Statistical Data in RDF

Knowledge Engineering Group
Seminar, November 4th 2010

Jindřich Mynarz
@jindrichmynarz

Scope of the talk

- not **microdata** (e.g., survey data)

DATA DATA DATA DATA
DATA DATA DATA DATA
DATA DATA DATA DATA

- but **aggregated** data (e.g., averages)

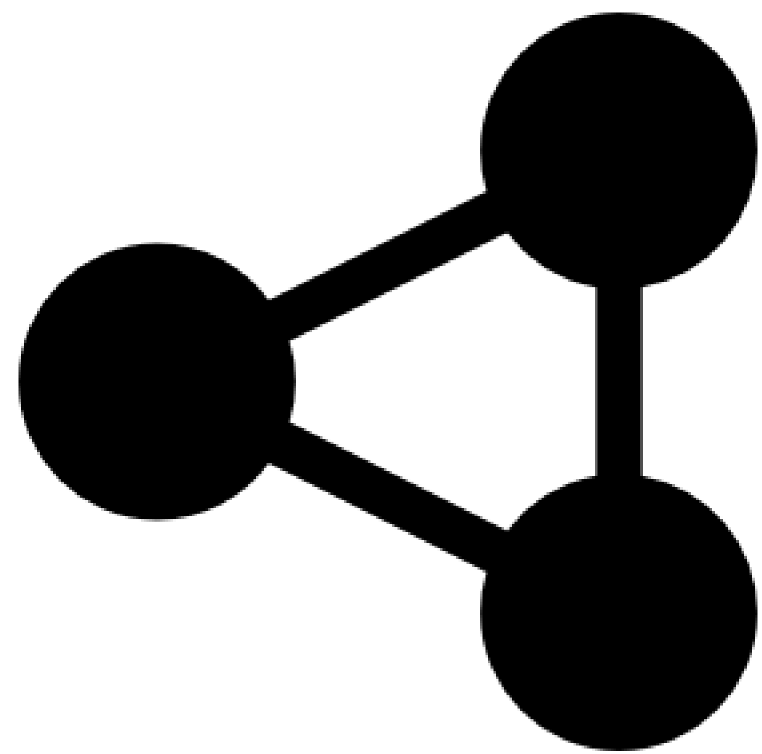
DATA

- only **RDF**

- overview of existing statistical **datasets**

RDF

- **separation** of content and layout
  - in *tabular data* table layout defines the way of interpretation
- **flexible**, schema-less data format
  - not overly inclusive, nor overly exclusive



Existing statistics in RDF

- CIA World Factbook
- U.S. Census 2000 dataset
- LOIUS - Italian linked university statistics
- Linked Environment Data
- EnAKTing datasets
- data.gov.uk datasets

Eurostat data

- **Freie Universität Berlin** - D2R Server
- **riese** (RDFizing and Interlinking the EuroStat Data Set Effort)
- **OntologyCentral** - real-time wrapper
- **Eurostat**'s own RDF datasets

Governmental statistics

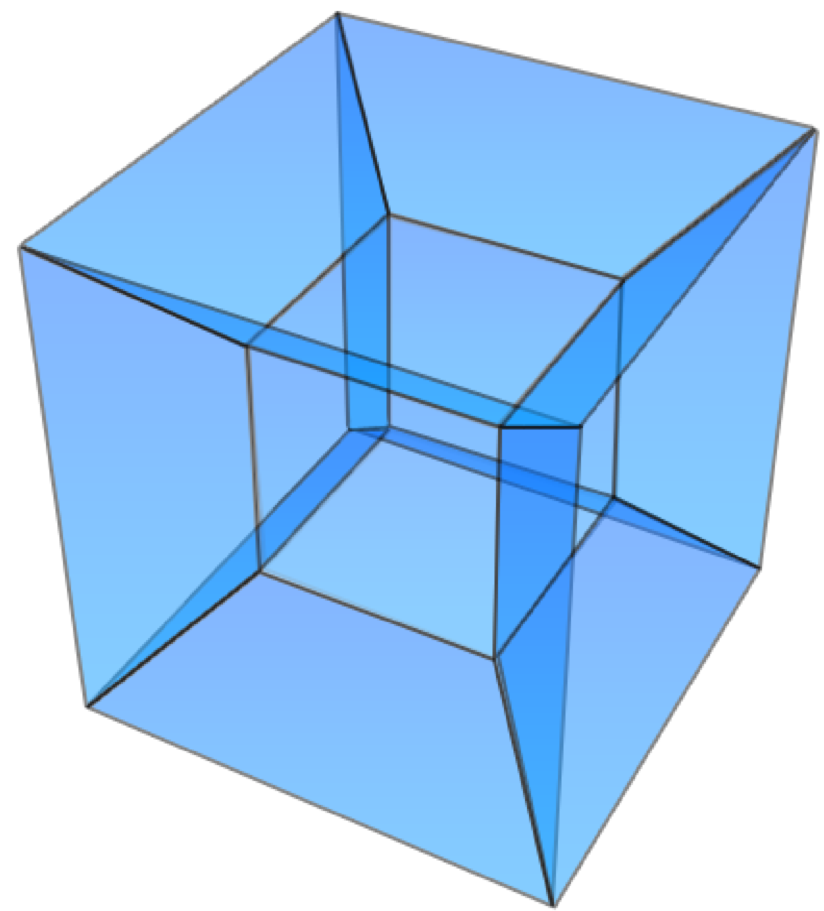
- **data.gov**
- **data.gov.uk**
    - *EnAKTing* mashups and data visualizations
    - population, crime, CO2 emissions, transport, agriculture, education...

Data modelling

- what is being modelled?
  - **the real world**
  - **a part** of the real world
  - **statistics**
- two parts of modelling
  - **structural** semantics
  - **domain** semantics

Structural semantics

- means of expression for the **cube's structure**
- groups, slices, time series
- addressed in **Data Cube** vocabulary

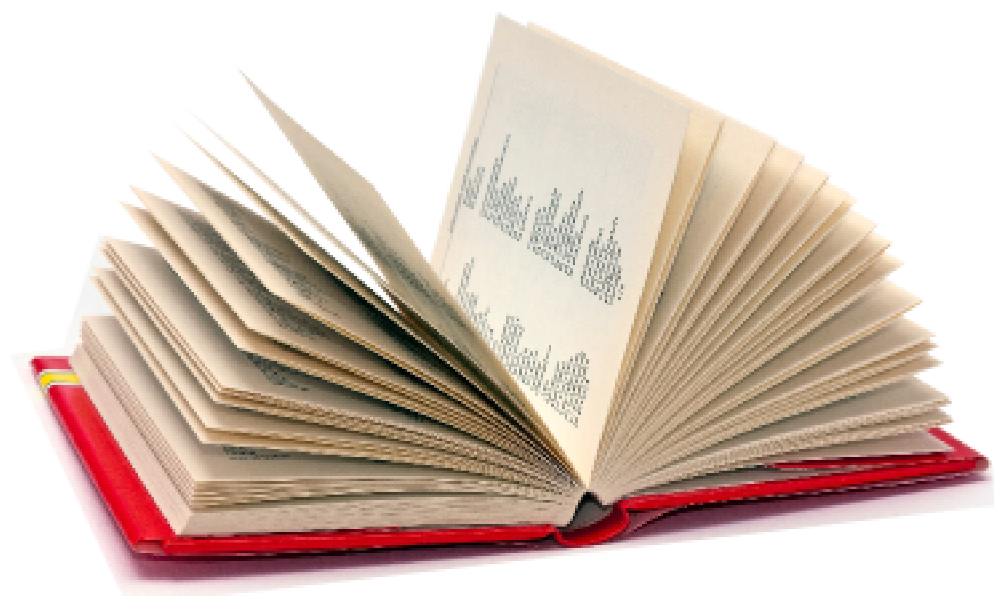


Domain semantics

- how a dataset refers to the things that it is **about**
- **connecting** statistical observations to the **model of the domain** described by them
- domain is a set of *non-information* resources

Vocabularies

- number of *ad hoc* vocabularies
- riese
- SCOVO
- SCOVOLink
- Data Cube
- SDMX/RDF

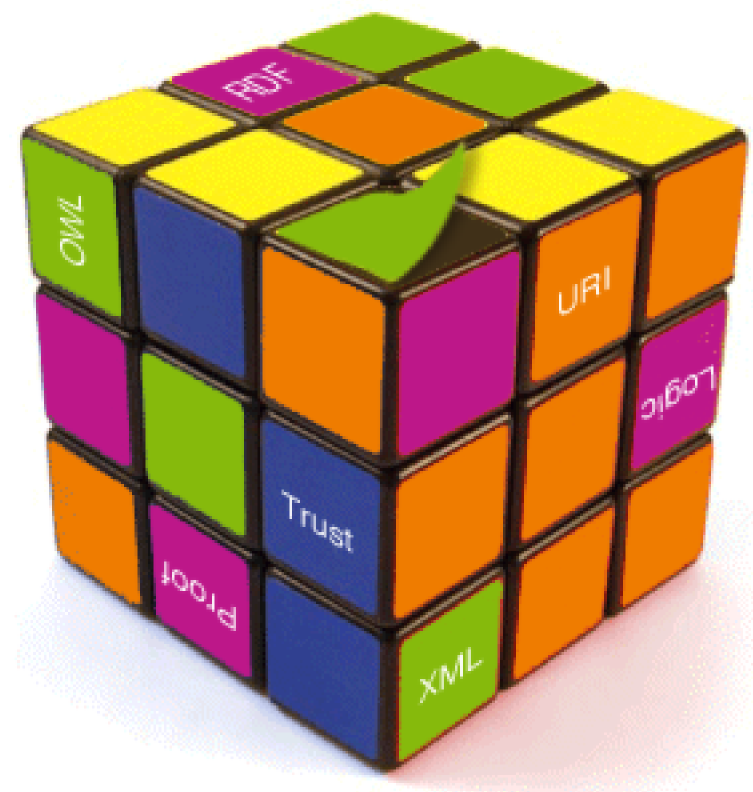


SCOVO

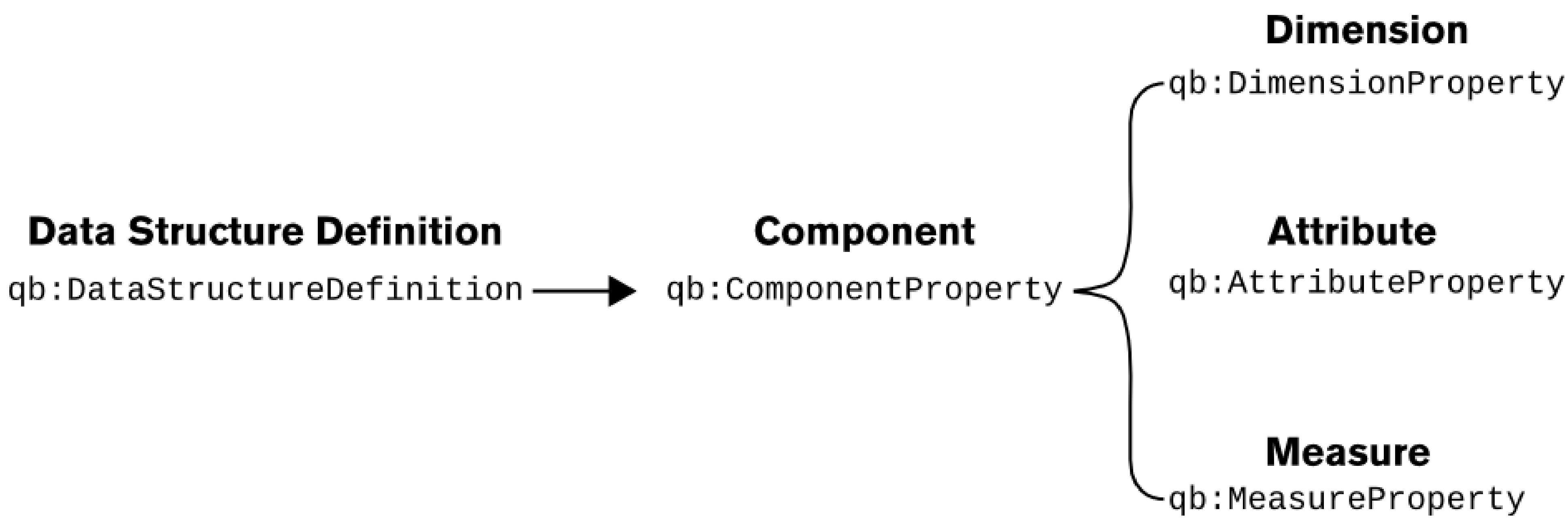
- The **S**tatistical **Co**re **Vo**cabulary
- inspired by **riese** vocabulary
- modelling of dimensions and observations as **separate resources**
- lightweight, easy to adopt
- **SCOVOLink** addresses *domain semantics*

Data Cube

- inspired by SCOVO
  - added expressive power
- generalization from SDMX/RDF
- re-use of SKOS for **codelists**

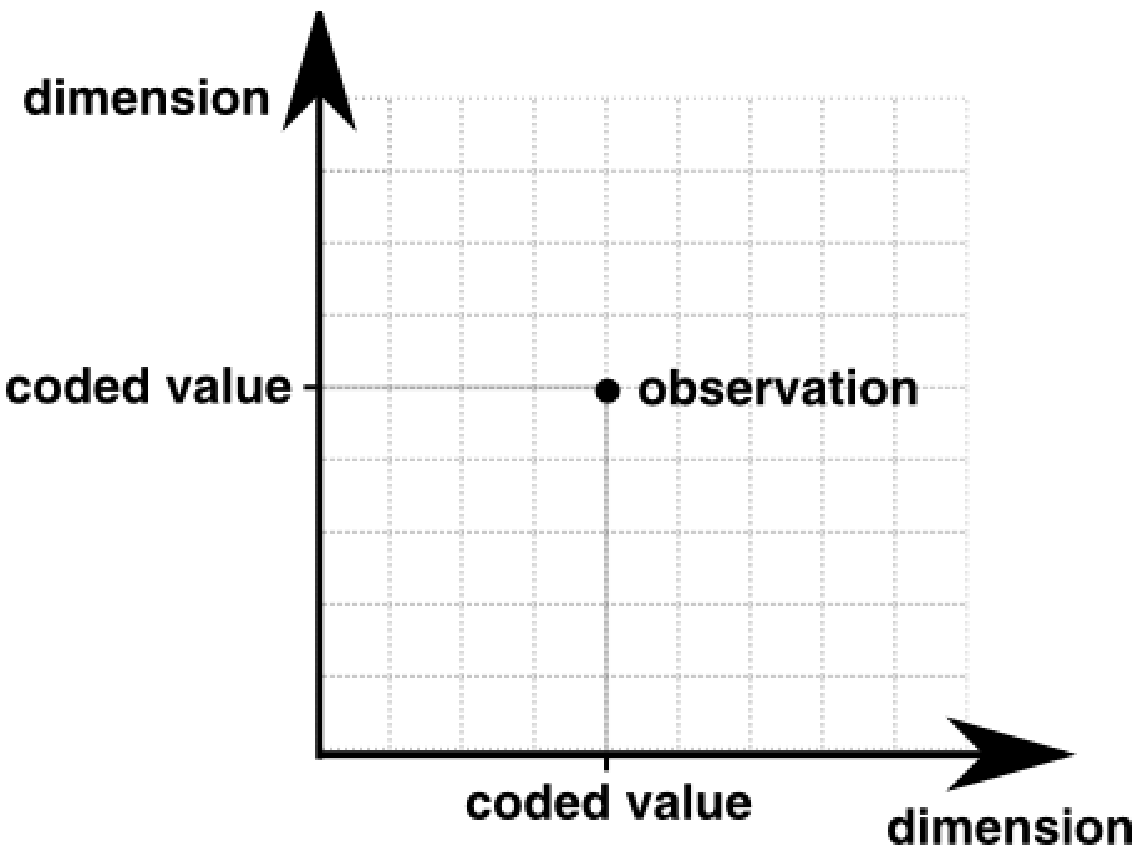


Data Cube

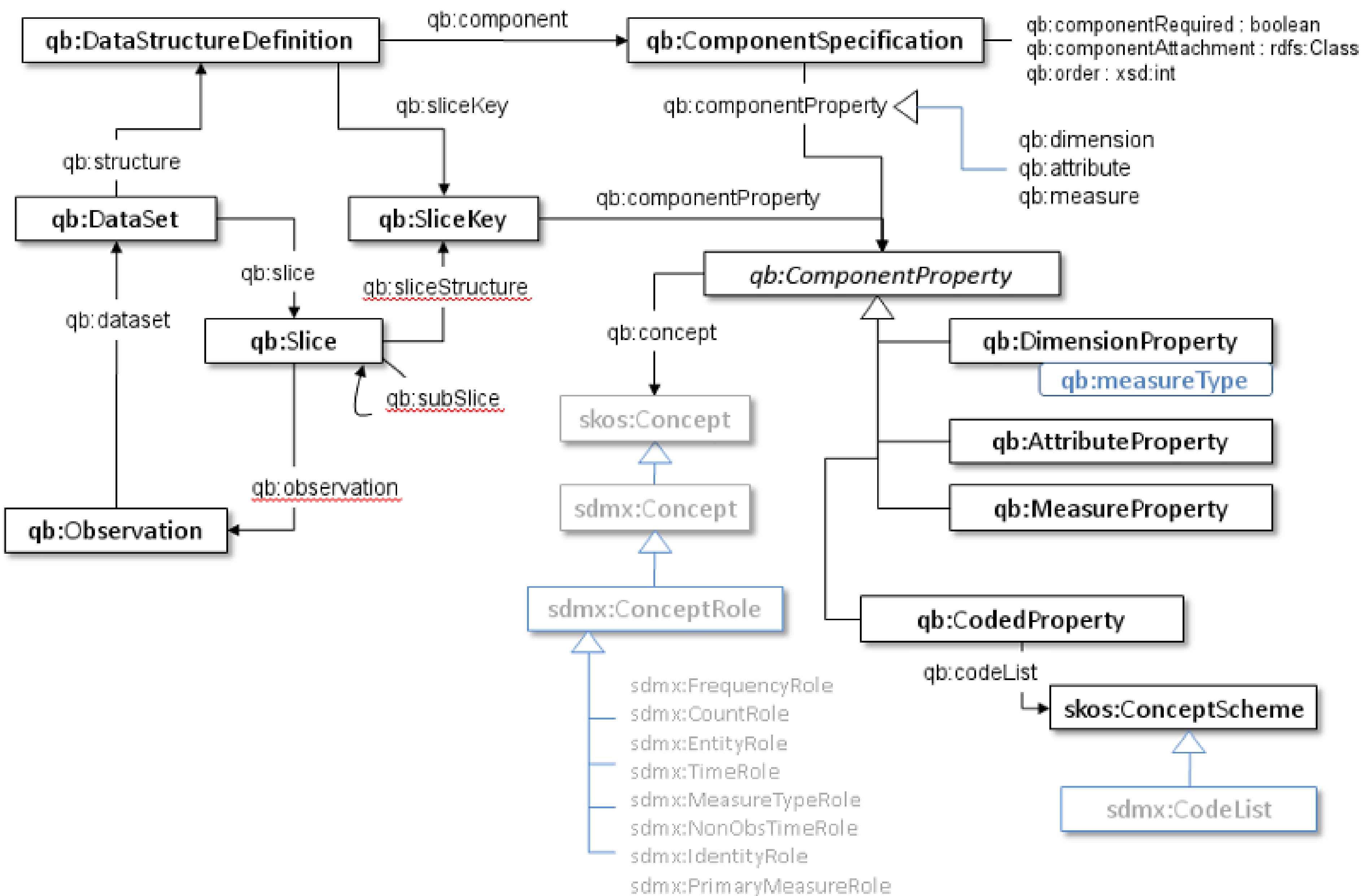

Data Structure Definition
qb:DataStructureDefinition
Component
qb:ComponentProperty
Dimension
qb:DimensionProperty
Attribute
qb:AttributeProperty
Measure
qb:MeasureProperty

Data Cube

- dimensions (`rdf:Property`)
- coded values (`skos:Concept`)



Data Cube


qb:DataStructureDefinition
qb:component
qb:ComponentSpecification
qb:componentRequired : boolean
qb:componentAttachment : rdfs:Class
qb:order : xsd:int
qb:sliceKey
qb:componentProperty
qb:dimension
qb:attribute
qb:measure
qb:structure
qb:DataSet
qb:SliceKey
qb:componentProperty
qb:slice
qb:sliceStructure
qb:ComponentProperty
qb:dataset
qb:Slice
qb:concept
qb:DimensionProperty
qb:measureType
qb:subSlice
skos:Concept
qb:AttributeProperty
qb:observation
sdmx:Concept
qb:MeasureProperty
qb:Observation
sdmx:ConceptRole
qb:CodedProperty
qb:codeList
sdmx:FrequencyRole
sdmx:CountRole
sdmx:EntityRole
sdmx:TimeRole
sdmx:MeasureTypeRole
sdmx:NonObsTimeRole
sdmx:IdentityRole
sdmx:PrimaryMeasureRole
skos:ConceptScheme
sdmx:CodeList

SDMX/RDF

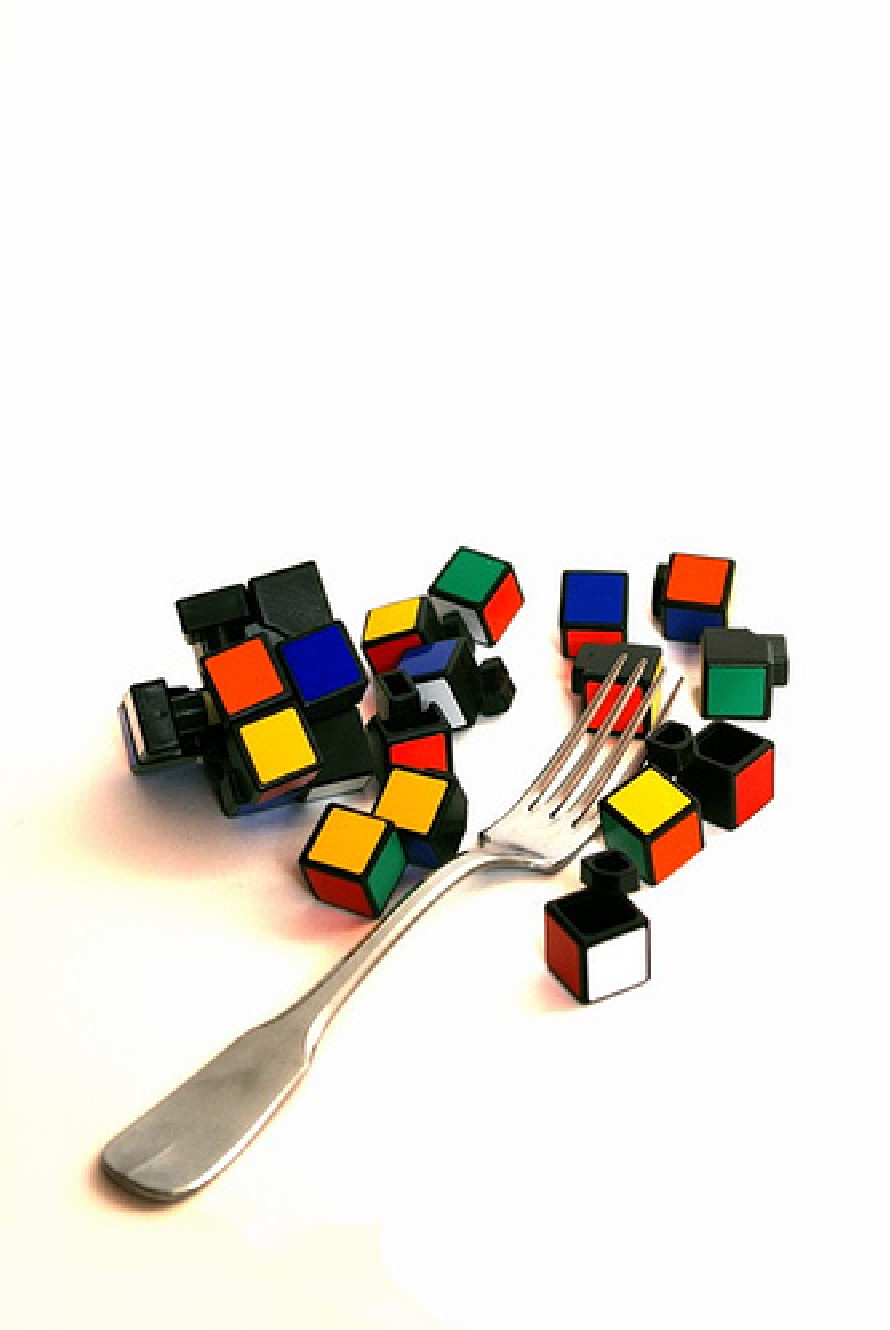
- **Statistical Data and Metadata eXchange** reformulated in RDF
- built on top of **Data Cube**
- contains:
  - `sdmx`
  - `sdmx-attribute`
  - `sdmx-code`
  - `sdmx-concept`
  - `sdmx-dimension`
  - `sdmx-measure`
  - `sdmx-metadata`
  - `sdmx-subject`

Important parts of modelling

- re-use
- units
- time
- identifiers
- URI patterns

Re-use oriented design

- **re-purposing** parts of the existing datasets
- re-using **shared vocabularies**
- vocabulary **hi-jacking** and **extension**

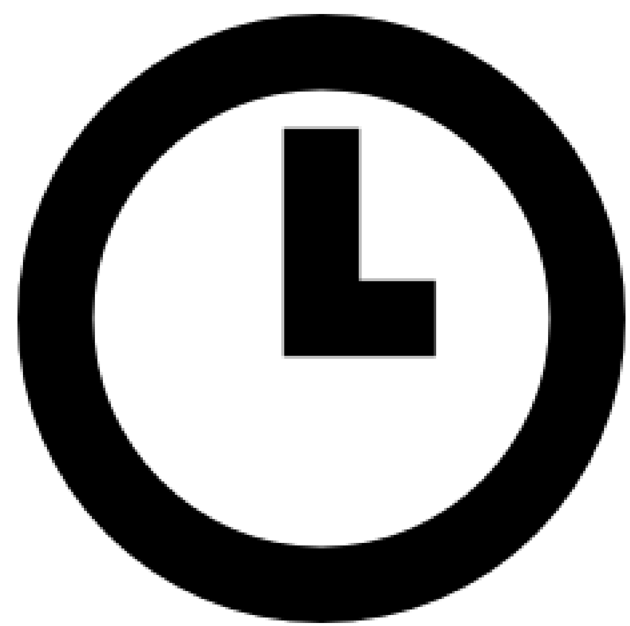


Units of measurement

- **implicit**
  - “78693011 m^2”, “117 b”
  - `eurostat:total_area_km2`
- **explicit**
  - `:unit`, `sdmx-attribute:unitMeasure`

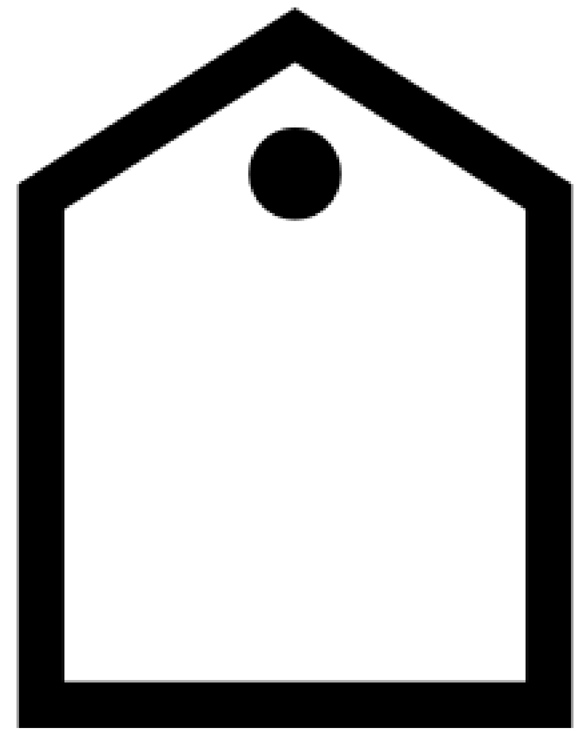
Modelling of time

- **exclusion** of the dimension of time (D2R Eurostat, U.S. Census 2000)
- time **dimension** (riese, SDMX/RDF)
    - `dimension:Time`, `sdmx:TimeRole`
    - time series



Identifiers

- blank nodes
- URIs
- HTTP URIs



URI design patterns

- on the Web
  - `http://`
- human-readable
  - `what/is/this/about`
- clustered by resource type
  - `type/unique-id`
- standardized
  - `{provider 1}/path/to/an/observation`
  - `{provider 2}/path/to/an/observation`
- hierarchical
  - `{broader}/{narrower}`
- reflecting the location of an observation in a data cube
  - `{dimension 1}/{dimension 2}`

Following steps

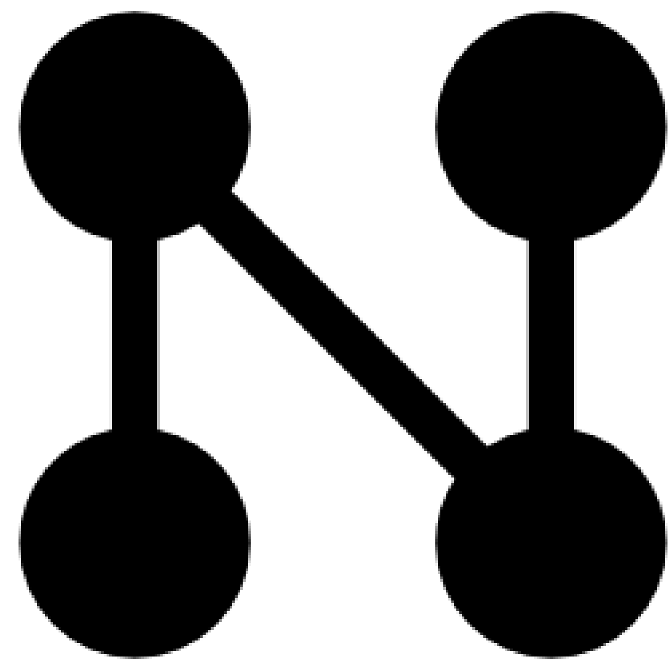
- data conversion
- interlinking dataset's resources
- linking external datasets
- publishing

Legacy datasets

- statistics-**specific** data formats
- **implicit** context of interpretation
- parsing, cleaning
- conversion **mechanisms**
  - SQL DB **wrappers** (e.g., D2R Server)
  - real-time **exporters** (e.g., OntologyCentral)
  - **RDFizers** (e.g. RDF123)
  - **custom-built** scripts

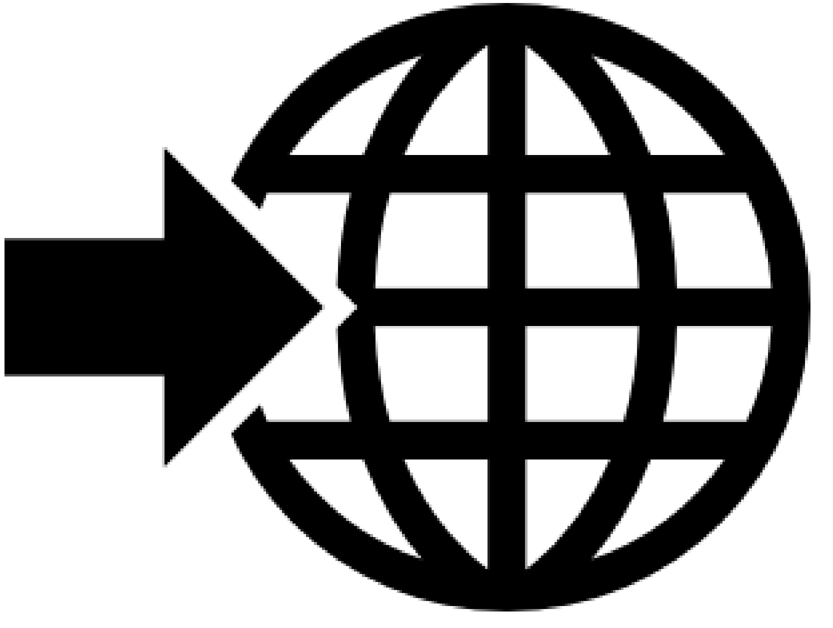
Linking

- **re-use** by reference
- lightweight **intergration**
- **linkable** data
- linking **properties**
    - **e.g.**, `owl:sameAs`, `skos:closeMatch`

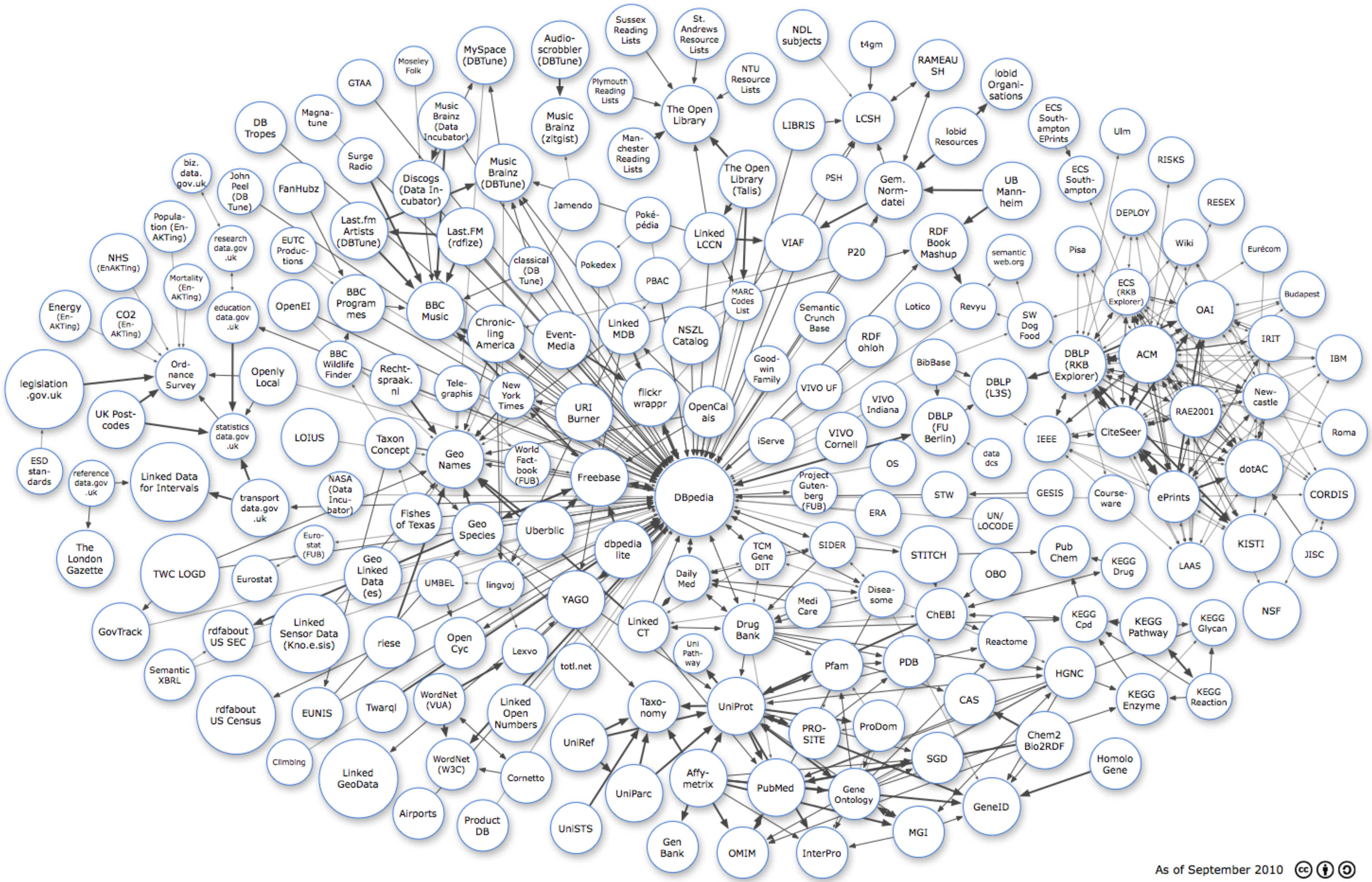


Publishing data

- new **dissemination** standards
- exchanging data with the Web
- RDF **dumps**
- **linked data** distribution
- **SPARQL**
- **RDFa**



Linked open data cloud

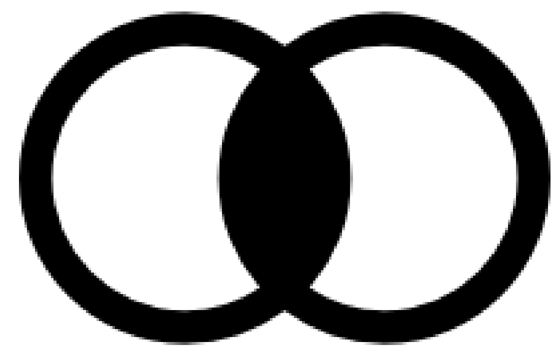


Benefits

- data can be intergrated
- open data
- re-usable data
- data available for applications

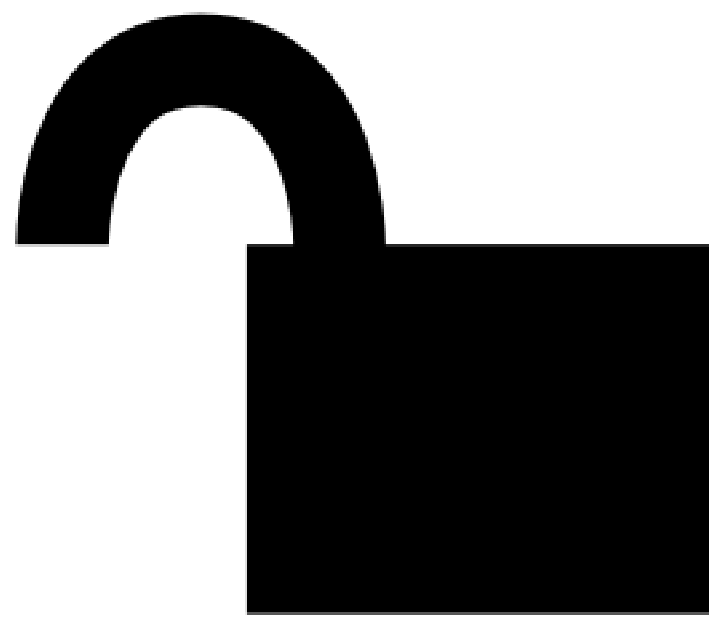
Integration

- **combining** and **merging** with other datasets
- **re-use** oriented design



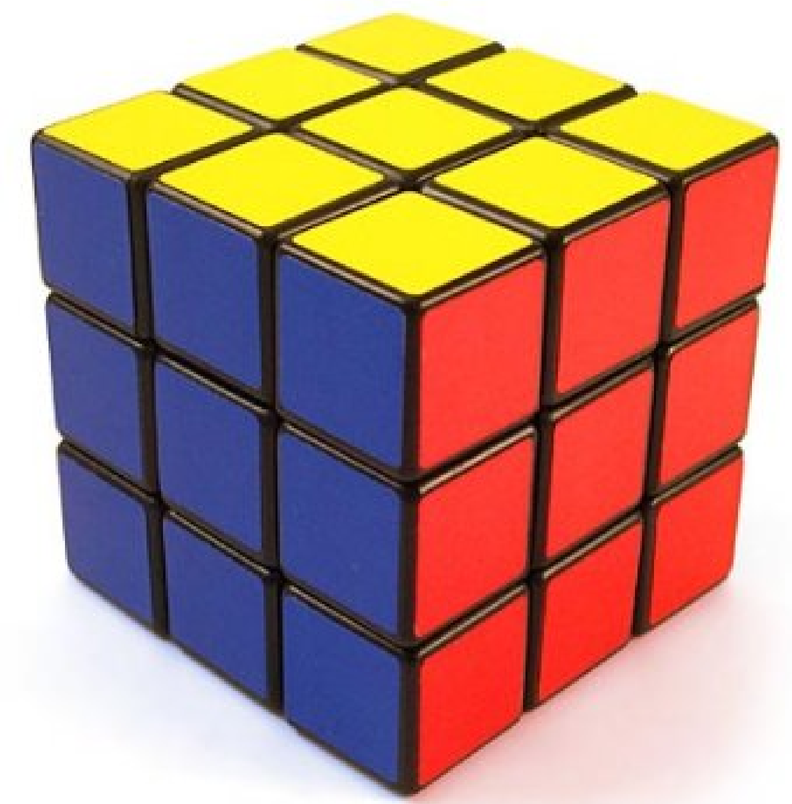
Open data

- **freedom of information** for public sector information
- open licences
  - Creative Commons, Open Government Licence...
- public domain



Anyone can solve the cube

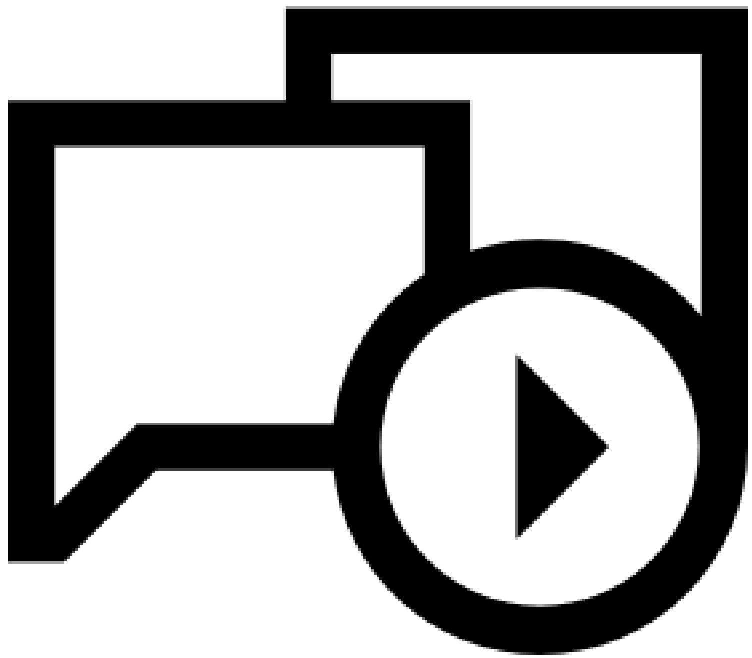
- data is **available** for individual analysis
- offices for national statistics still have the monopoly on data **collection**, but no longer on **interpretation** of that data
- data-driven **journalism**



Building on top of statistical data

- once the data is available useful **applications** can be built on top of it
- data **visualizations**
- data **analysis tools**

Questions!



Thank you for attention!

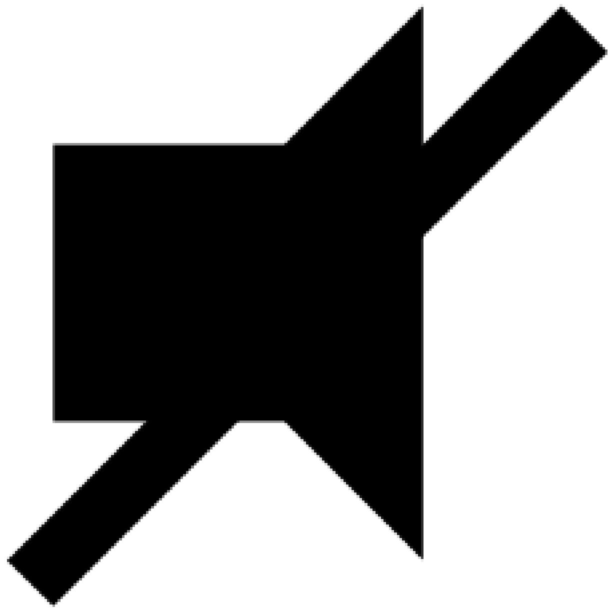


Image credits

Semantic Web Rubik's Cube. http://www.flickr.com/photos/dullhunk/3448804778/
Rubik's Cube. http://www.flickr.com/photos/bramus/3249196137/
Hypercube. http://commons.wikimedia.org/wiki/File:Hypercube.png
PICOL: Pictorial communication language. http://picol.org/
Dictionary. http://www.flickr.com/photos/horiavarlan/4268897748/
Oops! http://www.flickr.com/photos/rore/299375688/
Tape Measure. http://www.flickr.com/photos/wwarby/4915969081/
Rubik's Cube 1. http://www.flickr.com/photos/lifeontheedge/374960949/
Detroit's Skyline. http://www.flickr.com/photos/showmeone/4154861617/
Linked Oped Data Cloud. http://richard.cyganiak.de/2007/10/lod/
Cube. http://followtherhythm.deviantart.com/art/cube-128329792
Data Cube diagram. http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/qb-fig1.png