

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: 001010101010111010010 ...
- 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

“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

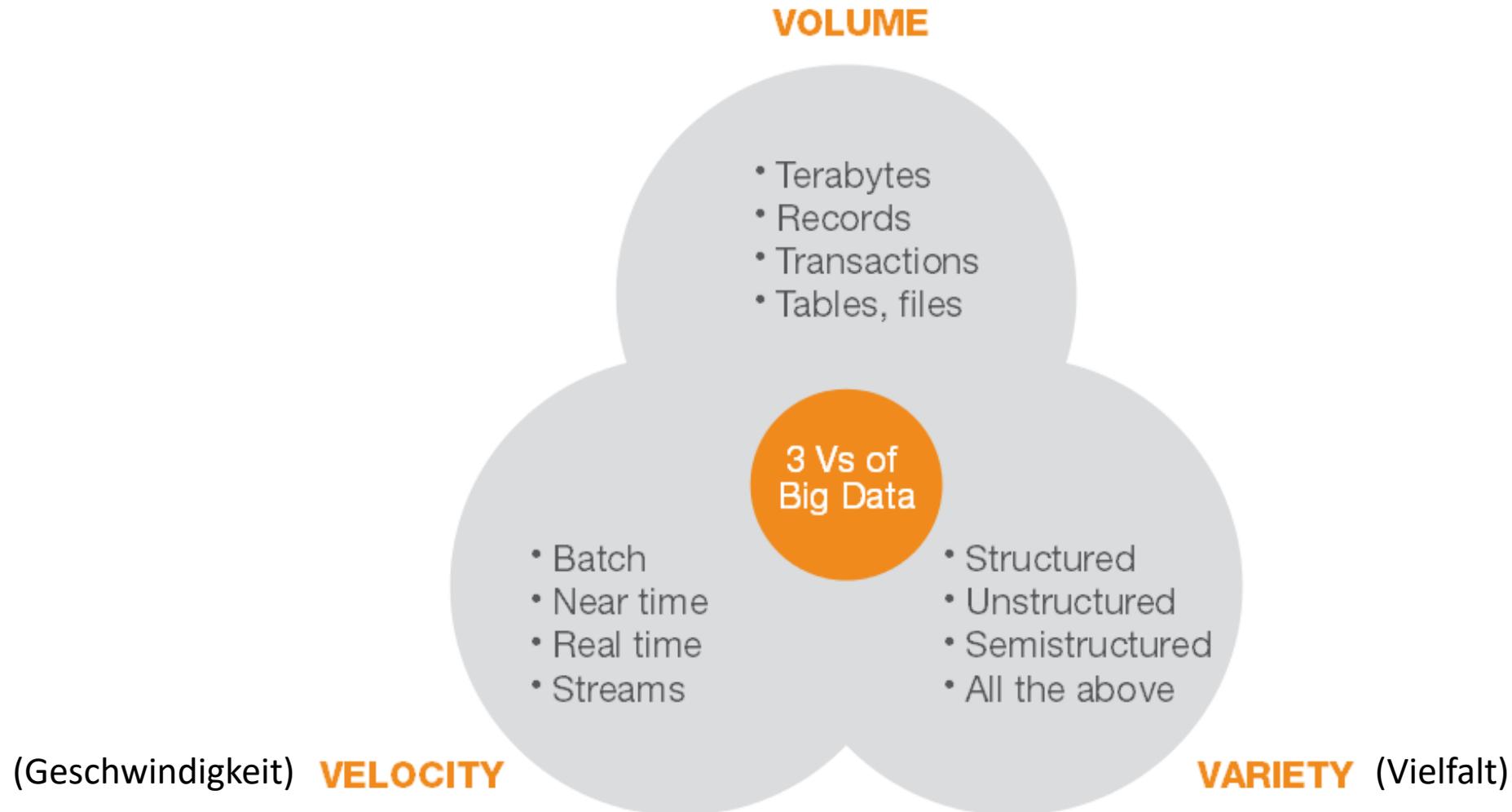
1 exabyte = 10^{18} bytes

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



[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

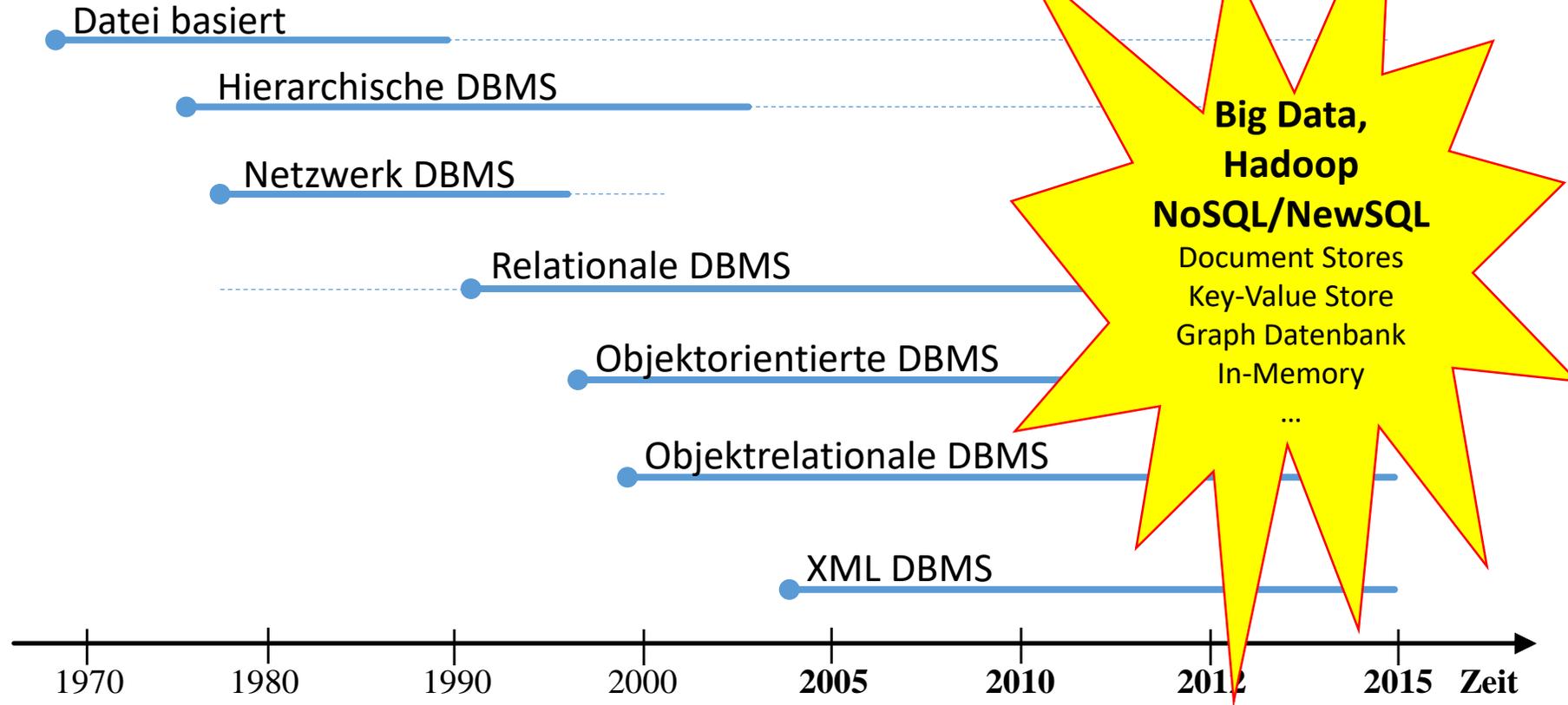
- Oracle
- IBM DB2
- Microsoft SQL Server
- ...

Aber was heißt das eigentlich?

Company	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



Big Data Landscape 2016 (Version 3.0)



<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	} Gründung 2004
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	

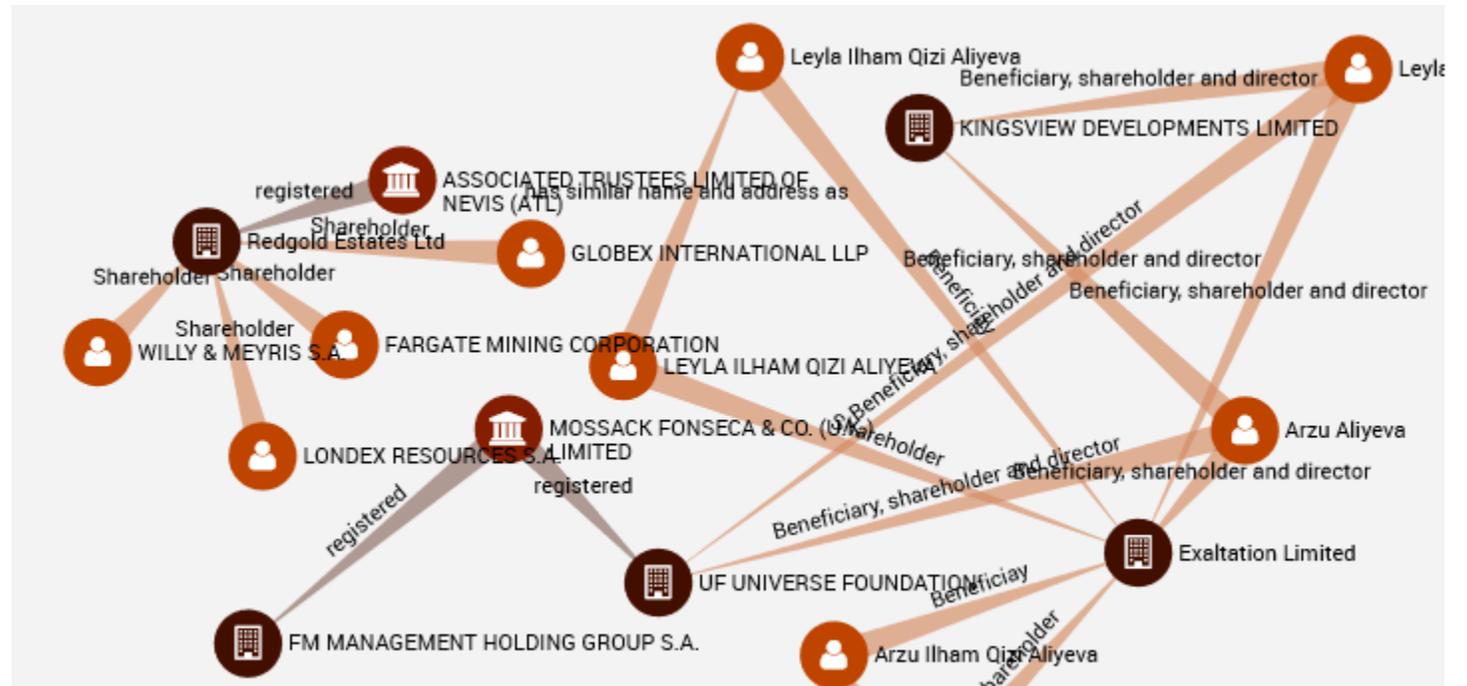
Positives Beispiel: Panama Paper

- „ ... 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 anonymes 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

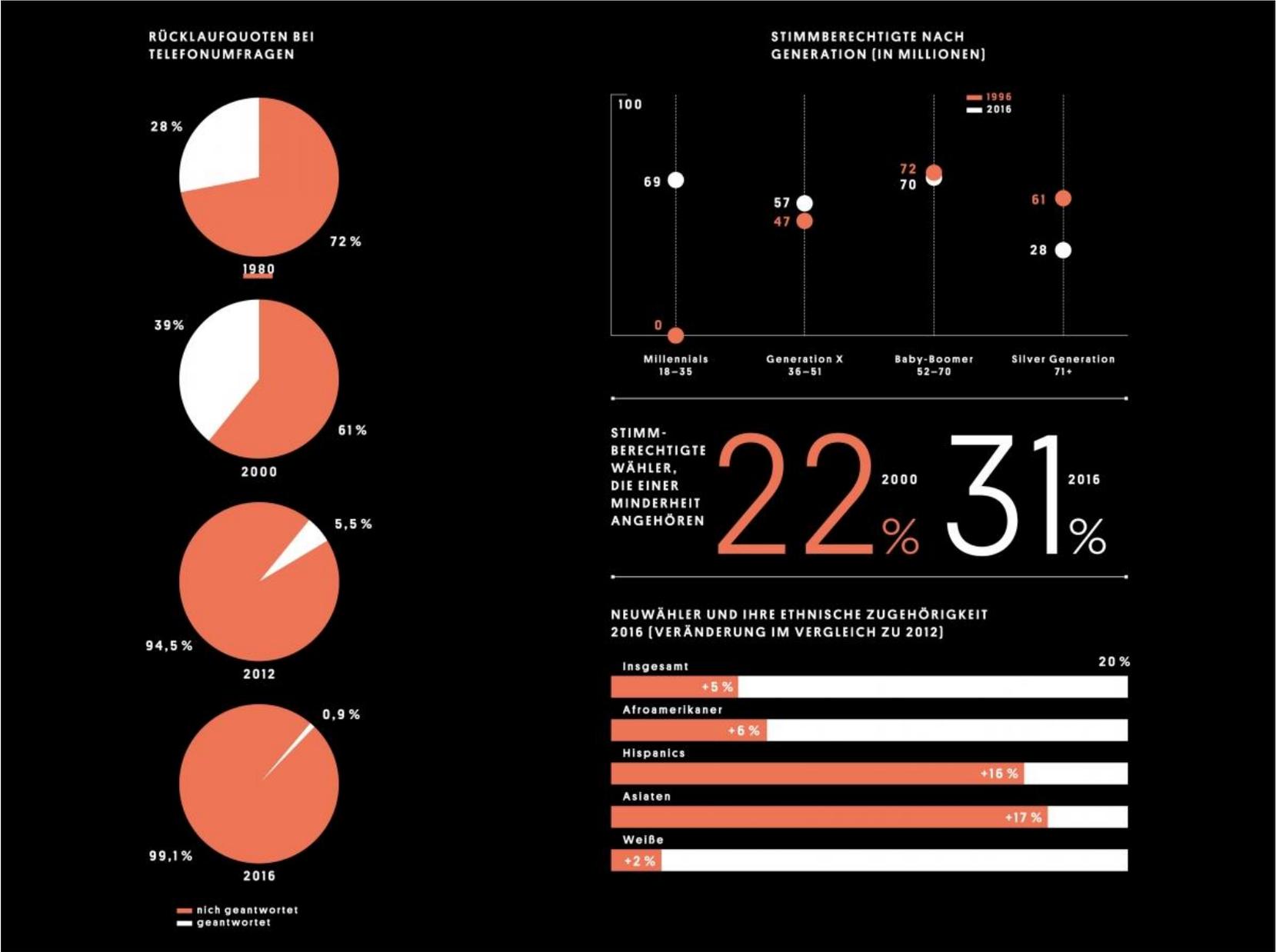
Panama Paper

- ... 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).



Digital Humanities „digitale Geisteswissenschaften“

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

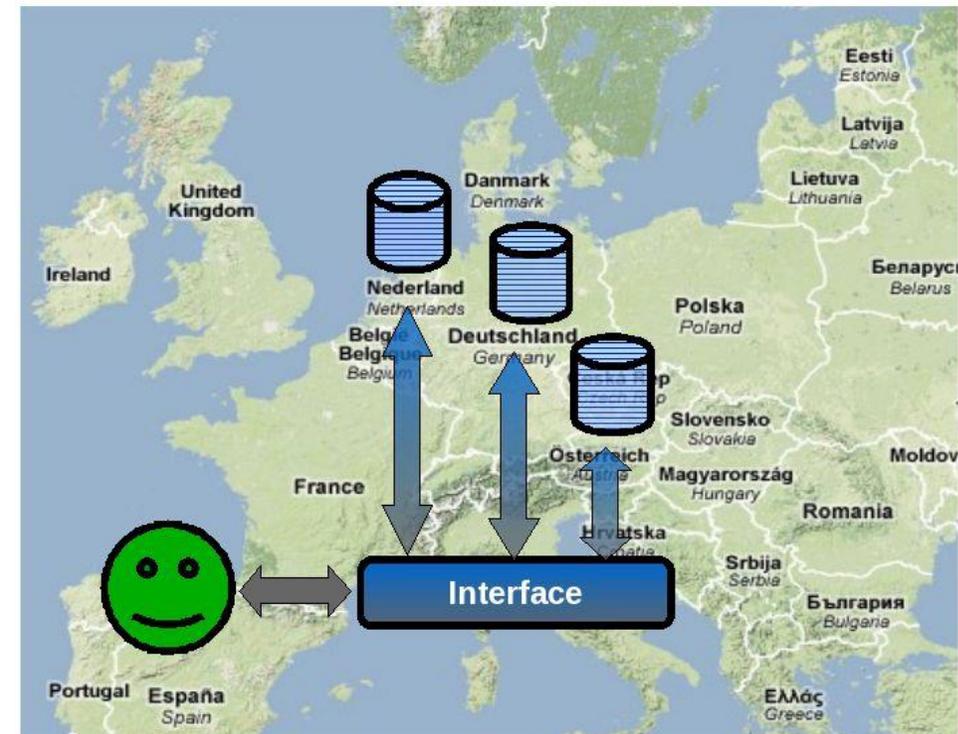
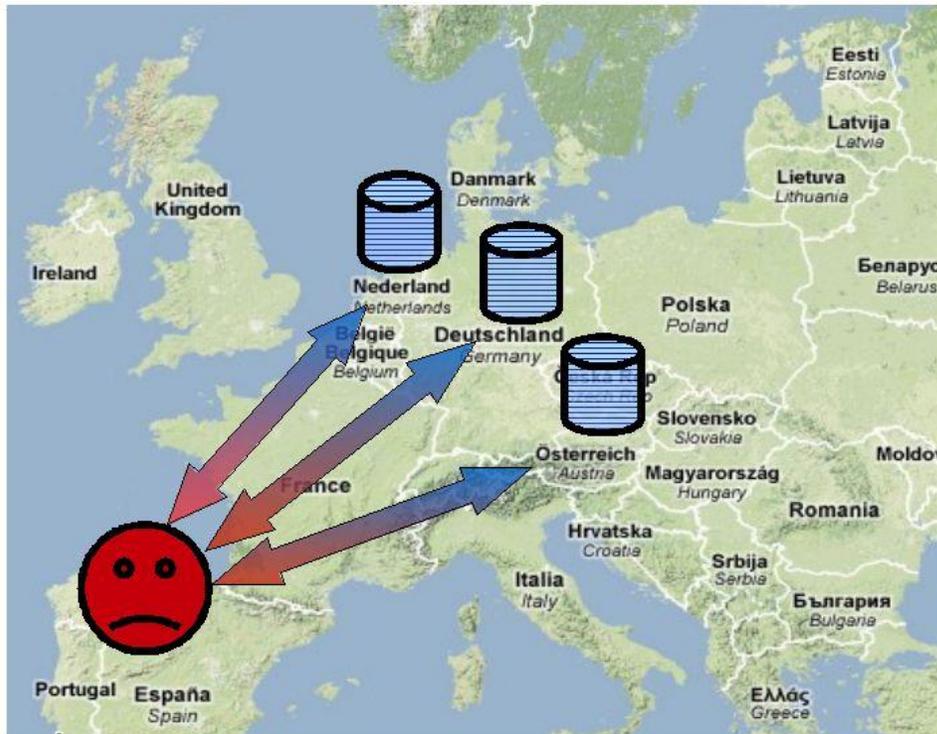


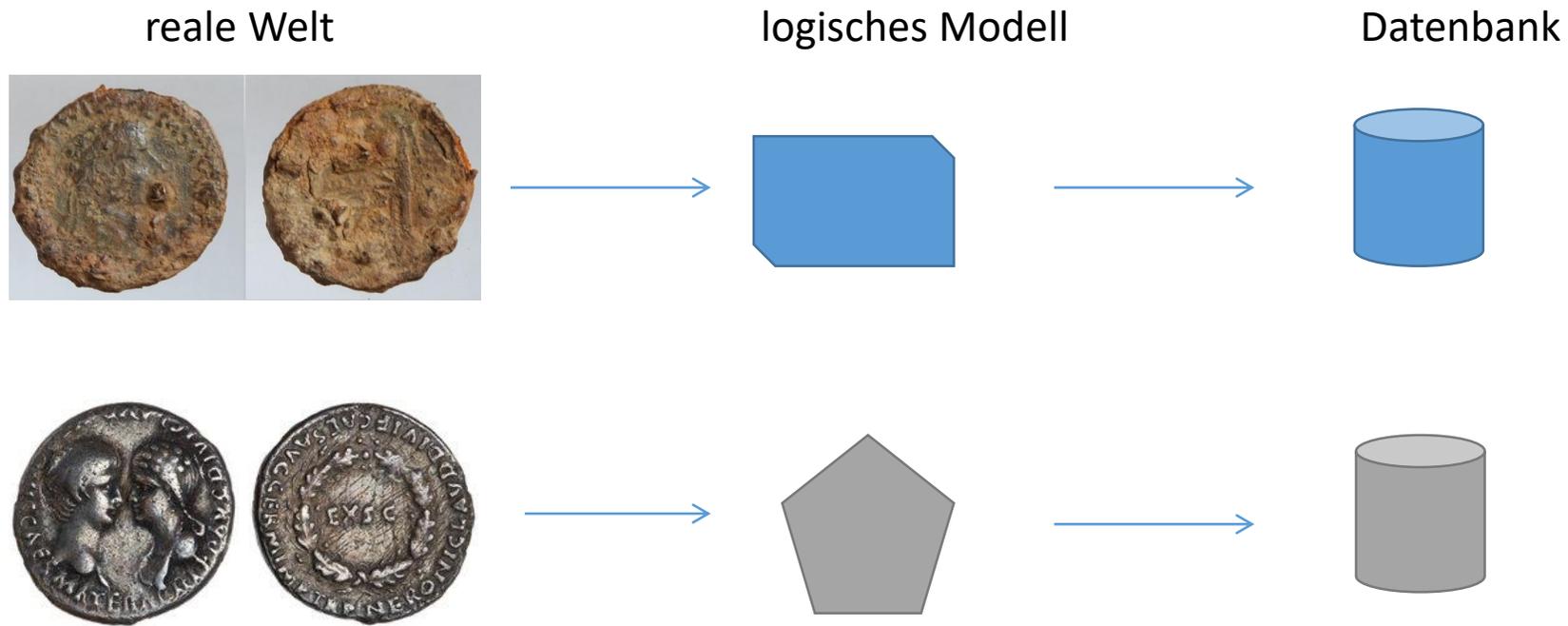
Wo wurden alles solche Münzen gefunden?



Vespasian -RIC(2)-Type 777

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





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**

Nomisma.org



Contributors

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



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!!!!

nomisma.org: augustus

nomisma.org/id/augustus

umrechnung meter

Aktuelle Lehrveranstalt... Fahrschüler - FAHRSC... bootstrap navbar lang... Gruppeneinteilungssys... Goethe-Universität ... Social Network Bench... OLAT - Online Learnin... Log In | Frankfurt Big ... User Authentication FOCUS Online - Nachr... BigData Reasearch Uni

nomisma.org Browse IDs APIs Documentation Ontology SPARQL Datasets Search

augustus (foaf:Person)

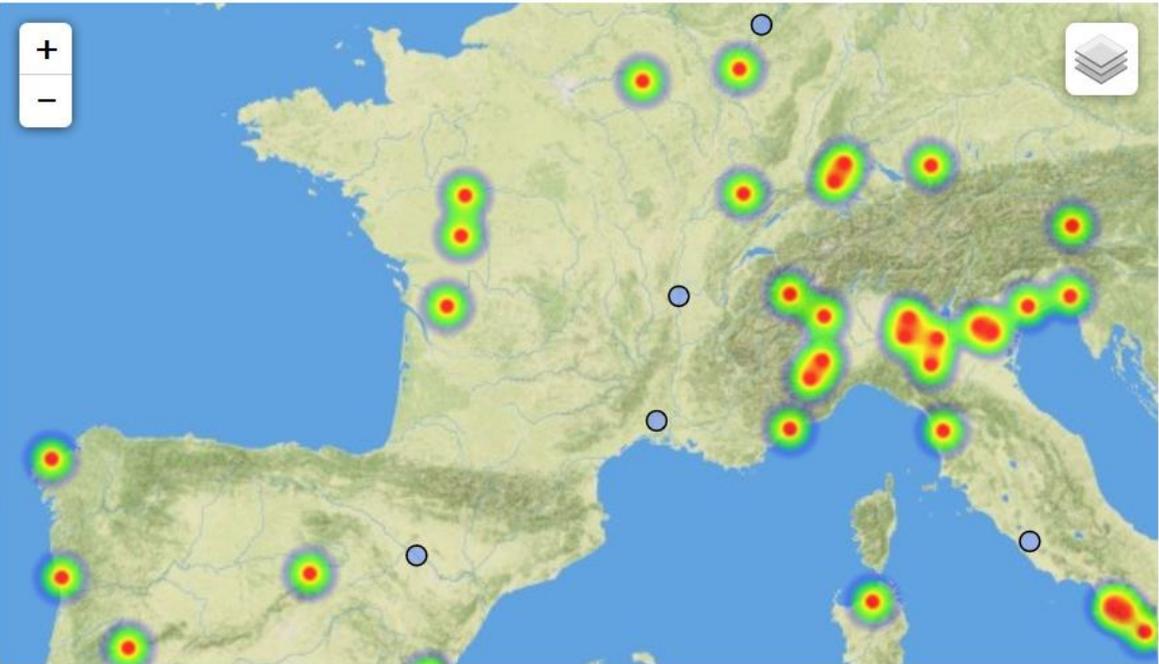
skos:prefLabel **skos:altLabel**

Октавиан Август (*ab*), Augustus Octavianus (*af*), አውግስጥስ (*am*), César Augusto (*an*), أغسطس (*ar*), César Augusto (*ay*), Oktavian Avqust (*az*), Октавиан Август (*ba*), Актавіян Аўгуст (*be*), Август (*bg*), Augustus (*bi*), আউগুস্তুস (*bn*), အာဂူစတုဇ် (*bo*), Aogust (*br*), Oktavijan August (*bs*), August (*ca*), Октавиан Август (*ce*), Augustus (*co*), Augustus (*cs*), Augustus (*cy*), Augustus (*da*), Augustus (*de*), Augustus (*dv*), Αύγουστος (*el*), Augustus (*en*), Aūgusto Cezaro (*eo*), César Augusto (*es*), Augustus (*et*), Zesar Augusto (*eu*), آگوستوس (*fa*), Augustus (*fi*), Augustus (*fo*), Auguste (*fr*), Augustus Oktavianus (*fy*), Ágastas (*ga*), Octavio Augusto (*gl*), Augustus (*gn*), Augustus (*ha*), אָקטאַװיאַן אָװגוסטוס (*he*), आगस्टस कैसर (*hi*), August (*hr*), Caius Octavianus Caesar Augustus (*hu*), Οκταβιανῆς Αὐγούστου (*hy*), Cesare Augusto (*ia*), Octavianus (*id*), Augustus (*ie*), Augustus (*io*), Ágústus (*is*), Augusto (*it*), אָװגוסטוס (*iu*), आउगुस्तुस (*ja*), Augustus (*ju*), အာဂူစတုဇ် (*ka*), Augustus (*ki*), Октавиан Август (*kk*), အာဂူစတုဇ် (*km*), ఆగస్టుస్ (*kn*), 아우구스투스 (*ko*), Augustus (*ku*), Октавиан Август (*ky*), Augustus (*la*), Keizer Augustus (*li*), Augustas (*lt*), Oktaviāns (*lv*), Auguste

Export

Linked Data [GitHub File](#) [RDF/XML](#) [RDF/TTL](#) [JSON-LD](#)

Geographic Data [KML](#) [geoJSON \(mints\)](#) [geoJSON \(hoards\)](#) [geoJSON \(finds\)](#)



DS9 Hervorheben Groß-/Kleinschreibung 1 von 3 Übereinstimmungen

VIAF - Virtual International Authority File

<http://viaf.org/viaf/18013086/>

Gaius Octavius, imperatore romano, 63 a.C.-14 d.C.

Octavius Caesar, imperatore romano, 63 a.C.-14 d.C.

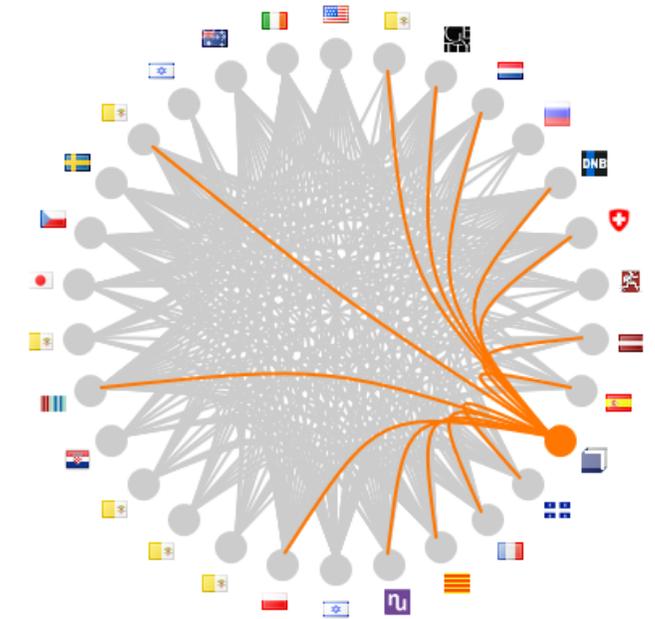
Augustus, Gaius Iulius Caesar Octavianus, imperatore romano, 63 a.C.-14 d.C.

VIAF ID:18013086 (Person)

Permalink:<http://viaf.org/viaf/18013086>

Vorzugsbezeichnungen

-  100 0 _ [ta August tc \(cesarz rzymski : td 63 a.C.-14\).](#)
-  100 0 _ [ta August tc emperador de Roma, td 63 aC-14 dC](#)
-  200 _ | [ta Auguste tc empereur romain tf 0063 av. J.-C.-0014](#)
-  100 0 _ [ta Auguste, tc empereur de Rome, td 63 av. J.C.-14 ap. J.C.](#)
-  100 0 _ [ta Auguste, tc empereur romain, td 0063 av. J.-C.-0014](#)
-  100 0 0 [ta Augusto, tc Emperador de Roma](#)
-  100 0 _ [ta Augusts, tc Romas imperators, td 63 p.m.ē.-14 m.ē.](#)
-  100 __ [ta Augustus tc Emperor of Rome td 63 B.C.-14 A.D](#)
-  100 0 _ [ta Augustus tc Römisches Reich, Kaiser td v63-14](#)
-  100 0 _ [ta Augustus tc Römisches Reich, Kaiser td v63-14](#)
-  200 _ 0 [ta Augustus tc император римский tf 63 до н.э.- 14 н.э.](#)
-  100 1 _ [ta Augustus, C. Julius Caesar Octavianus, tc Romeins keizer, td 63 v. Chr. - 14 n. Chr.](#)
-  100 0 _ [ta Augustus, Emperor of Rome tg Roman emperor, 63 BCE-14 CE](#)
-  100 1 _ [ta Augustus, Gaius Iulius Caesar Octavianus, tc imperatore romano, td 63 a.C.-14 d.C.](#)



SUDOC (Frankreich)

SUDOC-032190026

Nutzung im eigenen Projekt ...

<http://afe.fundmuenzen.eu/>

- AFE – Antike Fundmünzen Europa

The image shows two overlapping browser windows. The left window displays the AFE website's home page, which includes a header with the title 'Antike Fundmünzen' and logos for Germany and the UK. Below the header is a navigation menu with options like 'Item entry', 'Item search', and 'Locations / Context'. A central text block describes the project's goal: 'In a joint project, Databases and Information Systems (DBIS) and the Römisch-Germanische Kommission (RGK) are investigating the logical integration of different coin find databases using ontologies.' It lists two main benefits: problems are lifted to a content level and not additionally hindered by technical issues, and ontologies provide the possibility of viewing data from a new perspective. The footer features the Goethe University Frankfurt logo and mentions support from nomisma.org and AFE under the LGPL license.

The right window shows the AFE data entry interface for a coin. The URL is 'afe.dainst.org/AFE_RGK/show_coin_admin?coin_id=1678'. The interface is a form with various fields for entering coin details. A blue arrow points from the 'Link' field in the 'Issuer' row to the 'nomisma.org/id/vespasian' URL in the browser's address bar. The form fields include:

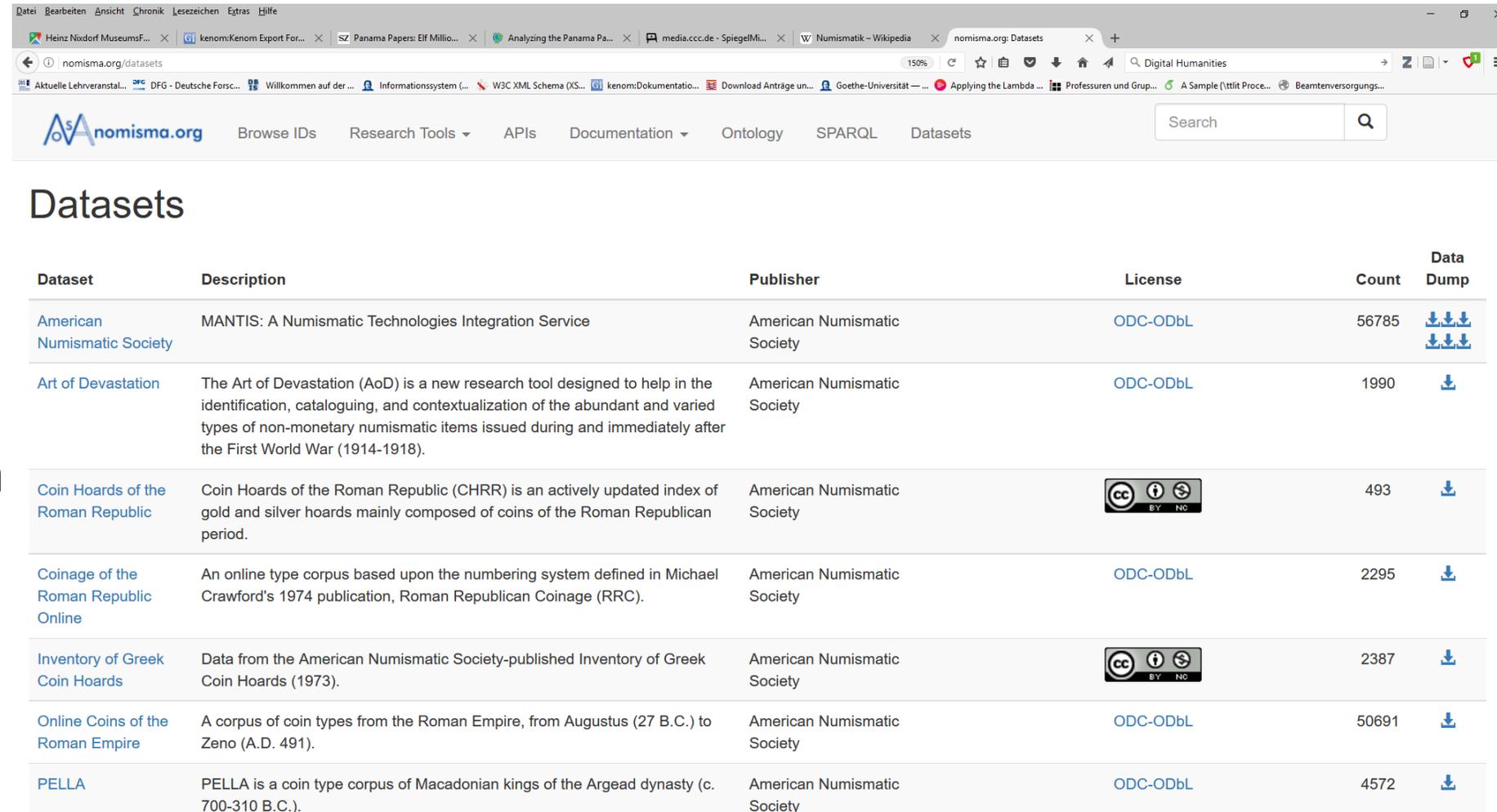
Material	Silver (silver)	Link	<input type="checkbox"/> uncertain
Issuer	Vespasianus (Vespasianus)	Link	<input type="checkbox"/> uncertain
Issuer alternative 1			
Issuer alternative 2			
Issuing for			<input type="checkbox"/> uncertain
Issuing for alternative			
Mint	Roma (Rom)	Link	<input type="checkbox"/> uncertain
Mint alternative			
Date from	75		<input type="checkbox"/> uncertain
Date to	75		<input type="checkbox"/> uncertain
Date written	75		
Reference	<input type="checkbox"/> RIC: 93 <input type="checkbox"/> RIC (2): 777		

At the bottom of the form is a 'Save Changes' button. The browser's address bar shows 'nomisma.org/id/vespasian' and the page title is 'vespasian (foaf:Person)'. The page content includes the nomisma.org logo and navigation links for 'Browse IDs', 'APIs', and 'Documentation'. Below the logo, the text 'vespasian (foaf:Person)' is followed by a list of properties and their values:

- skos:prefLabel**: Веспасиан (bg), Vespasià (ca), Vespasianus Vespasiano (it), Vespasianus (nl), Vespasianus (pt), Vespasianus (ro), Vespasianus (ru), Vespasianus (uk), Vespasianus (de)
- dcterms:isPartOf**: http://nomisma.org/id/roman_numismatics
- dcterms:isPartOf**: http://nomisma.org/id/roman_provincial_numismatics
- org:hasMembership**: http://nomisma.org/id/vespasian#roman_empire
- rdf:type**: skos:Concept

Datasets unter Nomisma.org

<http://nomisma.org/datasets>



The screenshot shows the nomisma.org website with a search bar and navigation menu. The main content is a table titled "Datasets" with the following columns: Dataset, Description, Publisher, License, Count, and Data Dump. The table lists several datasets, including "American Numismatic Society", "Art of Devastation", "Coin Hoards of the Roman Republic", "Coinage of the Roman Republic Online", "Inventory of Greek Coin Hoards", "Online Coins of the Roman Empire", and "PELLA".

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	↓

- Datasets von 19 Institutionen
- fast 220.000 Münzen

Datenexplosion ...

Big Data

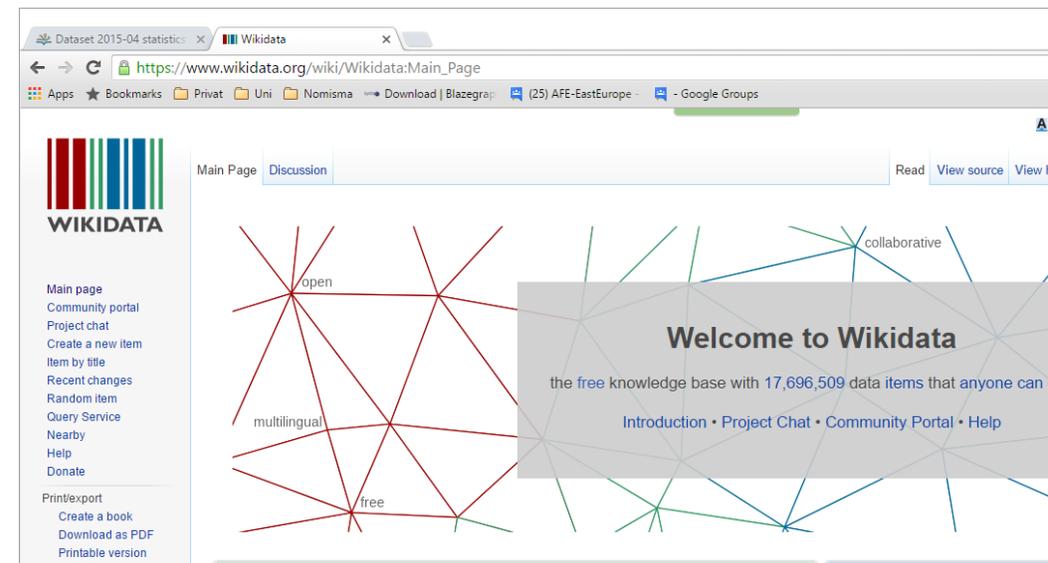


~220.000 Münzen



WIKIDATA

~17.700.000 Data Items

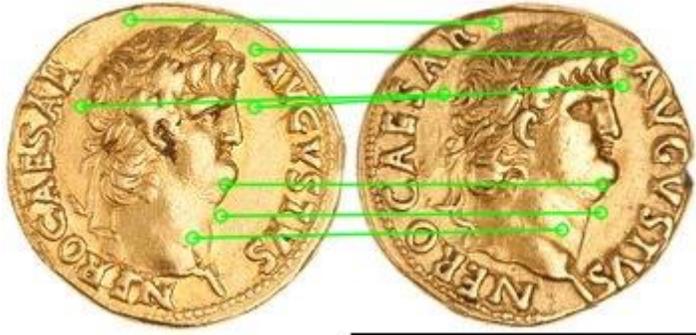


Herausforderungen und Ziele

- 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?



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



Vielversprechend!



... das wird schwierig!

Big Data im Alltag ... vielleicht sollten wir Bremsklötze sein?

SpiegelMining – Reverse Engineering von Spiegel-Online

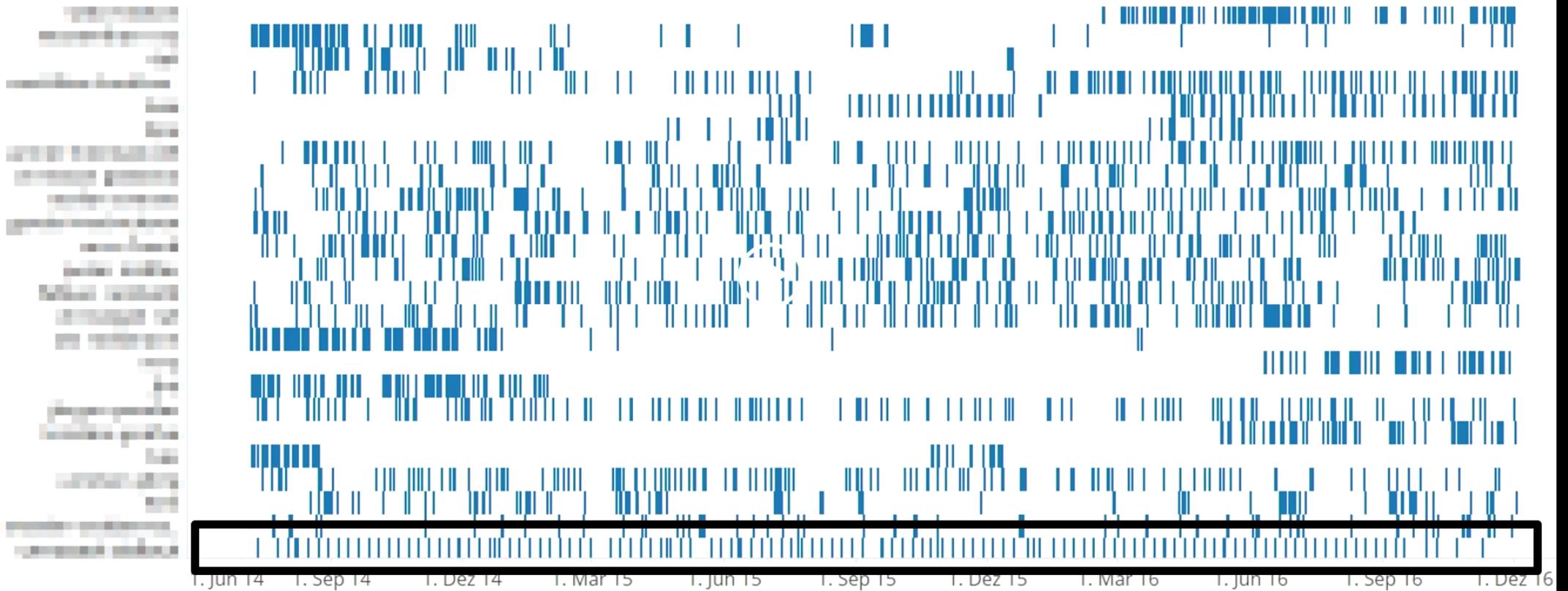
Wer denkt, Vorratsdatenspeicherungen und „Big Data“ sind harmlos, der kriegt hier eine Demo an Spiegel-Online.

von David Kriesel

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

SpiegelMining – von David Kriesel

Wer, wann, was, mit wem?



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 ...?

