Linked Data Quality

What can Linked Data learn from traditional DQM and on the contrary

Ing. D. Pejčoch

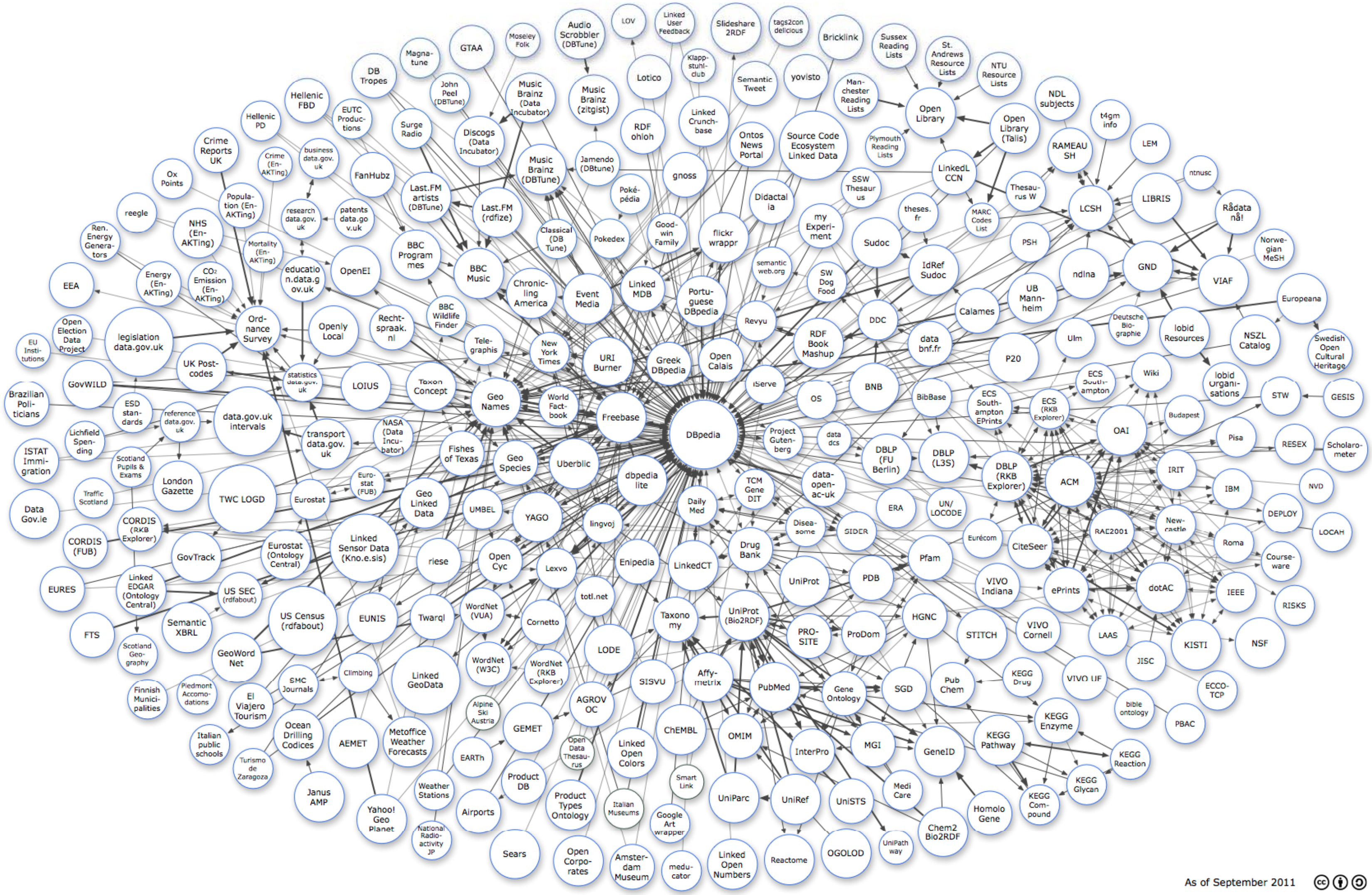
Outline

- What is Linked Data (technologies, relations to Big Data concept, ...)
- Data Quality Management
- Specific DQ problems of Linked Data
- Linked Data in context of global Data Quality Management

What do you first imagine when I say

"Linked Data"

... is it LOD* Cloud?



* LOD = Linking Open Data

Src:

What is Linked Data?

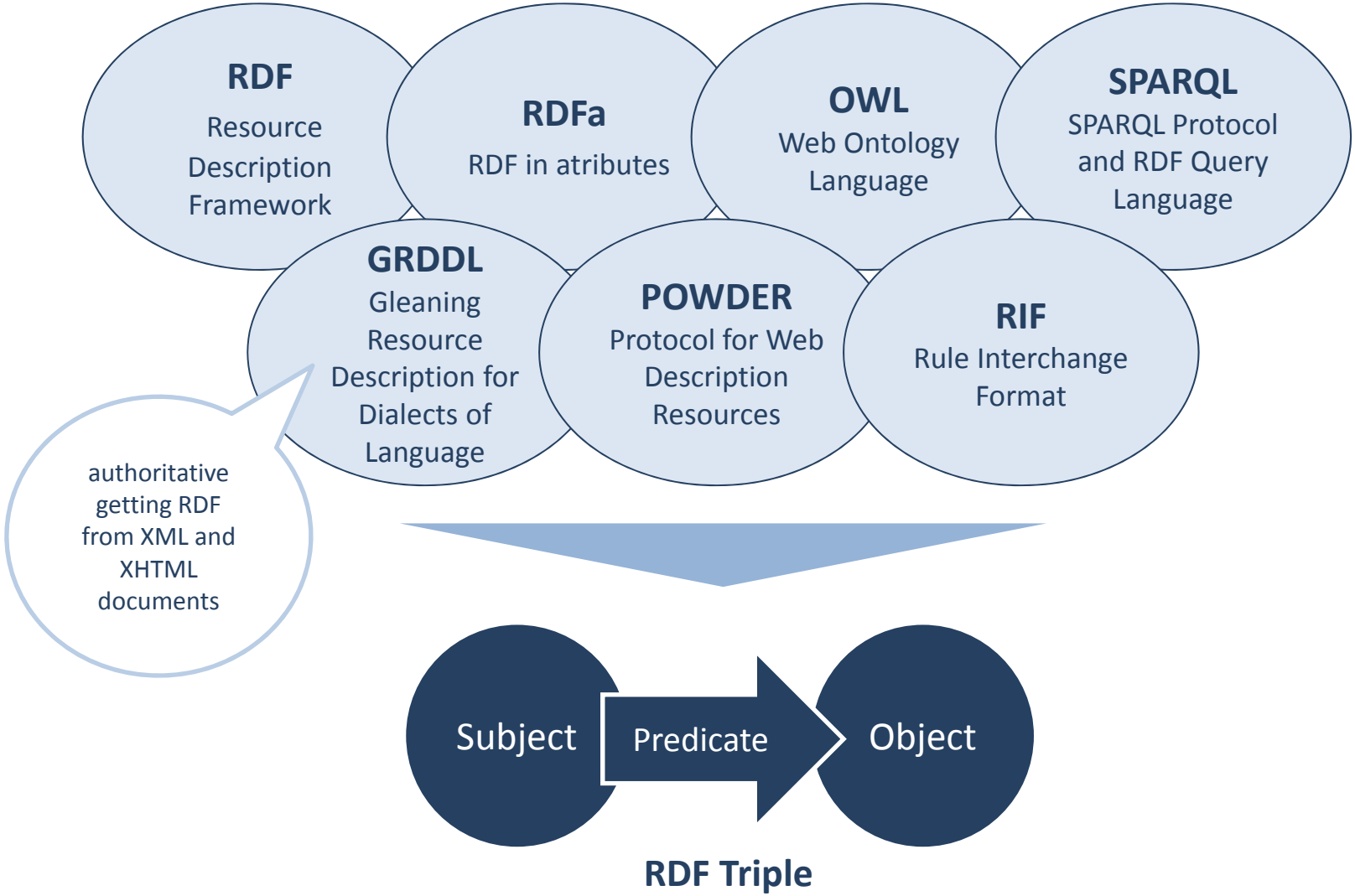
„Set of best practices to publish and interlink data on the web“

Principles (Tim Berners-Lee):

- URI as names
- Dereference using HTTP URI
- Useful information as a target
- Links to other URIs

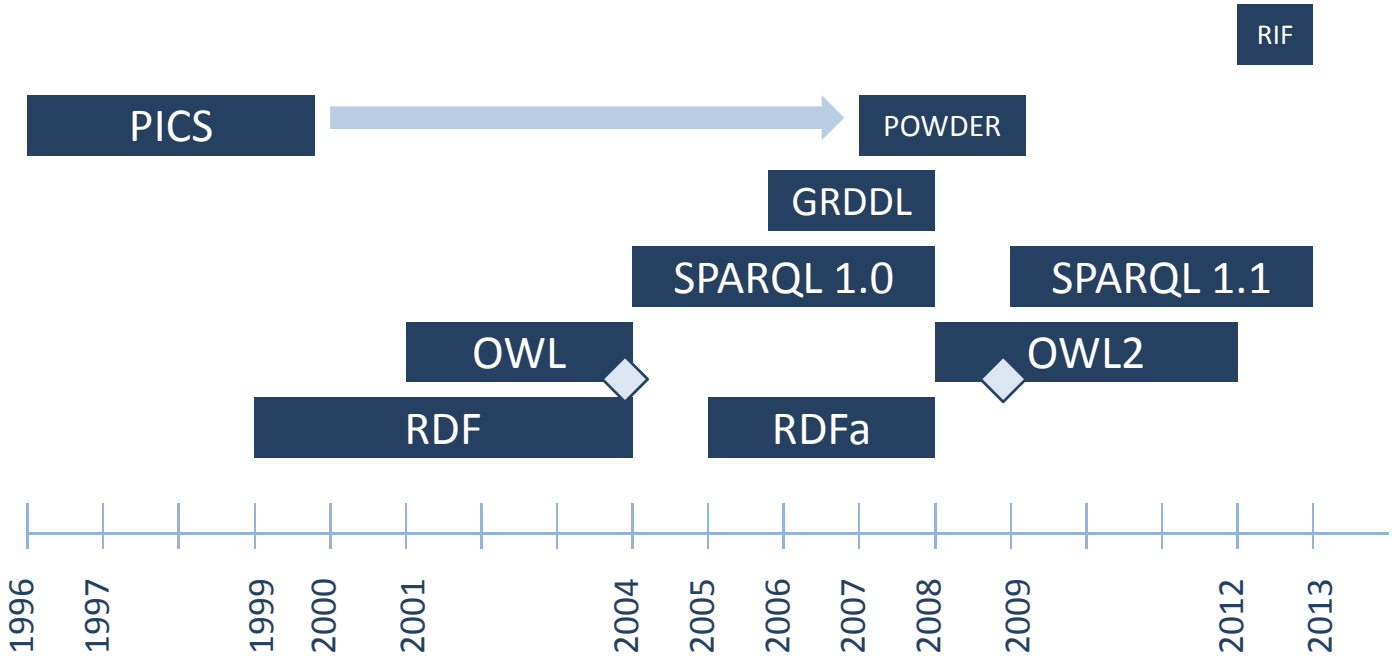
... related to the concept of Semantic Web (web of data)

Techniques of Semantic Web

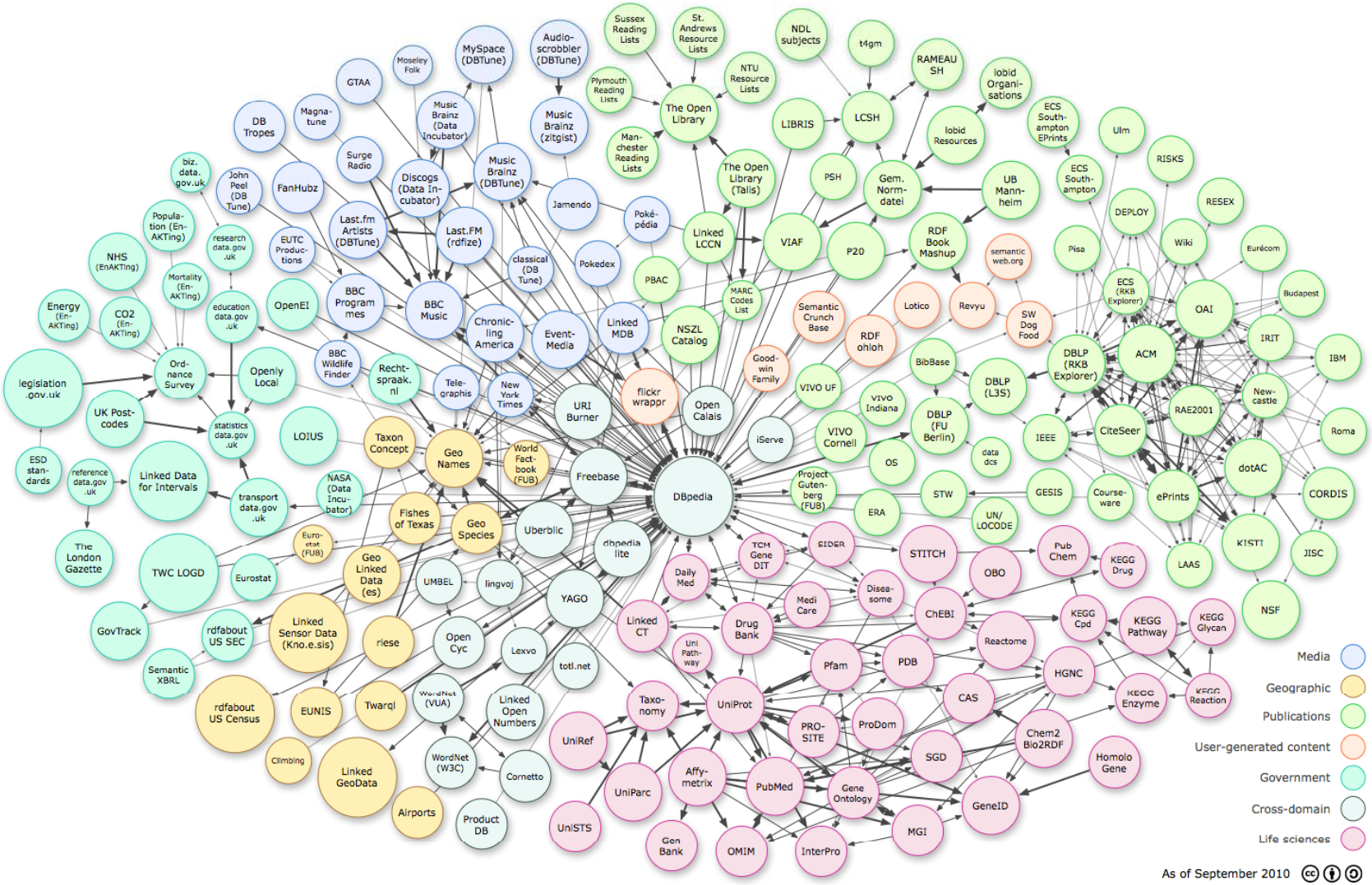


Src: [10]

Techniques of Semantic Web: Timeline

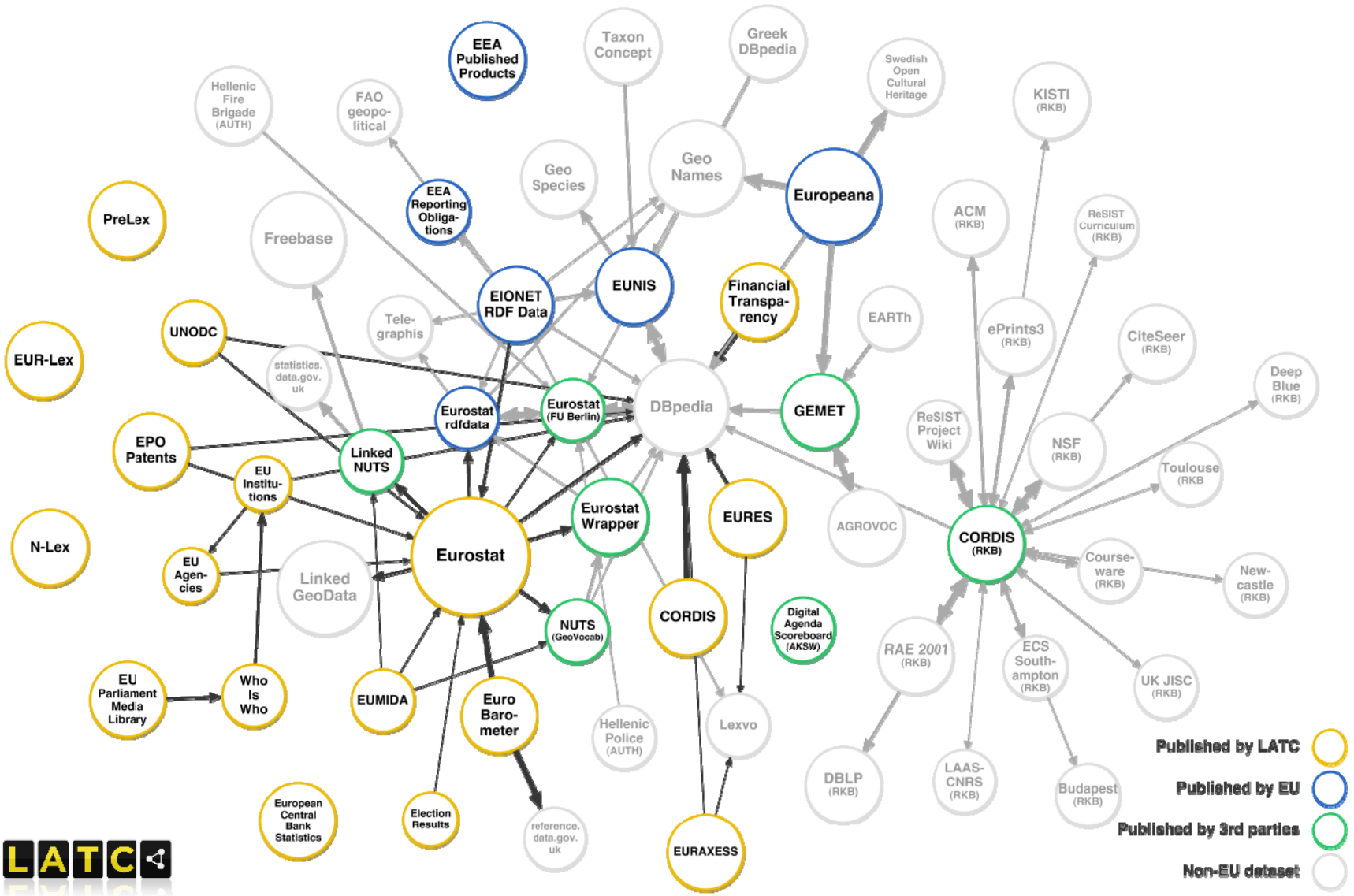


Src: [10]


Media
Geographic
Publications
User-generated content
Government
Cross-domain
Life sciences
As of September 2010

Src: [11]

Linked Open Data Around The Clocks



Another sources of RDF data

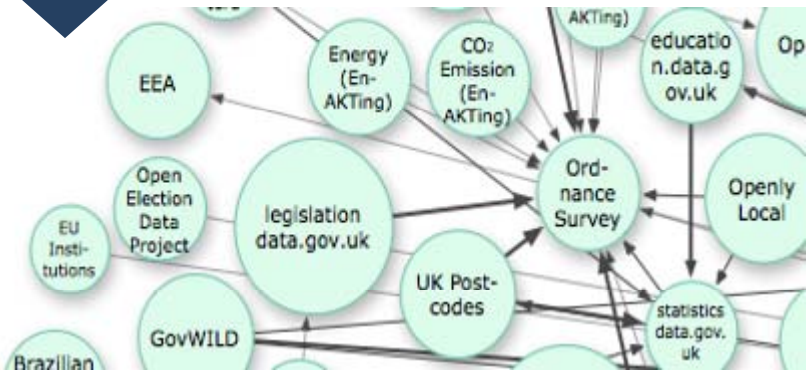
Over 6.4 billion Resource Description Framework (RDF)

geodata.gov U.S. Maps & Data → DATA.GOV EMPOWERING PEOPLE

DISCOVER POSSIBILITIES opengov — Open Government Platform (OGPL)
developed through India - US collaboration

OPENDATA CZ — 10 dsets

DATA.GOV.UK (beta) Opening up Government



publicdata.eu Europe's Public Data BETA

LOD2

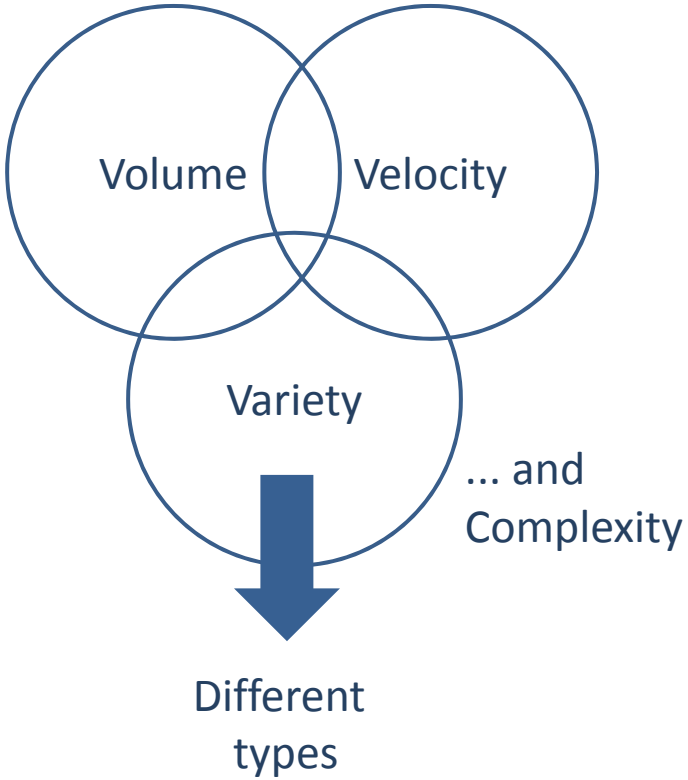
		RDF:
Finance and Budgeting (436)		
Social Questions (226)	Environment	uk (171)
Transportation (196)		it (3)
Education and Communication (188)		fr (2)
Agriculture, Fisheries, Forestry (176)		es (2)
Population (145)		nl (1)
Economy and Industry (114)	Health	at (1)

Is Linked Data big enough to be considered as a

„**Big Data**“

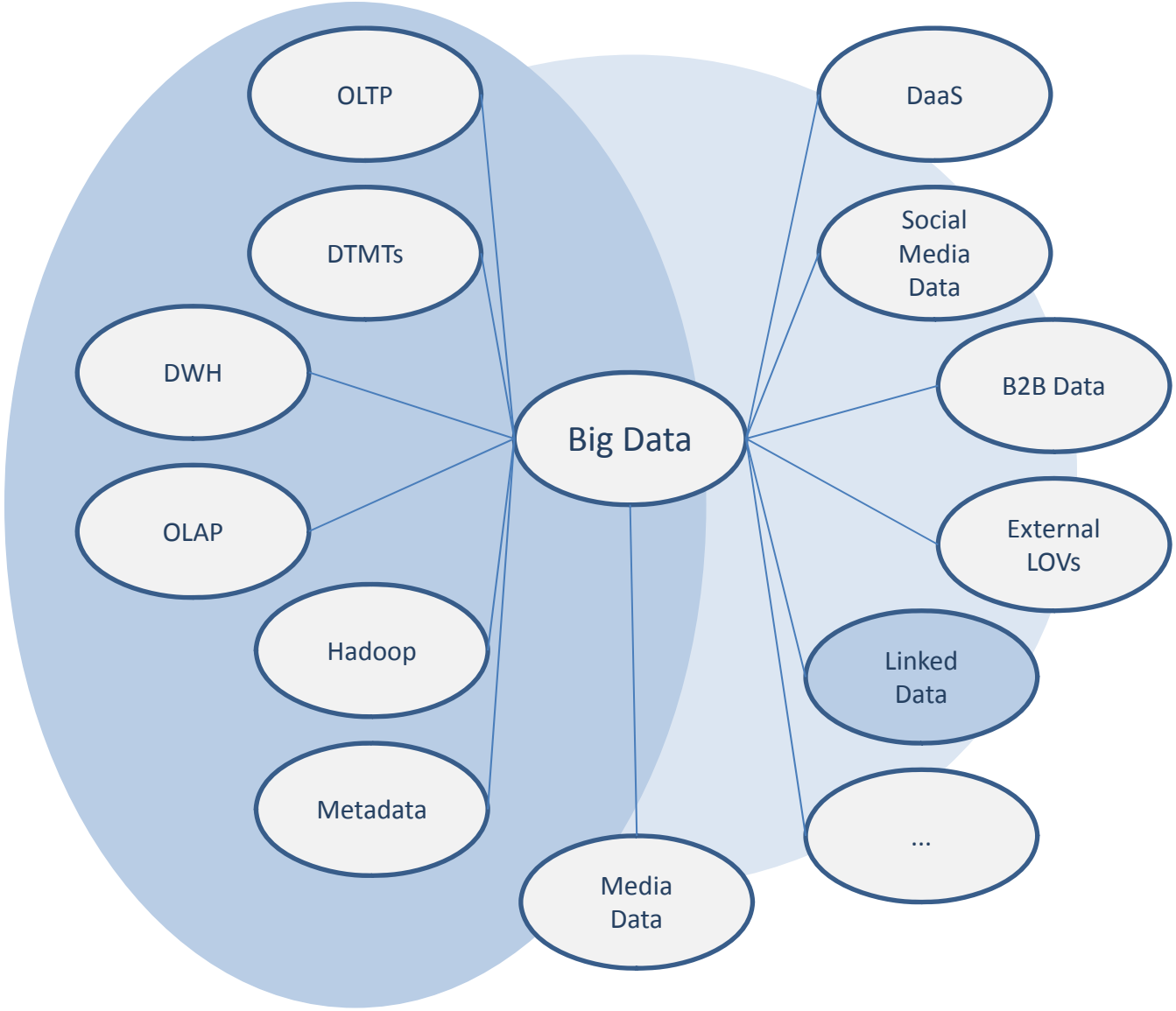
Big Data

Gartner: Big Data Definition



- In relation to cloud computing
- Structuralized / not structuralized data
- Not possible to capture and process using common SW tools
- Hadoop (paralell processing of data in No SQL structures)
- Gartner: *„Year 2013 is the year of Big Data“*
- SAS: different Variability and Complexity => Big Data or rather Huge / Wide Data?

Sources of Wide Data



LD Business Case: Insurance Company

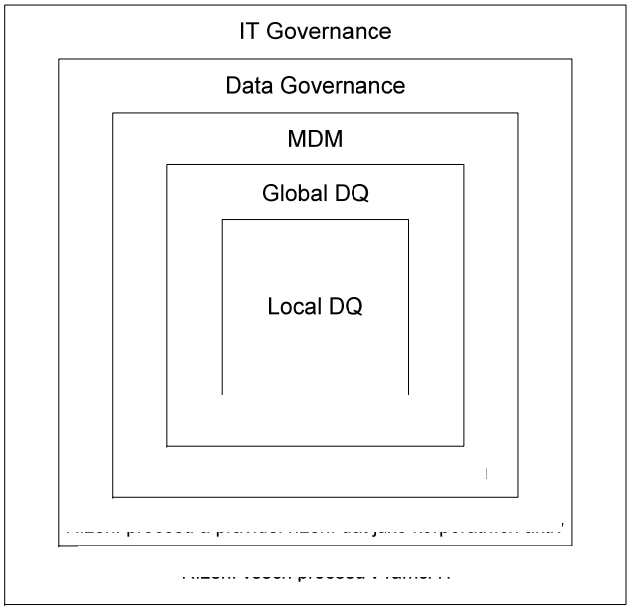
- Registry of Cars (Czech Insurance Office)
- Registry of „SPZ“ (car owner)
- Central Database of Claims
- Database of Frauds

One World of Big Data

- Single Data Governance Principles applied to whole data environment
- Can I manage characteristics of all data within Wide Data World?

Data Quality Management and Governance

- Data are of high quality "*if they are fit for their intended uses in operations, decision making and planning*" (J. M. Juran). ... a lot of definitions. Practicaly all of them refer to some characteristics which are measured.
- Data Governance: data as an asset, principles, politics, rules, ownership (stewardship), necesary condition for MDM
- Focus on data lineage
- Modern approach is proactive instead of reactive
- Process analyses instead of technical assessment



Common DQM Techniques

- **Data Quality Assessment**:
  - Technical Profiling (pattern analysis + EDA)
  - Verification / Validation: syntax, LOVs, checksums, business rules (consistency), constraints (integrity + allowed values)
  - Root-cause analysis
  - Analyses of implemented controls
  - Process Analysis
- **Unification / Standardization**: schemas, rules
- **Deduplication**: clustering, fuzzy / crisp match-merge
- **Imputation + Enrichment**: using models (explicit / implicit), single values or external data sources
- **Geocoding**: linking to external sources
- **Householding**: identification of relationships among entities
- **Stewardship**: setting up ownership of data
- **Implementation of policies, principles and controls**
- **Permanent Monitoring**: business rules

BP: Using DQ Knowledge Base

- Common **„Semantic Data Element“** => Grammars, Syntax rules, LOVs, expected values, ... , business rules, additional knowledges
- **Usage**: data profilling, monitoring, standardization, validation, ... practically all steps in DQM cycle
- **CDM** (Common Data Model, Canonical Data Model, ...) usually used in online integration as a data model independent on individual application
- **Examples of CDM**: ACORD (Association for Cooperative Operations Research and Development) for the insurance industry, SID for telecommunications, CIM (Common Information Model) for public services, PPDM and MMDM for energetic industry, OAGIS (Open Application Group Integration Specification) for production and supply chains, HL7 (Health Level Seven International) and HIPAA for healthcare, ARTS (The Association for Retail Technology Standards) for sales and finally FPML and SWIFT for capital markets.
- **Forrester:** 58% of respondents answered they use a conventional tool for Enterprise Architecture modelling, 21% of them use the modelling tool that is part of its SOA / BPM (Business Process Management) solution, 4% use the tool centred on XML schema and 17% don’t use any tools. No respondent considers semantic technologies such as RDF (Resource Description Format) or OWL (“Web Ontology Language “) as suitable solution for modelling and managing CDM.

... and Semantic Web?

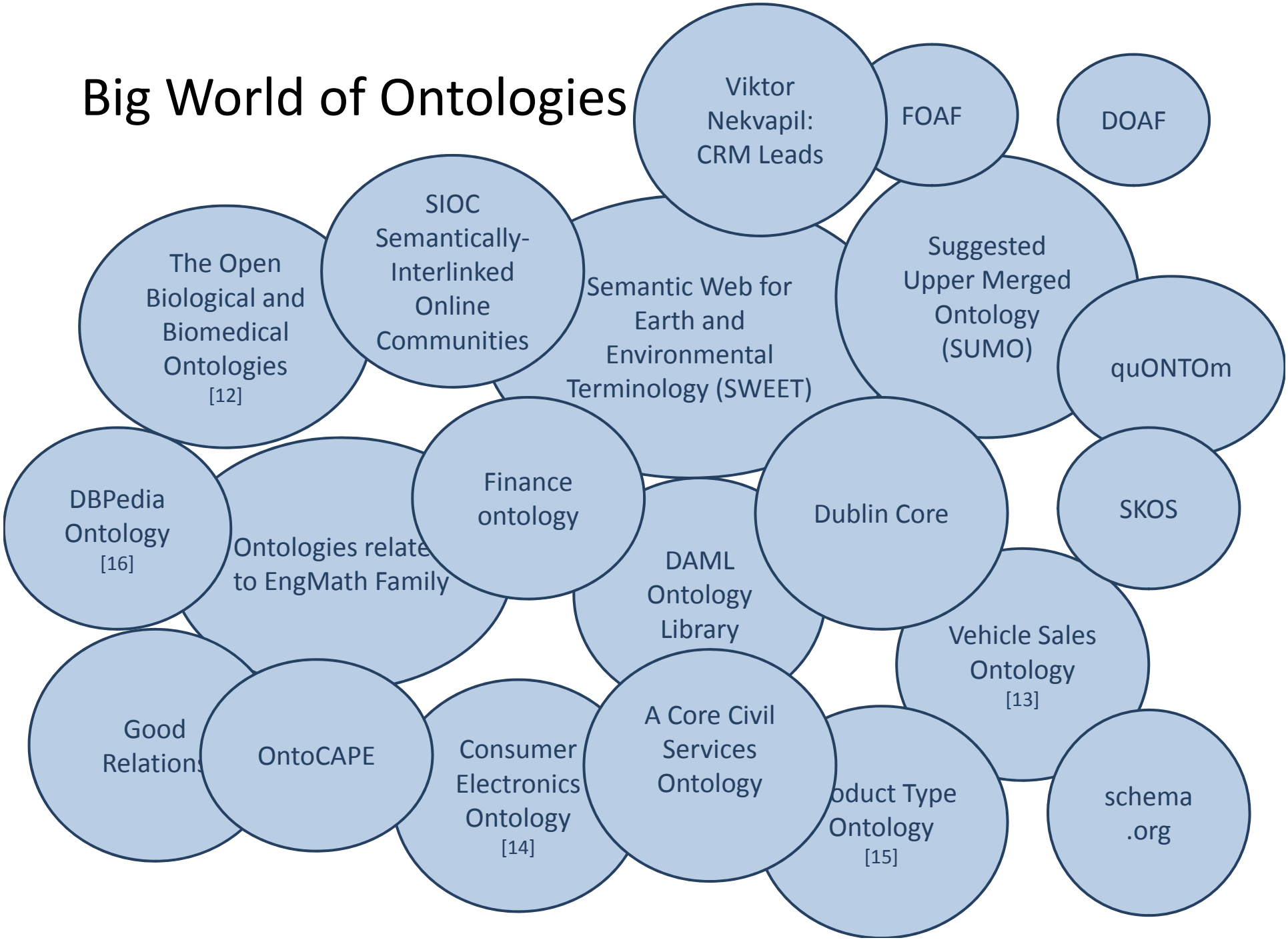
Principle of 3As:

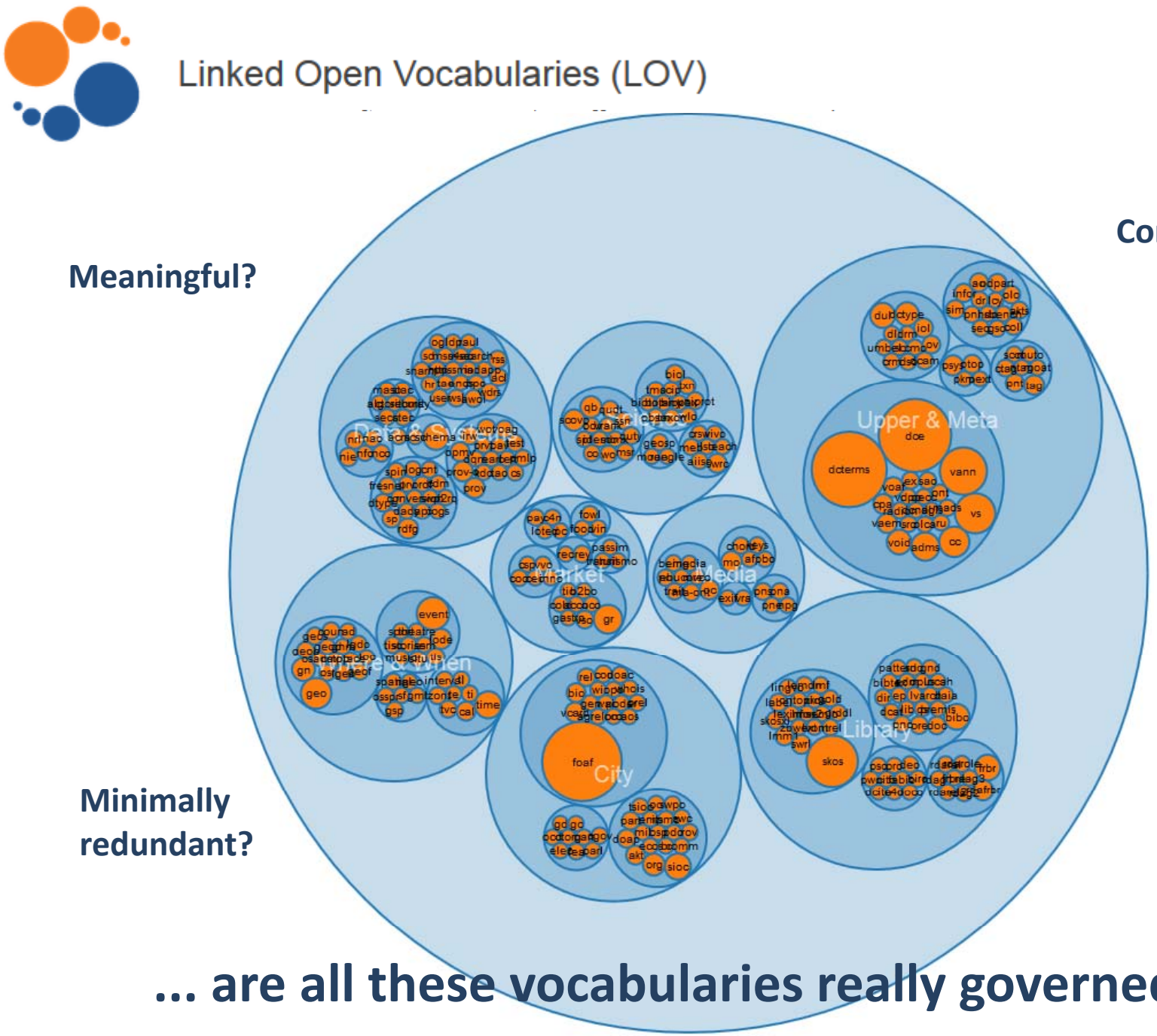
„Anybody can write Anything about Any topic“

.... but is it really true?

- Ontologies / Vocabularies
- Recommendations
- Best Practices
- Ex-post Validation

Big World of Ontologies





Linked Open Vocabularies (LOV)

Meaningful?

Correct?

Minimally redundant?

Opened?

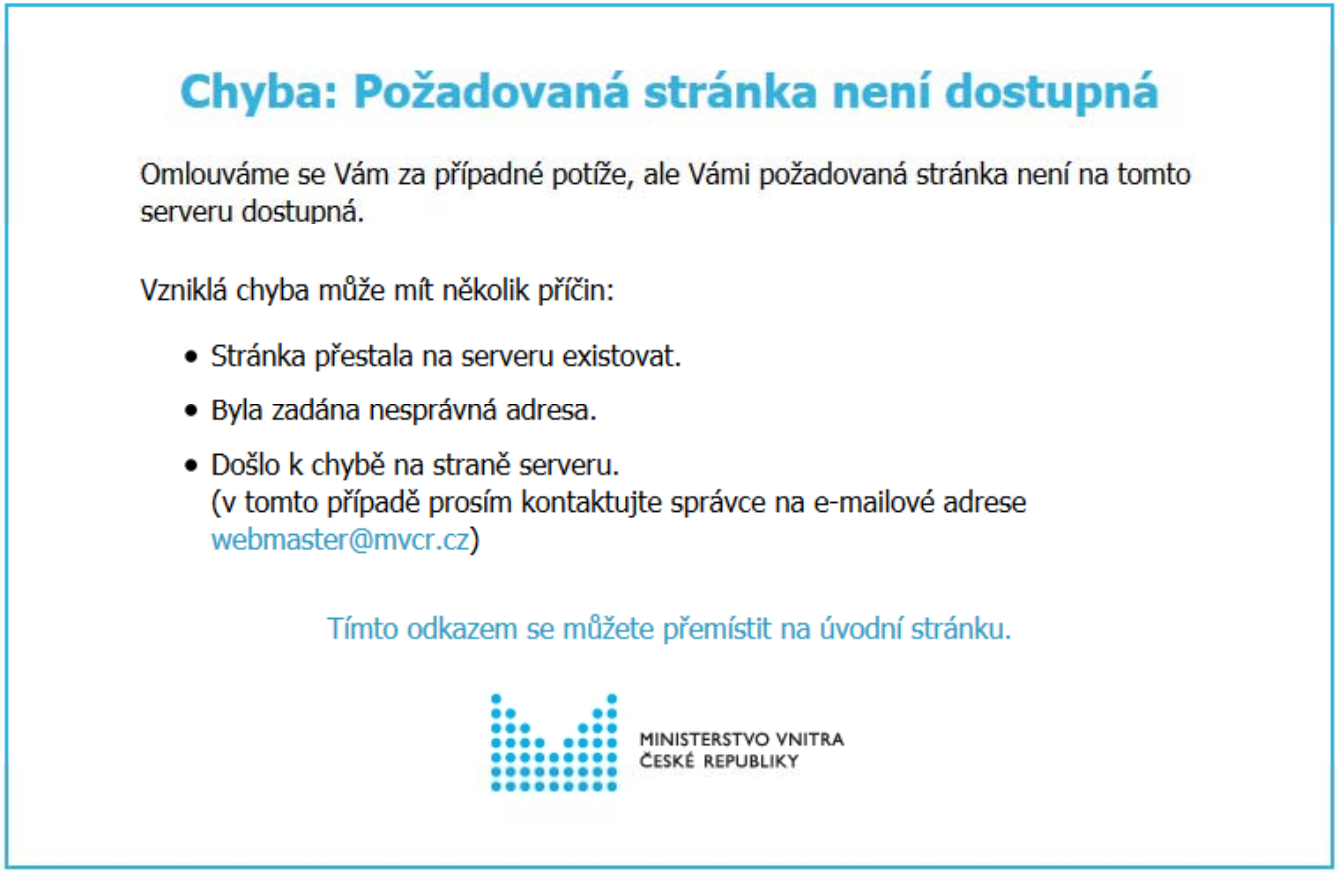
... are all these vocabularies really governed?

Src: [8],[10]

=> Big Data =? Small Data + Big Garbage

Src: [8]

CKAN Czech Republic

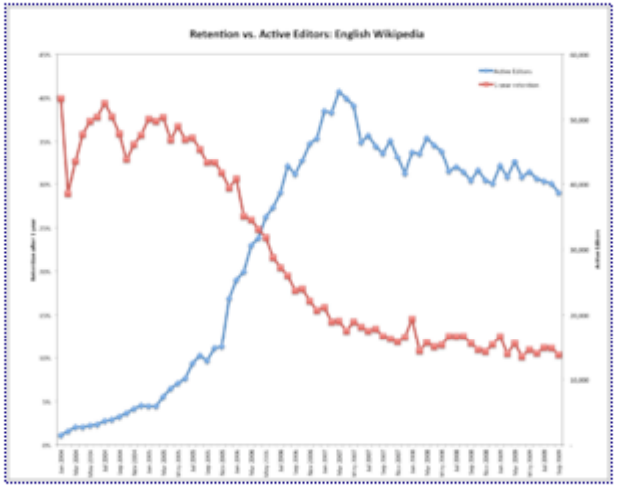


... are community data seriously Governed?!

General Problems of community webs demonstrated on Wikipedia example

Here is the problem

We call this the "oh shit" graph.



In early 2007, the number of active editors plateaued and then started decreasing. The rate at which new editors stick around plateaued even earlier, in early 2005, and then dropped sharply. These numbers haven't improved.

Why contribute?

71 / 100

71% of the editors contribute because they like the idea of volunteering to share knowledge. 69% believe that information should be freely available, and 63% pointed that contributing is fun. 7% edit Wikipedia for professional reasons.

Src: [1]

Vocabulary Mapping

- Using unique identifier
- Identification of links using simillarity metrics => a lot of publications from traditional DQM (e.g. Winkler)
- The Silk Framework
- LinQL Framework

SKOS (Simple Knowledge Organization System)

- Common data model for sharing and linking knowledge organization systems
- Relations among terms from different vocabularies
  - skos:exactMatch
  - skos:narrowMatch
  - skos:broadMatch
  - skos:closeMatch

DQ Topics relevant to Linked Data

1. „Rules“ and Best Practices for publishing Linked Data
2. Linked Data Quality Assessment an Quality Improvement
3. Linked Data Enrichment
4. Enrichment of common data using Linked Data
5. Validation using Linked Data

RULES AND BEST PRACTICES

Principles of Publication

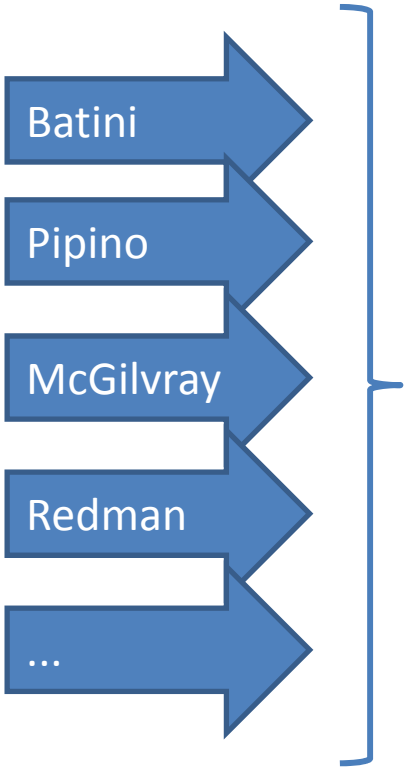
- Linked Data: Evolving the Web into a Global data Space
  - how to create „cool URIs“ => using natural keys, abstract from implementation details
  - publishing metadata (dc:creator, dc:publisher, cd:date or Open Provenance Model) and licences
  - reusing existing terms from vocabularies => vocabulary mapping: RDFS + OWL (owl:equivalentClass, owl:equivalentProperty, rdfs:subClassOf, rdfs:subPropertyOf)
  - Linked Data Publishing Checklist (8 points) = best practices
- Leigh Dodds, Ian Davis: Linked Data Patterns
- LATC: Initial Best Practices Guide [25]
- Pedantic Web Group => Frequently Observed Problems

Frequently Observed Problems

- Accessibility
  - Not retrievable (robots.txt, not published)
  - Incorrect CONTENT-TYPE
  - Content negotiation: different documents sent based on Accept header (RDF/XML, Turtle, ...), incorrect interpretation of Accept header
- Parsing and syntax of document
  - N3, N-triples, Turtle, RDFa, RDF/XML
- Naming and dereferencability
  - slash based / hash based URI
- Interpretation of datatype literals
  - not consistent values with datatype (e.g. datetime)
  - incompatibility with range
- Reasoning
  - Inconsistencies in word-views (Tomato = fruit / vegetable)
  - Inverse-functional properties (e.g. ISBN, RC, soc. ins. number, MAC address, IBAN, VIN, ...) – problem is with wrong / missing values when using for URI

LINKED DATA QUALITY ASSURANCE

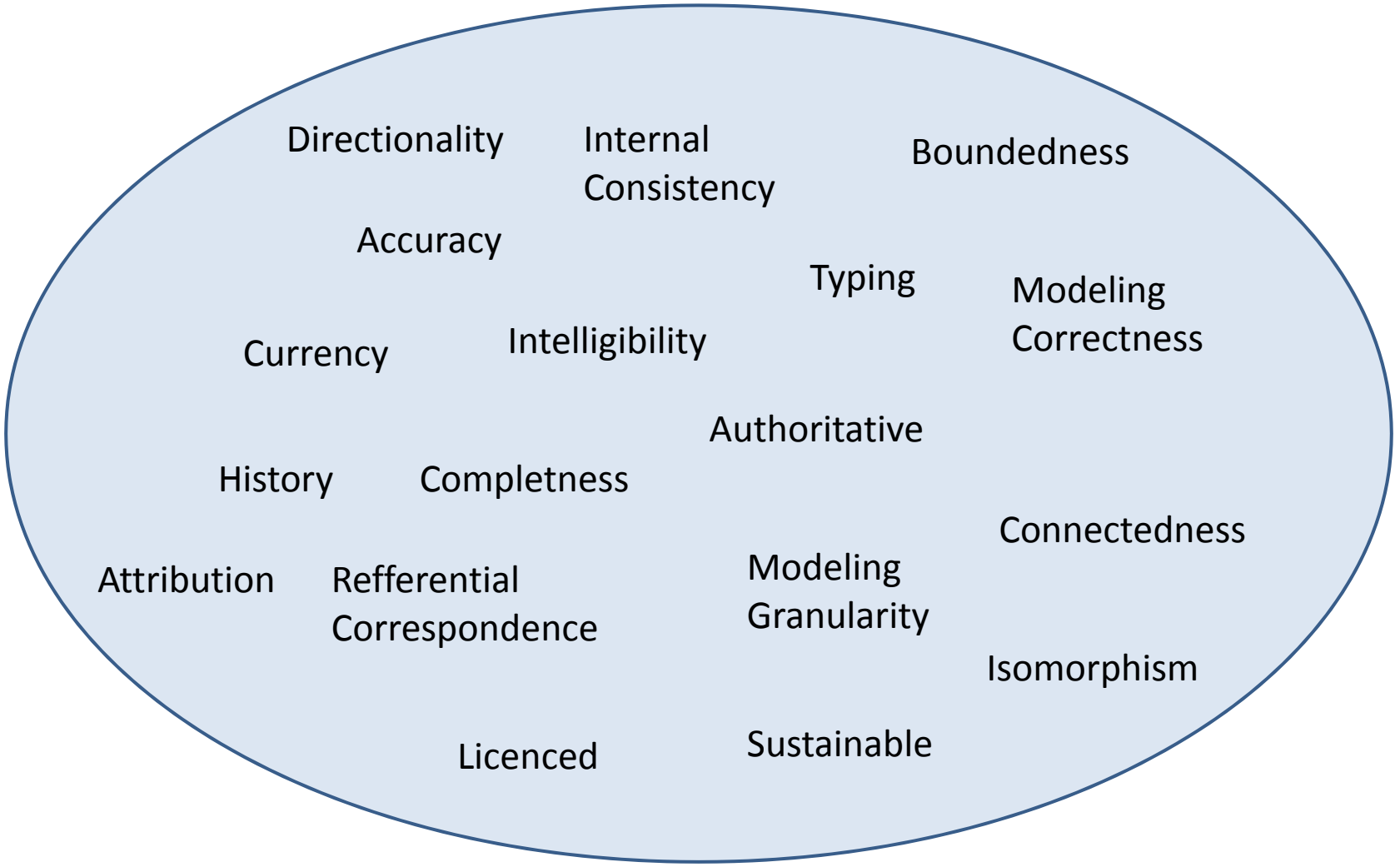
Characteristics of Conventional Data



	Data Characteristics
Intrinsic Dimension	Reliability
	Uniqueness
	Semantic Accuracy vs. Precision
	Syntactic Accuracy
Time Dimension	Currency
	Timeliness
	Volatility
	Time Synchronization
Contextual Dimension	Consistency
	Completeness
	Coverage
Dimension of Usage	Availability
	Comprehensibility
	Interoperability
	Security of Use
Economical Dimension	The cost of acquiring and updating data
	The costs of storing, sharing, distribution, backup and data archiving
	The costs of data protection

Different classifications, different meaning (accuracy / precision), ...

Specifics of Linked Data Quality



Src: Quality Indicators for Linked Data Datasets [22]

Quality Indicators for Linked Data Datasets

1. Accuracy - are facts actually correct?
2. Intelligibility - are there human readable labels on things?
3. Referential correspondence - are resources identified consistently without duplication?
4. Completeness - do you have all the data you expect?
5. Boundedness - do you have just the data you expect or is it polluted with irrelevant data?
6. Typing - are nodes properly typed as resources or just string literals?
7. Modeling correctness - is the logical structure of the data correct?
8. Modeling granularity - does the modeling capture enough information to be useful?
9. Connectedness - do combined datasets join at the right points?
10. Isomorphism - are combined datasets modeled in a compatible way?
11. Currency - is it up to date?
12. Directionality - is it consistent in the direction of relations?
13. Attribution - can you tell where portions of the data came from?
14. History - can you tell who edited the data and when?
15. Internal consistency - does the data contradict itself?
16. Licensed - is the license for use clear?
17. Sustainable - is there a credible basis for believing the data will be maintained?
18. Authoritative- is the provider of the data a credible authority on the subject?

Src: Quality Indicators for Linked Data Datasets [22]

Specifics of Linked Data Quality

Presentation Quality: => Information Quality?

Directionality

Boundedness

Intrinsic Dimension:

Accuracy

Refferential Correspondence

Authoritative

History

Attribution

Dimension of Usage:

Licenced

Typing

Intelligibility

Time Dimension:

Currency

Sustainable

Quality of the Model:

Modeling Correctness

Modeling Granularity

Isomorphism

Connectedness

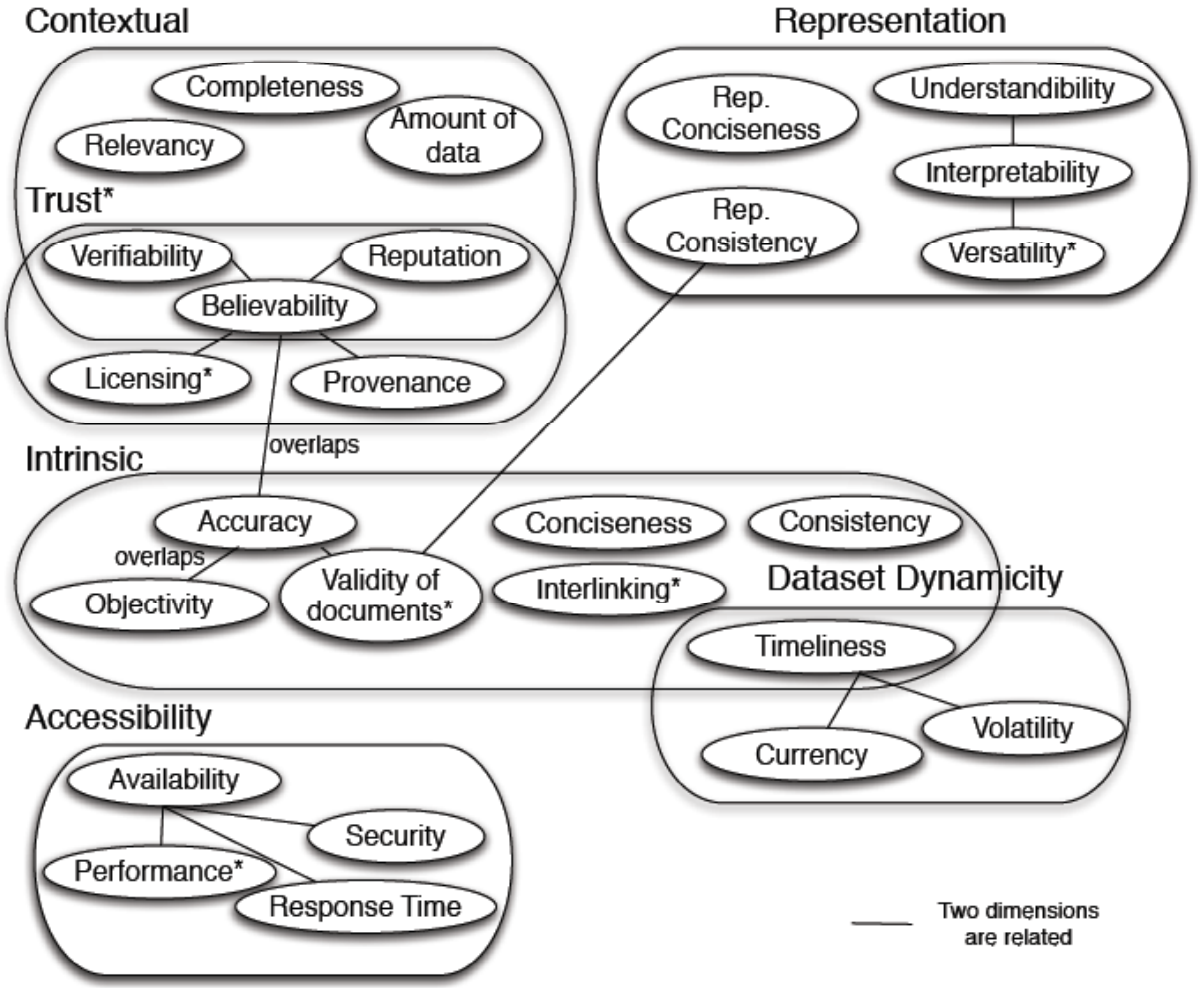
Contextual Dimension:

Completness

Internal Consistency

LD Characteristics: Zaveri

21 different sources



Src: [3]

More complex meaning of metrics in LD

Dimension	Metric	Description	Type
Completeness	degree to which classes and properties are not missing	detection of the degree to which the classes and properties of an ontology are represented [5,15,42]	S
	degree to which values for a property are not missing	detection of no. of missing values for a specific property [5, 15]	O
	degree to which real-world objects are not missing	detection of the degree to which all the real-world objects are represented [5,15,26,42]	O
	degree to which interlinks are not missing	detection of the degree to which instances in the dataset are interlinked [22]	O
Amount-of-data	appropria olume of data for a particular task	no. of triples, instances per class, internal and external links in a dataset [14,12]	O
	cover	scope and level of detail [14]	S
Rele	information attributes	counting the occurrence of relevant terms within these attributes or using vector space model and assigning higher weight to terms that appear within the meta-information attributes [5]	S
		sorting documents according to their relevancy for a given query [5]	S

In traditional DQM: The rate of false blank values

Table 2

list of data quality metrics of the contextual dimensions, how it can be measured and it's type - "S"ubjective or "O"bjective

Src: [3]

Results of Zaveri‘s Classification

Cathegory	No Dimension	No Metrics	No Subjective Metrics
Contextual Dimensions	3	8	4
Tust Dimensions	5	23	8
Intrinsic Dimensions	6	34	7
Accessibility Dimensions	4	15	1
Representinal Dimensions	5	20	3
Dataset Dynamicity Dimensions	3	9	0

=> 109 metrics !!!!

SPARQL

- ASK WHERE{...} => answer = Y/N
- HAVING conditions
- Rules
  - Completness: IF A son of B THEN b father of a
  - Logical rules: IF a = man AND man = mortal THEN a = mortal
- Definitions
- Business rules
- SPIN (SPARQL Inferencing Notation)
  - Own function and SPARQL templates
  - Store SPARQL queries as RDF triples
  - SPARQL rules created using CONSTRUCT statement and stored as triples
  - Constraints checking (ASK, CONSTRUCT, SPIN templates), conditional rules, calculating value of property based on other properties

```
# must be at least 18 years old
ASK WHERE {
    ?this my:age ?age .
    FILTER (?age < 18) .
}
```

RIF (Rule Interchange Format)

- Standard for exchanging rules
- Designs dialects = family of languages
  - Logic-based dialects: Basic Logic Dialect (RIF-BLD)
  - Dialects for rules with action => production rules: Production Rule Dialect (RIF-PRD)

```
Forall ?customer such that And( ?customer # ex1:Customer
                                ?customer[ex1:status->"Silver"] )
  (Forall ?shoppingCart such that And( ?shoppingCart # ex1:ShoppingCart
                                       ?customer[ex1:shoppingCart->?shoppingCart] )
     (If Exists ?value (And( ?shoppingCart[ex1:value->?value]
                             pred:numeric-greater-than-or-equal(?value 2000))
      Then Do( Modify( ?customer[ex1:status->"Gold"] ) ) ) )
```

A "Silver" customer with a shopping cart worth at least $2,000 is awarded the "Gold" status

Data Quality Constraints Library

- SPIN
- Christian Fürber
- Documentation, Reference, Constraints published in RDF
- Topics: Syntactical rules (EAN13, ZIP), Constraints for values, General dependencies, Uniqueness
- Web: http://semwebquality.org/mediawiki/index.php?title=SemWebQuality.org
- Constraints in RDF: http://semwebquality.org/ontologies/dq-constraints.rdf

DQ Frameworks: WIQA Policy Framework

- **WIQA** (Web Information Quality Assessment Framework)
- Filtering policies for information
- Representation -> Filtering -> Explaining decision
- **SWP** (The Semantic Web Publishing Vocabulary): terms for expressing different degrees of commitment and for representing digital signatures
- **WIQA-PL** (WIQA Information Quality Assessment Policy Language) for positive filtering; grammar based on SPARQL
- **DQ Heuristics**: *Content-based* (analyze content + compare with related information), *Context-based* (metadata, time dimensions), *Rating-based* (from consumers or engines)
- How to handling data conflict:
  - **Rank Data**: Show all data but evaluated by DQ rank
  - **Filter Data**: Show only successfully evaluated data
  - **Fuse Data**: Combine different data sources => DERI Pipes, KnoFuss, ...
- Web: http://www4.wiwiss.fu-berlin.de/bizer/wiqa/

Example of WIQA-PL

```
1.  NAME "Asserted by analysts with at least 3 positive ratings."
2.  DESCRIPTION "Accept only information that has been asserted by
3.           analysts who have received at least 3 positive ratings."
4.  PATTERNS {
5.
6.    GRAPH fd:GraphFromAggregator
7.          { ?GRAPH swp:assertedBy ?warrant .
8.            ?warrant swp:authority ?authority .
9.            EXPL "it was asserted by " ?authority " and " . }
10.
11.   GRAPH ?graph2
12.         { ?authority rdf:type fin:Analyst . }
13.
14.   GRAPH fd:GraphFromAggregator
15.         { ?graph2 swp:assertedBy ?warrant2 .
16.           ?warrant2 swp:authority ?authority2 .
17.           EXPL ?authority2 " claims that " ?authority
18.                " is an analyst." . }
19.
20.   GRAPH ANY
21.         { ?rater fin:positiveRating ?authority .
22.           FILTER (wiqa:count(?rater) > 2) .
23.           EXPL ?authority "has received positive ratings from" . }
24.
25.   GRAPH fd:BackgroundInformation
26.         { ?rater fin:affiliation ?company .
27.           EXPL ?rater "who works for" ?company . }
28.   }
```

Src: [17]

LATC Linked Data QA Framework

- Internal Quality Assurance: evaluating generated links = correct, not decided, incorrect
- External Quality Assurance:
  - number of links, duplicates, syntax errors
  - statistics about topology => detecting outliers
  - impact of newly created links
  - steps:
    - selection of resources
    - construct: creating local network
    - extend: new links
    - analyse
    - compare

Src: [24]

Another DQ Frameworks

- A lot of DQA methodologies from traditional DQM: e.g. AIMQ, combining measuring objective and subjective metrics
- SWIQA (Semantic Web Information Quality Assessment framework): Ch. Fürber, M. Hepp; framework using Semantic Web technologies

Examples of DQ Tools

- **Google-Refine**: correction of inconsistencies, distribution of attributes + definition of outliers, basic transformations (facets), data augmentation (comparation of keywords with external sources => linking)
- **ORE** (**O**ntology **R**epair and **E**nrichment): fixing inconsistencies in ontologies, DL-Learner for adding new axioms
- Pedantic Web Group
  - Online validators: Parsing and Syntax (W3C RDF/XML), Accessibility/Dereferencability, (Vapour, URI Debugger, RDF Triple-Checker) Vocabulary-specific Validators (QDOS FOAF Validator), Ontologies (Pellet OWL Reasoner Validator), General Validators (RDF:ALERTS)
  - Command-line Validators (Eyeball, The Validating RDF Parser)
- Zaveri: comparation of 9 tools:
  - Automated: [23] using SPARQL
  - Manual: WIQA, Sieve (metrics, scoring function, parameters in XML, component of Linked Data Integration Framework, Identity resolution => cannonical URI, vocabularies matching)
  - Semi-automated: Flemming‘s Data Quality Assessment Tool (interactions: questions about data + weights), RDFValidator, Trellis, tRDF
- Pellet Integrity Constraints: http://clarkparsia.com/pellet/icv

Ontology / Controlled Vocabulary

- A **controlled vocabulary** is a list of terms that have been enumerated explicitly. This list is controlled by and is available from a controlled vocabulary registration authority. All terms in a controlled vocabulary should have an unambiguous, non-redundant definition. A controlled vocabulary may have no meaning specified (it could be just a set of terms that people agree to use, and their meaning is understood), or it may have very detailed definitions for each term.
- A formal **ontology** *is a controlled vocabulary* expressed in an *ontology representation language*. This language has a grammar for using vocabulary terms to express something meaningful within a specified domain of interest. The grammar contains formal constraints (e.g., specifies what it means to be a well-formed statement, assertion, query, etc.) on how terms in the ontology's controlled vocabulary can be used together.

Src: [9]

VALIDATION AND ENRICHMENT

Using Linked Data for Validation

- Simple LOVs: no problem with Validation (if maintained)
- Adresses in Czech Republic – for validation it is more effective to use local DB with several generated matchcodes (for fuzzy match) + potential performance problems (batch validation) => for Enrichment not for Validation
- Trusted data source?

Problem: Trust

- Metadata for „easy“ evaluation (not only Trust): creator, date of publication, method, interlinkage + relations with other data sets, ...
- How to realize „rating“ of documents? Something like citation index?
- Tim Berners-Lee: „Oh yeah“ button
- *Why it wouldn‘t work: „Galileo‘s paradox“, selection of voters, ...*
- *For commercial purpose: rating by specialized vendors (independent authorities)*

Conclusion

- Generally it is possible to use Linked Data for Enrichment not for Validation
- Replication of DQ problems from original sources => Data Lineage
- Extension of traditional DQM tools by functionality useful for Linked Data Quality needed
- Missing large evaluation of data quality rules
- Missing large heuristics for automated repair
- Global approach to all processed data needed (LD are just another data) => Ontology (conceptional model) => CDM (logical model) => rules
- More „directive“ Semantic Web (controlls during publishing process provided by credible linked data source)

References

1. http://en.wikipedia.org/wiki/Wikipedia:Editor_engagement
2. http://www.w3.org/standards/semanticweb/ontology
3. ZAVERI, A., RULA, A., MAURINO, A. PIETROBON, R., LEHMANN, J., AUER, S. Quality Assessment Methodologies for Linked Open Data. [online] 2012. [cit. 6.4.2013]. Dostupné pod odkazem: http://www.semantic-web-journal.net/content/quality-assessment-methodologies-linked-open-data
4. GILPIN, M. From The Field: The First Annual Canonical Model Management Forum. *Forrester Mike Gilpin's Blog* [online]. 2010-03-15. Available on: http://blogs.forrester.com/mike_gilpin/10-03-15-field_first_annual_canonical_model_management_forum
5. http://shadowness.com/GoezCoz/tim-berners-lee-2
6. http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf
7. http://www.slideshare.net/fullscreen/mhepp/a-short-introduction-to-semantic-webbased-ecommerce-the-goodrelations-vocabulary-presentation/1
8. http://lov.okfn.org/dataset/lov/
9. http://answers.semanticweb.com/questions/12160/controlled-vocabulary-vs-ontology
10. http://www.slideshare.net/LeeFeigenbaum/semantic-web-landscape-2009

References

11. http://www.slideshare.net/LeeFeigenbaum/semantic-web-landscape-2009
12. http://lod-cloud.net/versions/2010-09-22/lod-cloud_colored.html
13. http://www.bioontology.org/bioportal
14. http://www.heppnetz.de/ontologies/vso/ns
15. http://www.ebusiness-unibw.org/ontologies/consumerelectronics/v1
16. http://productontology.org/
17. http://wiki.dbpedia.org/Ontology
18. http://www4.wiwiss.fu-berlin.de/bizer/wiqa/
19. Frequently Observed Problems on the Web of Data Aidan Hogan and Richard Cyganiak http://pedantic-web.org/fops.html
20. Fürber, Christian and Hepp, Martin, "SWIQA – A SEMANTIC WEB INFORMATION QUALITY ASSESSMENT FRAMEWORK" (2011). *ECIS 2011 Proceedings.* Paper 76. http://aisel.aisnet.org/ecis2011/76

References

21. http://web.cba.neu.edu/~ywlee/publication/AIMQ.pdf
22. http://answers.semanticweb.com/questions/1072/quality-indicators-for-linked-data-datasets
23. GUÉRET, C., GROTH, P., STADLER, C., AND LEHMANN, J.Assessing linked data mappings using network measures. In ESWC (2012).
24. http://latc-project.eu/sites/default/files/deliverables/latc-wp1-D141%20First%20deployment%20of%20QA%20module.pdf
25. http://latc-project.eu/sites/default/files/deliverables/latc-wp4-D411%20Initial%20best%20practice%20guide.pdf

W3C Specifications

- http://www.w3.org/standards/history/rdf-mt
- http://www.w3.org/TR/owl-features/
- http://www.w3.org/TR/owl2-overview/
- http://www.w3.org/2009/sparql/wiki/Main_Page
- http://www.w3.org/2001/sw/wiki/POWDER
- http://www.w3.org/PICS/
- http://www.w3.org/TR/rif-overview/
- http://www.w3.org/TR/2012/NOTE-rif-overview-20121211/
- http://www.w3.org/TR/rdf-schema/
- http://www.w3.org/standards/techs/skos#w3c_all

Tools

- http://code.google.com/p/google-refine/wiki/Screencasts
- http://ore-tool.net/Projects/ORE