Knowledge Graphs for Cultural Heritage and Digital Humanities

Victor de Boer
SUMAC 2023

With input from: Xander Wilcke, Sarah Shoilee, Jacco van Ossenbruggen, Go Sugimoto, Niels Ockeloen, Paul Groth, Oana Inel, Lora Aroyo, Jur Leinenga, Matthias van Rossum, Andrea Bravo Balado, Robin Ponstein, Ronald Siebes, Roderick van der Weerdt, Loan Ho…

More and more structured data available online

Government data

Social web data

Medical data

Museum data

Research data

Development data

Moverum.com

The Digital Turn

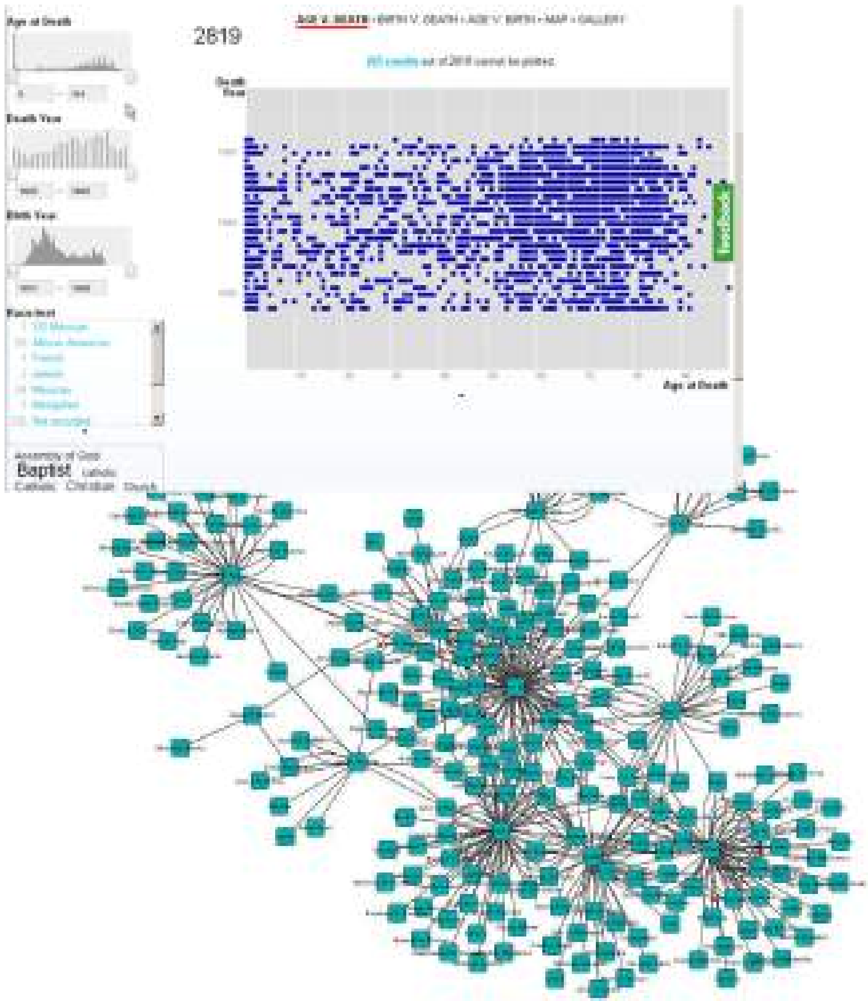
From the physical archives to digital ones

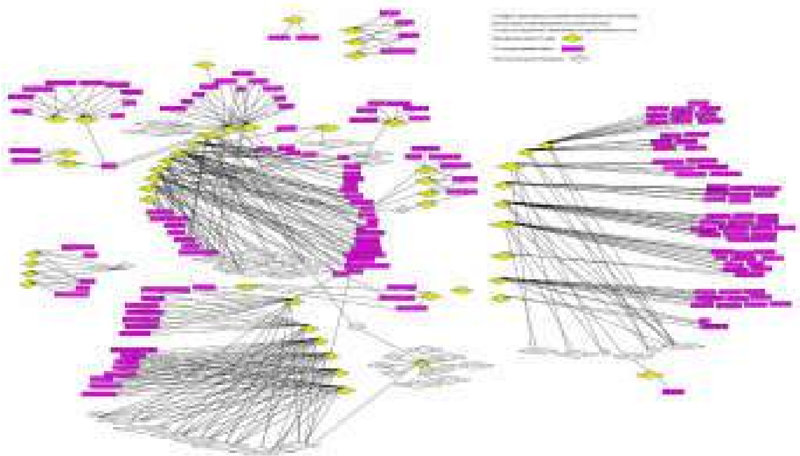
 

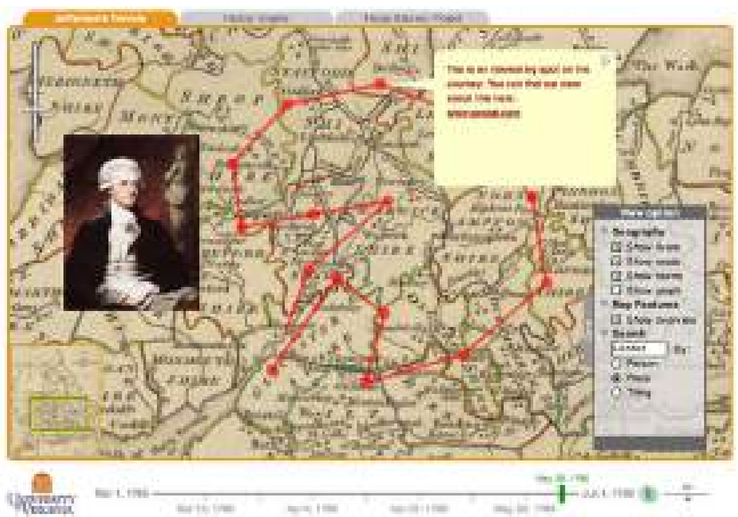
allows for new (types of) research, access

Img:www.doaks.org, www.dkrz.de

Tools and visualisations

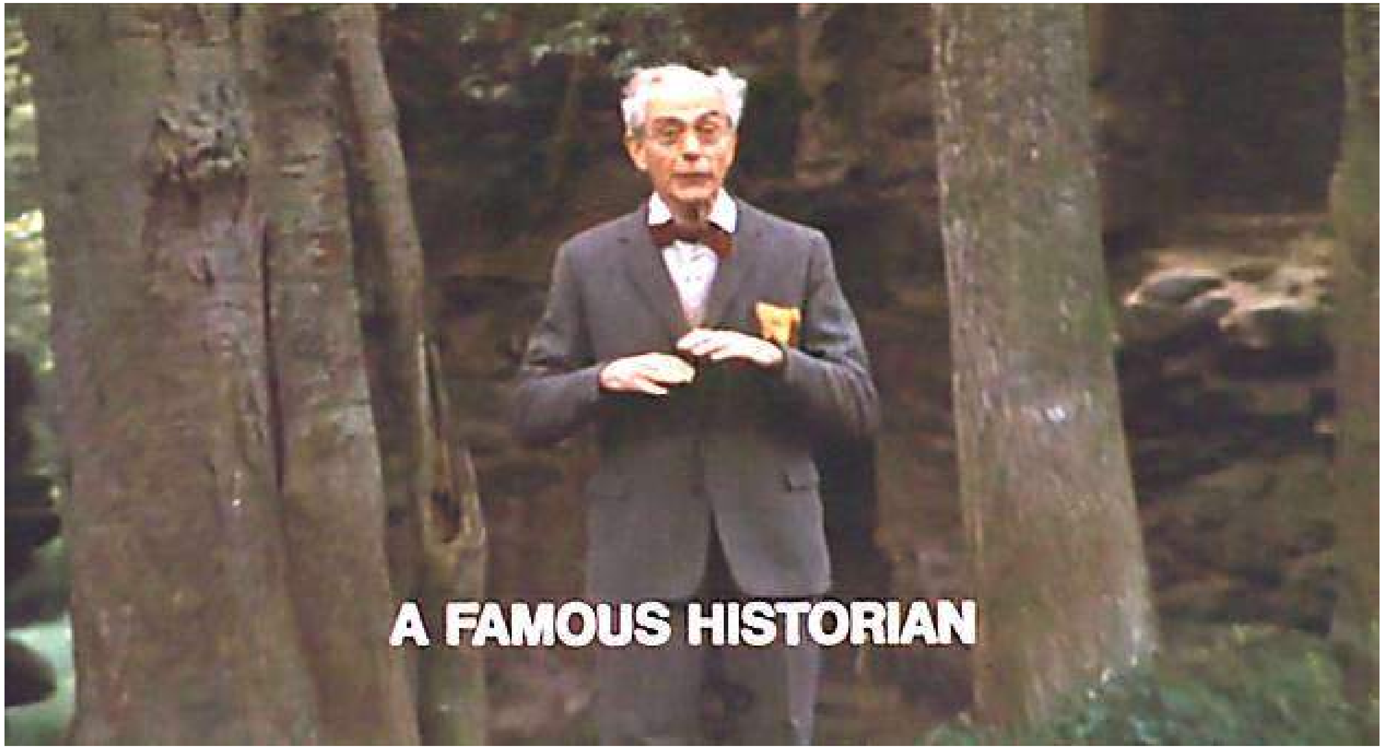






http://armstrongdigitalhistory.org/, http://www.vcdh.virginia.edu/courses/fall07/hius401-f/
http://digitalhistory.unl.edu/essays/thomasessay.php, http://www.philipvickersfithian.com/2013/05/gender-in-stacks-on-managing-small.htm

“That is great. I would love that…



…but my research questions are slightly different.”

Img:Monty Python

Aging

Data

Tool

C. Guéret based on http://redmonk.com/jgovernor/2007/04/05/why-applciations-are-like-fish-and-data-is-like0wine/

Data - Centric

Do not bake the data into the tool

Build tools on top of the data.

Allow for integration of various data

New ways of analyzing integrated data

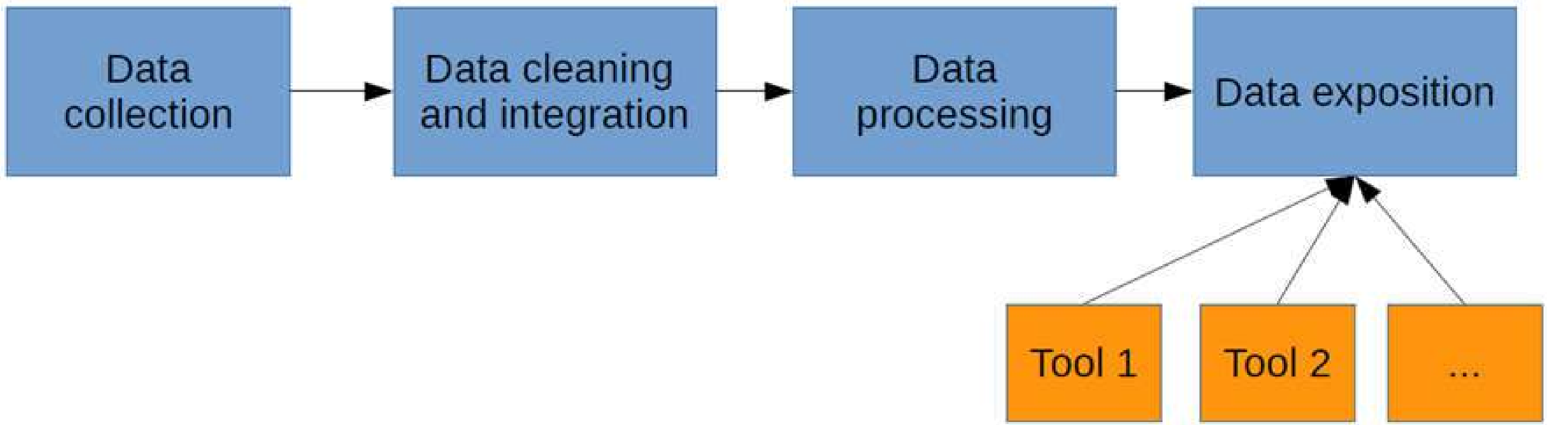
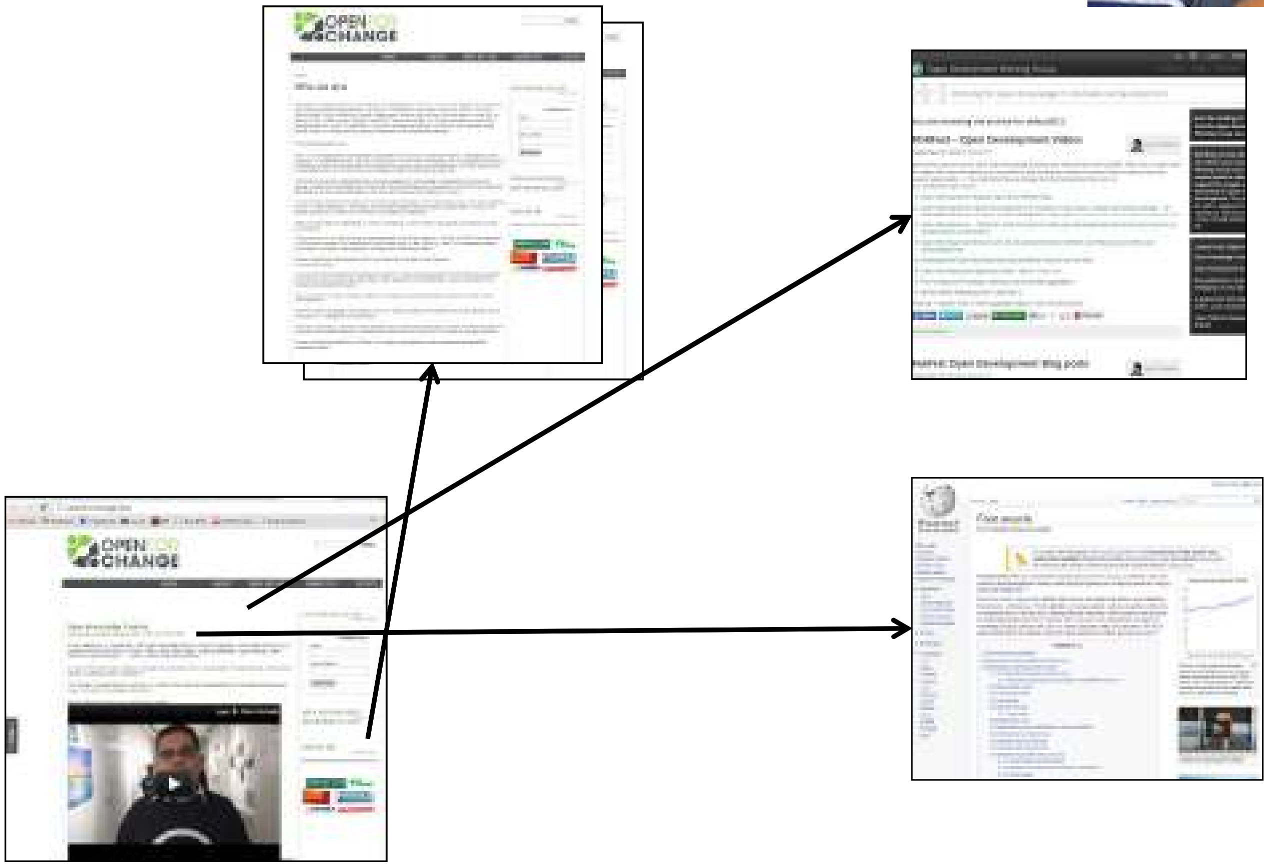


Fig: C. Guéret

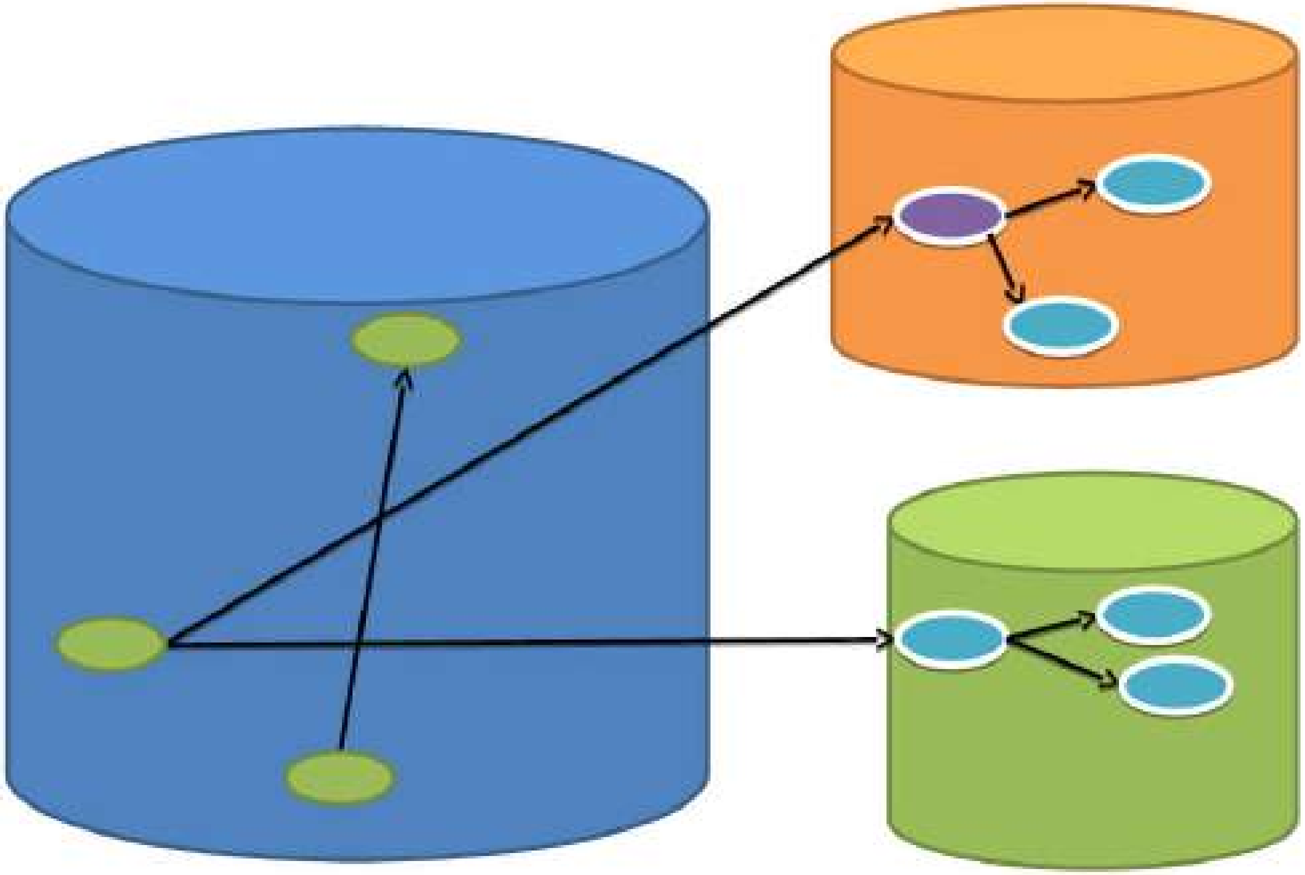
Moving away from silos

Acoustic Coupler
Source: "Games Aktuell Blog", http://www.gamesaktuell.de/Community/MySite/GenX3601966-2605282/Blogs/Cyberpunks-beim-Mauerfall-694794/

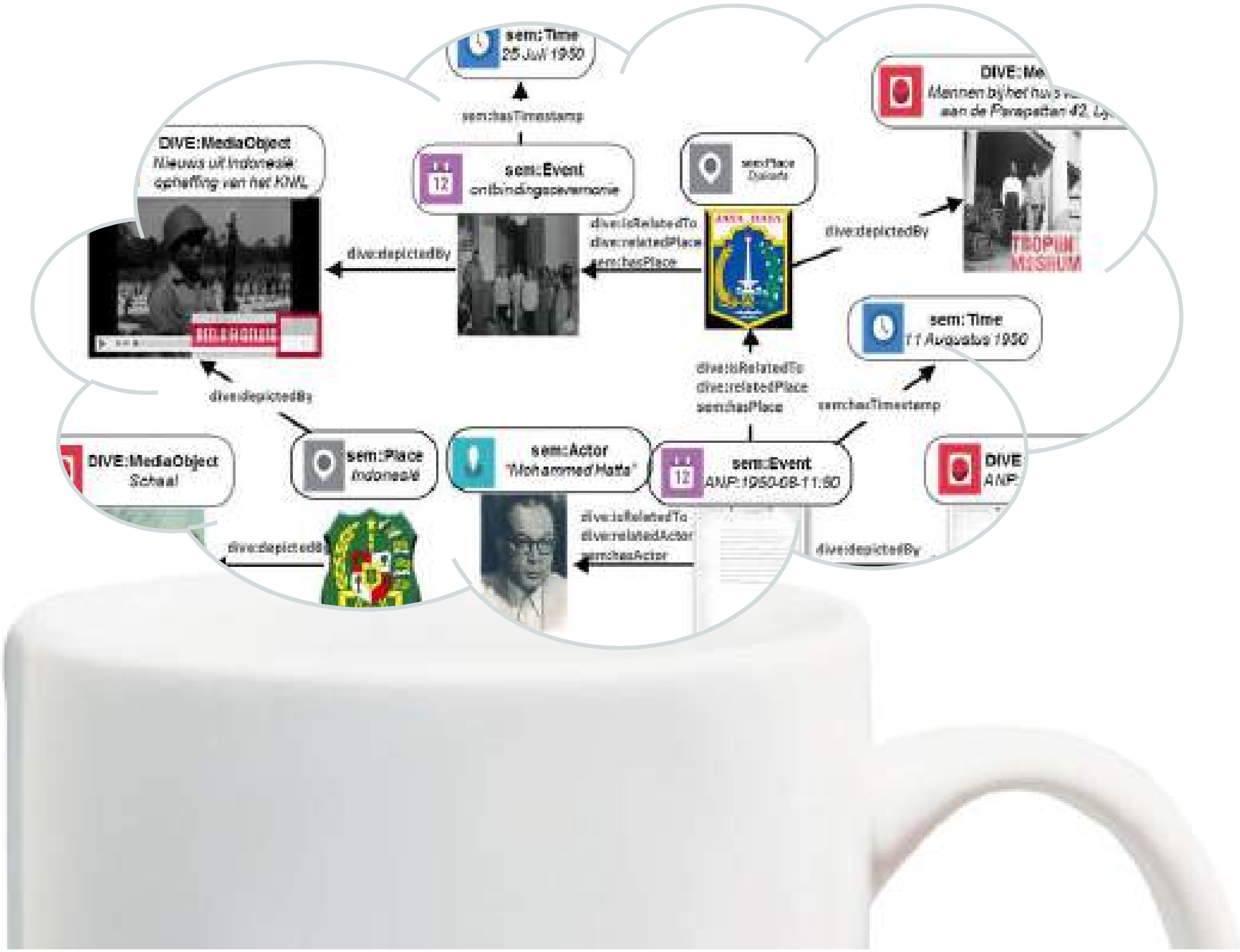
Web of Documents (WWW)
Linked Documents



Web of Data / Knowledge
Linked Data



Welcome Knowledge Graphs



Knowledge Graphs

Set of principles and technologies to represent *data, information* and *knowledge…*

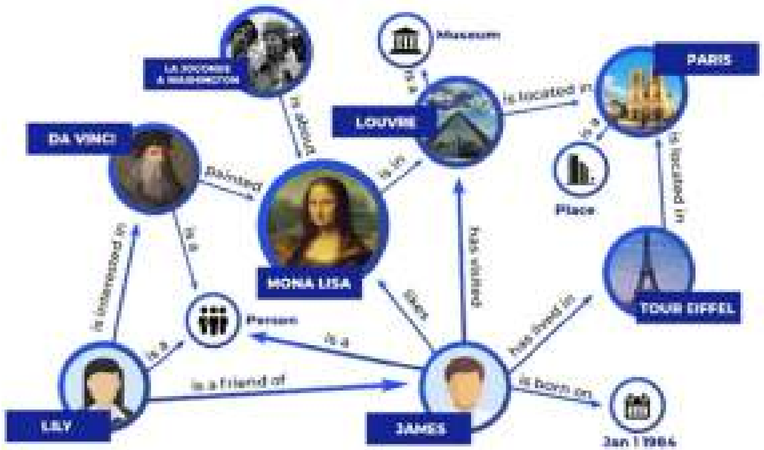
…allowing integration of heterogeneous and distributed data…

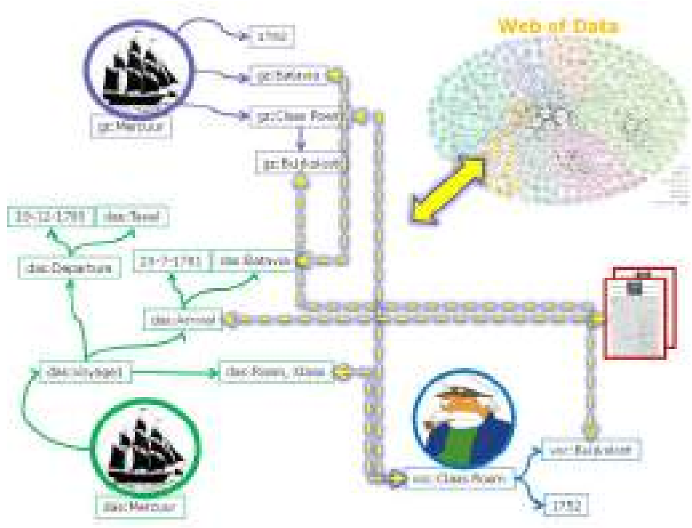
…using Semantic *Web standards* (RDF, OWL)…

…in the form of networks (graphs)…

…applicable in many domains…

…including Cultural Heritage.





4 proposals for knowledge graphs

1. Give all things a name
2. Names are addresses on the Web
3. Relations between things form Graphs of Data
4. Add explicit semantics (formal knowledge) to allow for predictable inferencing

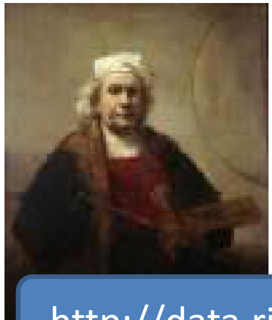
P1 Give all things* a name

"Now! *That* should clear up a few things around here!"

*) That you want to / can talk about

P2: Names are addresses on the Web (HTTP URIs)

Uniform **R**esource **I**dentifier (URI) is
a string of characters used to identify a
name of a resource



http://data.rijksmuseum.nl/person/Rembrandt

http://data.rijksmuseum.nl/person/Painting001

Often can be locally abbreviated or (rijks:painting1)

P3: Resource Description Framework (RDF)

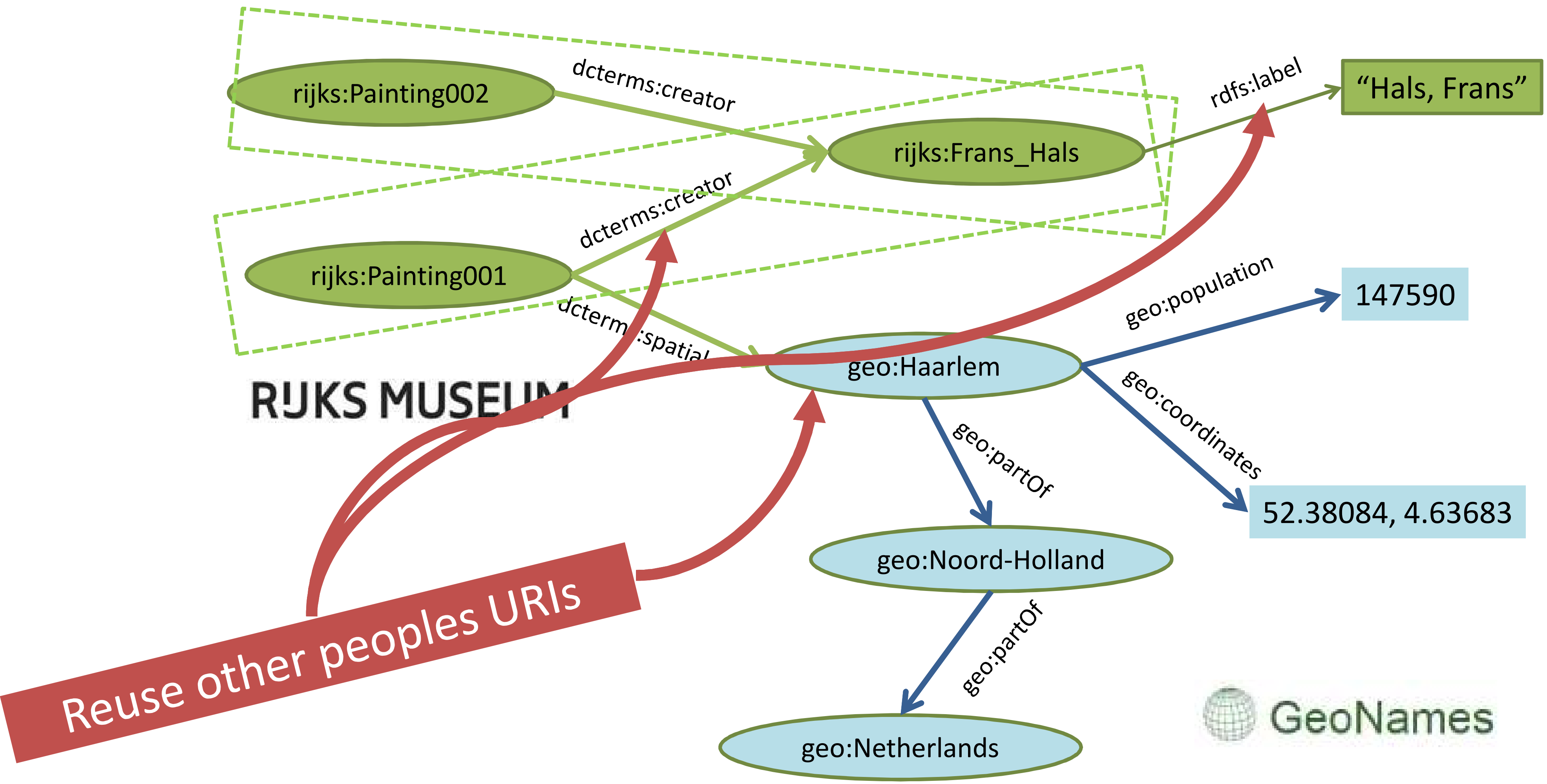
Semantic Web standard for writing down data, information

Triples!!

ject, Relation, Object)

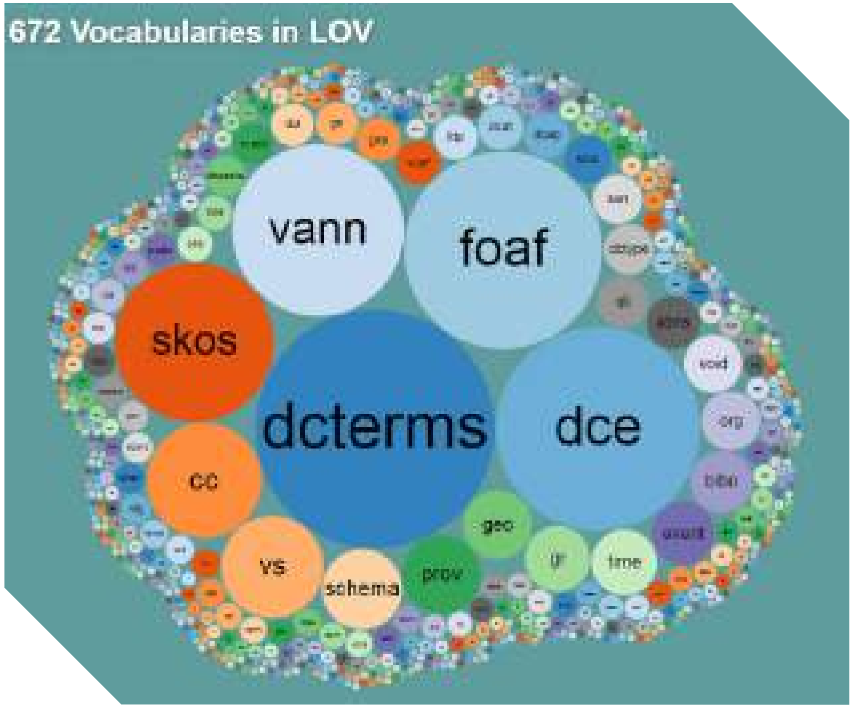
<Painting001> < has_location> <Amsterdam> .

P3: Triples form **Graphs**



Reuse other URIs: Examples

- **RDF and RDFS:** basic definitions of objects, properties, class-relations
- **OWL:** Description logics
- **FOAF** (Friend of a Friend): People, Organisations, Social Networks
- **schema.org** (Google, Yahoo!, Bing, Yandex): cross-domain, what search engines are interested in
- **Dbpedia/Wikidata** (Wikipedia as LOD): cross-domain
- **Dublin Core** (Bibliographic): publications, authors, media, etc.
- **CIDOC-CRM**:event-based model for cultural heritage.
- **PROV** to describe provenance of data



https://lov.linkeddata.es/

P4: Explicit semantics for predictable inferencing

All **Greek** are **men**
All **men** are **mortal**
All Greek are mortal

All **A** are **B**
All **B** are **C**
All A are C

Inferencing is algorithmic manipulation of **knowledge** to derive new knowledge

... where the **meaning** of words is **not needed**!

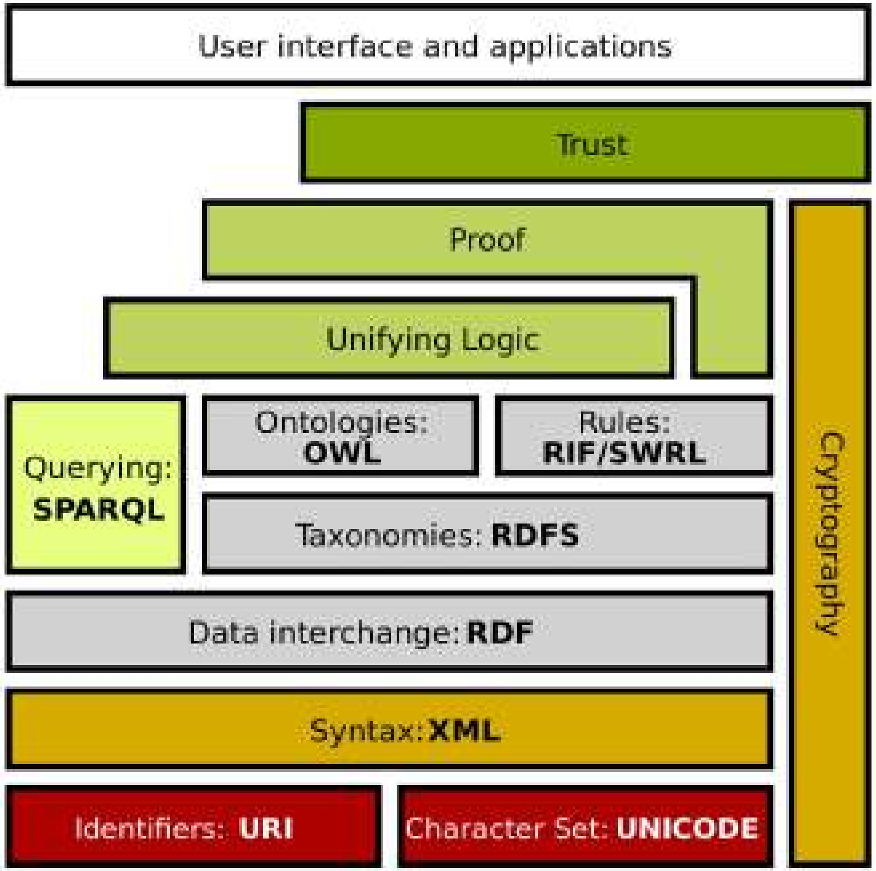
Calculating with Knowledge = **inferencing** or **reasoning**

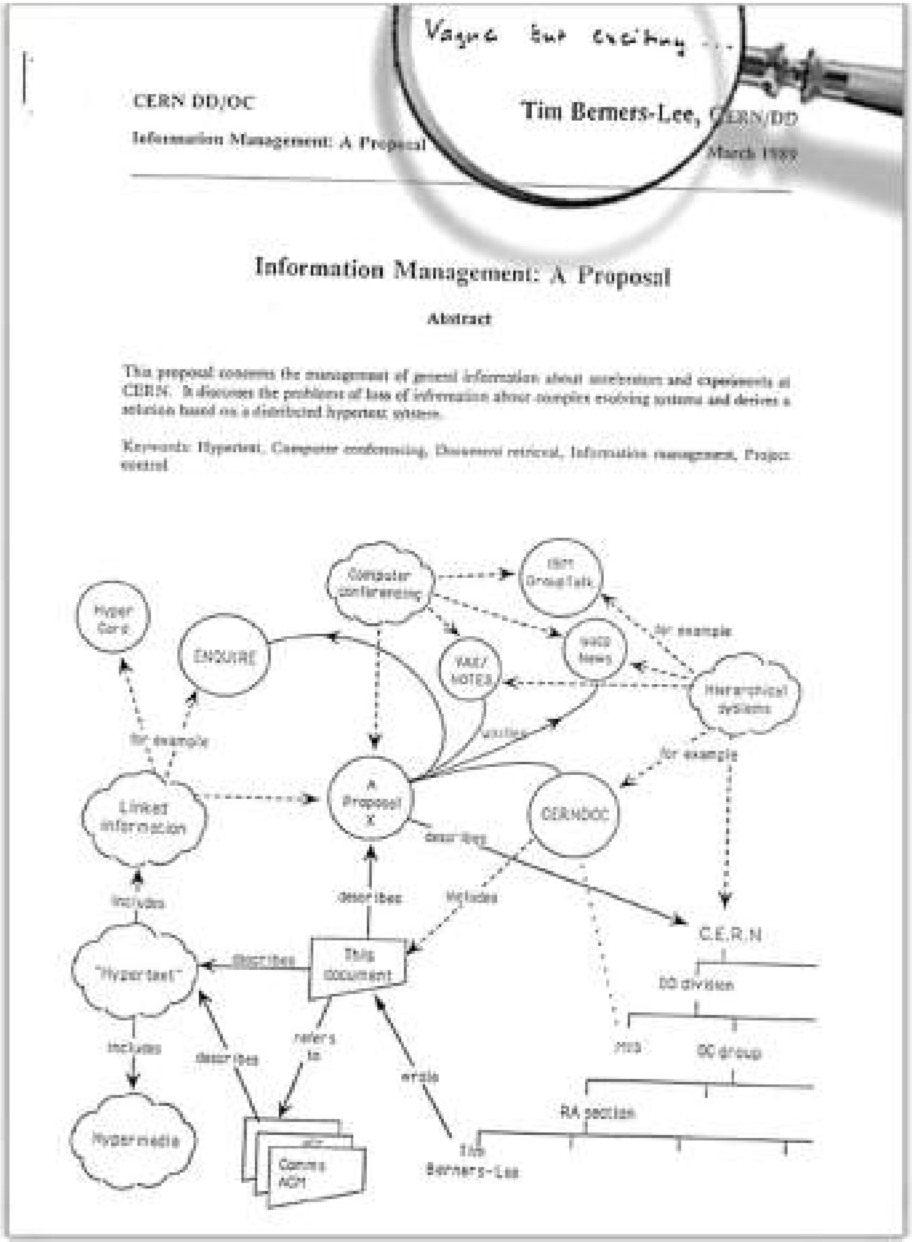
P4: Reserved, standardized symbols with clear *formal* semantics

- rdf:type

- rdfs:subClassOf
- rdfs:range

- owl:TransitiveProperty
- owl:disjointWith
- owl:sameAs



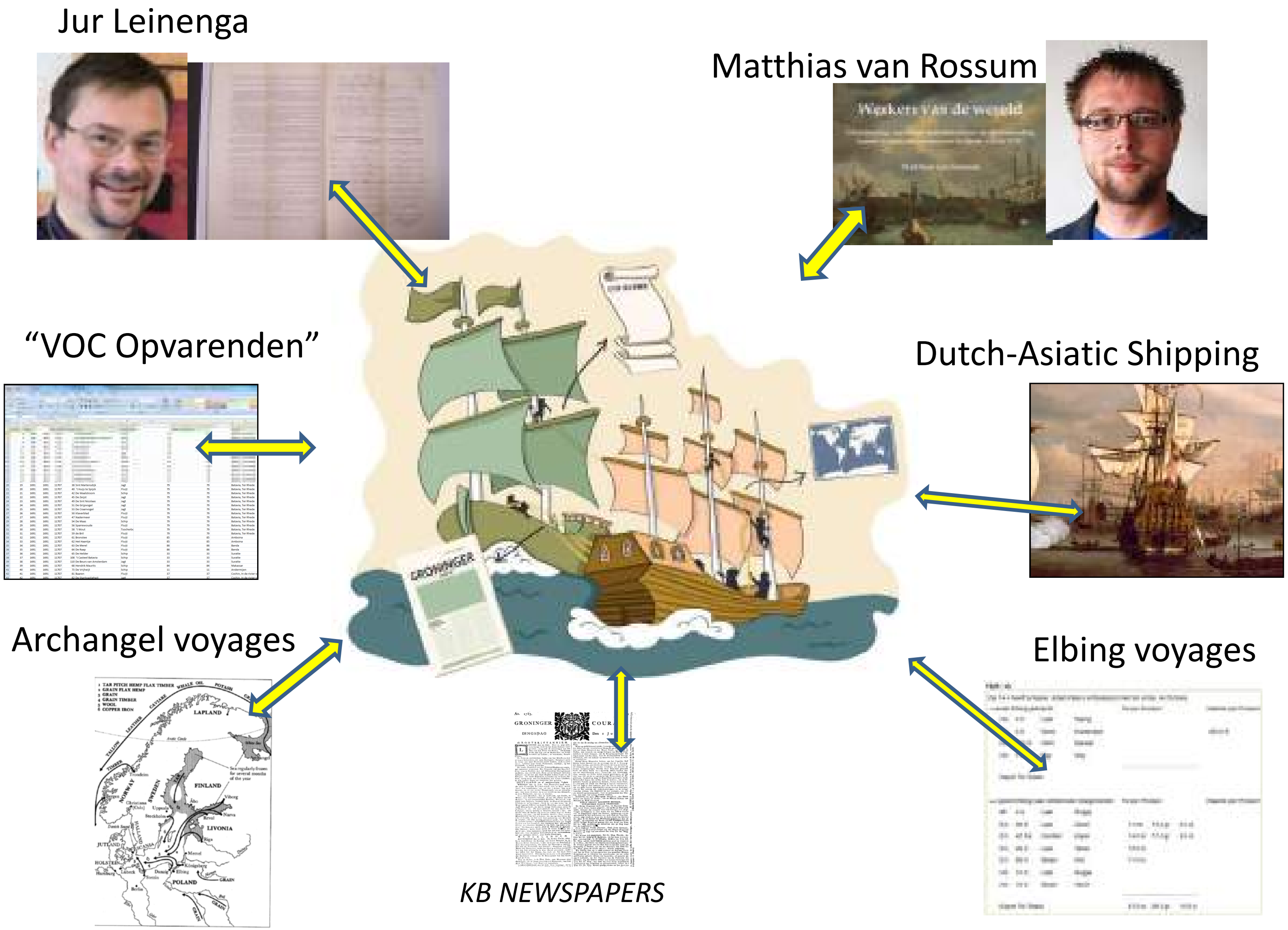


Tim Berners-Lee
The inventor of the (Semantic) Web

http://info.cern.ch/Proposal.html

How do you construct it and
what would we do with them?

Example: Dutch Ships and Sailors



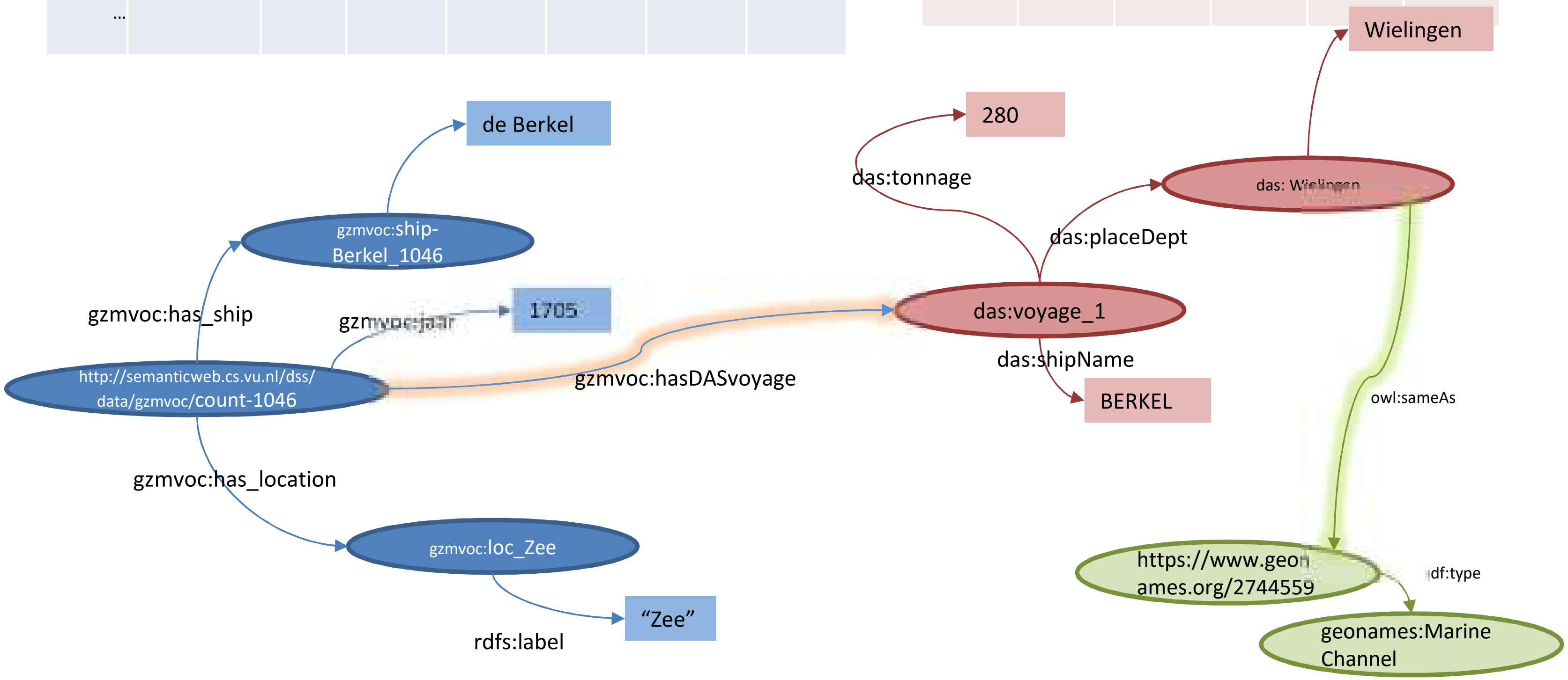
Building a maritime datahub

Generale Zeemonsterrollen VOC (Muster rolls)

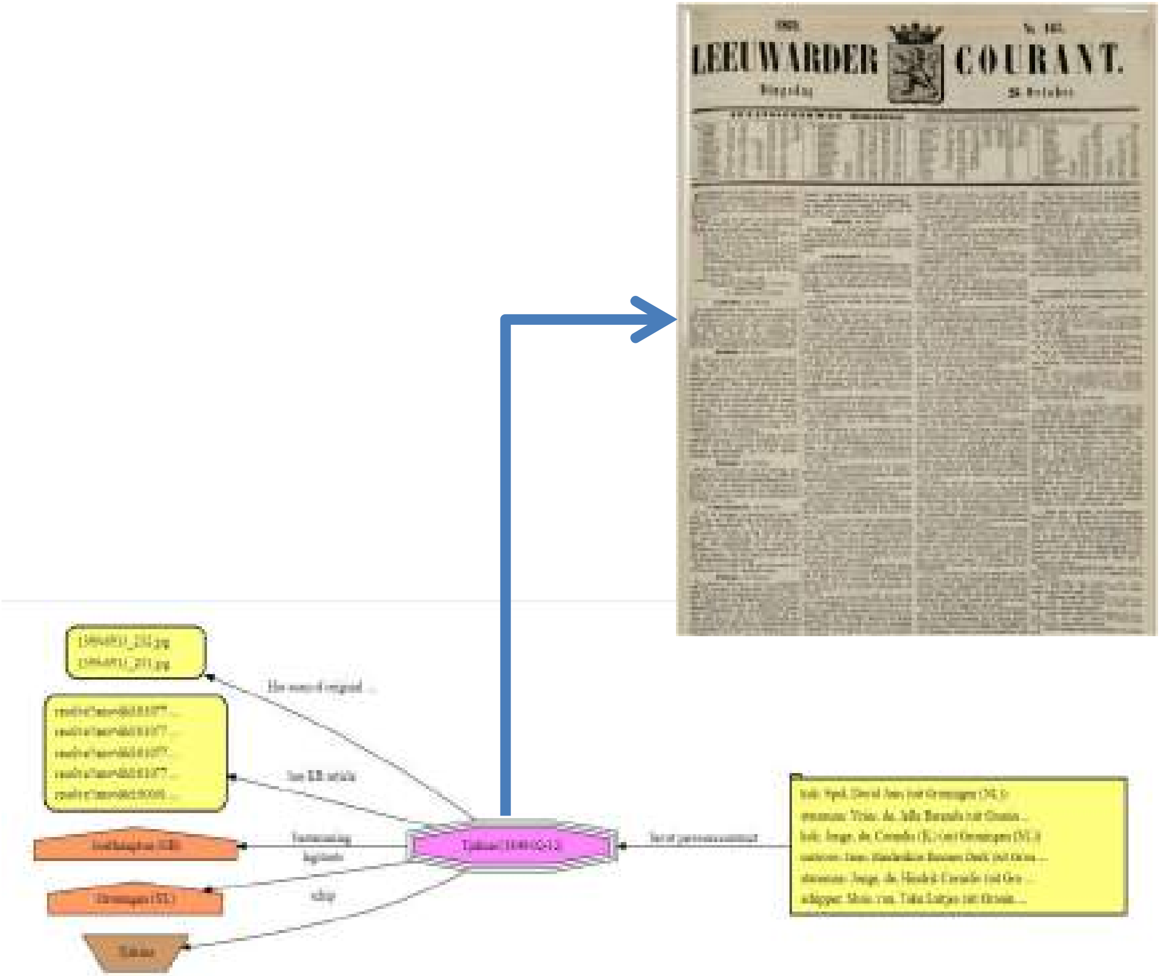
ID	Schip	Schipper	Jaar	aantalZeevarendenAz	aantalZeevarendenEur	DAS Heen	Location
1046	de Berkel	Phillipsen	1705	10	22	1918.6	Zee
1077	Den Boogaart	Pavije	1705	16	35	1933.1	Padang
1048	Waarden	Geleijns	1706	26	67	1915.1	Chiribon
...							

DAS Voyages

NR	ID	Name	PlaceDept	PlaceArr	Tonnage
1	1915.1	WAARDE	Wielingen	Batavia	830
2	1918.6	BERKEL	Texel	Batavia	280
3	1933.1	BOGAARD	Goeree	Batavia	200
...					



Use ML + background knowledge to identify Links to Historical Newspapers published by National Library



Only a few examples to learn from, re-use of background knowledge helps accuracy

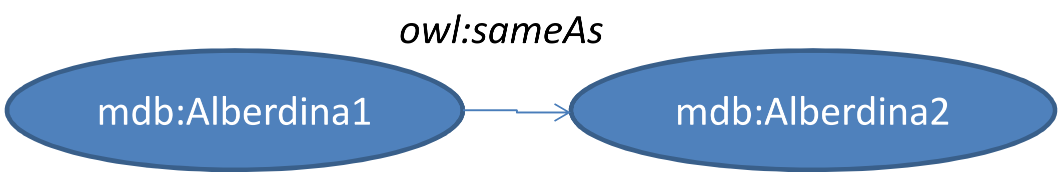
- Andrea Bravo Balado

Use ML+ background knowledge to identify ships

Date	ShipName	ShipType	ShipSize	HomePort	CurrentPort	Captain	
1852-02-27	Alberdiena	kof	NULL	NULL	Noorwegen (N)	Wolkammer	Albert Augustinus
1852-07-31	Alberdina	kof	NULL	Farmsum	Friedrichstadt (D)	Wolkammer	Albert A.
1861-09-30	Alberdina	kof	98	NULL	Gdansk, Danzig (PL)	Wolkammer	Albert Augustinus
1870-03-08	Alberdina	brik	222	NULL	NULL	Wolkammer	Albert Augustinus
1875-09-22	Alberdina	bark	309	NULL	Oostzee	Wolkammer	Augustinus

Only a few examples to learn from, re-use of background knowledge helps accuracy.

Results are clusters of same-as links, retain provenance.



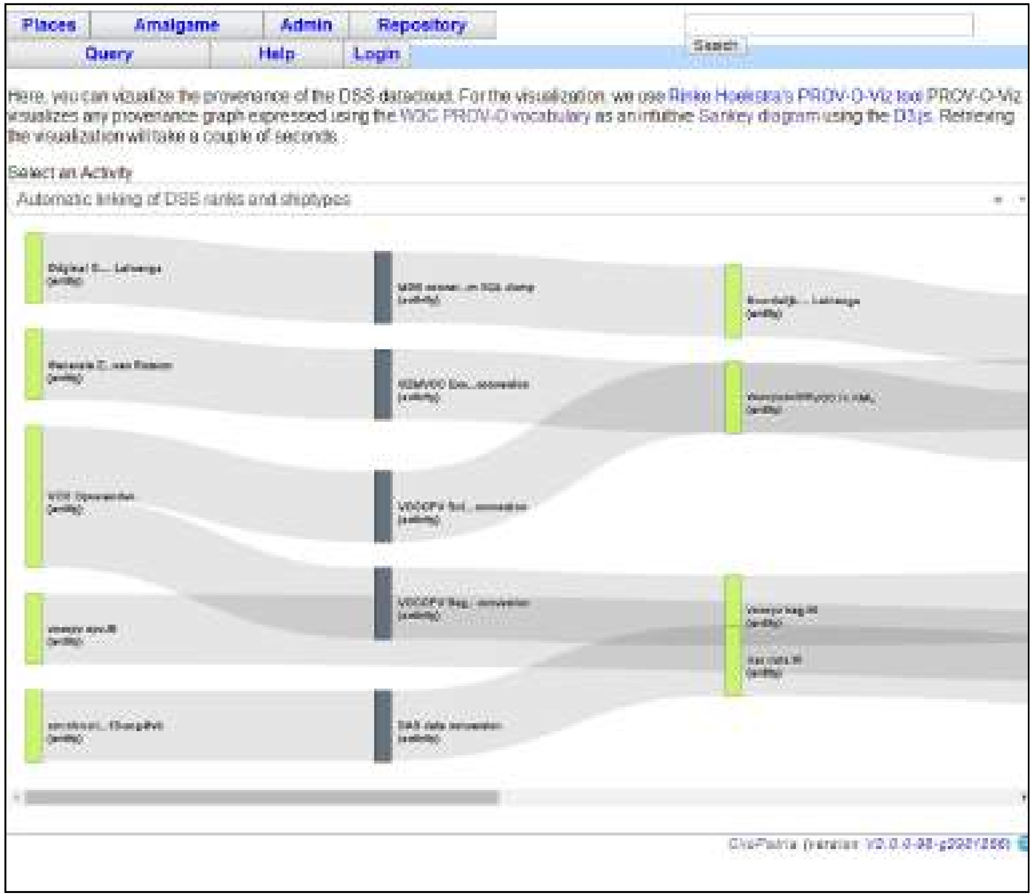
– Robin Ponstein

Provenance (1)

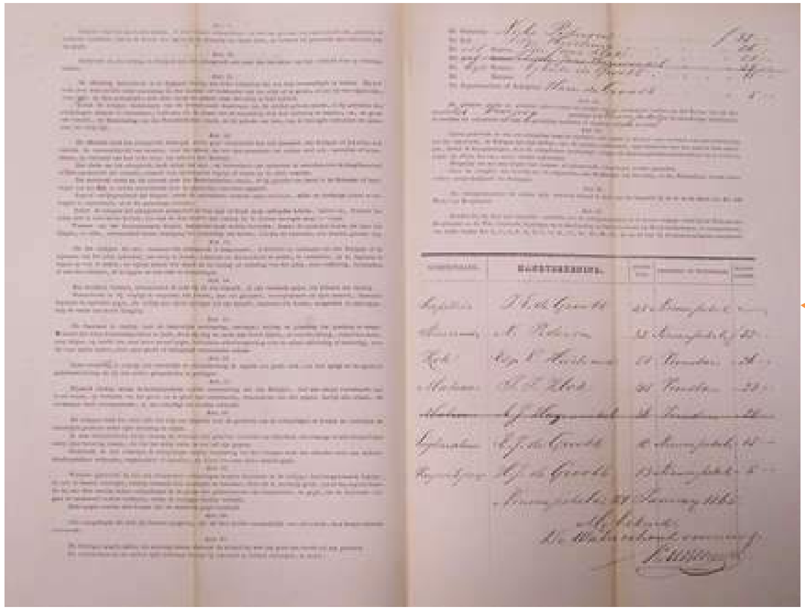
Individual *named graphs* have provenance information

- Who made it
  - Human: Me, historian, crowd
  - Algorithm
  - Hybrid?
- Based on what source
- Content confidence

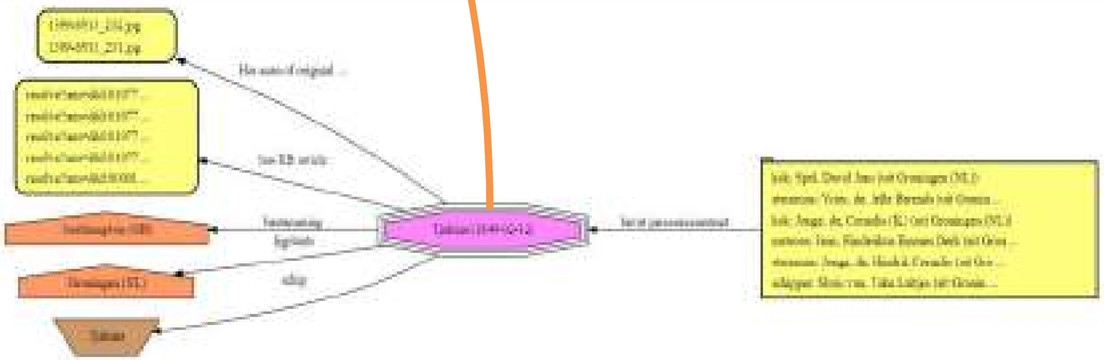
- Prov vocabulary



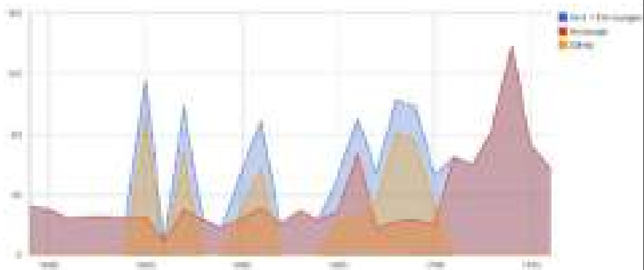
Provenance (2)

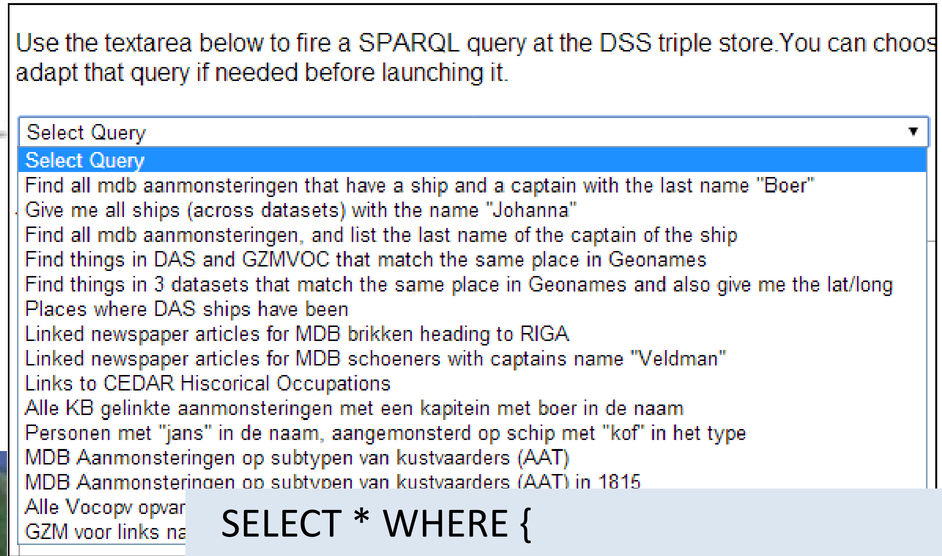


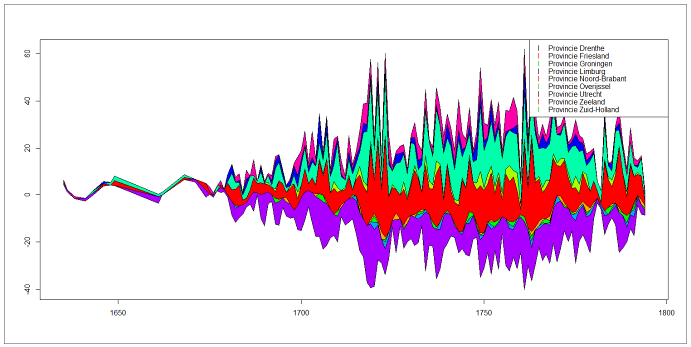
hasOriginalScan



Novel data access, analysis and visualisation







```
SELECT * WHERE {
    ?record dss:hasOriginalScan ?scan.
    ?record dss:has_kb_link ?kblink.
    ?record mdb:schip ?schip.
    ?schip mdb:scheepstype ?shiptype.
    ?shiptype skos:exactMatch ?em.
    ?em skos:broader* aat:kustvaarders.
}
```

Lessons learned

- KGs are great for integrating datasets
  - Without the need to force everything into one datamodel
  - Guided by domain experts
  - Enriched by hybrid methods
  - Retain original model and intent, reuse another day
  - New research questions
- Re-use background knowledge
- Provenance fits very well to make source, enrichments transparent
  - Accessible to end-users
- Linked Data is the (technically) best way to do FAIR data publishing

Knowledge graphs for heterogenous,
multimodal heritage data

DIVE+


Presser Breivik knallhardt om
«Knights Templar»-nettverk
Dagbladet
SLIK ER
REDD
FOR I RETTEN

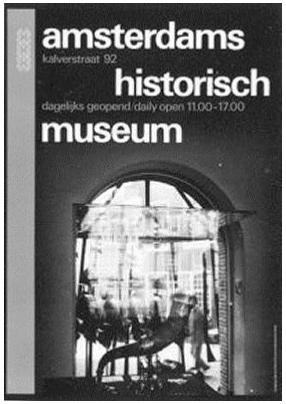
Collections and Vocabularies

OPENIMAGES.EU

3,220 news broadcasts

Netherlands Institute for Sound & Vision

GTAA thesaurus

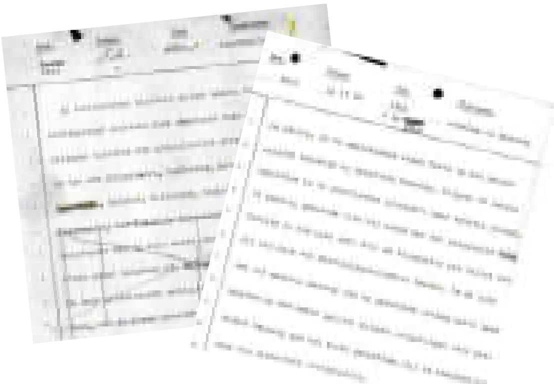


AMSTERDAM MUSEUM

73,447 cultural heritage objects

AM T

AMSTERDAM MUSEUM



KB National library of the Netherlands

DELPHER.NL

197,199 Scans of Radio bulletins

1937 – 1984

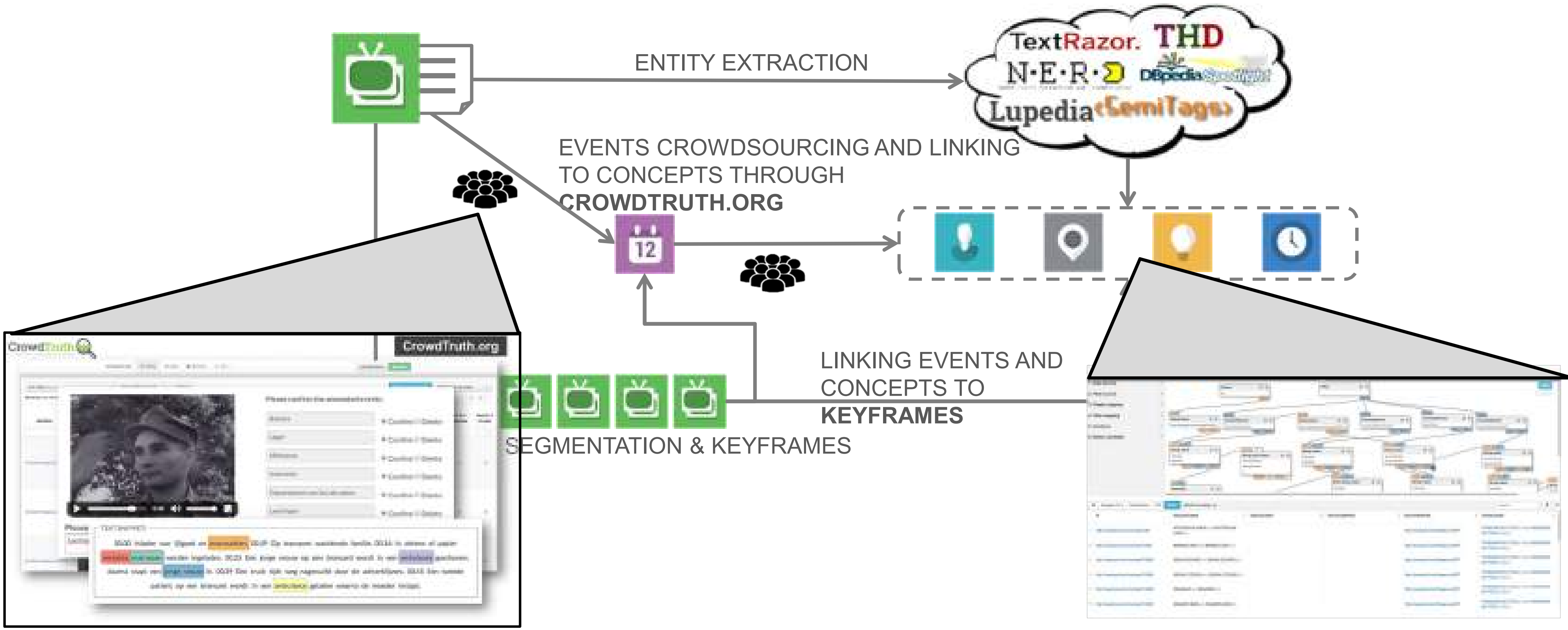
TROPENMUSEUM

78,270 cultural heritage objects

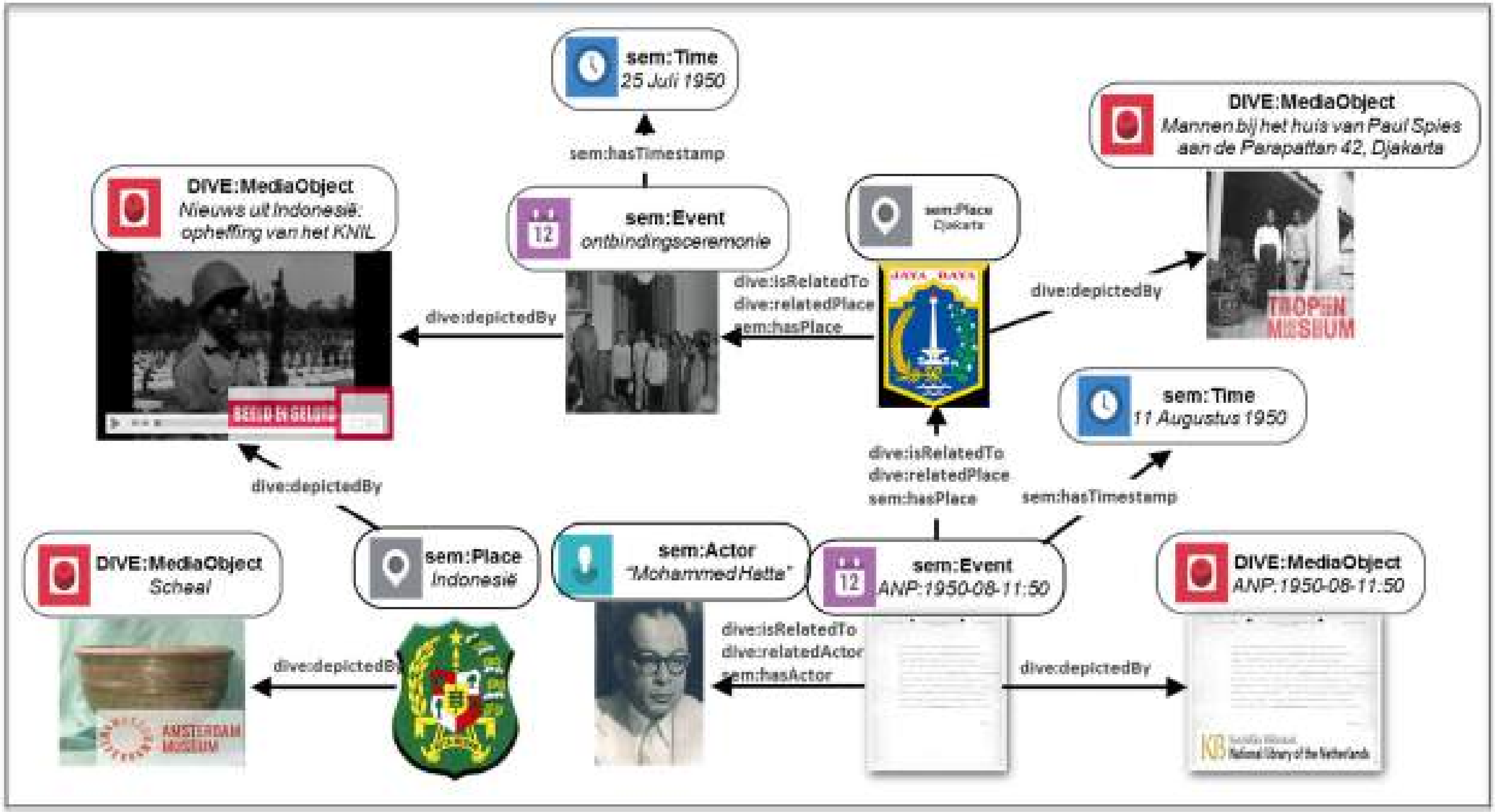
SVNC us

Tropenmuseum

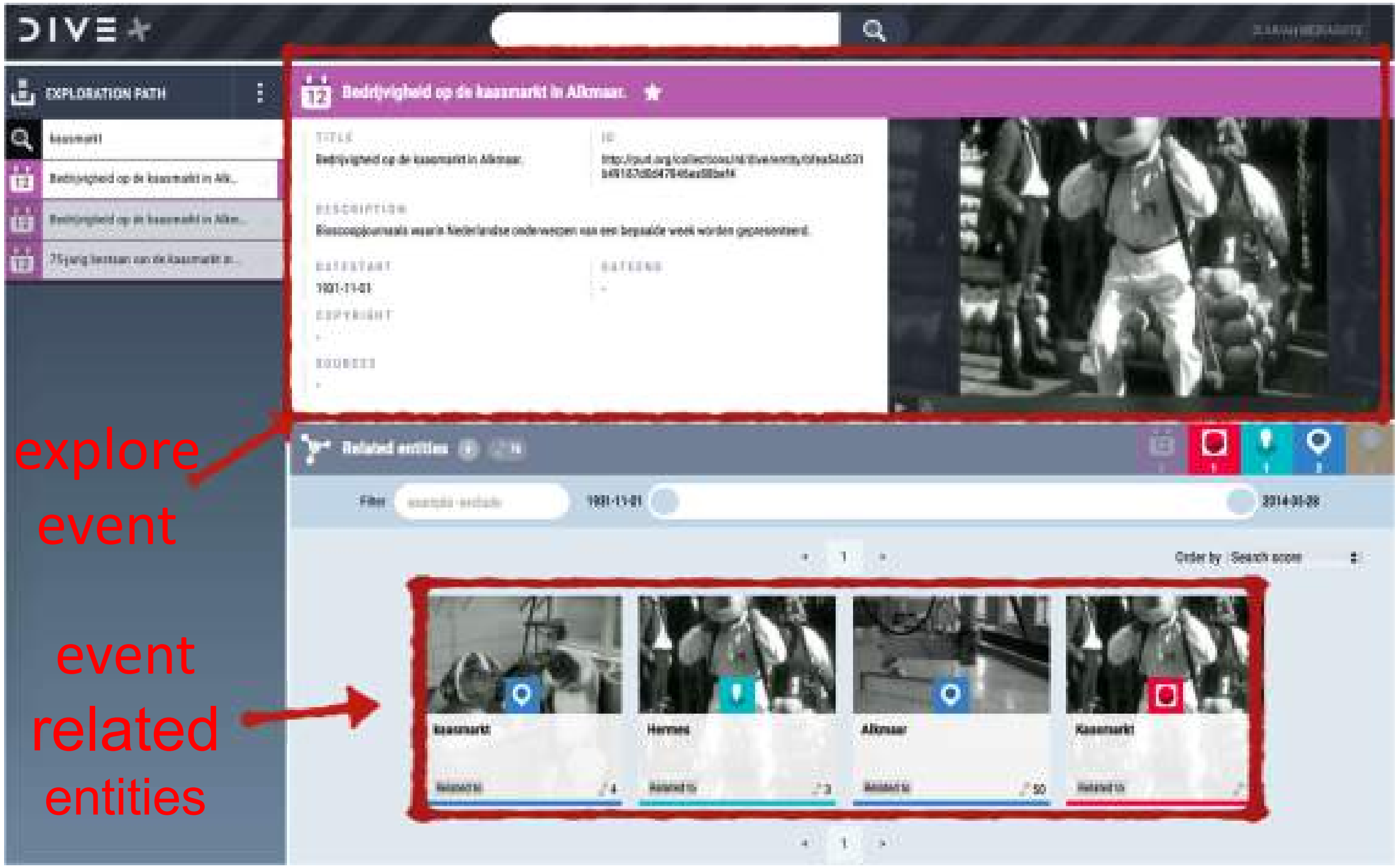
Hybrid Enrichment Pipeline: Text Analysis, Crowdsourcing, Alignment



Cultural Heritage/ Media **Knowledge Graph**



Exploratory event-based browsing



https://www.youtube.com/watch?v=nSJNHdqiTgM

Lessons learned

- KGs are great for integrating **multimodal** datasets
  - **Use RDFS to map to one shared** datamodel
  - Guided by domain experts (**interactive alignment**)
  - Enriched by hybrid methods (**ML + Crowdsourcing**)
  - Retain original model and intent, reuse another day
  - New research questions **and exploratory browsing**

- Re-use background knowledge

- Provenance fits very well to make source, enrichments transparent
  - Accessible to end-users

How about Machine Learning and Knowledge Graphs

Learning and Reasoning

Reasoning

Deductive

Based on formal *logic*al rules

If *<x rdf:type A>* and *<A rdfs:subClassOf B>* then *<rdf:type B>*

RDFS, OWL, other

Learning

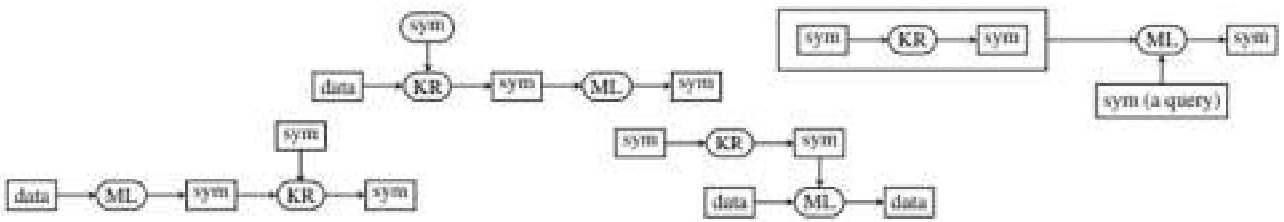
Inductive

Based on *statistics*

If *<x rdf:type A>* and *<x hasSize 100> and <y rdf:type A>* then maybe *<y hasSize ~100>*?

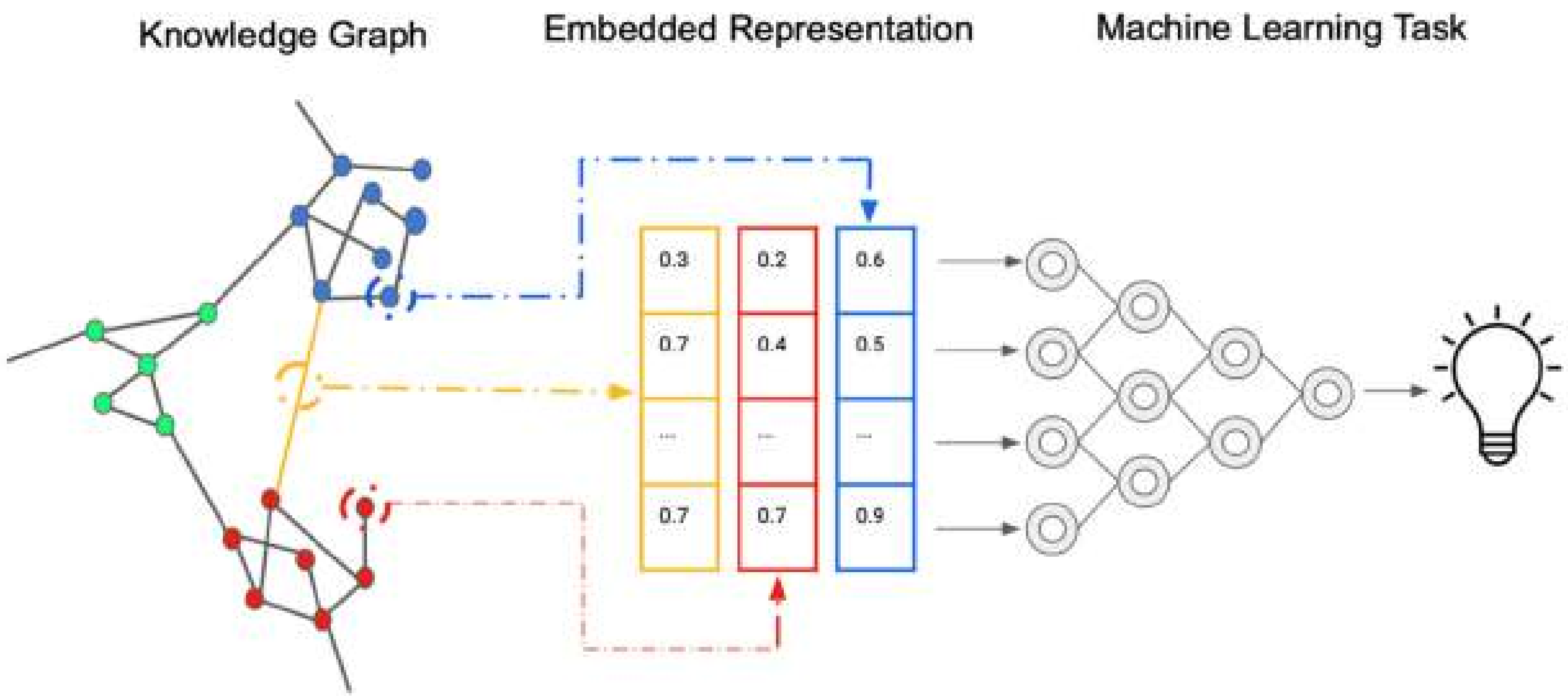
Rule mining, Embedding methods, Deep methods

Many Hybrid Approaches



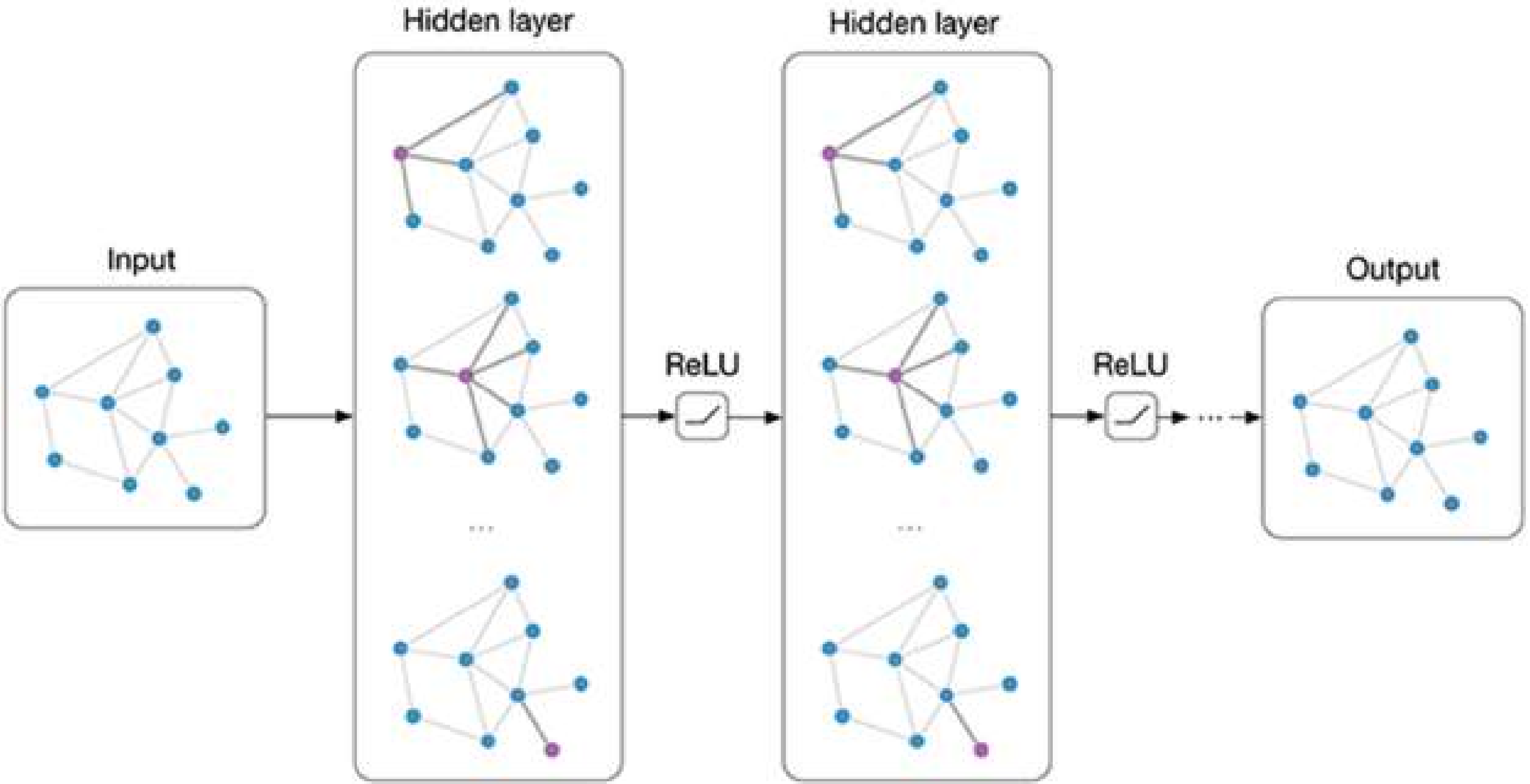
A Boxology of Design Patterns for Hybrid Learning and Reasoning Systems- van Harmelen, ten Teije (2019)

Knowledge Graph Embedding methods (RDF2VEC, TransE, DistMult,…)



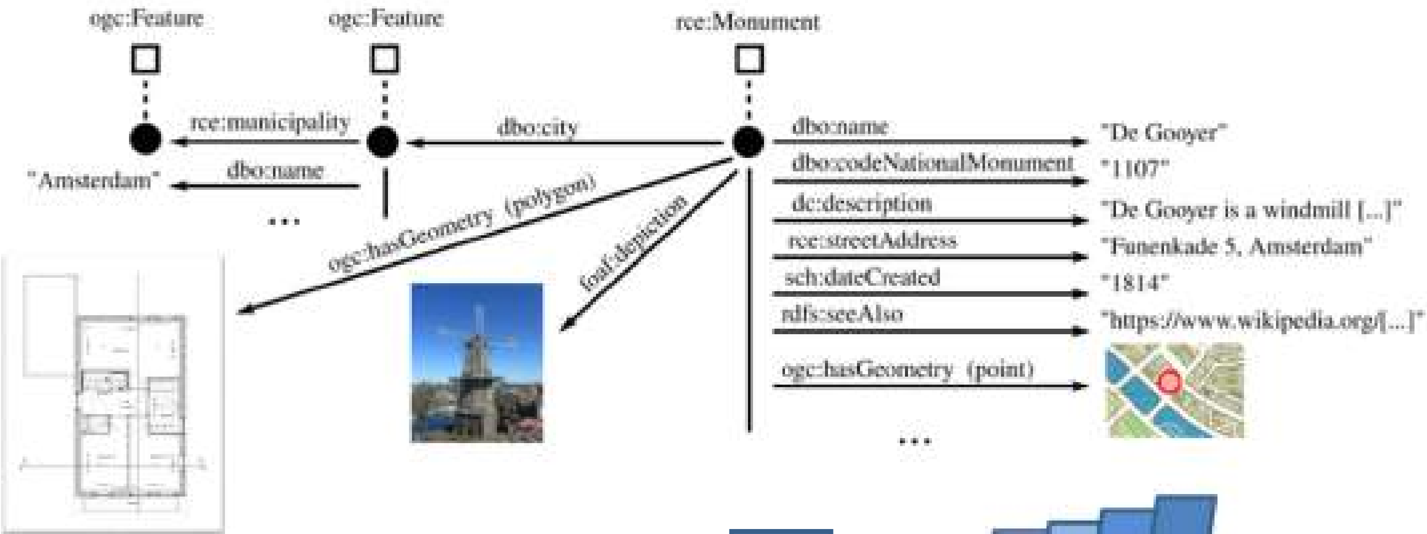
https://en.wikipedia.org/wiki/Knowledge_graph_embedding

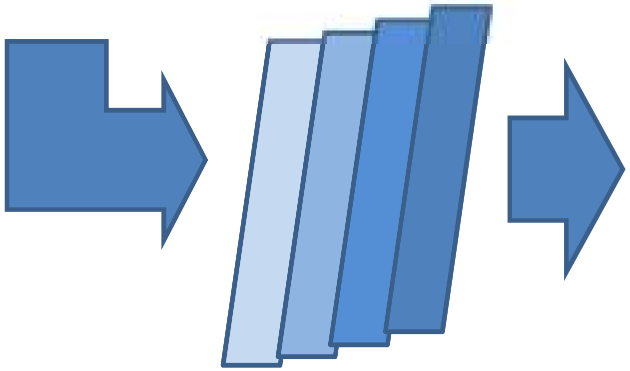
(Knowledge) Graph convolutional methods



https://tkipf.github.io/graph-convolutional-networks/

End-to-end learning on **multimodal** knowledge graphs



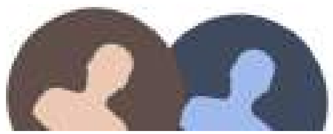


Church

Barn

house

...

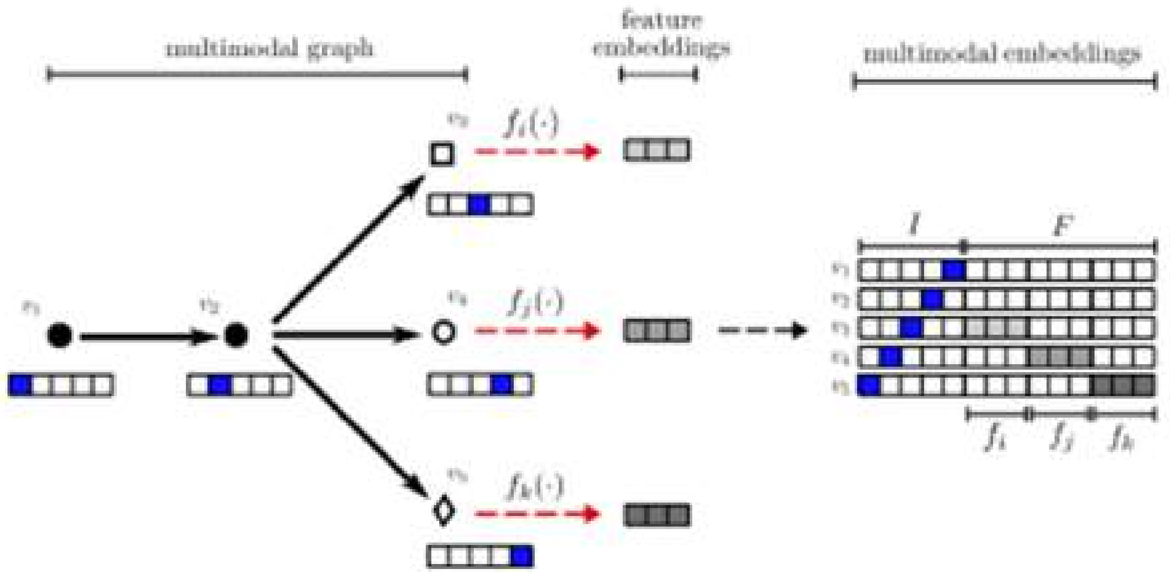


Xander Wilcke

[redacted]

Can be derived from RDF literal datatypes

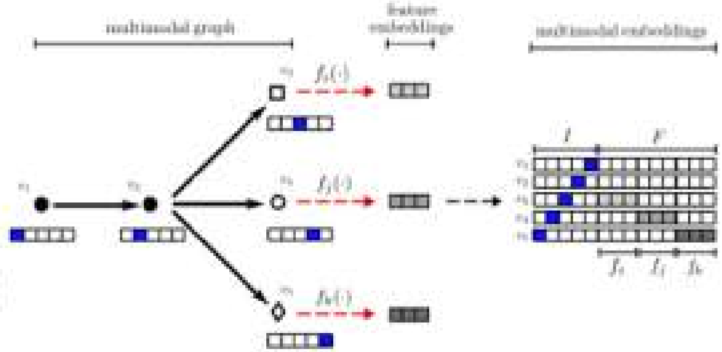
- Extend GCNs [redacted]



Modules for:
- Numbers (int/float)
- Temporal information (months, days etc)
- Text (CNN)
- Visual information (CNN)
- Spatial information (van het Veer et al)
- RGCN message passing for nodes

gitlab.com/wxwilcke/mrgcn

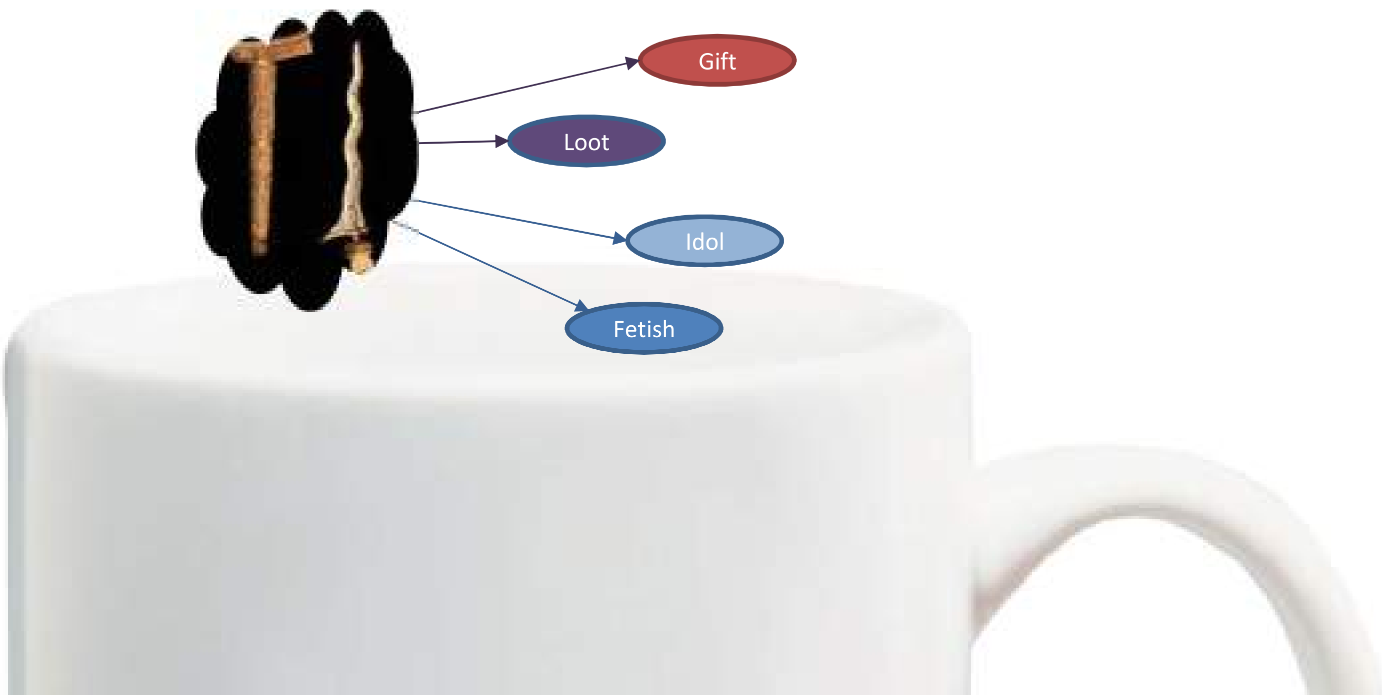
Lessons learned



- KGs are great for integrating **multimodal** datasets
  - Excellent as “default” data model for ML
- End-to-end (Deep) Machine Learning offers great op
  - Link prediction
  - Classification
  - Message passing methods to learn embedding of graph extended with modality-specific modules
- Various challenges to resolve

Dealing with 1) *implicit* 2) *incomplete 3) differently-structured 4) multi-modal* knowledge

Knowledge Graphs and Polyvocality



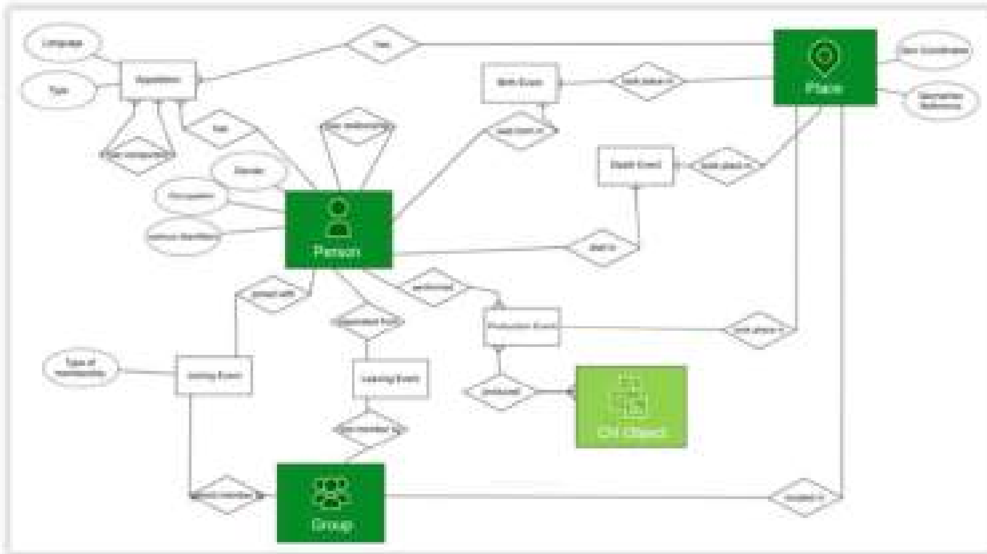
Integration of four national biographical KGs (Austrian, Dutch, Finnish, Slovenian)

Shared data model (CIDOC-CRM + BIOCRM)

But individual richness intact

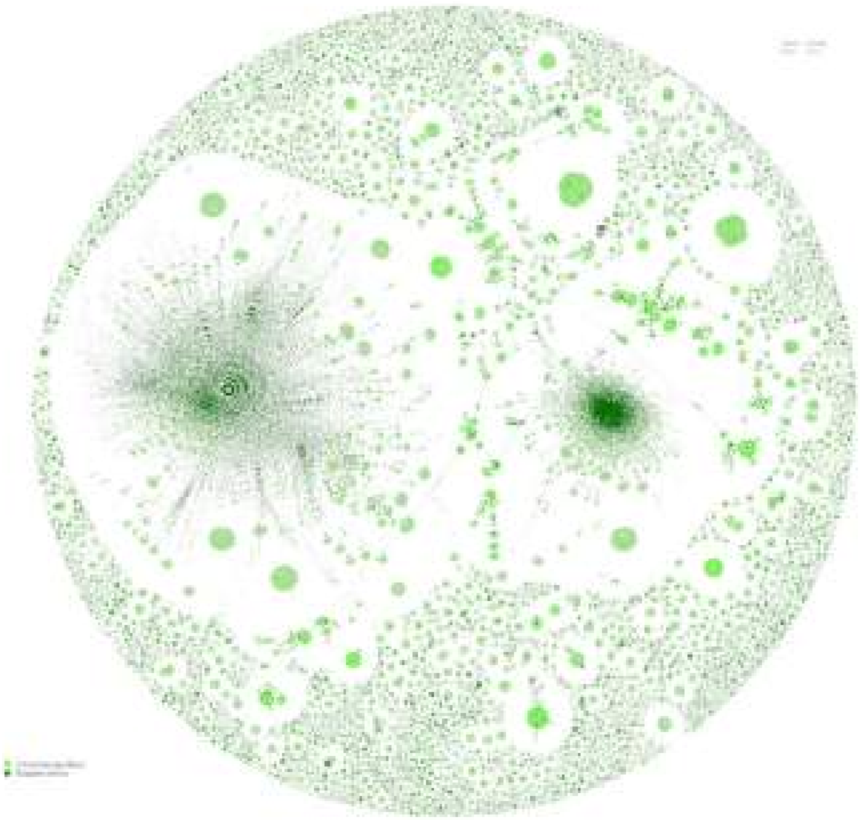
Tool suite for DH researchers, educators etc.

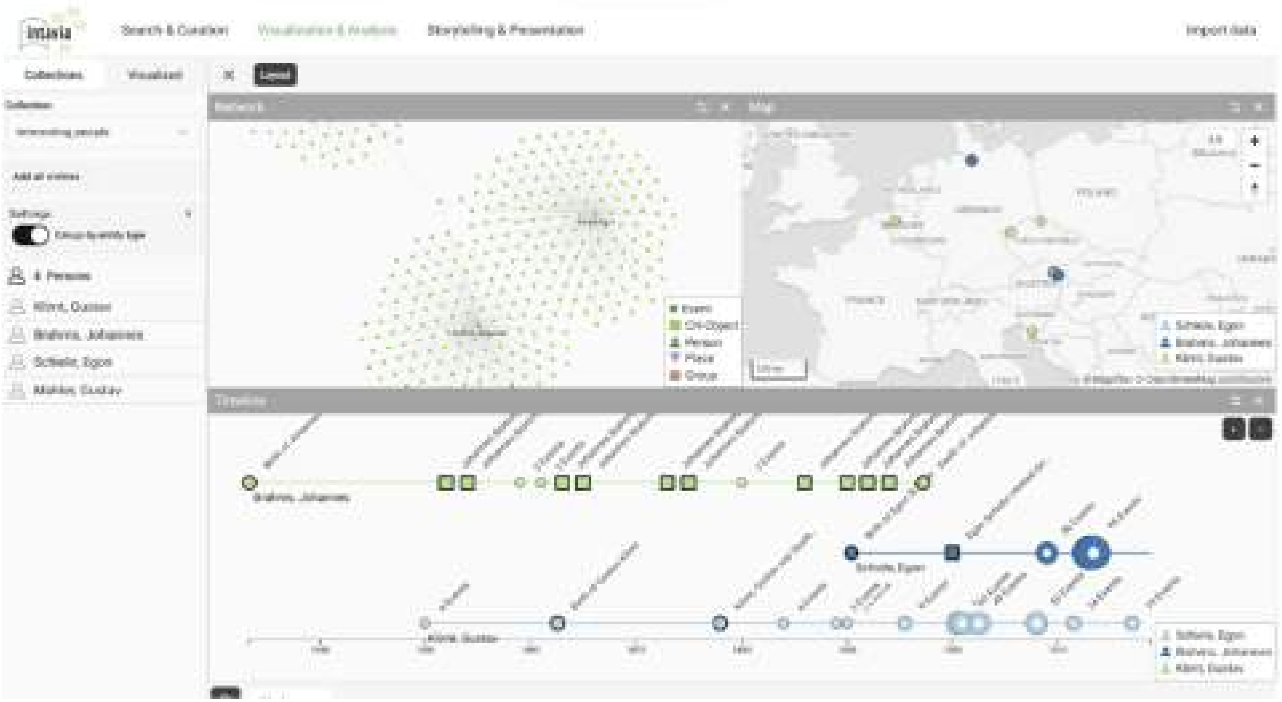
https://intavia.eu/



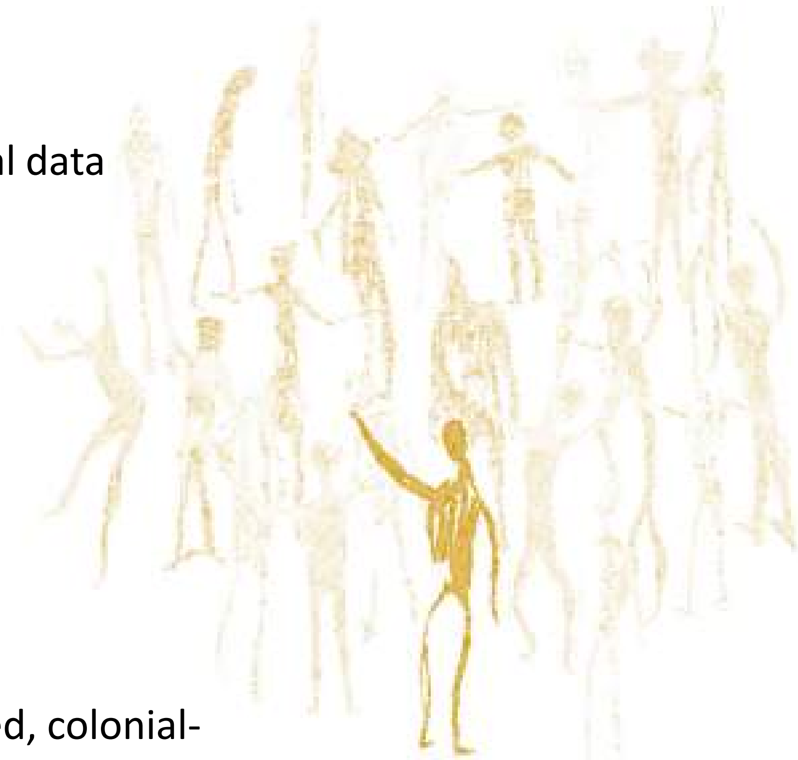
Response: Development of the IDM-RDF data model intavia

In/Tangible European Heritage
Visual Analysis, Curation & Communication





Polyvocality



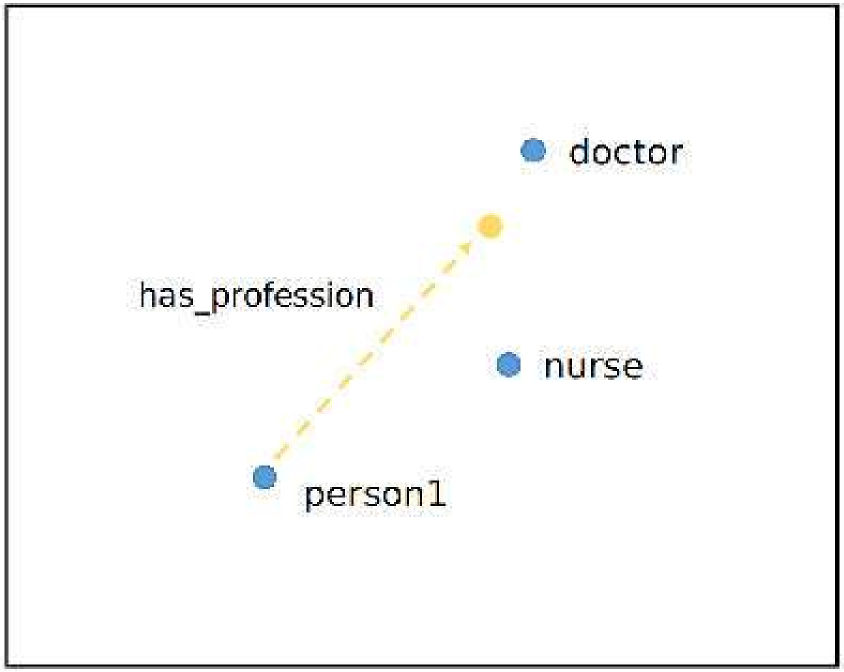
Knowledge graphs, especially those based on historical, cultural data are sure to contain

Biased
Univocal
Single-view
Culturally sensitive

information, based on the majority-view.

This poses the danger of perpetuating gender-/class- etc. biased, colonial-view data.

Bias



https://www.amazon.science/blog/mitigating-social-bias-in-knowledge-graph-embeddings

Perspective

Metadata enrichment and bias detection of colonial architecture

Roz Sabir

Towards *polyvocal* Knowledge Graphs

Identifying and acquiring polyvocality knowledge

- Identify existing voices
- Elicit information from polyvocal sources

Representation of polyvocality: models, patterns

- Represent disagreement on categorisation, provenance, etc.

Presentation of polyvocal knowledge

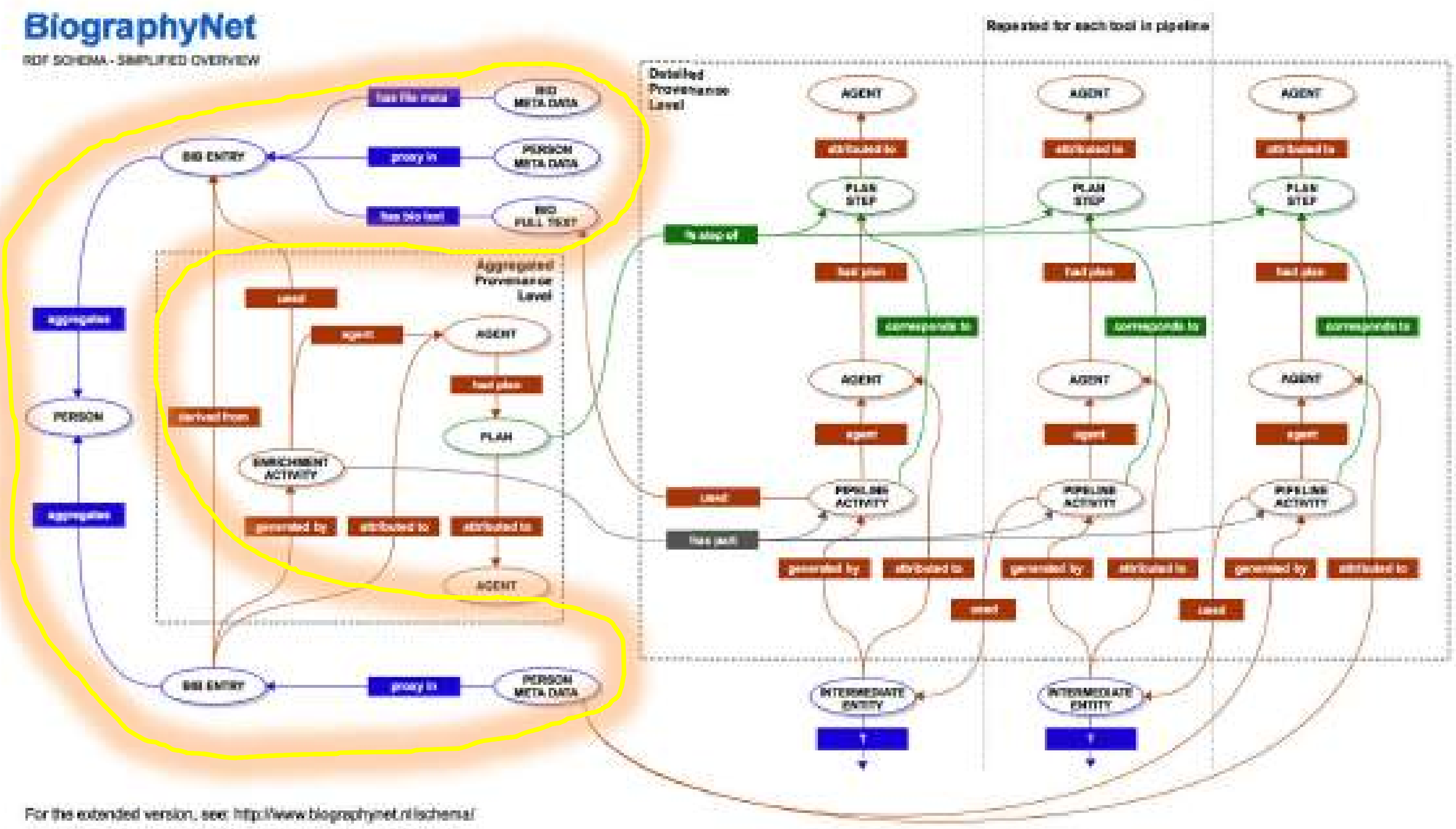
- present it to variety of users, including

researchers
heritage professionals
general public
'source communities'


A Polyvocal and Contextualised Semantic Web*

1 Introduction

Polyvocality and **cultural context** in Biographical knowledge graphs



Niels Ockeloen

Go Sugimoto

Ockeloen, Niels, et al. "BiographyNet: Managing Provenance at Multiple Levels and from Different Perspectives." *LISC@ ISWC*. 2013.

Argumentation for **explainable** inconsistency resolving in **polyvocal** knowledge graphs

Loan Ho

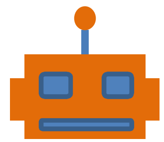
Consider $\mathscr{K}_1 = \{\mathscr{R}_1, \mathscr{C}_1, \mathscr{F}_1\}$ where:

$$\mathscr{R}_1 = \{R : \forall x Person(x) \rightarrow \exists y hasDeathdate(x,y)\},$$
$$\mathscr{C}_1 = \{C : \forall x,y,z Person(x) \wedge hasDeathdate(x,y) \wedge hasDeathdate(x,z) \rightarrow y = z\},$$
$$\mathscr{F}_1 = \{f_1 : Person(Thorbecke), f_2 : hasDeathdate(Thorbecke, 14/10/1860),$$
$$f_3 : hasDeathdate(Thorbecke, 10/10/1860)\}$$

$$A_2 = (\{Person(Thorbecke)\}, \{hasDeathdate(Thorbecke, 10/10/1860)\})$$
$$\uparrow \; \forall x,y,z Person(x) \wedge hasDeathdate(x,y) \wedge hasDeathdate(x,z) \rightarrow y = z$$
$$A_1 = (\{Person(Thorbecke)\}, \{hasDeathdate(Thorbecke, 14/10/1860)\})$$

User: Why not $hasDeathdate(Thorbecke, 10/10/1860)$ given that A_2? [8]

Reasoner: Because $hasDeathdate(Thorbecke, 10/10/1860)$ [9] the following constraint is violated: $\forall x,y,z Person(x) \wedge hasDeathdate(x,y) \wedge hasDeathdate(x,z) \rightarrow y = z$.



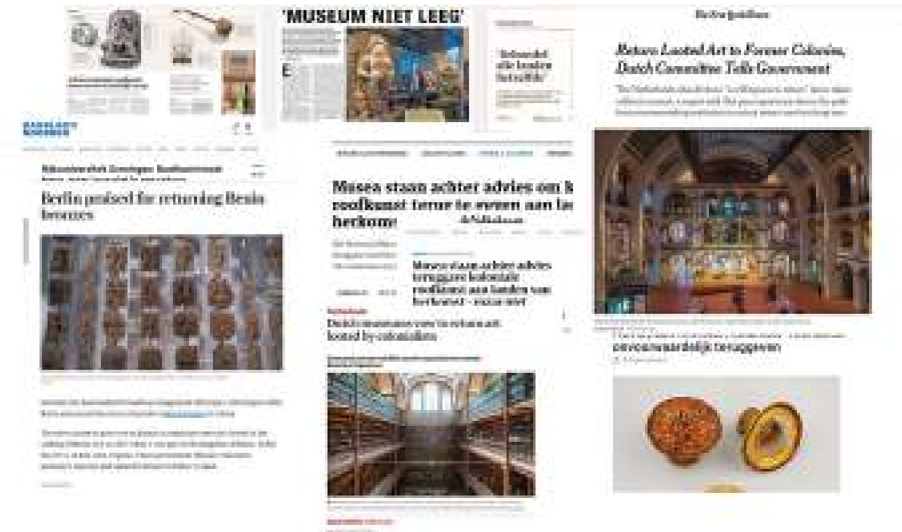
User: I understood "why 10/10/1860 is not Thorbecke's death date"

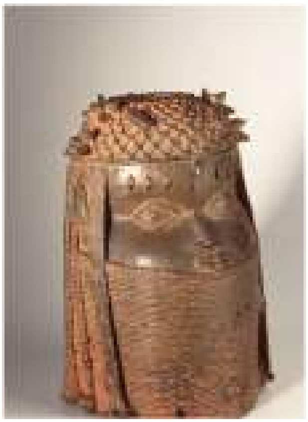
PRESSING MATTER

Rethinking colonial heritage collections

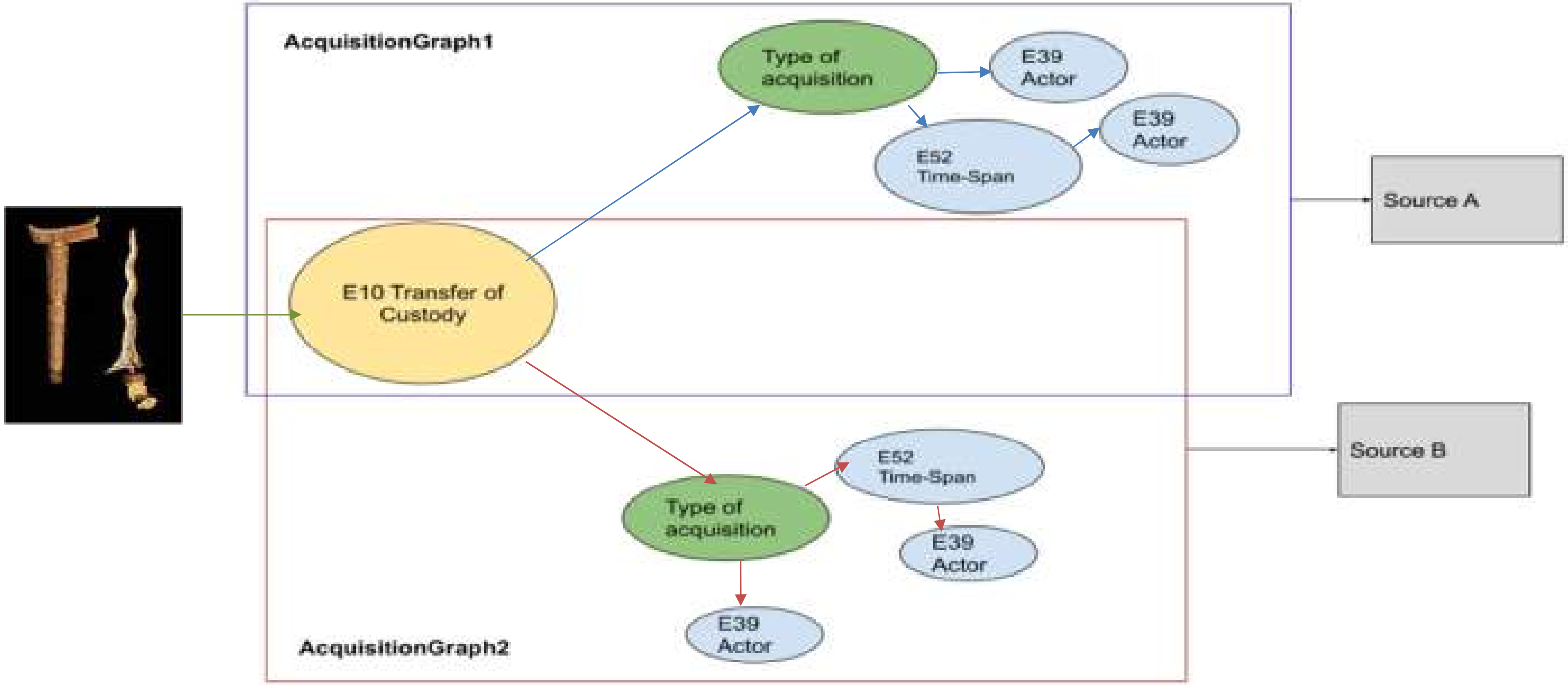
Modes of Acquisition

1. Scientific (including Expeditions)
2. Involuntary dispossession (violent)
3. Trade (diplomatic exchanges and legal sales)
4. Voluntary dispossession (Missionary)





Using provenance to represent multiple views in colonial heritage knowledge graphs

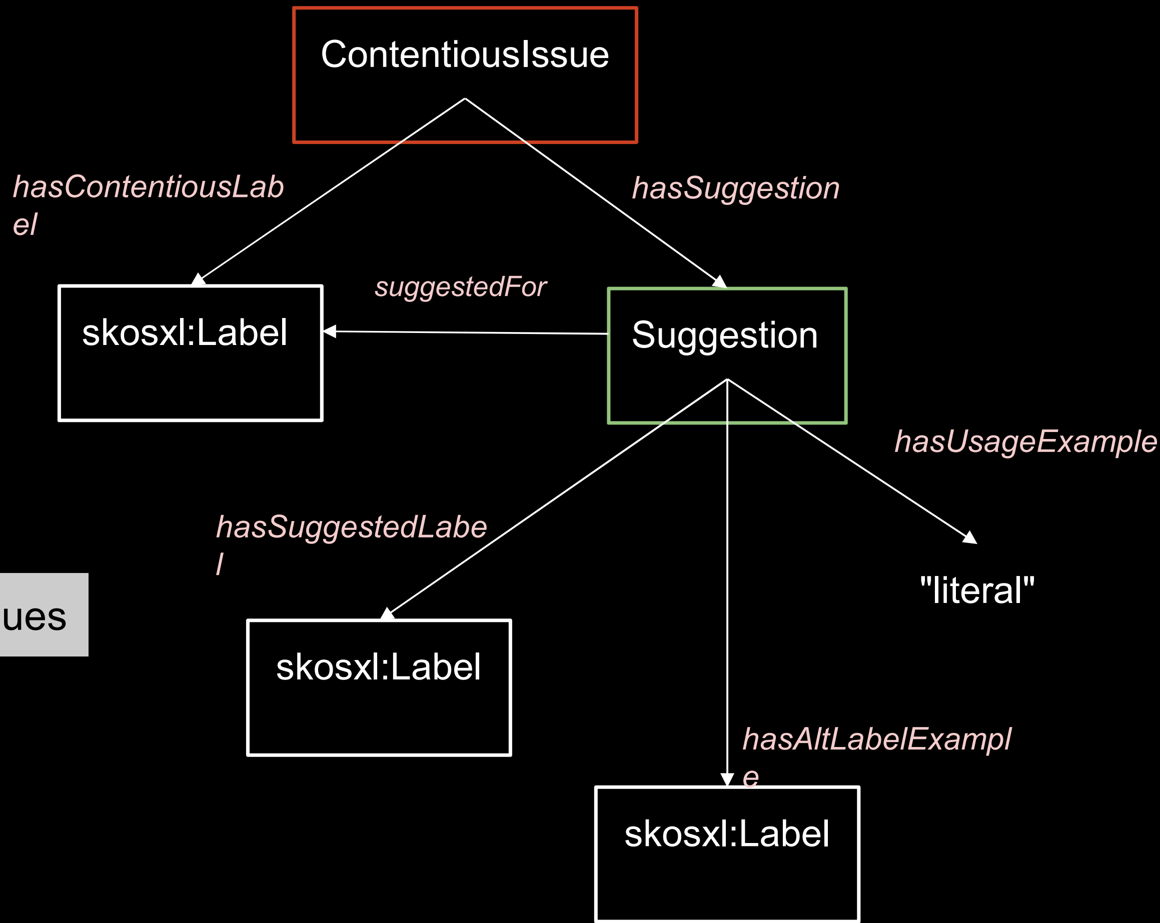


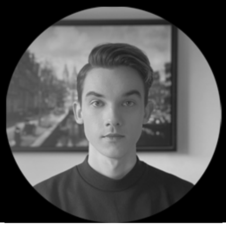

Sarah Shoilee

Contentious terms in the cultural sector: From expert knowledge to linked data

111 Contentious Issues

296 Labels





Andrei Nesterov
PhD student
CWI, Human-Centered
Data Analytics

Cultural AI Lab (culturalai.nl)

AI

Cultural
Heritage

Bias
Ethics
Cultural
differences
Perspectives

Lessons learned

- KGs provide means for integrating datasets
  - Keeping **multiple perspectives** on data in tact
  - Guided by domain experts

- Provenance is key
- Patterns allow for transparent analysis by end-users

Take home

- Cultural Heritage Organisations are becoming more ***Open, Smart, Connected***
- **Knowledge Graphs** to integrate heterogeneous and multimodal knowledge, information and data, with attention for provenance and transparency
  - For a variety of users (internal / experts, external experts, DH experts, toolbuilders, data scientists, laypersons)
  - Through query environments, raw-data access, purposeful tools
- **Re-use** and **re-usability**
- Logical **Reasoning** and Statistical **Learning** to both enrich and analyse the KGs. Including simple methods, deep learning,
- Challenges around **polyvocality**, **bias**, **provenance**

Thank you

v.de.boer@vu.nl
@victordeboer
victordeboer.com
cultural-ai.nl