constitutes a **knowledge base**.



Ontology

Ontology is a precise explanation of terms and reasoning in a subject area.

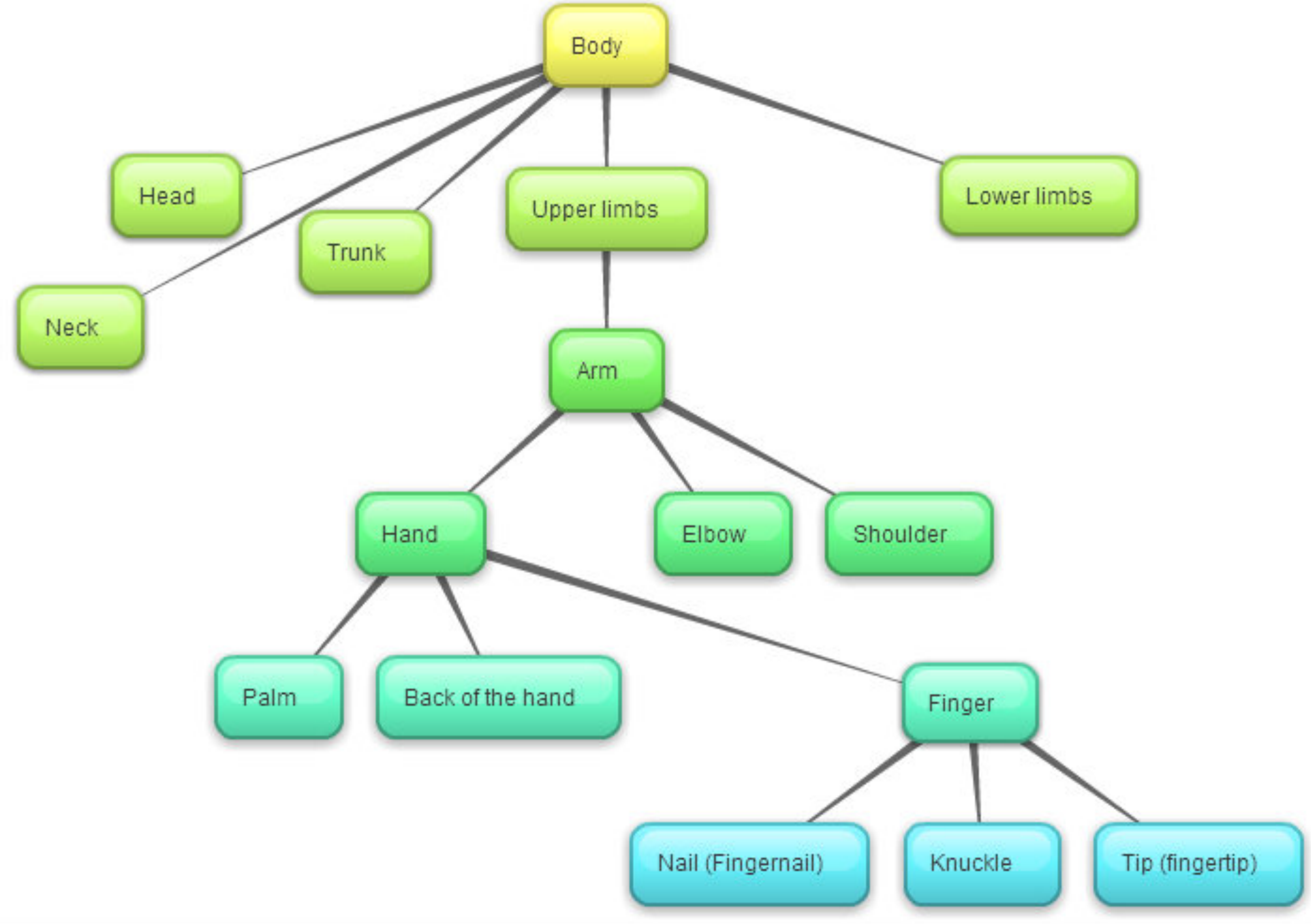
- Computers can act as if the "understand" the information they are handling.

Semantic

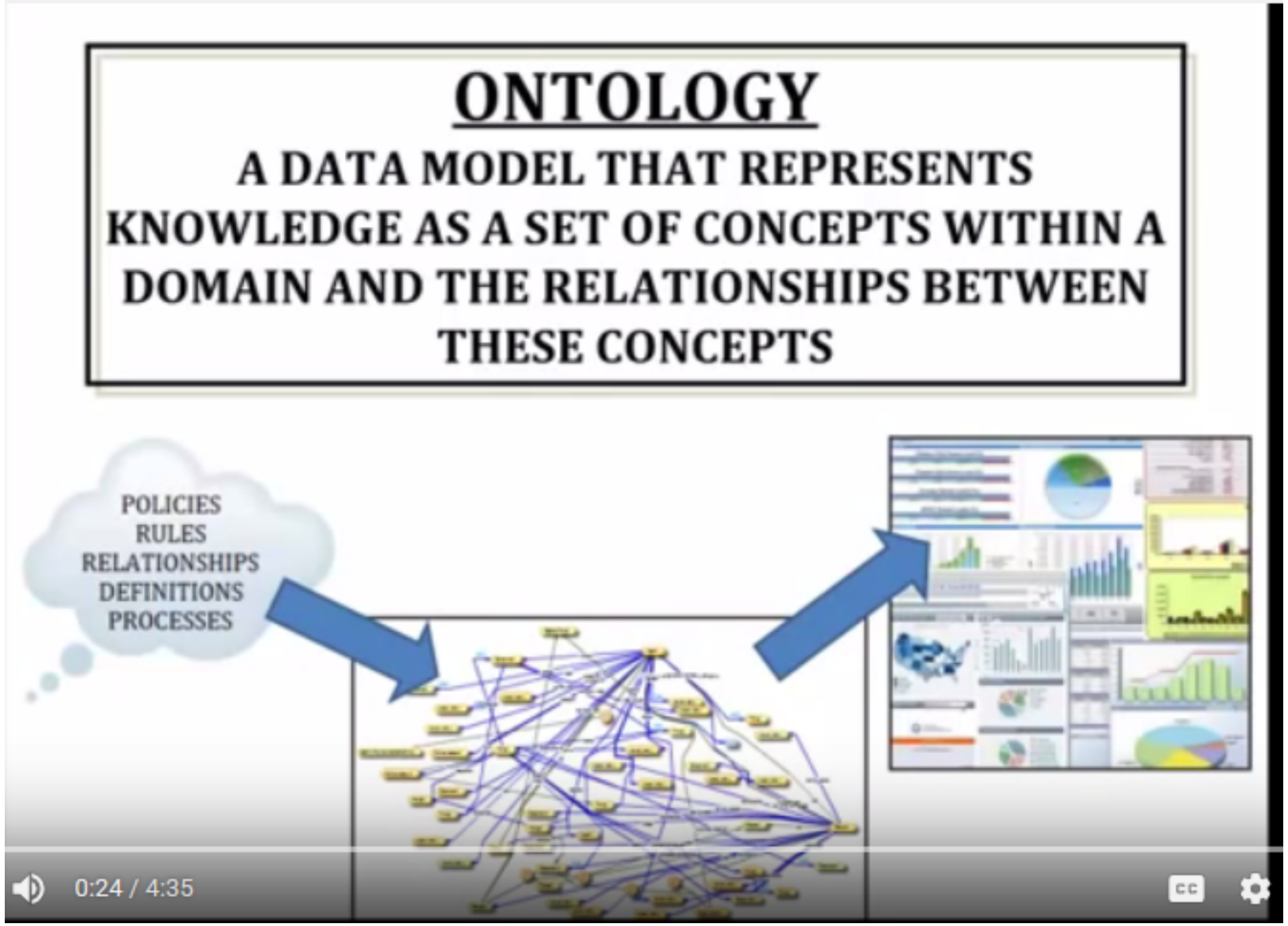
- Making the meaning so clear a computer can understand it, or at least utilize it.

(Source: https://www.youtube.com/watch?v=rhgUDGtT2EM)

Ontology Example



(Source: Google Images)



ONTOLOGY

A DATA MODEL THAT REPRESENTS KNOWLEDGE AS A SET OF CONCEPTS WITHIN A DOMAIN AND THE RELATIONSHIPS BETWEEN THESE CONCEPTS

POLICIES
RULES
RELATIONSHIPS
DEFINITIONS
PROCESSES

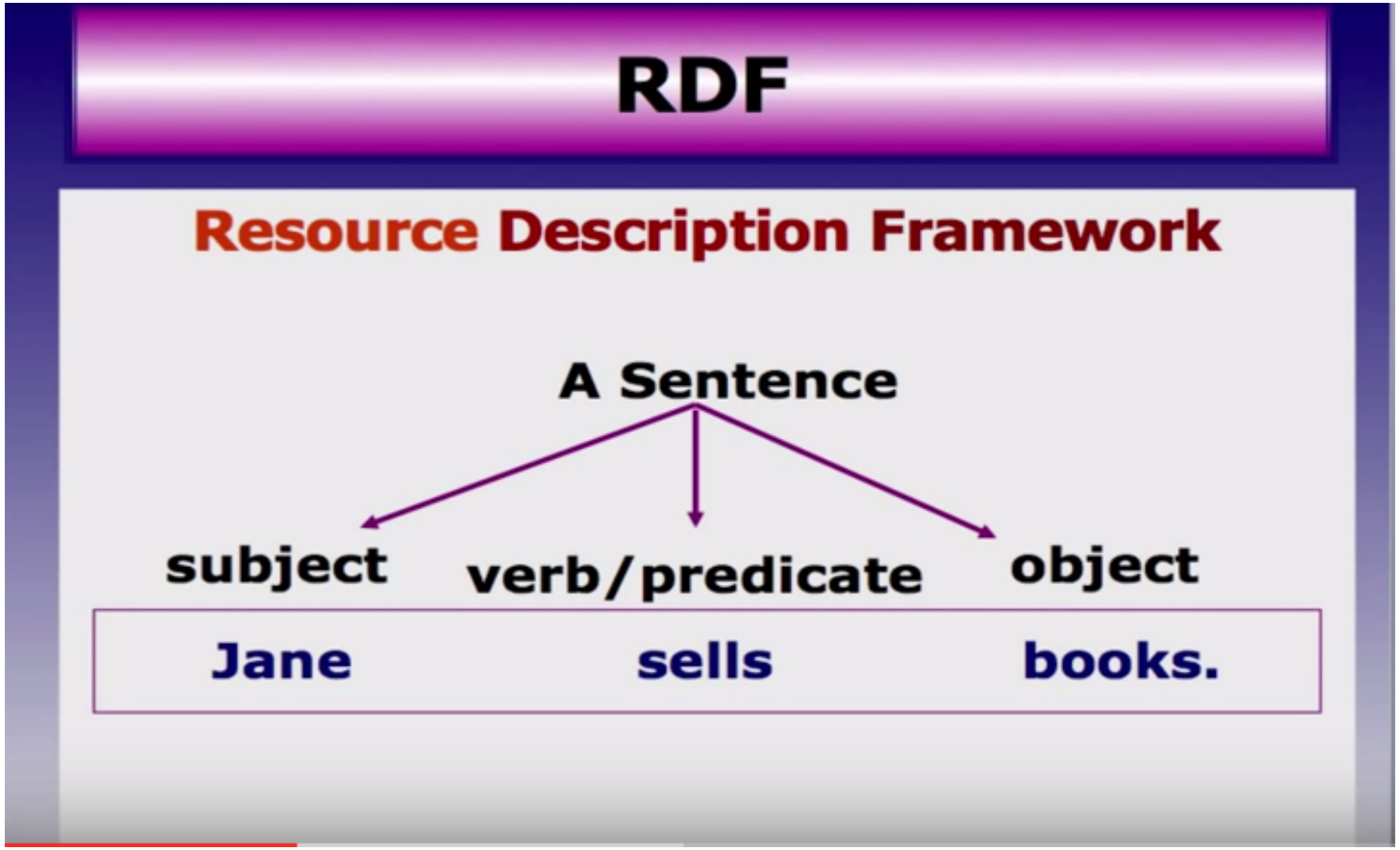
(Source: https://www.youtube.com/watch?v=jfUPLuPL3Ho)

What is OWL?

The W3C Ontology Web Language is a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things.

```
<?xml version="1.0"?>
<rdf:RDF
    xmlns:shop="http://www.workingontologist.org/Examples/Chapter5/Shopping.owl#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:daml="http://www.daml.org/2001/03/daml+oil#"
  xml:base="http://www.workingontologist.org/Examples/Chapter5/Shopping.owl">
  <owl:Ontology rdf:about="">
    <owl:versionInfo rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >Created with TopBraid Composer</owl:versionInfo>
  </owl:Ontology>
<owl:Class rdf:ID="Oxfords">
    <rdfs:subClassOf rdf:resource="#Shirts"/>
  </owl:Class>
  <shop:Oxfords rdf:ID="ClassicOxford">
    <rdf:type rdf:resource="#Shirts"/>
  </shop:Oxfords>
  <shop:Henleys rdf:ID="ChamoisHenley"/>
  <shop:Tshirts rdf:ID="BikerT">
    <rdf:type rdf:resource="#MensWear"/>
  </shop:Tshirts>
</rdf:RDF>
```

Typical OWL File



RDF

Resource Description Framework

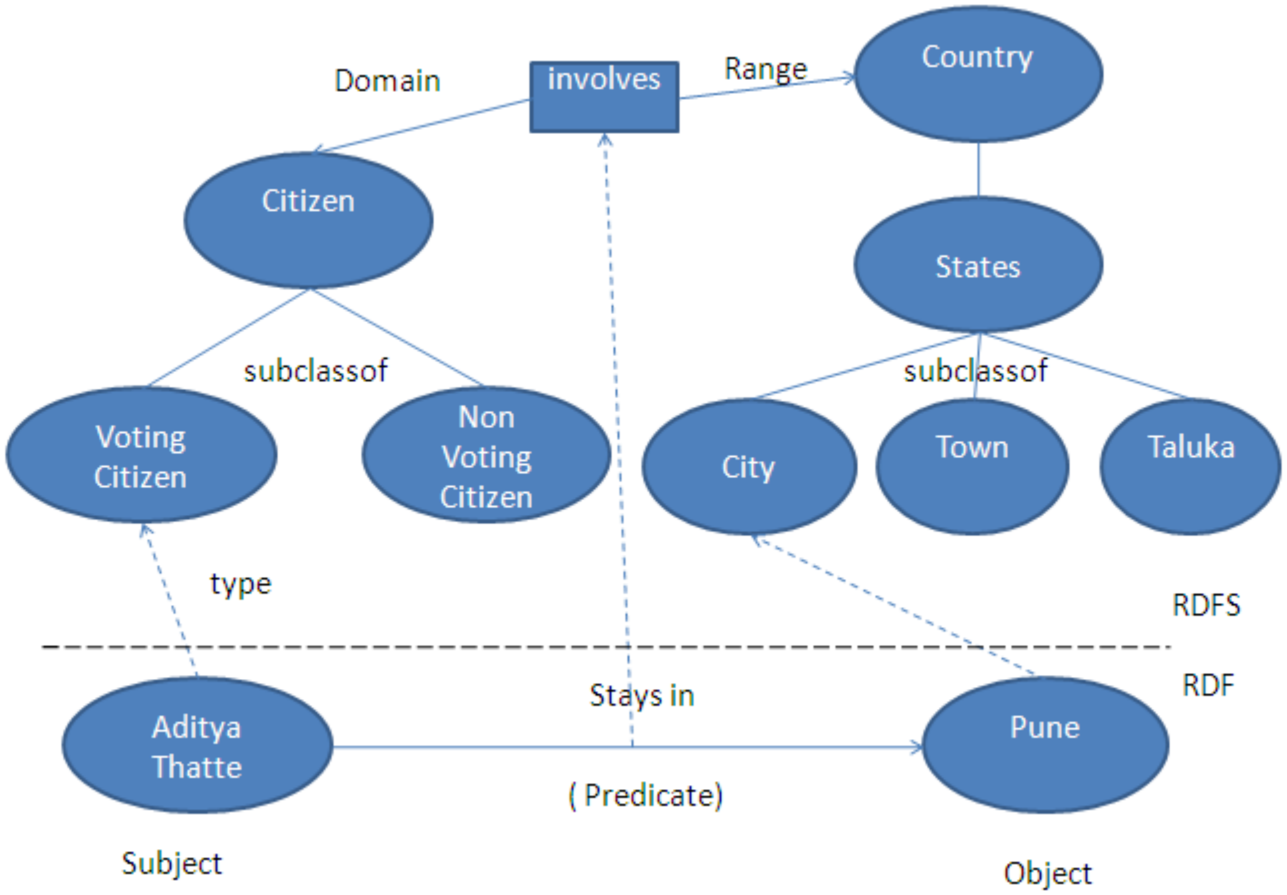
A Sentence

subject verb/predicate object

Jane sells books.

(Source: https://www.youtube.com/watch?v=rhgUDGtT2EM)

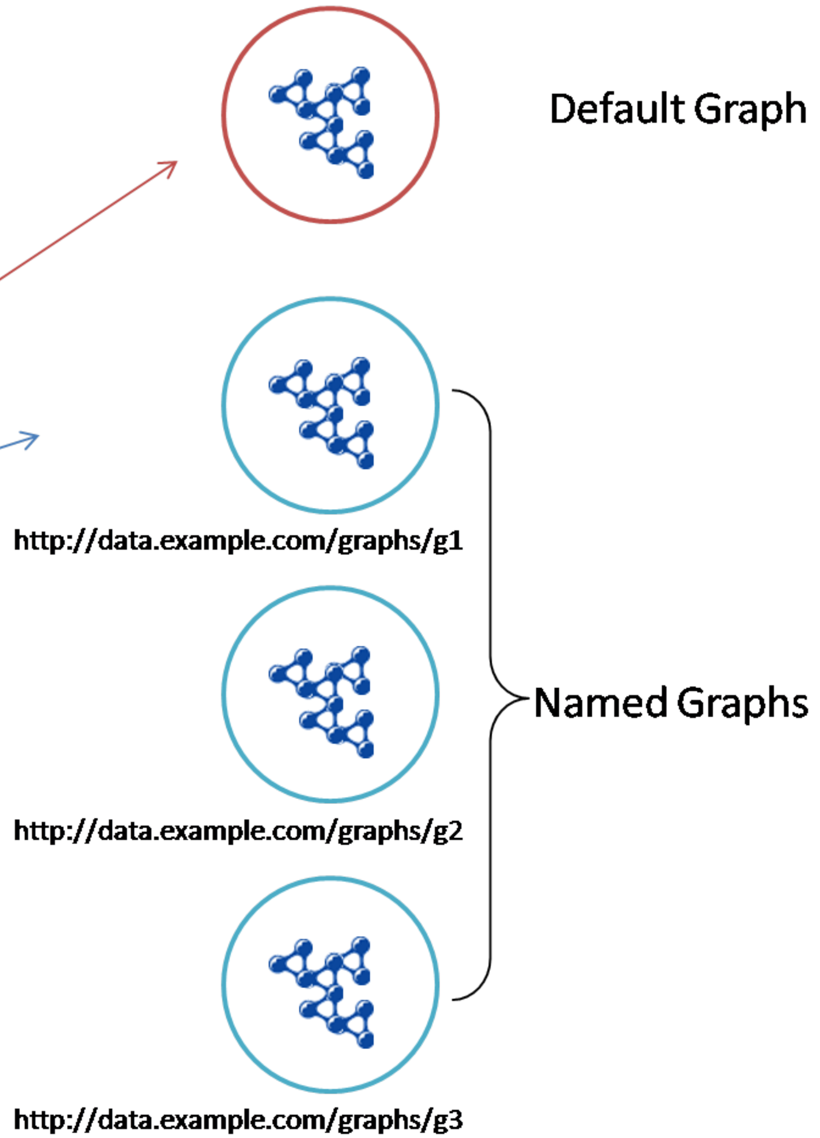
RDF Example



(Source: Google Images)

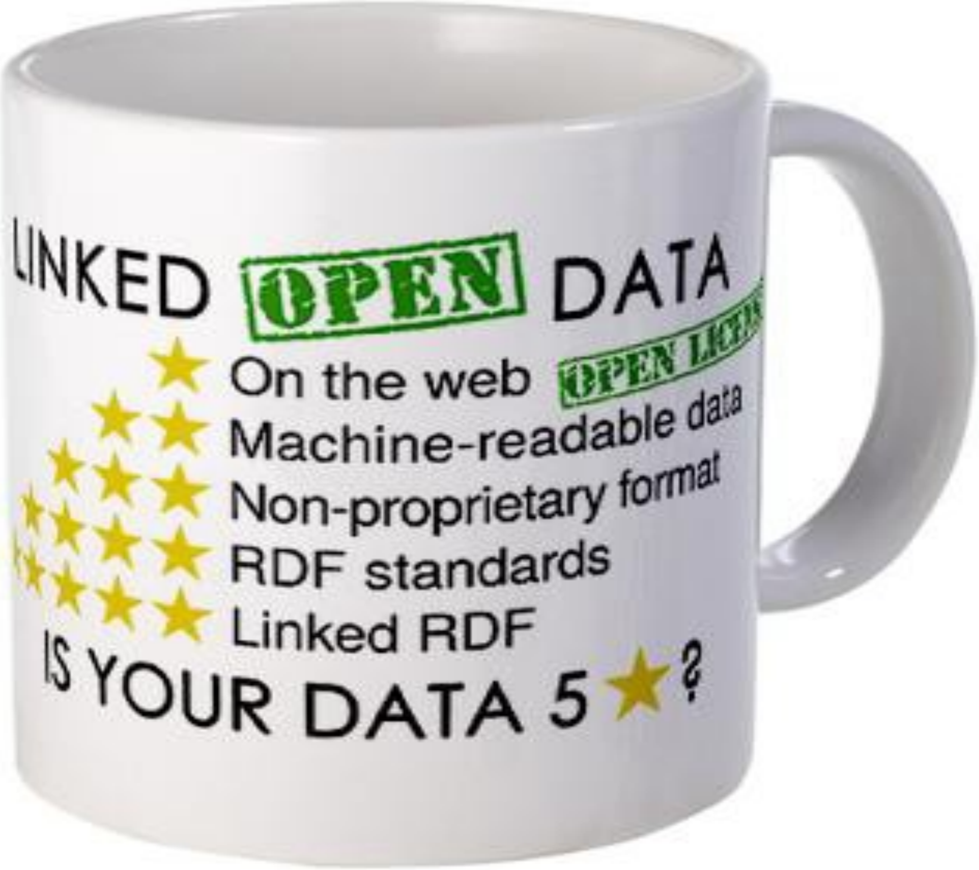
SPARQL Example

```
PREFIX g: <http://data.example.com/graphs/>
PREFIX ex: <...>
SELECT *
FROM <...>
FROM NAMED g:g1
FROM NAMED g:g2
FROM NAMED g:g3
WHERE {
    ?s ex:p1 ex:o1 ; ex:p2 ex:o2 .
    GRAPH g:g1 { ?s ex:p3 ex:o3 }
    GRAPH ?g {
        ex:s1 ex:p4 ?s .
        ex:s1 ex:p5 ex:o5 .
    }
}
```


Default Graph
http://data.example.com/graphs/g1
Named Graphs
http://data.example.com/graphs/g2
http://data.example.com/graphs/g3

(Source: Google Images)

What is Linked Data ?



LINKING OPEN DATA

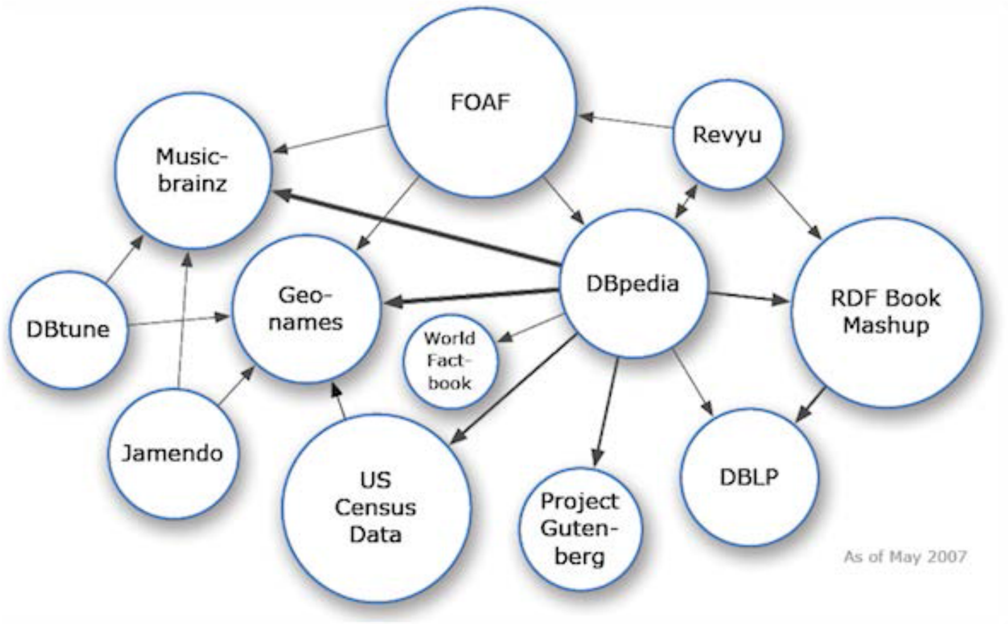


Diagram maintained by Richard Cyganiak (DERI, NUI Galway) and Anja Jentzsch (Freie Universität Berlin)

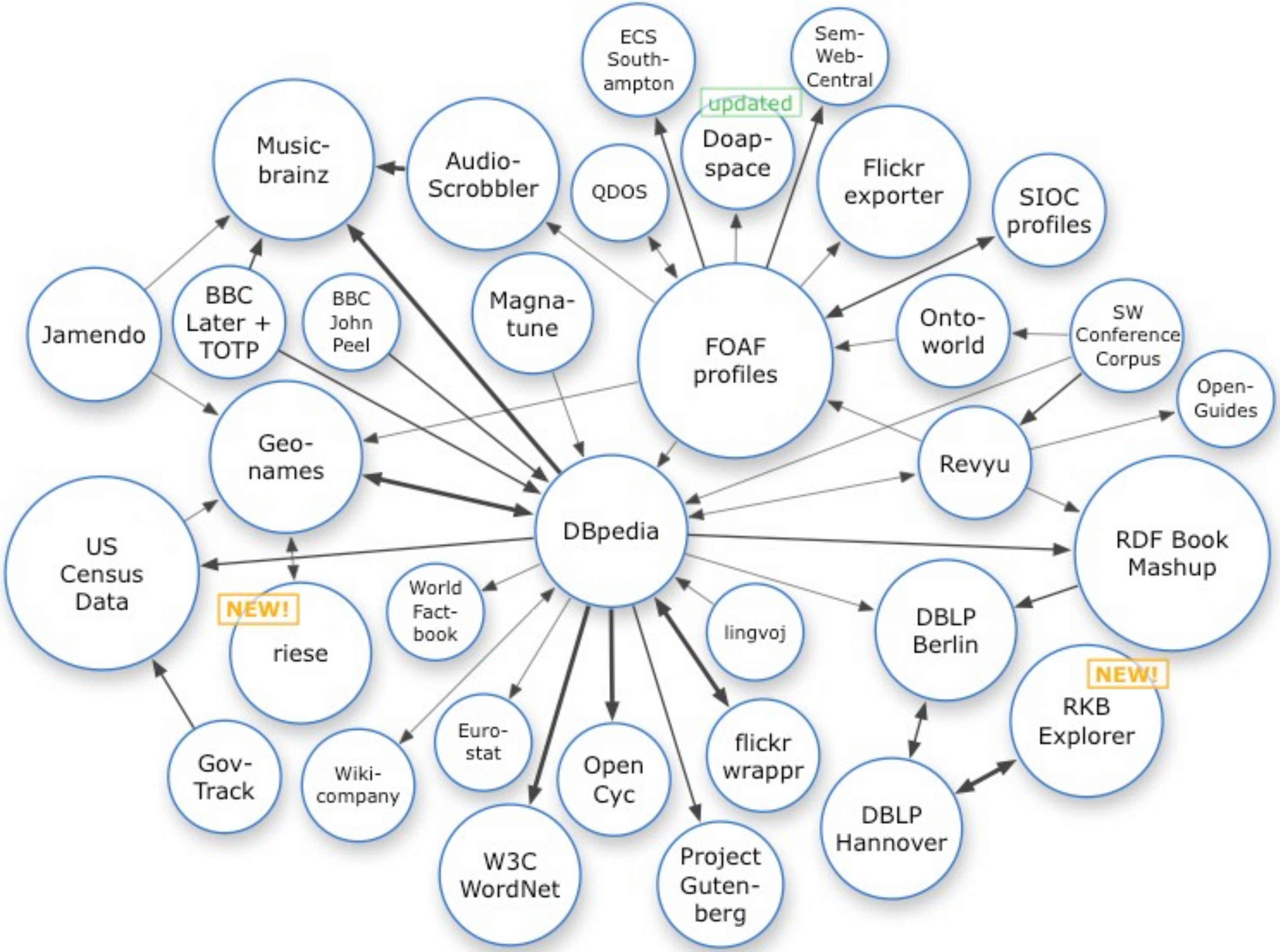
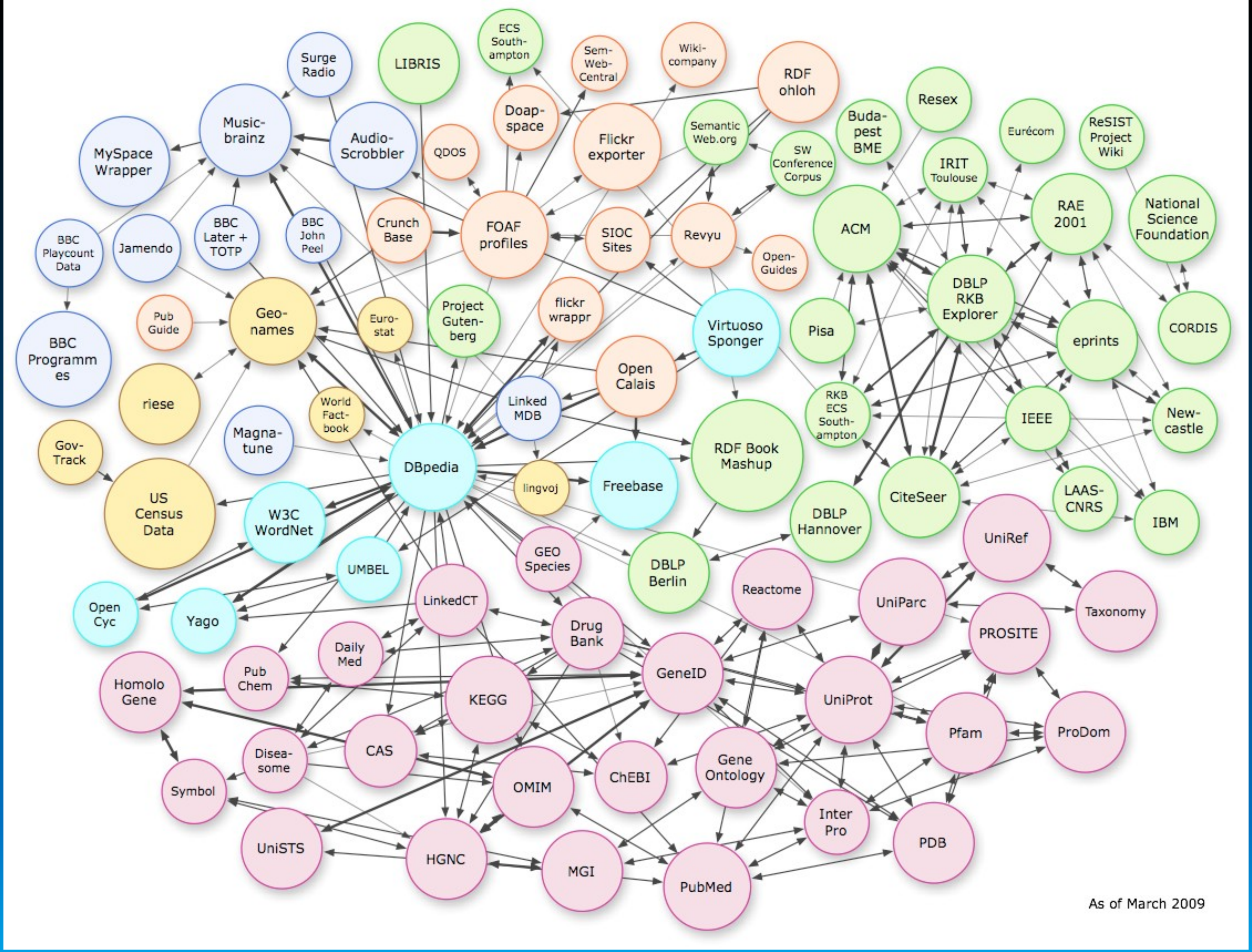


Diagram maintained by Richard Cyganiak (DERI, NUI Galway) and Anja Jentzsch (Freie Universität Berlin)



As of March 2009

Diagram maintained by Richard Cyganiak (DERI, NUI Galway) and Anja Jentzsch (Freie Universität Berlin)

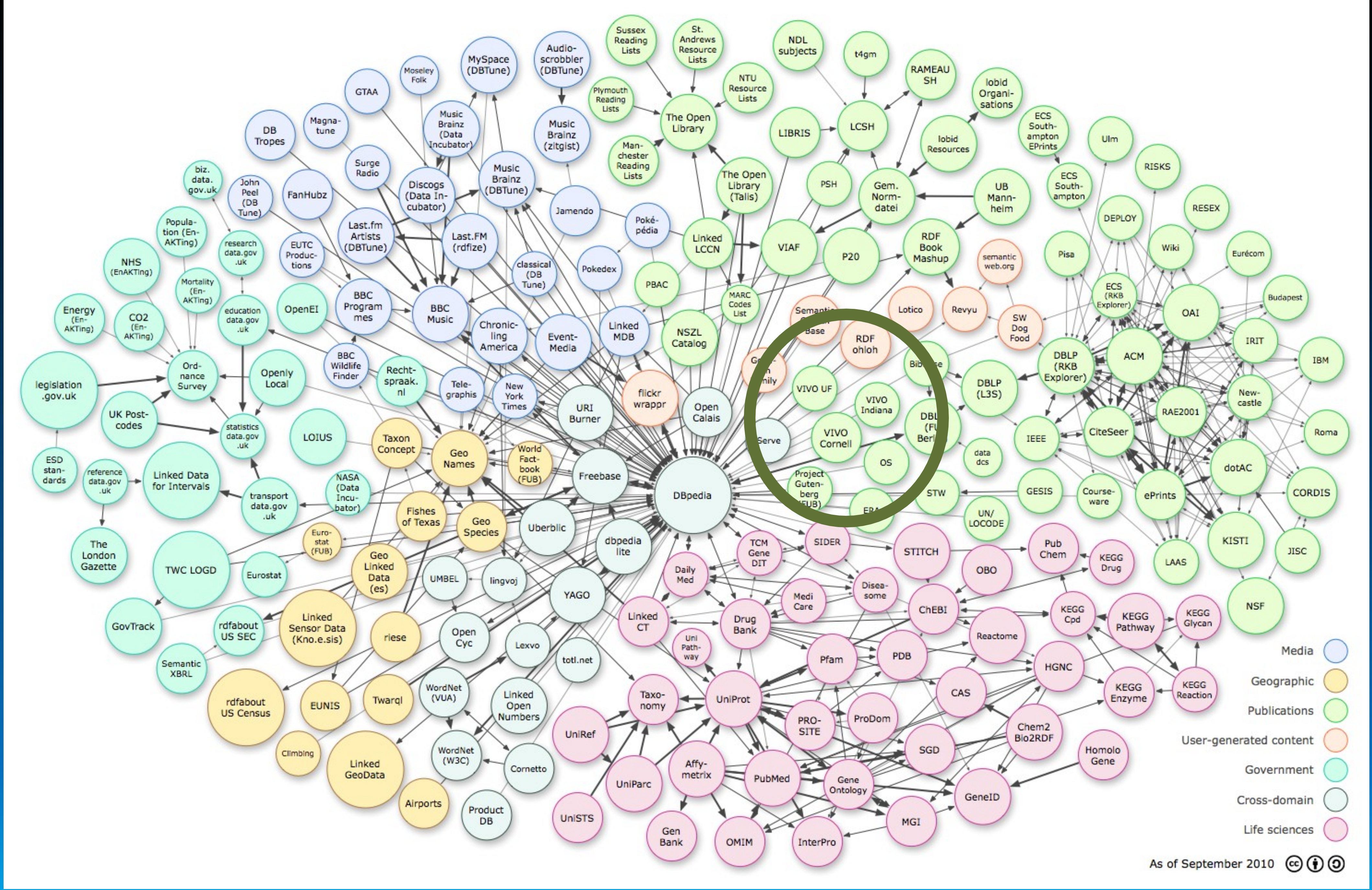
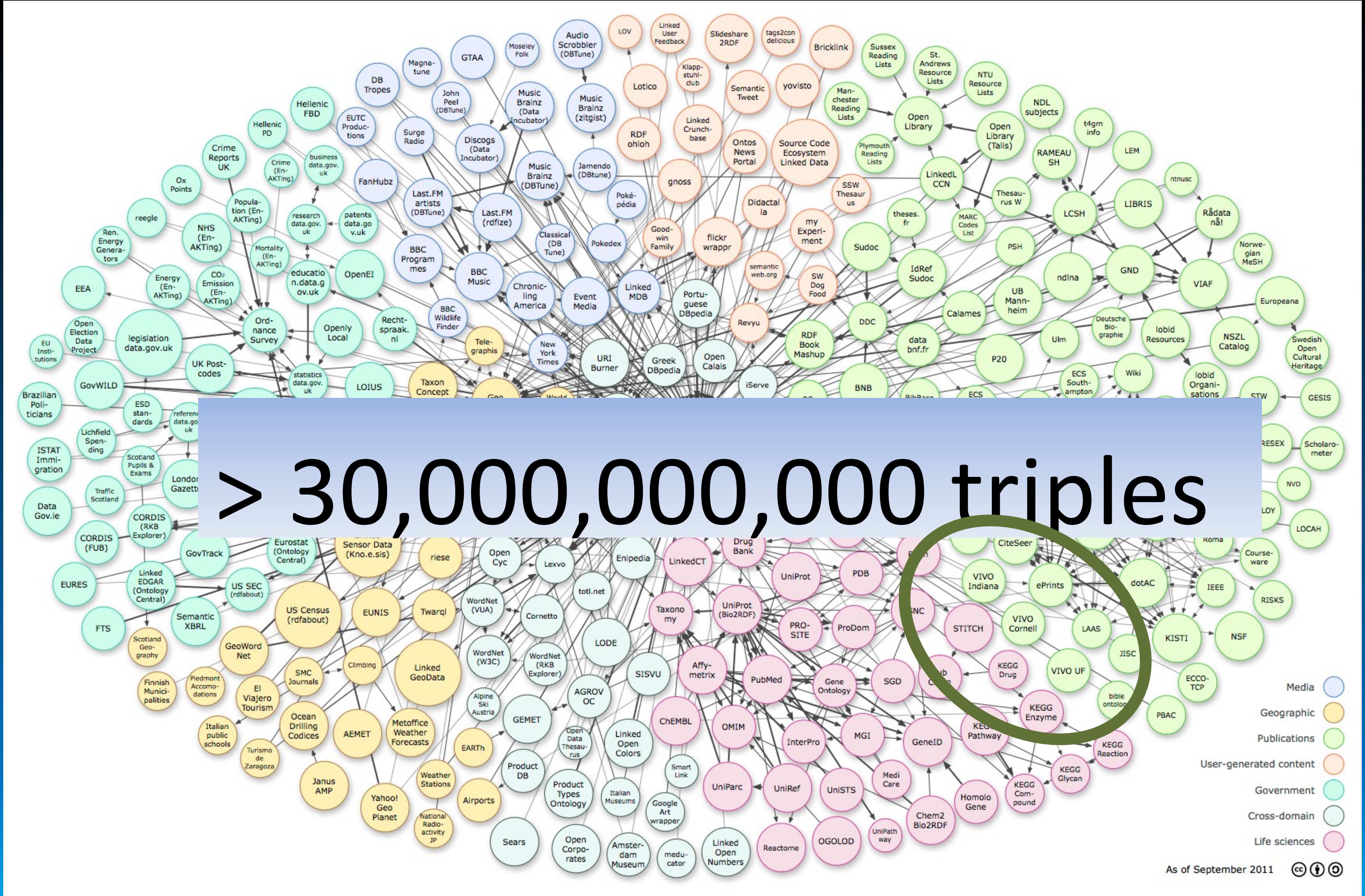
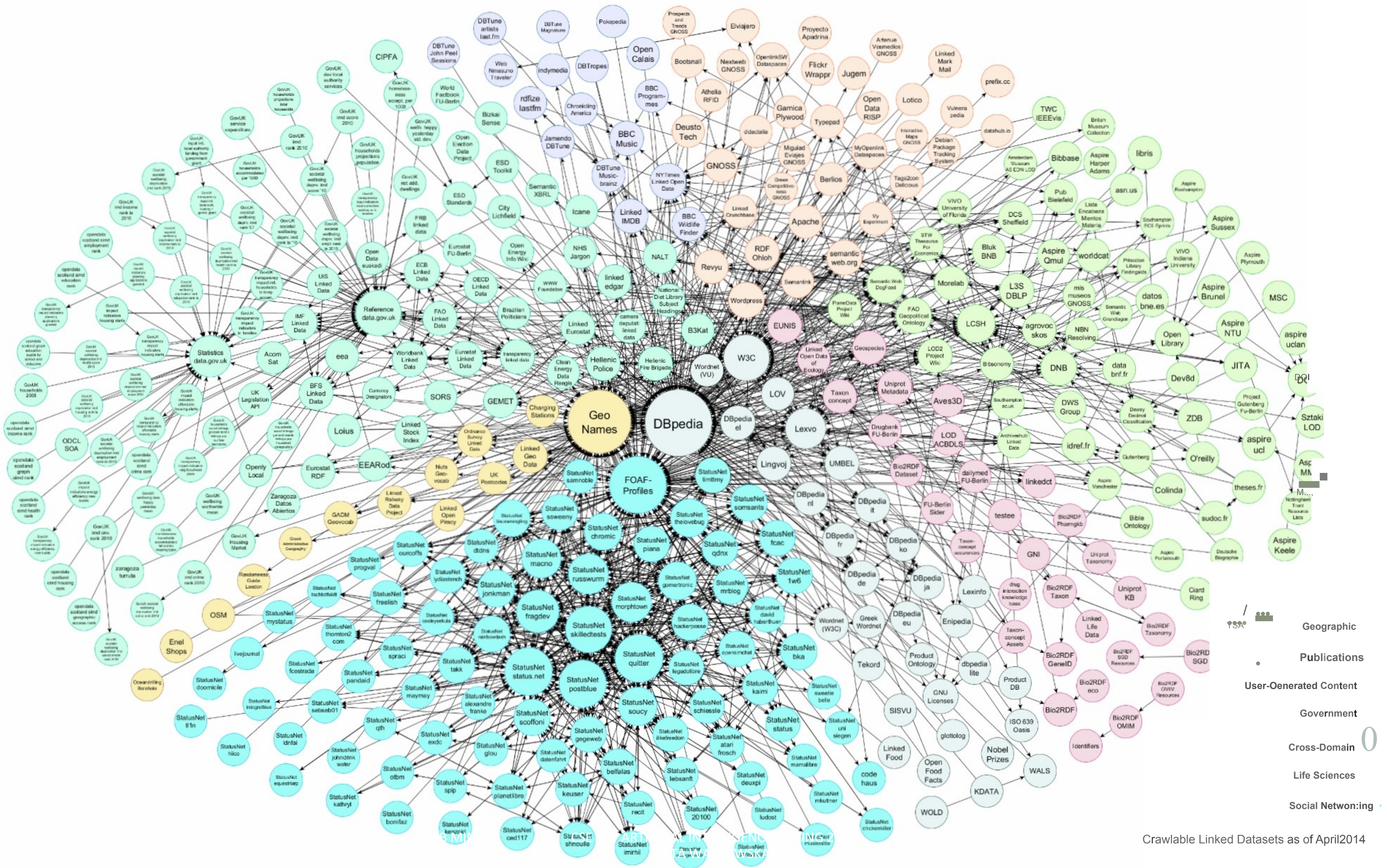


Diagram maintained by Richard Cyganiak (DERI, NUI Galway) and Anja Jentzsch (Freie Universität Berlin)



> 30,000,000,000 triples

Diagram maintained by Richard Cyganiak (DERI, NUI Galway) and Anja Jentzsch (Freie Universität Berlin)



Crawlable Linked Datasets as of April2014

Semantic web mining

Semantic web mining combines semantic web methodology and web mining technology. Better semantic web ontology can refine web mining algorithm and enhance web mining result. The web mining result can also extend the scope of semantic web ontology (domain knowledge). This is a win – win situation.

Semantic web perspective

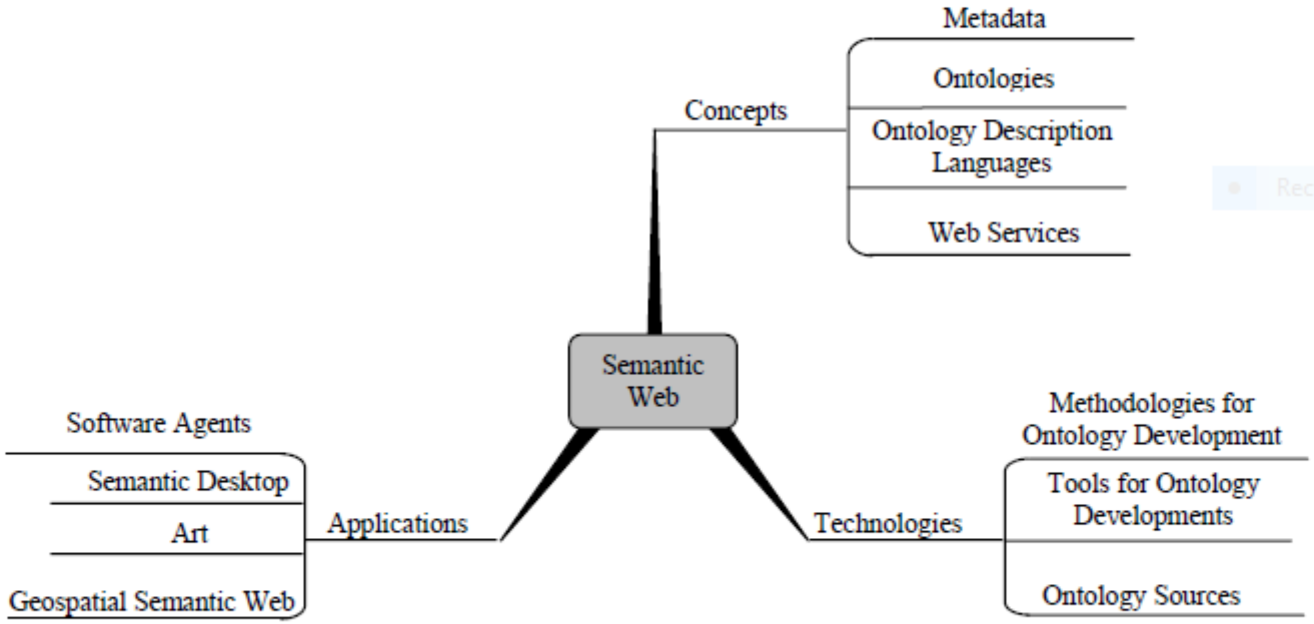
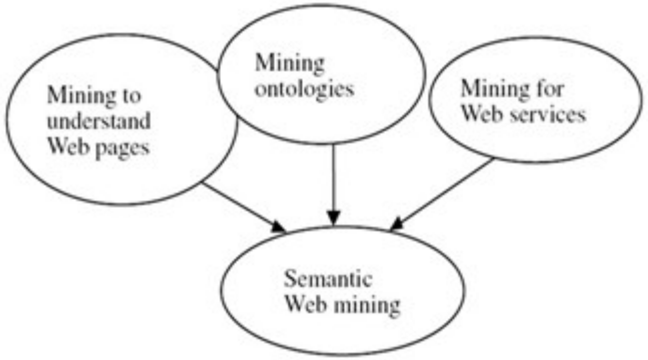


Fig. 2. Semantic web perspectives [9].

(Source: Hamed Hassanzadeh and Mohammad Reza Keyvanpour;*International Journal of Computer Theory and Engineering, Vol. 4, No. 4, August 2012)*

Semantic web mining components



(Source: Hamed Hassanzadeh and Mohammad Reza Keyvanpour;*International Journal of Computer Theory and Engineering, Vol. 4, No. 4, August 2012)*

Semantic web major requirements and tasks

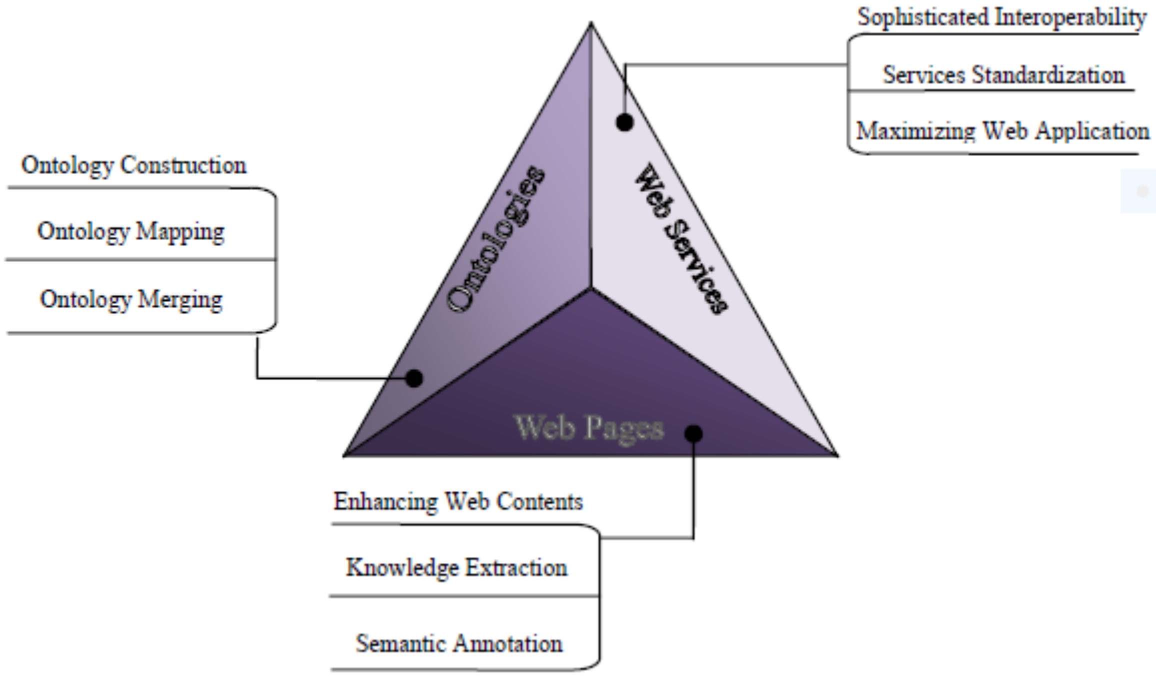
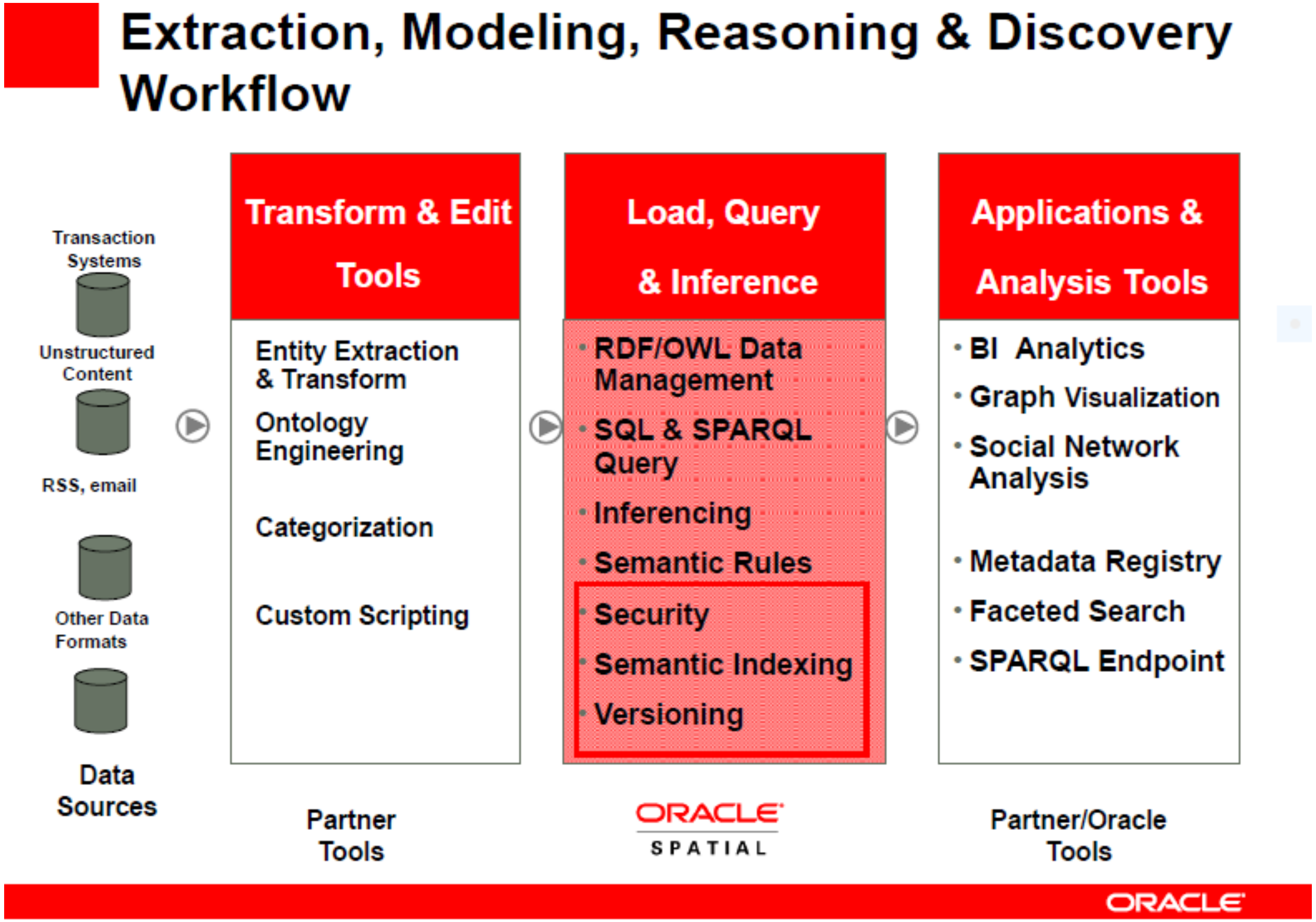


Fig. 3. Semantic web major requirements and tasks.

(Source: Hamed Hassanzadeh and Mohammad Reza Keyvanpour;*International Journal of Computer Theory and Engineering, Vol. 4, No. 4, August 2012)*



Extraction, Modeling, Reasoning & Discovery Workflow

Transaction Systems

Unstructured Content

RSS, email

Other Data Formats

Data Sources

Transform & Edit Tools

Entity Extraction & Transform

Ontology Engineering

Categorization

Custom Scripting

Partner Tools

Load, Query & Inference

- RDF/OWL Data Management
- SQL & SPARQL Query
- Inferencing
- Semantic Rules
- Security
- Semantic Indexing
- Versioning

ORACLE SPATIAL

Applications & Analysis Tools

- BI Analytics
- Graph Visualization
- Social Network Analysis
- Metadata Registry
- Faceted Search
- SPARQL Endpoint

Partner/Oracle Tools

ORACLE

Commercial companies are ready for semantic web

Semantic web mining applications

- **Faculty Report Card project:**

Local Project for Stony Brook University School of Medicine Dean's office.
Data source is from Triple store of PubMed.

- **Bio2RDF Project**

The Bio2RDF project aims to transform silos of life science data into a globally distributed network of linked data for biomedical knowledge translation and discovery.
(source: https://datahub.io/dataset/bio2rdf-pubmed)
378 datasets found

- **BestBuy use of GoodRelations/RDFa Markup to increase site traffic and promote better search result for their users**

Future direction of semantic web mining:

Semantic web mining is a new area in web mining. The combination of these two areas will bring a great success to World Wide Web. But due to the lack of global standards and lack of rugged database management system to manage semantic web mining opens up new avenues for the researchers to develop KIMS (Knowledge extraction management system) for unstructured data available on the web this area is slowly developing. If these fields explored in a right manner it will provide unlimited opportunities to extract knowledge from the goldmine of unstructured data available across the globe.

(Source: Amruta Arun Joshi et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (1) , 2015, 431-433)

Thank You !

Q /A ?