GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Big Data, Digital Humanities und der Mensch als Bremsklotz?!

Dr. Karsten Tolle

Frankfurt Big Data Lab

-understanding and applying technologies for Big Data-

Big Data *slogans*

"***Big Data: The next frontier for innovation, competition, and productivity***"

(McKinsey Global Institute – May 2011)

"***Data is the new gold***"

Open Data Initiative, European Commission (aim at opening up Public Sector Information) - 12th December 2011.

Wissen ist Macht …

- Unter **Daten** verstehen wir (Informatiker) die Bits, Bytes oder Zeichenketten: 00101010101010111010010 ….
- Mit den entsprechenden ***Kontext*** entstehen daraus **Informationen**, z.B. zu verstehen, was Personen in einem Gespräch gesagt haben.
- Zusammen mit weiteren Fakten und Regeln entsteht aus Informationen Wissen. … also z.B. dass jemand mit seinem Gespräch gegen Gesetze verstoßen hat.

Flynn-Affäre

http://www.focus.de/politik/ausland/unerlaubte-kontakte-mit-russland-flynn-affaere-bringt-trump-in-erklaerungsnot_id_6646964.html

“*Data is the new gold* ”

… eigentlich falsch, richtiger:

… (Big) Data ist der Berg, welcher das Gold/Wissen enthält …

Was ist Big Data?

1 megabyte = 1,000,000 =10^6 bytes

1 gigabyte = 10^9 bytes

1 terabyte = 1,000,000,000,000 bytes = 10^{12} bytes

---

1 *petabyte* is 1,000 terabytes (TB) =10^{15}bytes

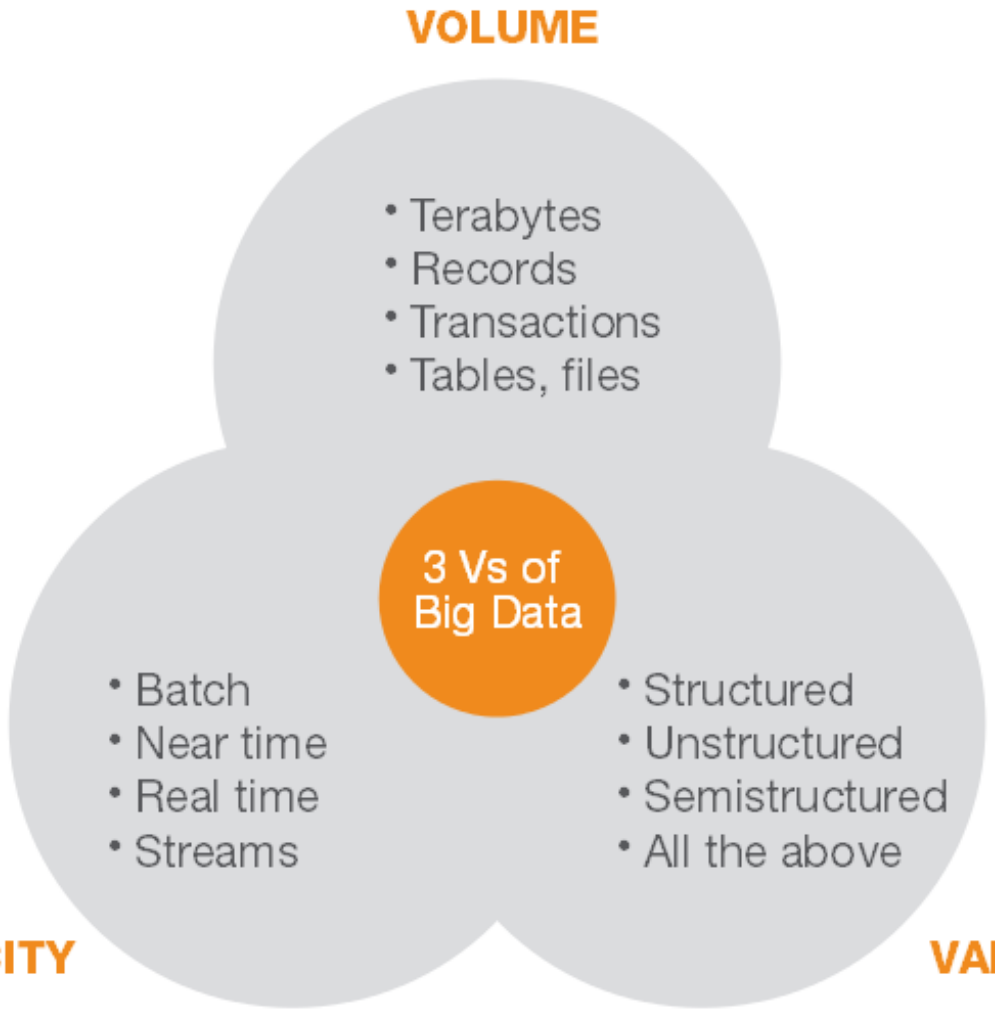
1 exabyte = 10^{18}bytes

1 *zettabyte* is 1,000 000,000,000,000,000,000== 10^{21}bytes

jährlich generierten digitalen Datenmenge weltweit in 2015 geschätzt laut einer EMC-Studie:
~8,6 zettabyte*

*siehe: https://de.statista.com/statistik/daten/studie/267974/umfrage/prognose-zum-weltweit-generierten-datenvolumen/
bzw. https://germany.emc.com/leadership/digital-universe/index.htm

Big Data Characteristics



(Geschwindigkeit) **VELOCITY** **VARIETY** (Vielfalt)

[1] D. Laney, "3D data management: Controlling data volume, velocity and variety," Appl. Deliv. Strateg. File, vol. 949, 2001.

What Is Big Data Question Finally Settled?

http://www.datanami.com/2014/10/29/big-data-question-finally-settled/

- “Data is big when data size becomes part of the problem.”
- “Big data is an **umbrella term** … doing extraordinary things using modern **machine learning techniques** on digital data.”
- “Many features and signals … **would not be detected using smaller samples**. Processing large datasets in this manner was often difficult, time consuming, and error prone before the advent of technologies like **MapReduce and Hadoop** … .”

Eigene Beschreibung …

- … **verschiedene Arten von Daten verbinden**, um Antworten auf eine Frage zu erhalten (beinhaltet also alles bis zur Analyse … Statistik!).
- … durch **große Datenmengen** sollen einzelne Fehler überdeckt werden.
- … zum **Erkennen** seltener aber sich wiederholender/abhängiger Ereignisse benötigt man **große Datenmengen**.

→ neue Verarbeitungstechnologien: Verarbeitung dort wo die Daten liegen

→ neue Visualisierungsmethoden, um große Datenmengen anzuzeigen

Anwendungsbeispiele:

- **Verbrechensbekämpfung**
- **Vorhersage von Krankheitsausbrüchen**
- **Vorhersagen für Wartungen**
- **Konsumentenverhalten/Kunden verstehen**
- **Machteinfluss von großen Datenmengen in der Politik (Wahlprognosen, PolitBarometer, …)**
- …

Wichtige DBMS 2007

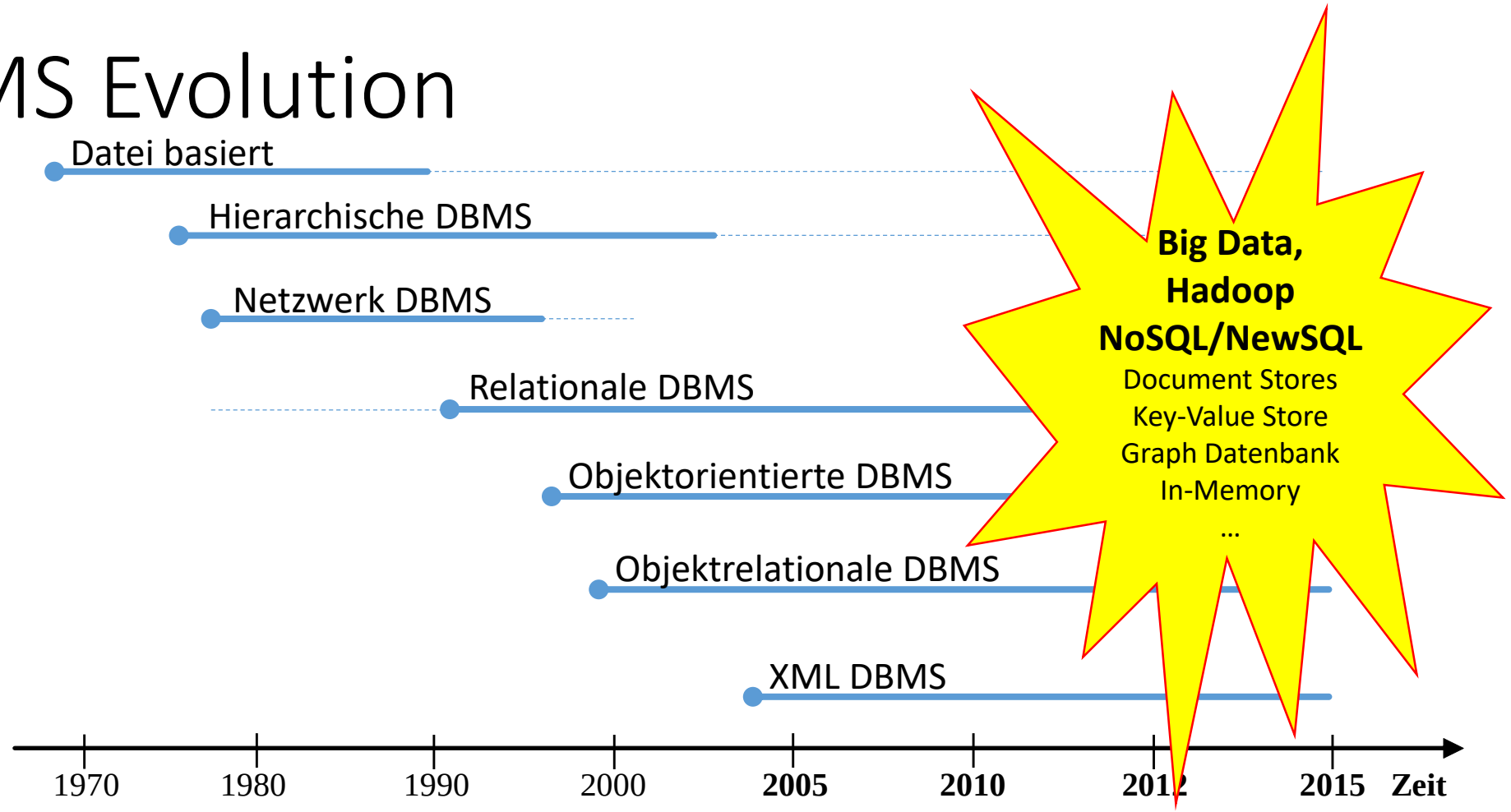
- Oracle
- IBM DB2
- Microsoft SQL Server
- …

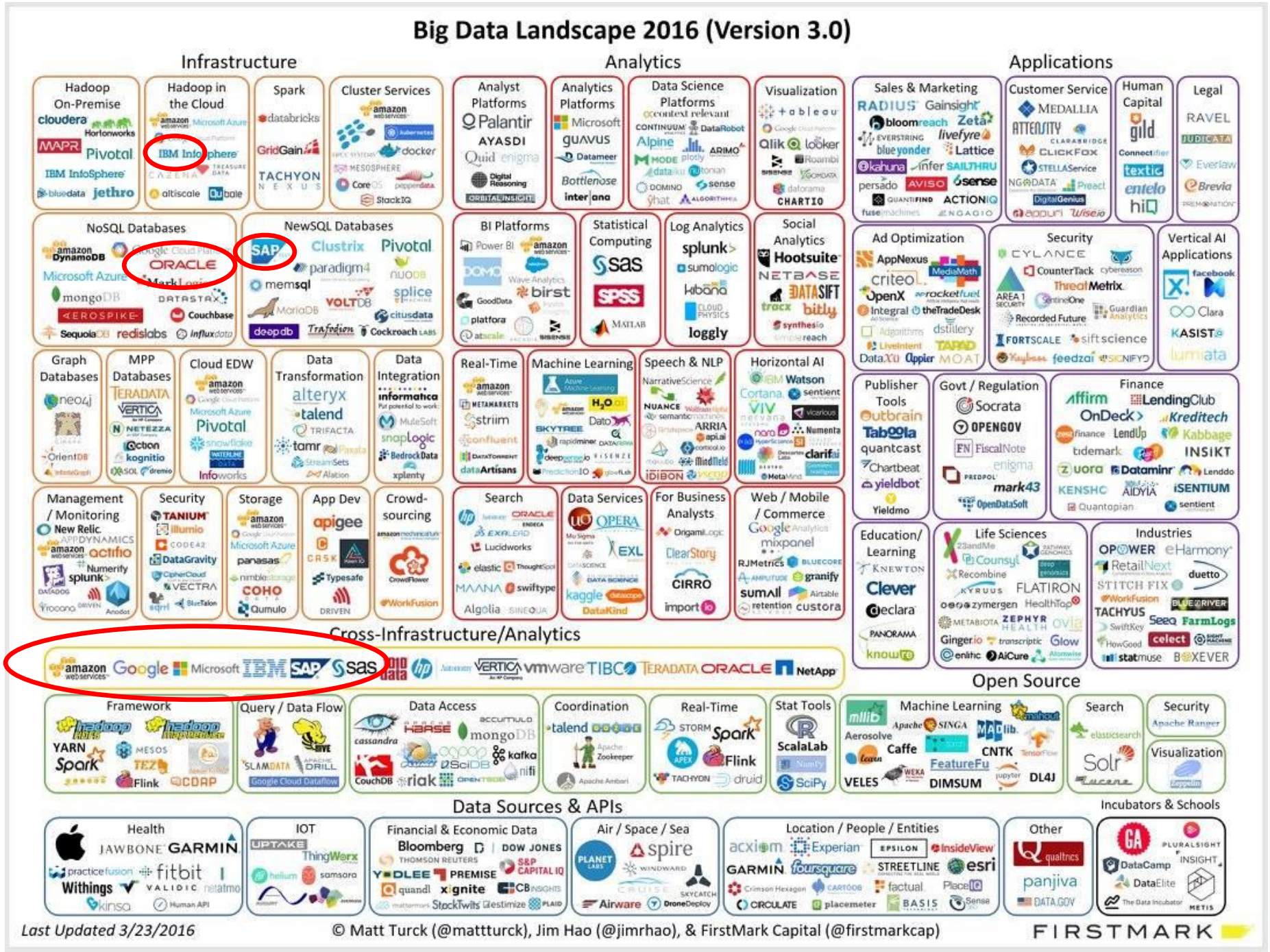
Aber was heißt das eigentlich?

mpany	Revenue 2007	Market share 2007
Oracle	8,343 Mrd. Dollar	37,6%
IBM	4,879 Mrd. Dollar	22,0%
Microsoft	4,670 Mrd. Dollar	21,0%

aus Computerwoche Nr. 3 vom 16. Januar 2009
Zahlen beziehen sich nur auf DBMS-Geschäft

DBMS Evolution





http://mattturck.com/big-data-landscape-2016-v18-final/

Top 10 Big Data Unternehmen …

http://www.information-management.com/gallery/big-data-in-2016-the-10-biggest-big-data-companies-by-revenue-10028947-1.html

1. IBM - $2,104M 9.3% market share
2. SAP - $890M 3.9% market share
3. Oracle - $745M 3.3% market share
4. HPE - $680M 3.0% market share
5. Palantir - $672M 3.0% market share
6. Splunk - $644M 2.8% market share
7. Accenture - $507M 2.2% market share
8. Dell - $489M 2.2% market share
9. Teradata - $432M 1.9% market share
10. Microsoft - $396M 1.8% market share

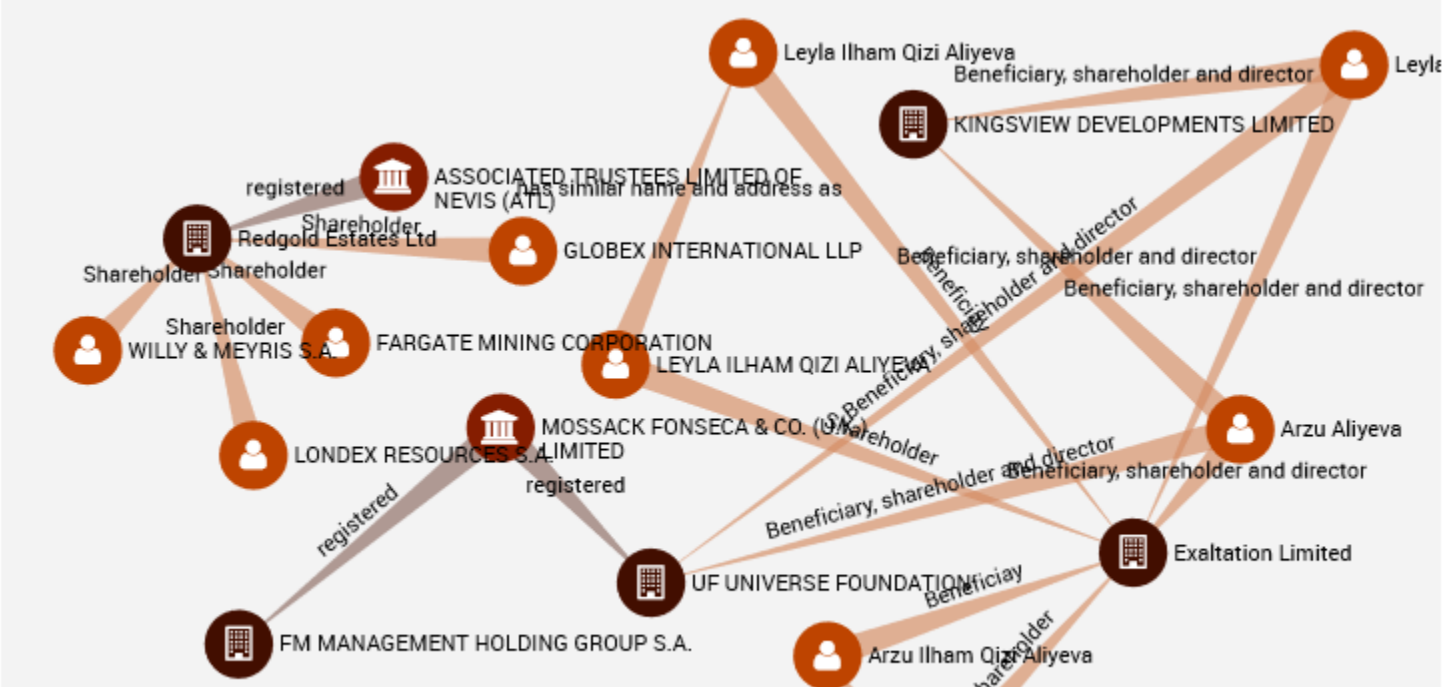
Gründung 2004

Positives Beispiel: Panama Paper

- „ … infolge eines **2,6 Terabyte großen Datenlecks** …“
- „ … rund 11,5 Millionen E-Mails, Briefe, Faxnachrichten, Gründungsurkunden, Kreditverträge, Rechnungen und Bankauszüge als PDF-, Text- sowie Bilddateien aus den Jahren 1977 bis 2016. Ein anonymer Whistleblower hatte sie 2015 zunächst dem deutschen Journalisten Bastian Obermayer von der *Süddeutschen Zeitung* zugespielt. Anschließend koordinierte das International Consortium of Investigative Journalists (ICIJ) die **einjährige Datenauswertung** und weiteren Recherchen … „ (Wikipedia)
  https://de.wikipedia.org/wiki/Panama_Papers

Panama Paper

- … Visualisierung war ein Schlüssel zum Erfolg … Darstellung als Graphen und nicht als Tabellen.
- Nutzung einer Graph-Datenbank (Neo4J): https://neo4j.com/blog/analyzing-panama-papers-neo4j/

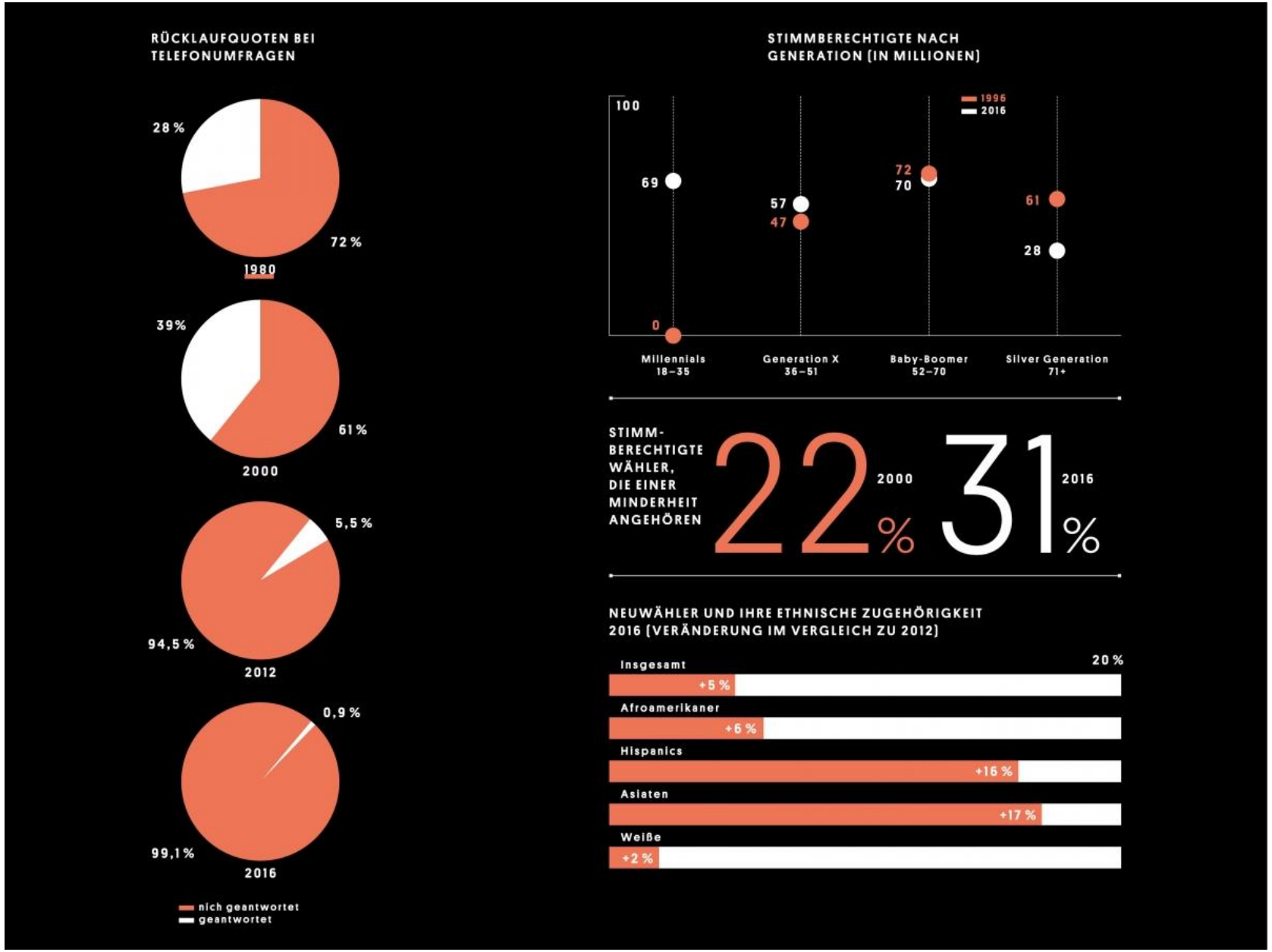


https://www.merkur.de/politik/trump-clinton-letzte-umfragen-us-wahlen-2016-akutell-usa-prognose-zr-6768167.html

- Letzte Zahlen vor der Wahl (von Real Clear Politics):
  Hillary Clinton 46,8 Prozent (vorher: 47,2),
  Donald Trump 43,6 Prozent (vorher: 44,2).

- Die letzten Prognosen für Wahlmänner:
  203 für Hillary Clinton (keine Veränderung),
  164 für Donald Trump (keine Veränderung).

https://www.wired.de/collection/science/wie-big-data-startups-die-prognosen-zur-us-wahl-revolutionieren



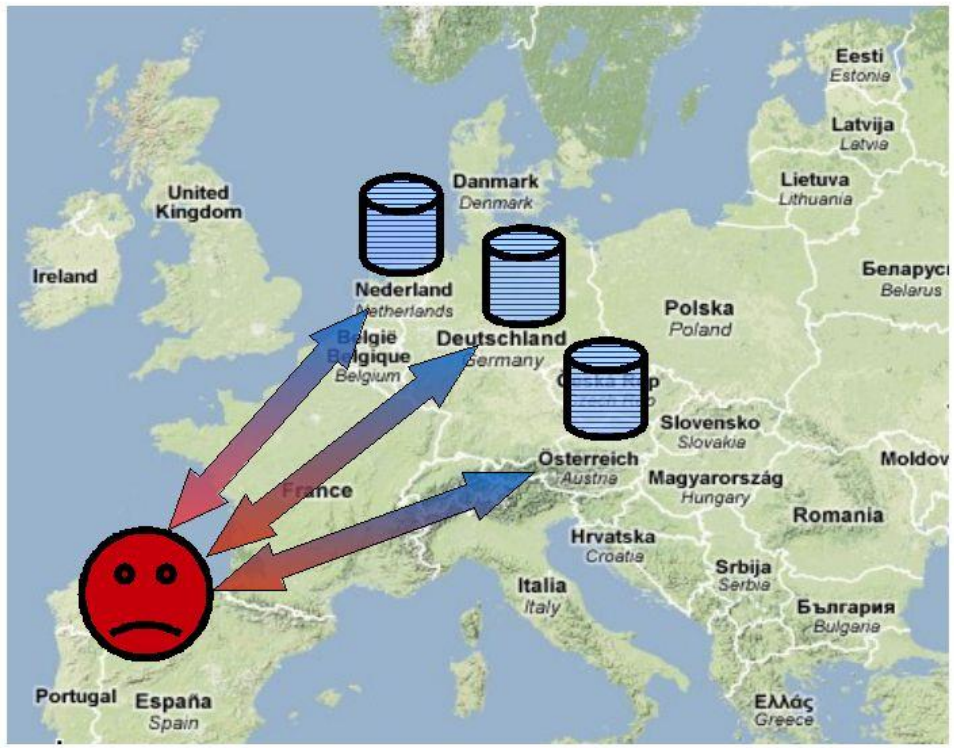
Digital Humanities „digitale Geisteswissenschaften“

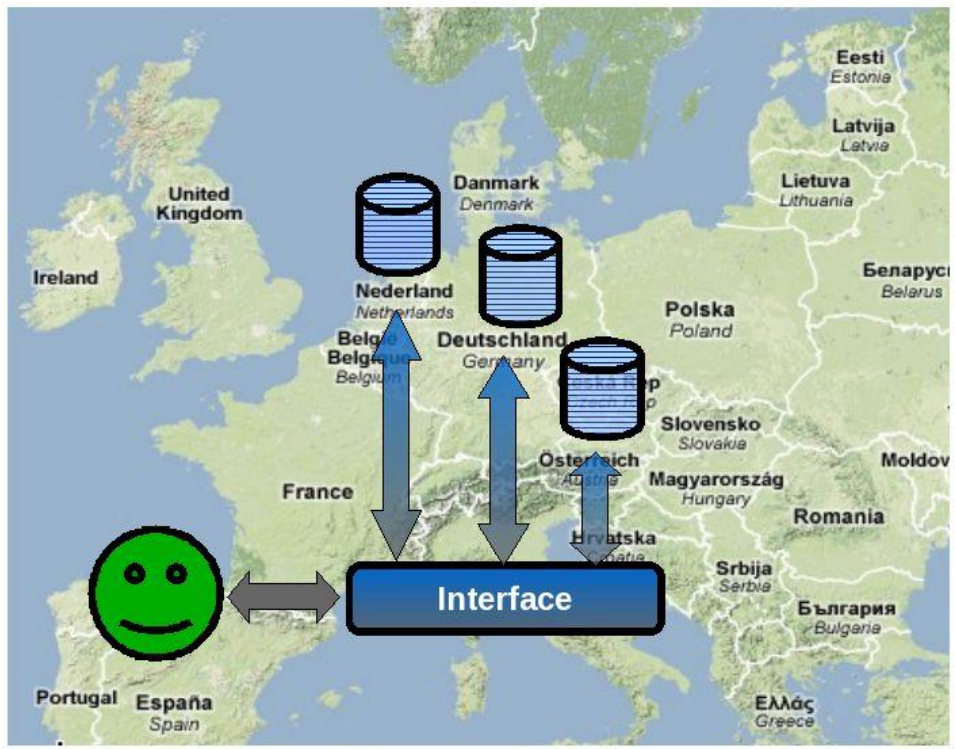
Archäologie → Numismatik „Münzkunde“:
Wo wurden alles solche Münzen gefunden?

Wo wurden alles solche Münzen gefunden?

Vespasian -RIC(2)-Type 777

Wo wurden alles Münzen von Vespasian -RIC(2)-Type 777- gefunden?





reale Welt

logisches Modell

Datenbank

Unterschiede in:

- Logischem Modell
- Datenbankmodell (rel. DB, OODB, …)
- Abbildung zwischen Log. und DB-Modell
- Datenbanksystem (MySQL, DB2, …)
- Sprache (Fachsprache, Deutsch, Englisch, …)
- …

Чайковский
Čajkovskij
Tschaikowsky
Tschaikowski
Tschaïkowsky
Tchaikovsky
Czajkowski
Tsjaïkovskiej

Linked Open Data
(Semantic Web)

★	**Available** on the web (whatever format) *but with an open licence, to be Open Data*
★★	Available as **machine-readable structured data** (e.g. excel instead of image scan of a table)
★★★	as (2) plus **non-proprietary format** (e.g. CSV instead of excel)
★★★★	All the above plus, Use open standards from W3C (**RDF** and SPARQL) to identify things, so that people can point at your stuff
★★★★★	All the above, plus: **Link your data to other people's data to provide context**

By Tim Berners-Lee (2006) online unter: http://www.w3.org/DesignIssues/LinkedData.html

Nomisma.org

nomisma.org

Contributors

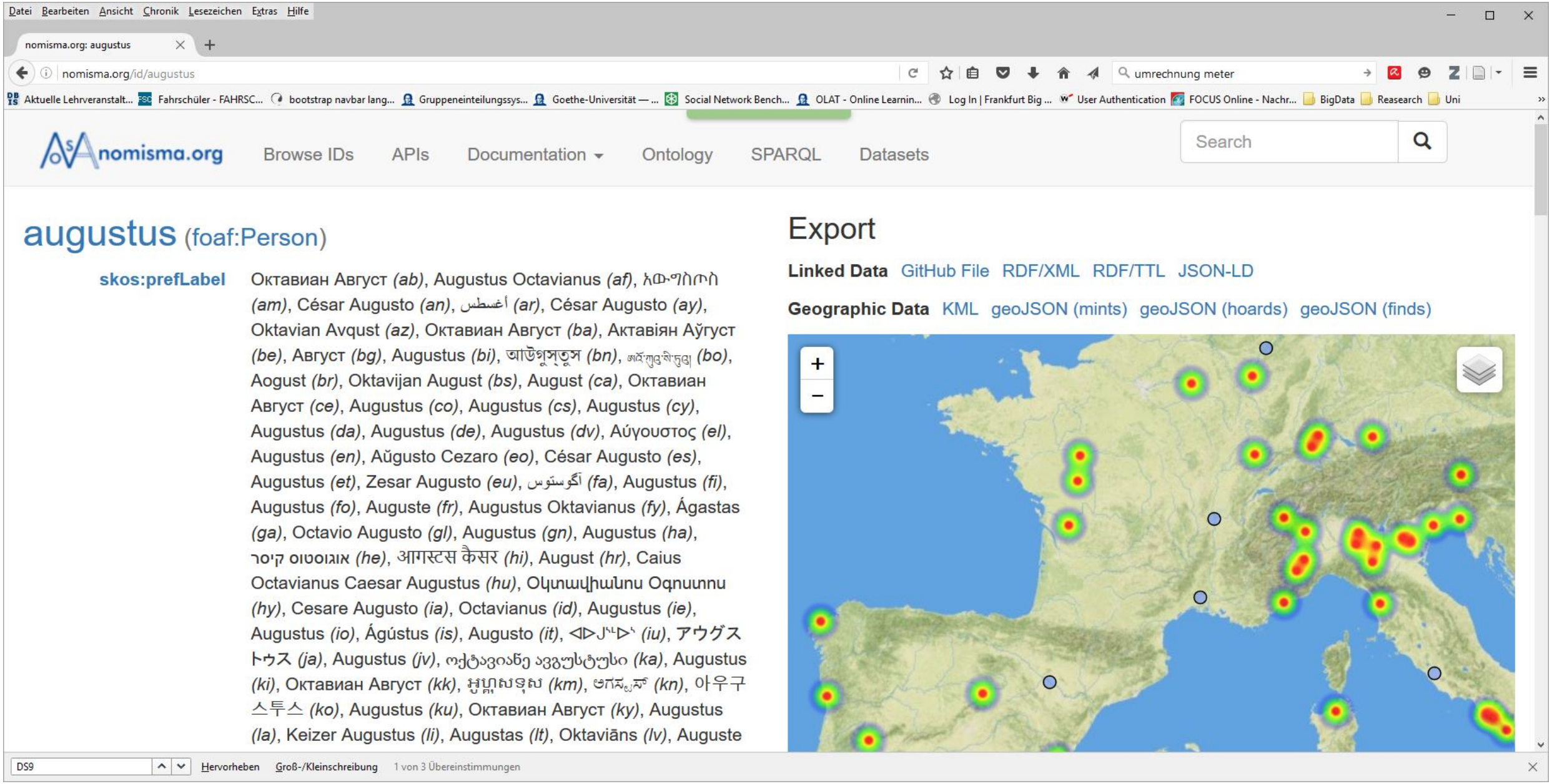
The following institutions have contributed data, specialist advice and/or financial support to the Nomisma project:

- Portable Antiquities Scheme (www.finds.org.uk)
- The British Museum
- Arts & Humanities Research Council
- Münzkabinett, Staatliche Museen zu Berlin
- Goethe Universität Frankfurt am Main
- Alexander von Humboldt Stiftung/Foundation

- Ziele:
  - Definition der Domain Konzepte durch verschiedene Experten
  - Bereitstellung der Konzepte in maschinenverständlicher Form

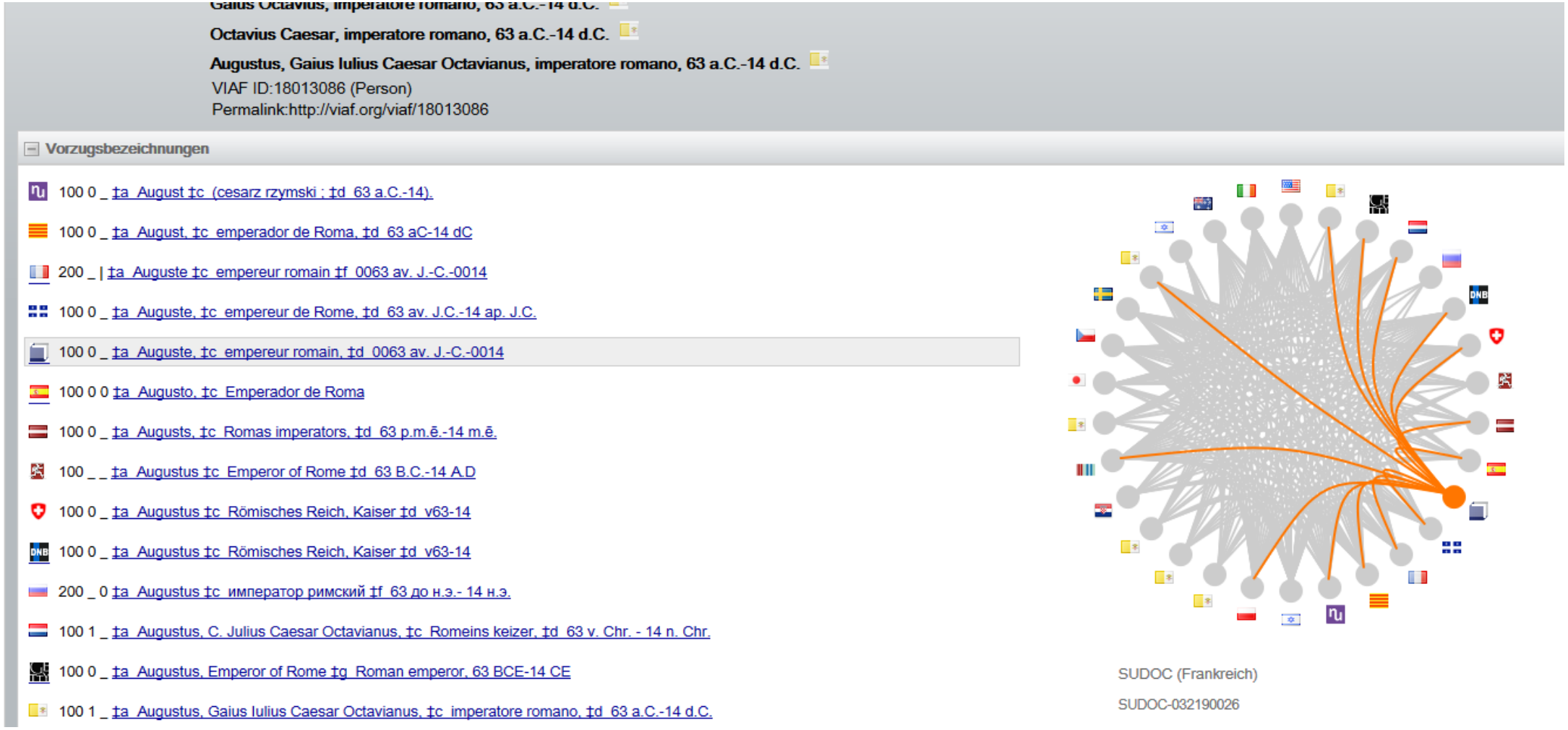
Start ~ 2011!!!!

http://nomisma.org/id/augustus



VIAF - **Virtual International Authority File**

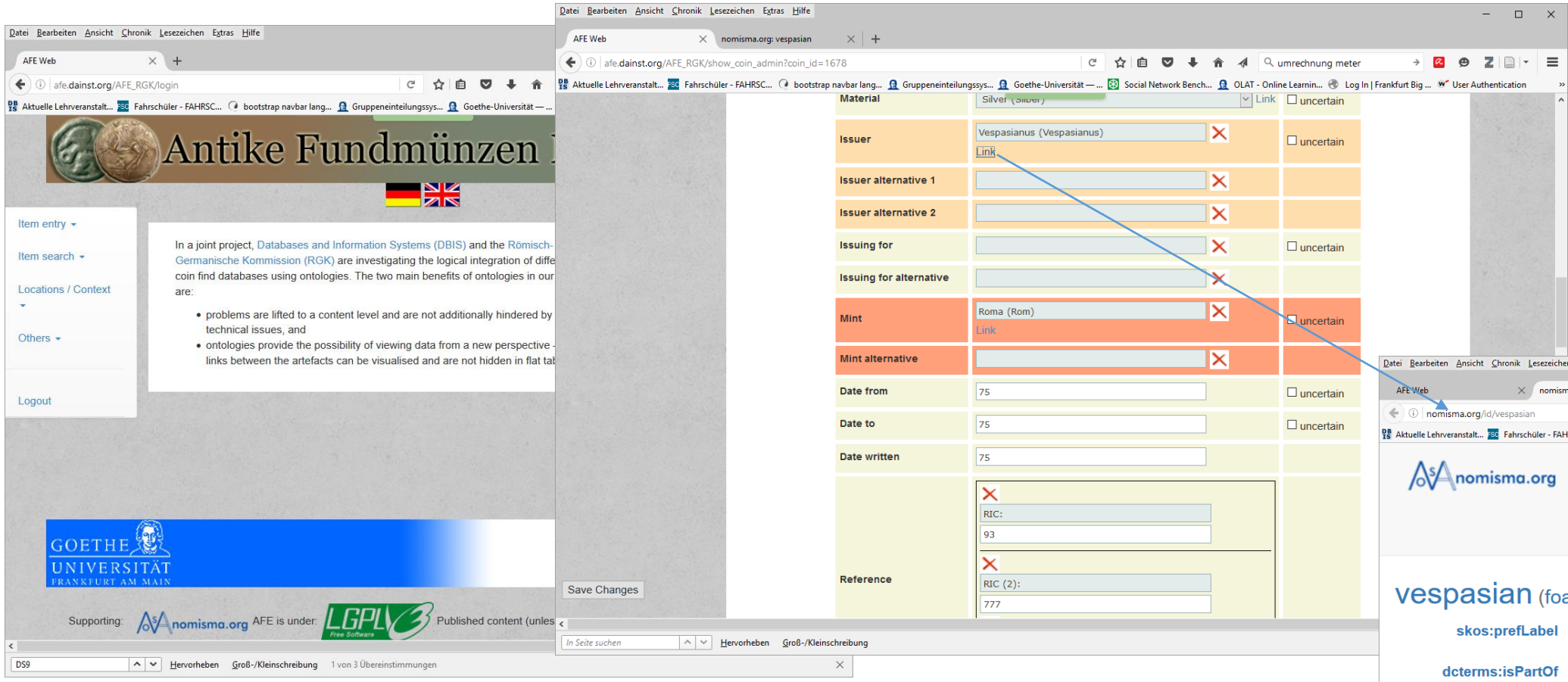
http://viaf.org/viaf/18013086/



Nutzung im eigenen Projekt …

http://afe.fundmuenzen.eu/

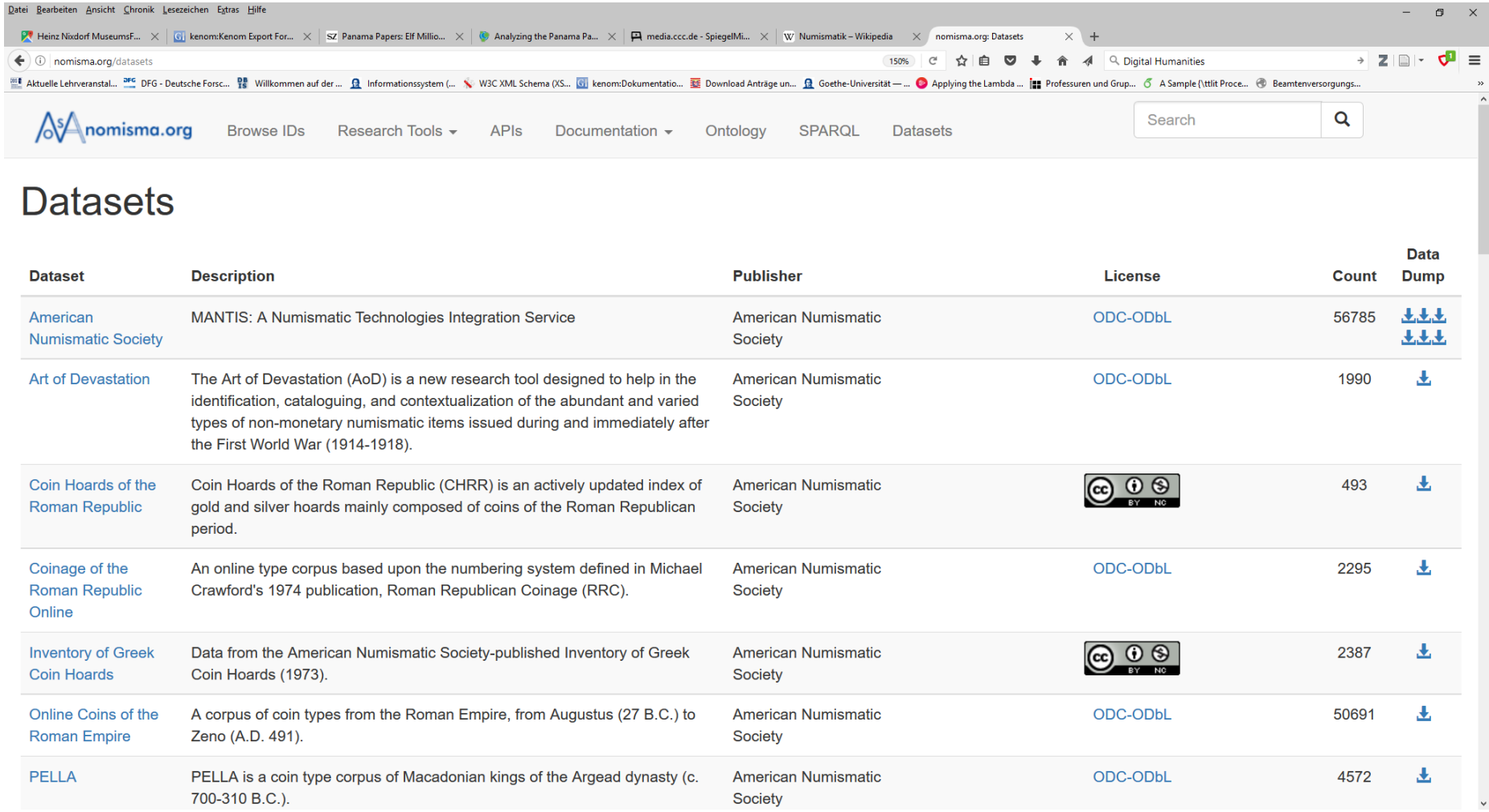
- AFE – Antike Fundmünzen Europa



Datasets unter Nomisma.org

http://nomisma.org/datasets

- Datasets von 19 Institutionen
- fast 220.000 Münzen



nomisma.org Browse IDs Research Tools APIs Documentation Ontology SPARQL Datasets Search

Datasets

Dataset	Description	Publisher	License	Count	Data Dump
American Numismatic Society	MANTIS: A Numismatic Technologies Integration Service	American Numismatic Society	ODC-ODbL	56785	
Art of Devastation	The Art of Devastation (AoD) is a new research tool designed to help in the identification, cataloguing, and contextualization of the abundant and varied types of non-monetary numismatic items issued during and immediately after the First World War (1914-1918).	American Numismatic Society	ODC-ODbL	1990	
Coin Hoards of the Roman Republic	Coin Hoards of the Roman Republic (CHRR) is an actively updated index of gold and silver hoards mainly composed of coins of the Roman Republican period.	American Numismatic Society		493	
Coinage of the Roman Republic Online	An online type corpus based upon the numbering system defined in Michael Crawford's 1974 publication, Roman Republican Coinage (RRC).	American Numismatic Society	ODC-ODbL	2295	
Inventory of Greek Coin Hoards	Data from the American Numismatic Society-published Inventory of Greek Coin Hoards (1973).	American Numismatic Society		2387	
Online Coins of the Roman Empire	A corpus of coin types from the Roman Empire, from Augustus (27 B.C.) to Zeno (A.D. 491).	American Numismatic Society	ODC-ODbL	50691	
PELLA	PELLA is a coin type corpus of Macadonian kings of the Argead dynasty (c. 700-310 B.C.).	American Numismatic Society	ODC-ODbL	4572	

Datenexplosion …

Big Data

WIKIPEDIA

DBpedia

WIKIDATA

~17.700.000 Data Items

The British Museum

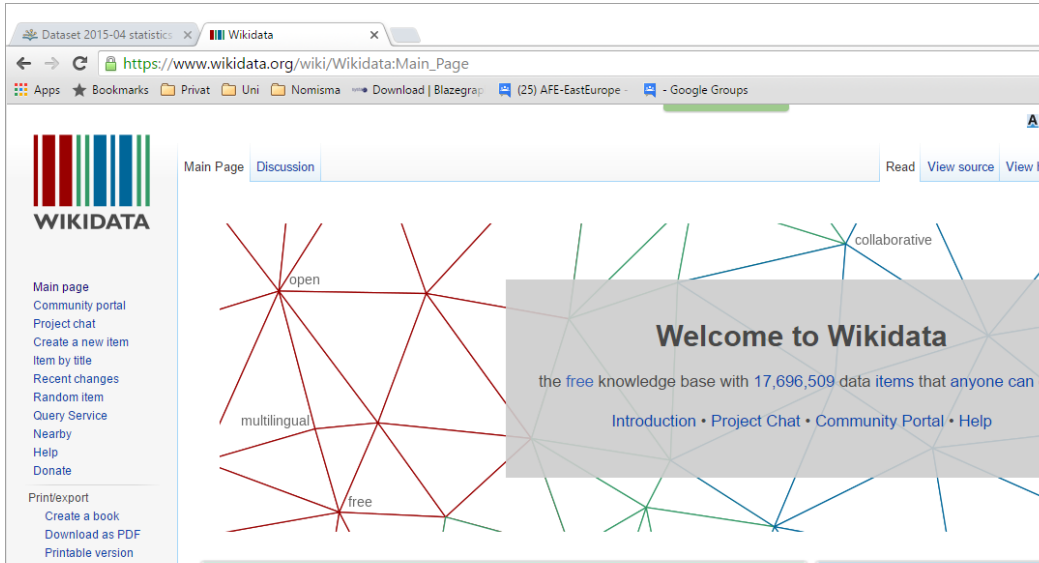
Arts & Humanities Research Council

Münzkabinett Staatliche Museen zu Berlin

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Alexander von Humboldt Stiftung/Foundation

~220.000 Münzen



Herausforderungen und Ziele

- Wie kann man die Nachhaltigkeit gewährleisten/unterstützen?
- Der Umgang mit Datenqualität: Will ich auch Daten berücksichtigen, denen ich nicht traue? → regelbasierte Ansätze, so dass Fehler autom. erkannt werden.
- Viele Daten enthalten Unsicherheiten (wo hören die Fakten auf, wo fangen die Interpretationen an?). Wie kann dies modelliert werden? … insbesondere Fundmünzen!

… wenn Facebook Gesichter erkennen kann …
… können wir dann auch Münzen erkennen?

Vielversprechend!

Bachelorarbeit unter Verwendung
des OpenCV-Frameworks (http://opencv.org/)

… das wird schwierig!

Big Data im Alltag … vielleicht sollten wir Bremsklötze sein?

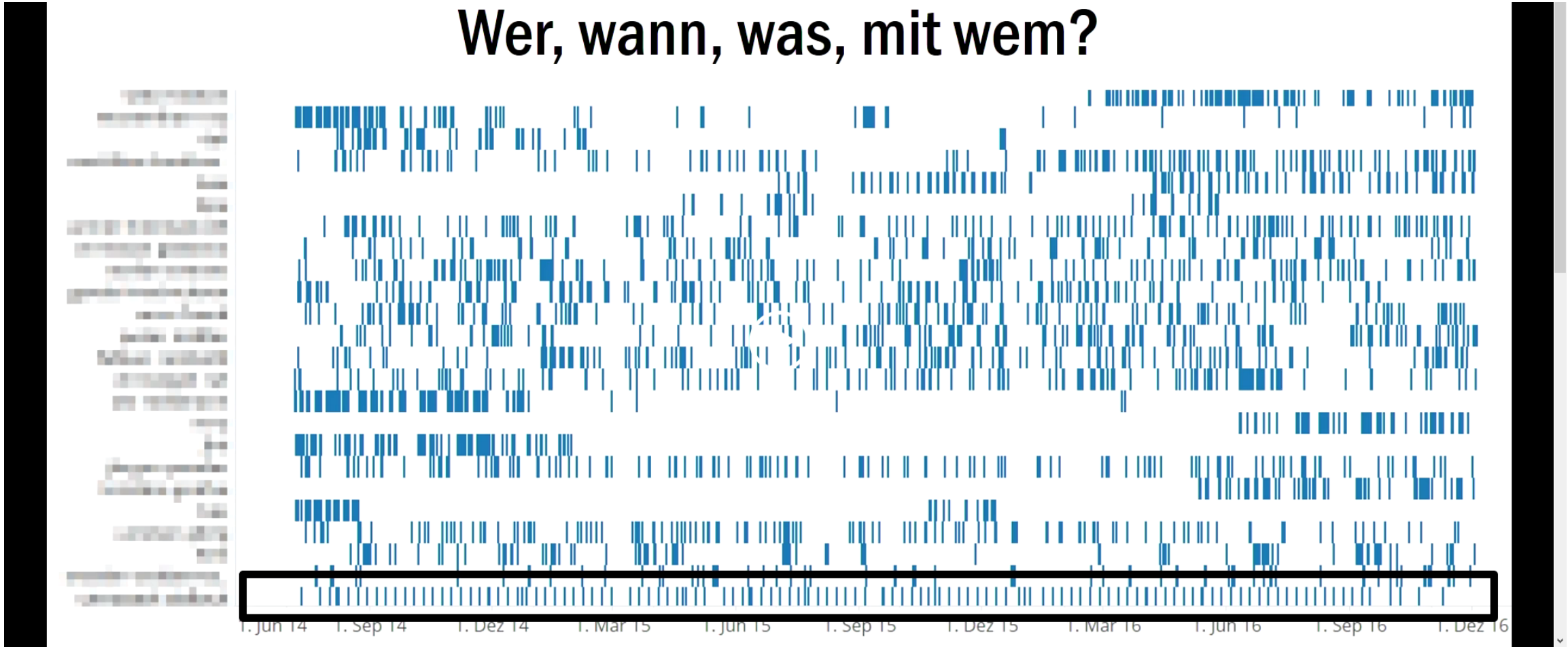
SpiegelMining – Reverse Engineering von Spiegel-Online

Wer denkt, Vorratsdatenspeicherungen und „Big Data“ sind harmlos, der kriegt hier eine Demo an Spiegel-Online.

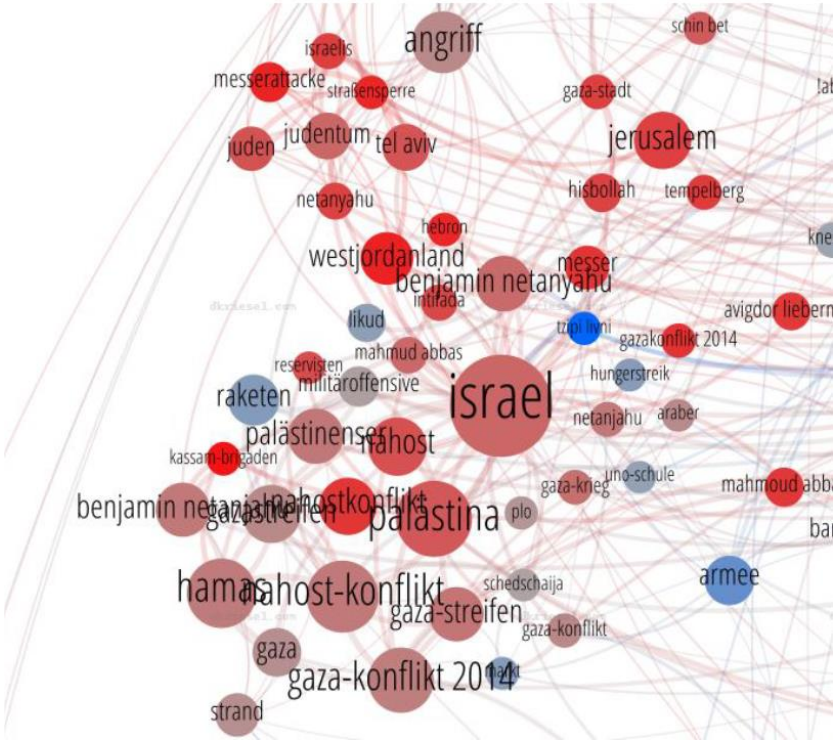
von David Kriesel

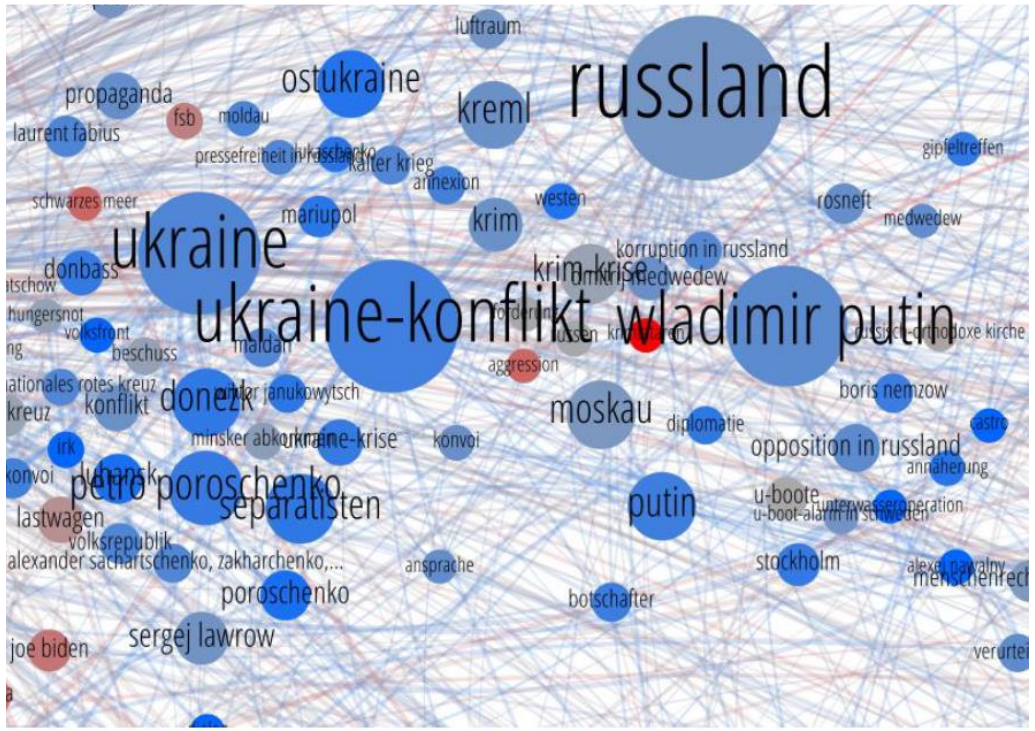
https://media.ccc.de/v/33c3-7912-spiegelmining_reverse_engineering_von_spiegel-online#video&t=1857

SpiegelMining – von David Kriesel



Kommentierbatkeit nach Themen von David Kriesel





Werden wir (und vor allem unsere Kinder) richtig ausgebildet mit neuen Entwicklungen wie Big Data umzugehen?

- Informatik in der Schule …?
- Informatik in den Geisteswissenschaften …?
- Lebenslanges Lernen auch nach Schule und Studium …?