



GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Digital Humanities for Computer Scientists ... or: How I became infected with the Indiana Jones virus

Marco Büchler

eTRAP Research Group

Göttingen Centre for Digital Humanities

Institute of Computer Science

Georg August University Göttingen, Germany



Who am I?

- 2001/2 Head of Quality Assurance department in a software company
- 2006 Diploma in Computer Science on big scale co-occurrence analysis
- 2007- Consultant for several SME in IT sector
- 2008 Technical project management of eAQUA project
- 2011 PI and project manager eTRACES project
- 2013 PhD in „Digital Humanities“ on Text Reuse
- 2014- Head of Early Career Research Group eTRAP at Göttingen Centre for Digital Humanities

The Indiana Jones virus?





Overview

- Big Humanities Data
- Filling gaps in inscriptions
- Uncovering unexpected relations
- Text Reuse

Big (Humanities) Data

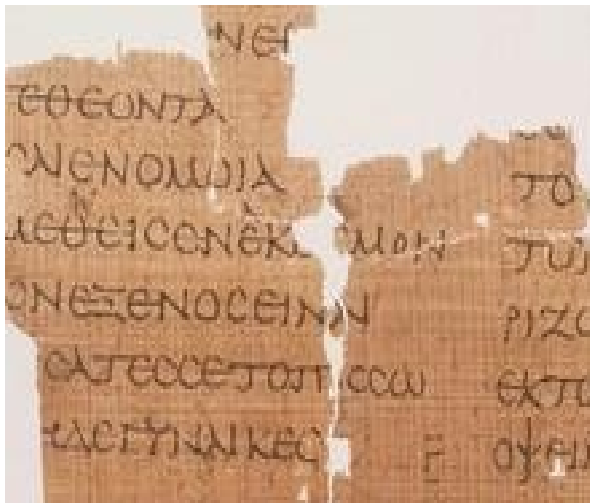
- **3 aspects** (by Ulrike Rieß, Big Data bestimmt die IT-Welt):
 - **Huge amount of data** that can't be processed and analyzed manually
 - **Less structured data**; e. g. in comparison to databases and data warehouse systems
 - Linked data between **heterogeneous and distributed resources**
- The fastest growing sources of Big Data are text and images.
- Researchers easily get lost in the **information overload** (Big Data) and in the **information poverty** (Humanities Data).



GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Filling (Guessing) gaps in inscriptions and Papyri

The data



Textcorrection

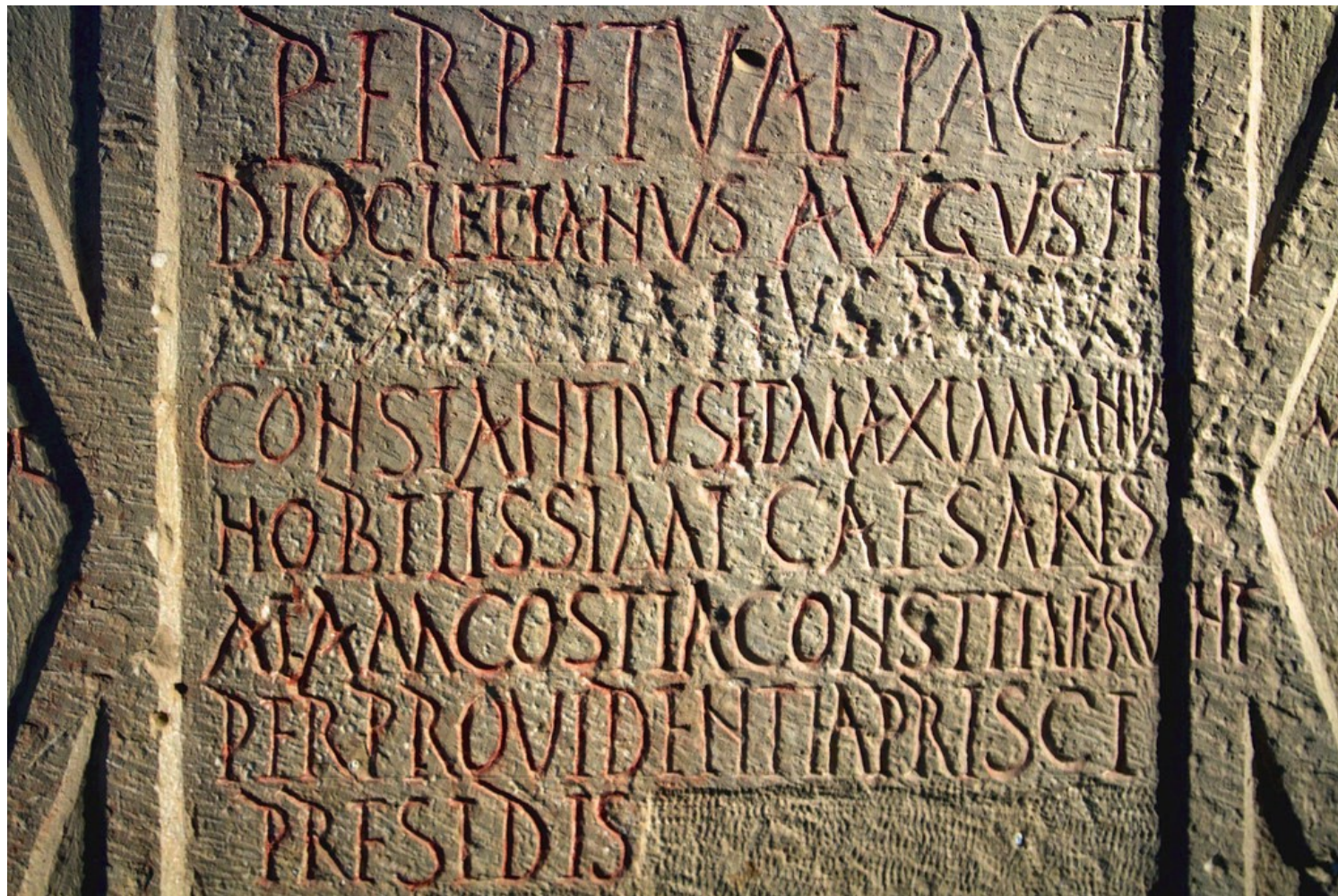
Possible passage in the text:

Platon Timaios, 38c7 bis 38d4 (from TLG-Online):

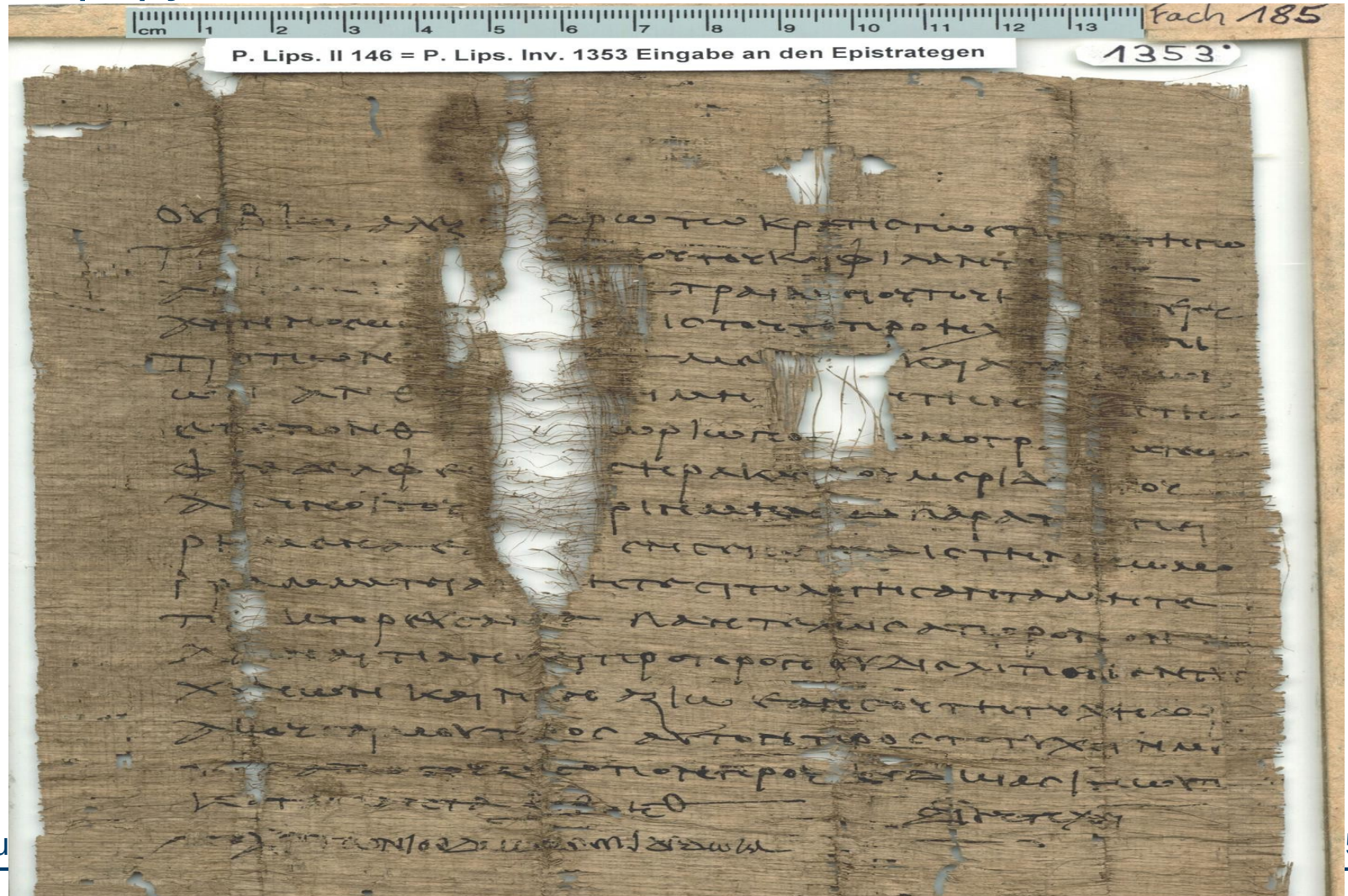
σώματα δὲ αὐτῶν ἐκάστων ποιήσας ὁ θεὸς ἔθηκεν εἰς τὰς
 @1 περιφορὰς ὥς ἡ θατέρου περίοδος ἦεν, ἐπτὰ οὖσας ὄντα
 (d.) ἐπτὰ, σελήνην μὲν εἰς τὸν περὶ γῆν πρῶτον, ἥλιον δὲ εἰς
 τὸν δεύτερον ὑπὲρ γῆς, ἑωσφόρον δὲ καὶ τὸν ἱερὸν Ἑρμοῦ
 λεγόμενον εἰς [τὸν] τάχει μὲν ἰσόδρομον ἡλίῳ κύκλον ἰόντας, τὴν δὲ ἐναντίαν εἰληχότας αὐτῷ δύναμιν·



Damnatio Memoriae



The papyrus



Transcribed data

Οὐίβίω **Ἀλεξά[ν]δρῳ** τῷ κρατίστῳ ἐπιστρατήγῳ
παρὰ Ἀντ[ωνίου Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου
Ἀντωνίου[υ Ρ]ωμανοῦ Τραιανείου τοῦ κα[ὶ Σ]τρα[τείου]
Ἀντινοέως. [οὐκ ἂν] εἰς τοῦτο προήχθ[η]ν, ἐπι-
τρόπων [μέγιστ]ε, μέ[τριος] καὶ ἀπράγμων
ὢν ἀνθρ[ωπος,] εἰ μὴ [ὑ]βρι[ν] τὴν μ[εγ]ίστην
ἐπεπόνθ[ειν ὑπὸ] Ὁρίωνος κ[ω]μογρα[μ]ματέως
Φ[ι]λαδελφεί[ας τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ
Ἀρσινοΐτου. [οὐ χά]ριν μην[ύ]ω παρὰ τ[ὰ ἀ]πει-
ρημένα ἐα[υτὸ]ν ἐνσείσαντα εἰς τὴν κωμο-
γραμματείαν [μ]ήτε σιτολογήσαντα μήτε
πρ[α]κτορεύσαντα παντελῶς ἄπορον ὄν[τ]α.
δι' ἣν αἰτίαν καὶ πρότερον οὐ διέλιπον ἐντυγ-
χάνων καὶ νῦν ἀξιῶ, ἐάν σου τῇ τύχῃ δόξ[η],
ἀκοῦσαί μου π[ρ]ὸς αὐτὸν πρὸς τὸ τυχεῖν με
τῆς ἀπὸ σοῦ [μ]ισοπονήρου ἐγδ[ι]κίας, ἵν' ὦ ὑπὸ [σ]οῦ
κατὰ πάντα βεβοηθ[η]μένος). διευτύχει
Ἀντώνιος Δόμνος ἐπιδέδωκα.

Input form

Text

Οὐβίῳ [_]λεξά[____] τῷ κρατίστῳ ἐπιστρατήγῳ
παρὰ Ἀντ[ωνίου Δ]όμονου τοῦ καὶ Φιλαντι[νό]ου.
Ἀντωνίου[υ Ῥωμανο]ῦ Τραιανείου τοῦ κα[ὶ Στρα]τείου
Ἀντινοέως. [οὐκ ἄν] εἰς τοῦτο προήχθ[η]ν, ἐπι-
τρόπων [μέγιστ]ε, μέ[τριος] καὶ ἀπράγμων
ὢν ἄνθρ[ωπος,] εἰ μὴ [ὑβρι]ν τὴν μ[εγ]ίστην.
ἐπεπόνθ[ειν ὑπὸ] Ὀρίωνο[ς κ]ωμογρα[μ]ματέως
Φ[ι]λαδελφεί[ας τῆς Ἡρακλείδου μερίδο[ς] τοῦ
Ἀρσινόιτου. [οὐ χά]ριν μην[ύ]ω παρὰ τ[ὰ ἀ]πει-
ρημένα ἑαυτὸν ἐνσείσαντα εἰς τὴν κωμο-

☐ TLG ☐ PHI7 ☒ Epiduke

Send

#

Sentence

Parsed input (parsed for Leiden conventions)

Text

Οὐιβίω [_]λεξά[____] τῷ κρατίστῳ ἐπιστρατήγῳ
παρὰ Ἀντ[ωνίου Δ]όμονου τοῦ καὶ Φιλαντι[νό]ου.
Ἀντωνίου[υ] Ῥωμανοῦ Τραιανείου τοῦ κα[ὶ Στρα]τείου
Ἀντινοέως. [οὐκ ἂν] εἰς τοῦτο προήχθ[η]ν, ἐπι-
τρόπων [μέγιστ]ε, μέ[τριος] καὶ ἀπράγμων
ὦν ἄνθρ[ωπος,] εἰ μὴ [ὑβρι]ν τὴν μ[εγ]ίστην.
ἐπεπόνθ[ειν ὑπὸ] Ὀρίωνο[ς κ]ωμογρα[μ]ματέως
Φ[ι]λαδελφεί[ας τῆς] Ἡρακλείδου μερίδο[ς] τοῦ
Ἀρσινόιτου. [οὐ χά]ριν μην[ύ]ω παρὰ τ[ὰ ἀ]πει-
ρημένα ἑαυτοῦ ἐνσεύσαντα εἰς τὴν κωμο-

☐ TLG
☐ PHI7
☒ Epiduke

Send

Sentence

- 0 Οὐιβίω [_]λεξά[____] τῷ κρατίστῳ ἐπιστρατήγῳ παρὰ Ἀντ[ωνίου] [Δ]όμονου τοῦ καὶ Φιλαντι[νό]ου Ἀντωνίου[υ] Ῥωμανοῦ Τραιανείου τοῦ κα[ὶ] [Στρα]τείου Ἀντινοέως
- 1 [οὐκ] [ἂν] εἰς τοῦτο προήχθ[η]ν ἐπι- τρόπων [μέγιστ]ε μέ[τριος] καὶ ἀπράγμων ὦν ἄνθρ[ωπος] εἰ μὴ [ὑβρι]ν τὴν μ[εγ]ίστην ἐπεπόνθ[ειν] [ὑπὸ] Ὀρίωνο[ς] [κ]ωμογρα[μ]ματέως Φ[ι]λαδελφεί[ας] [τῆς] Ἡρακλείδου μερίδο[ς] τοῦ Ἀρσινόιτου
- 2 [οὐ] [χά]ριν μην[ύ]ω παρὰ τ[ὰ] [ἀ]πει- ρημένα ἑαυτοῦ ἐνσεύσαντα εἰς τὴν κωμο- γραμματείαν [μ]ήτε σπολογήσαντα μήτε πρ[α]κτορεύσαντα παντελῶς ἄπορον ὄν[τι]α
- 3 δι ἦν αἰτίαν καὶ πρότερον οὐ διέλιπον ἐντυγ- χάνων καὶ νῦν ἀξιῶ ἐάν σου τῇ τύχῃ δόξ[η] ἀκοῦσαί μου π[ρ]ὸς αὐτὸν πρὸς τὸ τυχεῖν με τῆς ἀπὸ σοῦ [μ]ισοπονῆρου ἐγδ[ι]κίας ἵν ᾧ ὑπὸ [σ]οῦ κατὰ πάντα βεβοηθ[η]μένος
- 4 διευτύχει Ἀντώνιος Δόμονος ἐπιδεδωκα



Strategy 1: use only information of the word

Word length + „survived“ pattern

[]λεξά[]

Interpreted word : _λεξά_

Length : 9

Candidate	Score	<input type="checkbox"/> Word length	<input type="checkbox"/> Neighbourhood letter bigrams	<input type="checkbox"/> Word similarity (letters)	<input type="checkbox"/> Named Entity	<input type="checkbox"/> Word bigram	<input type="checkbox"/> Semantic context	<input type="checkbox"/> Classification	Show
Ἀλεξάνδρα	2	1.0		1.0					
Ἀλεξάνρου	2	1.0		1.0					
Ἀλεξάνδρα	2	1.0		1.0					
Ἀλεξάνδρω	2	1.0		1.0					
Ἀλεξάρχου	2	1.0		1.0					
Ἀλεξάνδου	2	1.0		1.0					
ἐνεχάραξα	1	1.0							
ὁμολογεῖ	1	1.0							
ὁλοκλήρον	1	1.0							



GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Strategy 2: use only of context information

Word bigrams + co-occurrences + classification

[]λεξά[]

Interpreted word : _λεξά_

Length : 9

Candidate	Score	<input type="checkbox"/> Word length	<input type="checkbox"/> Neighbourhood letter bigrams	<input type="checkbox"/> Word similarity (letters)	<input type="checkbox"/> Named Entity	<input type="checkbox"/> Word bigram	<input type="checkbox"/> Semantic context	<input type="checkbox"/> Classification	Show
νομοῦ	3					0.5	0.4	0.000	
Ἀλεξάνδρω	3					1.0	0.8	0.003	
ἀπόδος	2					0.5		0.001	
Δόμνου	2						0.8	0.040	
ἀνέτεινα	2						0.2	0.025	
Αὐρηλίου	2						0.4	0.000	
Ἀχιλλεῖ	2					0.5	0.2		
Ἑπτὰ	2						0.2	0.010	
διαδεχομένω	2						0.2	0.250	
Σεουηριανῶ	2					0.5	0.2		
Αἰγύπτου	2						0.4	0.001	
ἡγεμόνι	2						0.2	0.001	
Λικννιανῶ	2						0.2	0.200	



The „real“ Strategy 2: use only of context information and removing any information about the damaged word

Reparsing

Text

Οὐβίω [_] τῷ κρατίστῳ ἐπιστρατήγῳ
παρὰ Ἀντ[ωνίου Δ]όμονου τοῦ καὶ Φιλαντι[νό]ου.
Ἀντωνίου [υ Ρωμανο]ῦ Τραιανείου τοῦ κα[ὶ Στρα]τείου
Ἀντινοέως. [οὐκ ἄν] εἰς τοῦτο προήχθ[η]ν, ἐπι-
τρόπων [μέγιστ]ε, μέ[τριος] καὶ ἀπράγμων
ὦν ἄνθρ[ωπος,] εἰ μὴ [ὑβρι]ν τὴν μ[εγ]ίστην.
ἐπεπόνθ[ειν ὑπὸ] Ὠρίωνο[ς κ]ωμογρα[μ]ματέως
Φ[ι]λαδελφεί[ας τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ
Ἀρσινοίου. [οὐ χά]ριν μην[ύ]ω παρὰ τ[ὰ ἀ]πει-
ρημένα ἑαυτῶν ἐνσεύσαντα εἰς τὴν κωμο-

☐ TLG

☐ PHI7

☒ Epiduke

Send

Sentence

- 0 Οὐβίω [_] τῷ κρατίστῳ ἐπιστρατήγῳ παρὰ Ἀντ[ωνίου] [Δ]όμονου τοῦ καὶ Φιλαντι[νό]ου Ἀντων[ο]ῦ [Ρωμανο]ῦ Τραιανείου τοῦ κα[ὶ]
[Στρα]τείου Ἀντινοέως
- 1 [οὐκ] [ἄν] εἰς τοῦτο προήχθ[η]ν ἐπι- τρόπων [μέγιστ]ε μέ[τριος] καὶ ἀπράγμων ὦν ἄνθρ[ωπος] εἰ μὴ [ὑβρι]ν τὴν μ[εγ]ίστην ἐπεπόνθ[ειν]
[ὑπὸ] Ὠρίωνο[ς] [κ]ωμογρα[μ]ματέως Φ[ι]λαδελφεί[ας] [τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ Ἀρσινοίου
- 2 [οὐ] [χά]ριν μην[ύ]ω παρὰ τ[ὰ] [ἀ]πει- ρημένα ἑαυτῶν ἐνσεύσαντα εἰς τὴν κωμο- γραμματεῖαν [μ]ήτε
πρ[ο]σ[κ]ορεύσαντα παντελῶς ἄπορον ὄν[τα]
- 3 δι ἦν αἰτίαν καὶ πρότερον οὐ διέλιπον ἐντυγ- χάνων καὶ νῦν ἀξιῶ ἂν σου τῇ τύχῃ δόξ[η] ἀκοῦσαί μου π[ρ]ὸς αὐτὸν πρὸς τὸ τυχεῖν με
τῆς ἀπὸ σοῦ [μ]ισοπον[η]ρίου ἐγδ[ι]κίας ἢ ὧ ὑπὸ [σ]οῦ κατὰ πάντα βεβωθ[η]μένος
- 4 διευτύχει Ἀντώνιος Δόμνος ἐπιδέδωκα

Word bigrams + co-occurrences + classification



Interpreted word : _

Length : 1

Candidate	Score	<input type="checkbox"/> Word length	<input type="checkbox"/> Neighbourhood letter bigrams	<input type="checkbox"/> Word similarity (letters)	<input type="checkbox"/> Named Entity	<input type="checkbox"/> Word bigram	<input type="checkbox"/> Semantic context	<input type="checkbox"/> Classification	Show
νομοῦ	3					0.5	0.4	0.000	
Ἀλεξάνδρῳ	3					1.0	0.8	0.003	
ἀπόδος	2					0.5		0.001	
Δόμνου	2						0.8	0.040	
ἀνέτεινα	2						0.2	0.025	
Αὐρηλίου	2						0.4	0.000	
Ἀχιλλεῖ	2					0.5	0.2		
Ἑπτὰ	2						0.2	0.010	
διαδεχομένῳ	2						0.2	0.250	
Σεουηριανῳ	2					0.5	0.2		
Αἰγύπτου	2						0.4	0.001	
ἡγεμόνι	2						0.2	0.001	
Λικνινιανῳ	2						0.2	0.200	



GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Strategy 3: Complete strategy

Multiple options

[]λεξά[]

Interpreted word : _λεξά_

Length : 9

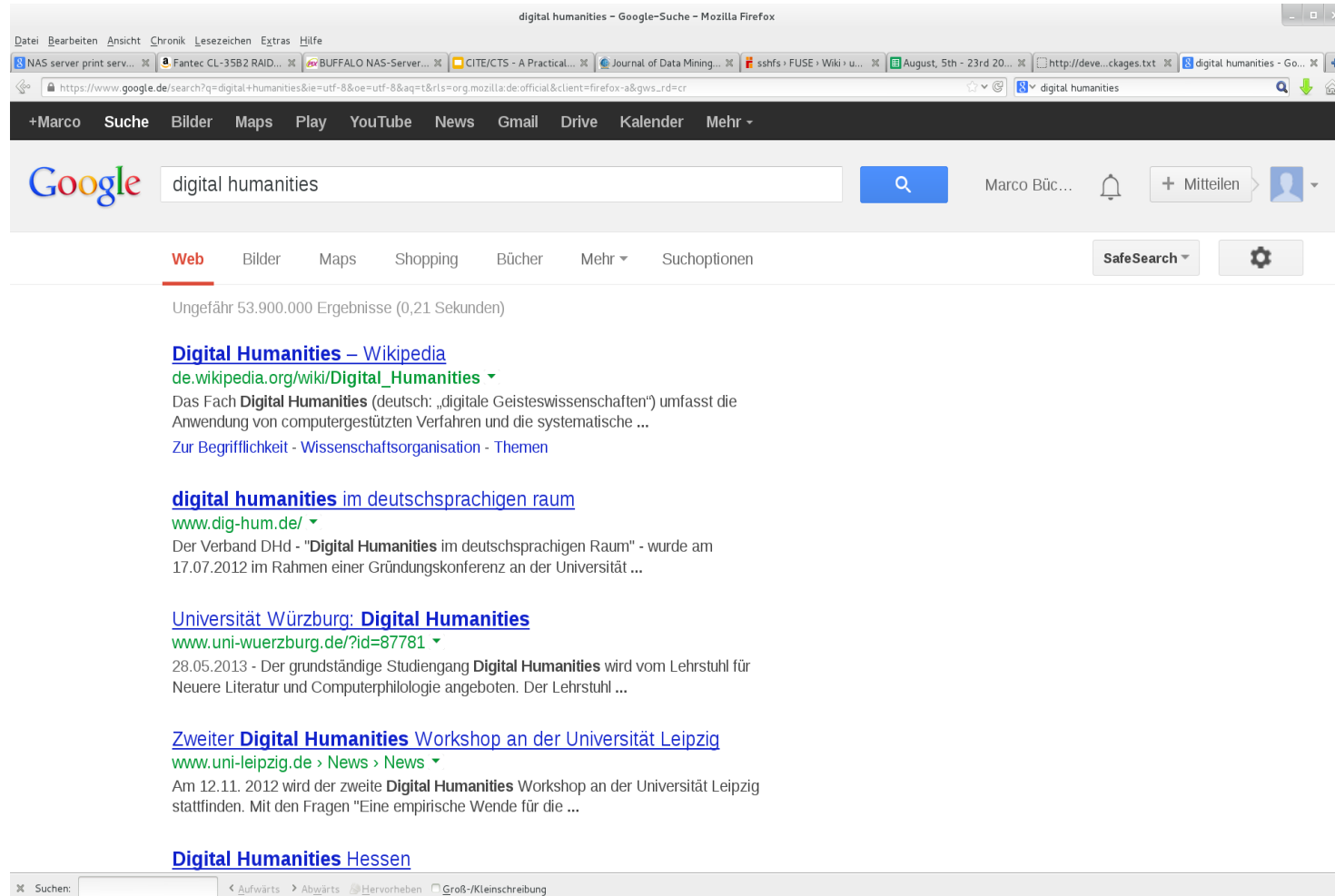
Candidate	Score	<input type="checkbox"/> Word length	<input type="checkbox"/> Neighbourhood letter bigrams	<input type="checkbox"/> Word similarity (letters)	<input type="checkbox"/> Named Entity	<input type="checkbox"/> Word bigram	<input type="checkbox"/> Semantic context	<input type="checkbox"/> Classification	Show
Ἀλεξάνδρω	5	1.0		1.0		1.0	0.8	0.003	
γενομένην	3	1.0				0.5	0.2		
διοίκησιν	3	1.0					0.2	0.008	
νομοῦ	3					0.5	0.4	0.000	
Ἀντιοέως	3	1.0					0.8	0.011	
βιβλίδου	3	1.0					0.4	0.003	
ἐππρόπων	3	1.0					0.2	0.005	
Ἀντιοέων	3	1.0					0.4	0.002	
Δημητρίωι	3	1.0					0.2	0.002	
βιβλίδων	3	1.0					0.2	0.002	
Στρατείου	3	1.0					0.6	0.214	
στρατηγός	3	1.0					0.2	0.001	
Ἀλεξάρχου	2	1.0		1.0					



GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN


Search & Find

How to find relevant information in massive data?



The screenshot shows a Google search results page for the query "digital humanities". The browser window title is "digital humanities - Google-Suche - Mozilla Firefox". The search bar contains "digital humanities" and the search button is a blue magnifying glass. The results show approximately 53,900,000 results in 0.21 seconds. The first result is "Digital Humanities - Wikipedia" with a link to de.wikipedia.org/wiki/Digital_Humanities. The description states: "Das Fach **Digital Humanities** (deutsch: „digitale Geisteswissenschaften“) umfasst die Anwendung von computergestützten Verfahren und die systematische ...". Below this is a link to "Zur Begrifflichkeit - Wissenschaftsorganisation - Themen". The second result is "digital humanities im deutschsprachigen raum" with a link to www.dig-hum.de/. The description states: "Der Verband DHD - 'Digital Humanities im deutschsprachigen Raum' - wurde am 17.07.2012 im Rahmen einer Gründungskonferenz an der Universität ...". The third result is "Universität Würzburg: Digital Humanities" with a link to www.uni-wuerzburg.de/?id=87781. The description states: "28.05.2013 - Der grundständige Studiengang **Digital Humanities** wird vom Lehrstuhl für Neuere Literatur und Computerphilologie angeboten. Der Lehrstuhl ...". The fourth result is "Zweiter Digital Humanities Workshop an der Universität Leipzig" with a link to www.uni-leipzig.de > News > News. The description states: "Am 12.11. 2012 wird der zweite **Digital Humanities** Workshop an der Universität Leipzig stattfinden. Mit den Fragen 'Eine empirische Wende für die ...'. The fifth result is "Digital Humanities Hessen". At the bottom of the browser window, there is a search bar with the text "Suchen:" and a magnifying glass icon, and a link to "Groß-/Kleinschreibung".

How to find relevant information in massive data?



Kundenservice 24/7
+49 221 2077 600

Ihre Reisedaten

Ihr Ziel
Köln (Nordrhein-Westfalen)

Anreise
19.07.13

Abreise
21.07.13

Einzelzimmer
1

Doppelzimmer
0

Erwachsene
1

Kinder
0









▼ Weitere Suchkriterien

Hotel suchen

143 freie Hotels in Köln (Nordrhein-Westfalen)

Ergebnisse sortiert nach:

HRS empfiehlt NEU Günstigster Preis Rabatte Kunden-Bewertung HRS Sterne Entfernung zu

Nr.	Hotelname Ort/Region	HRS Sterne	Preis pro Zimmer / Nacht	Verpflegung pro Person	Entfernung (km)	
1	 Savoy Köln - Altstadt Nord	★★★★ 8,7 / 10	EZ 80,27 EUR HOT DEAL	exkl. Frühstück (+18,00 EUR)	0,5 0,2 15 3,0	Zur Buchung
2	 Maria Suites Apartments Köln - Zentrum	★★★★ 9,2 / 10	EZ 115,00 EUR	exkl. Frühstück (+20,00 EUR)	0,3 0,3 18 1,0	Zur Buchung <small>Nur noch 2 Zimmer verfügbar!</small>
3	 Lint Köln - Altstadt	★★★ 8,3 / 10	EZ 74,00 EUR EXKLUSSIV PREIS	exkl. Frühstück (+10,00 EUR)	0,5 0,3 16 2,7	Zur Buchung
4	 Ihr Hotel Garni Köln - Dellbrück	★★★ 8,5 / 10	EZ 64,50 EUR	inkl. Frühstück	9,6 8,0 12 3,5	Zur Buchung <small>Nur noch 3 Zimmer verfügbar!</small>
5	 7 Wege Garni Köln - Dellbrück	★★★ 7,9 / 10	EZ 77,00 EUR	inkl. Frühstück	8,5 8,0 12 3,0	Zur Buchung <small>Nur noch 1 Zimmer verfügbar!</small>
6	 Thai Royal Köln - Zentrum	★★★★ 6,5 / 10	EZ 59,00 EUR	exkl. Frühstück (+12,50 EUR)	1,2 1,5 10 0,0	Zur Buchung
7	 Burns Art Cologne Köln	★★★★ 8,1 / 10	EZ 100,50 EUR	exkl. Frühstück (+11,00 EUR)	2,5 2,5 14 2,0	Zur Buchung
8	 Kunibert der Fiese Köln - Altstadt	★★★ 6,7 / 10	EZ 74,00 EUR	exkl. Frühstück (+8,50 EUR)	0,5 0,3 15	Zur Buchung <small>Nur noch 1 Zimmer verfügbar!</small>

Suchergebnis verfeinern

Preis 30 - 190 EUR

Frühstück einrechnen

Internet kostenlos im Zimmer (75)

HRS Top Quality Hotels anzeigen

HRS Sterne ★ - ★★★★★

Bewertung 0 - 10

H Haustiere erlaubt (94)

Definition of co-occurrence

Definition of co-occurrences:

- Common occurrence of at least two objects/events within a dedicated window
- Possible windows in Humanities: line, sentence, paragraph, document, author, century

Motivation:

- Psycholinguistic experiments: Given a word: What is the first word that comes to mind for test subjects?

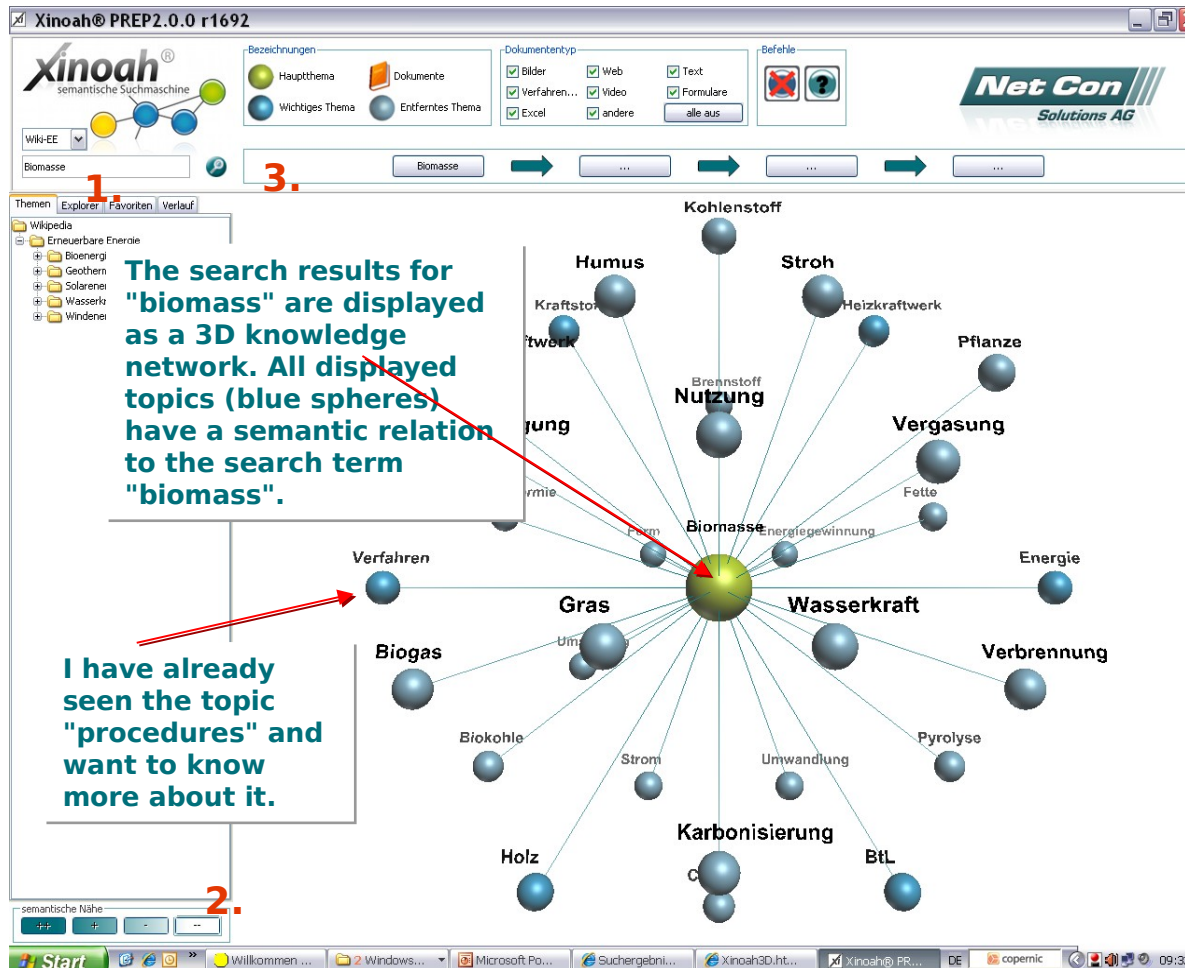
Definition of co-occurrence

Motivation:

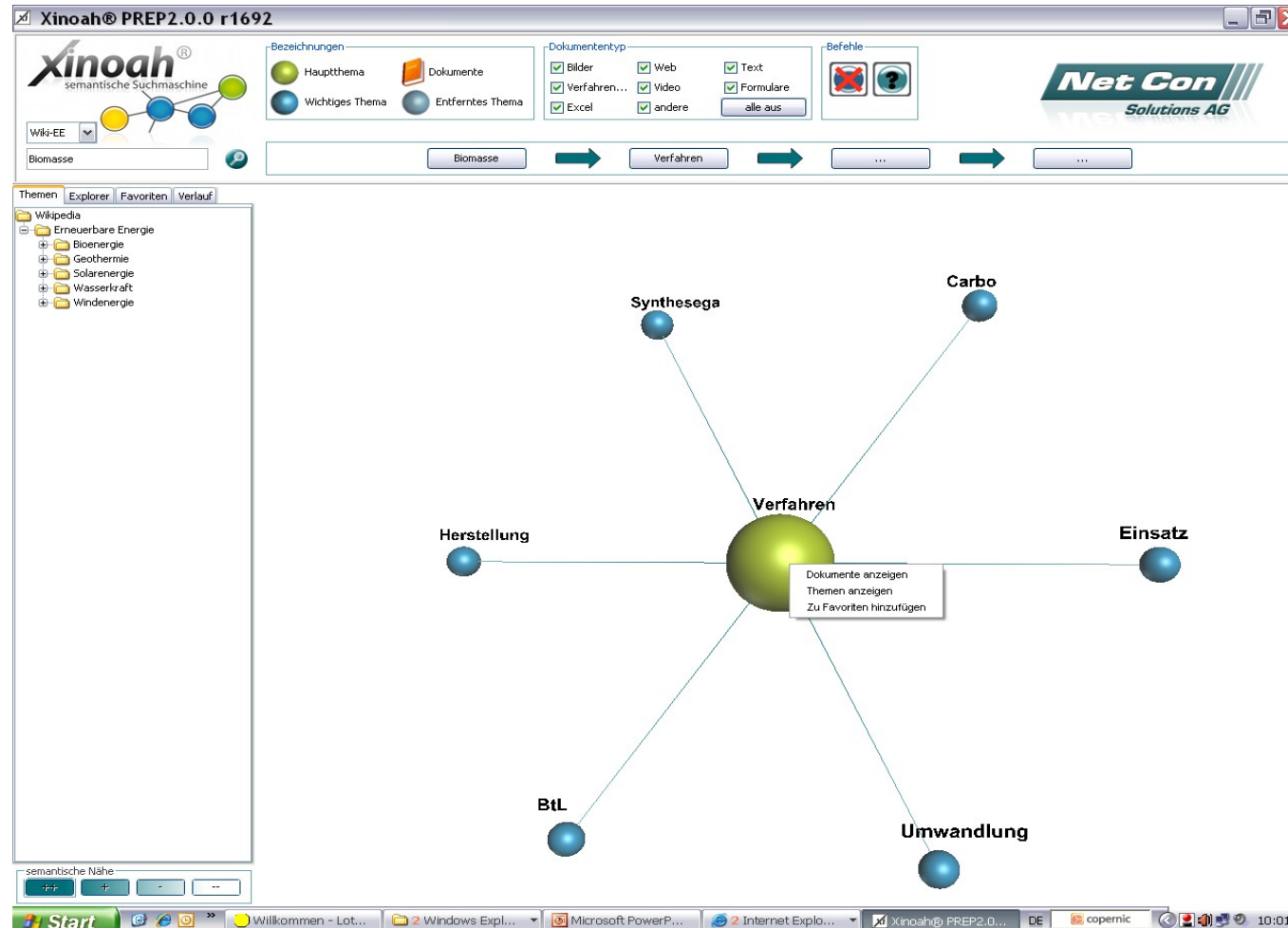
- Psycholinguistic experiments: Given a word: What is the first word that comes to mind for test subjects?

Stimulus	Response Prob.	# of Prob.'s	Co-occurrence	Significance
butter	Bread	60	Bread	51
	soft	40	Cheese	49
	Milk	32	Sugar	29
	Margarine	27	Milk	23
	Cheese	20	Margarine	22
	Fat	16	Farina	18
	yellow	14	Eggs	16
	Bread and butter	8	Pound	14
	Box / can	6	Meat	13
	eat	6		

How to find relevant information in massive data?

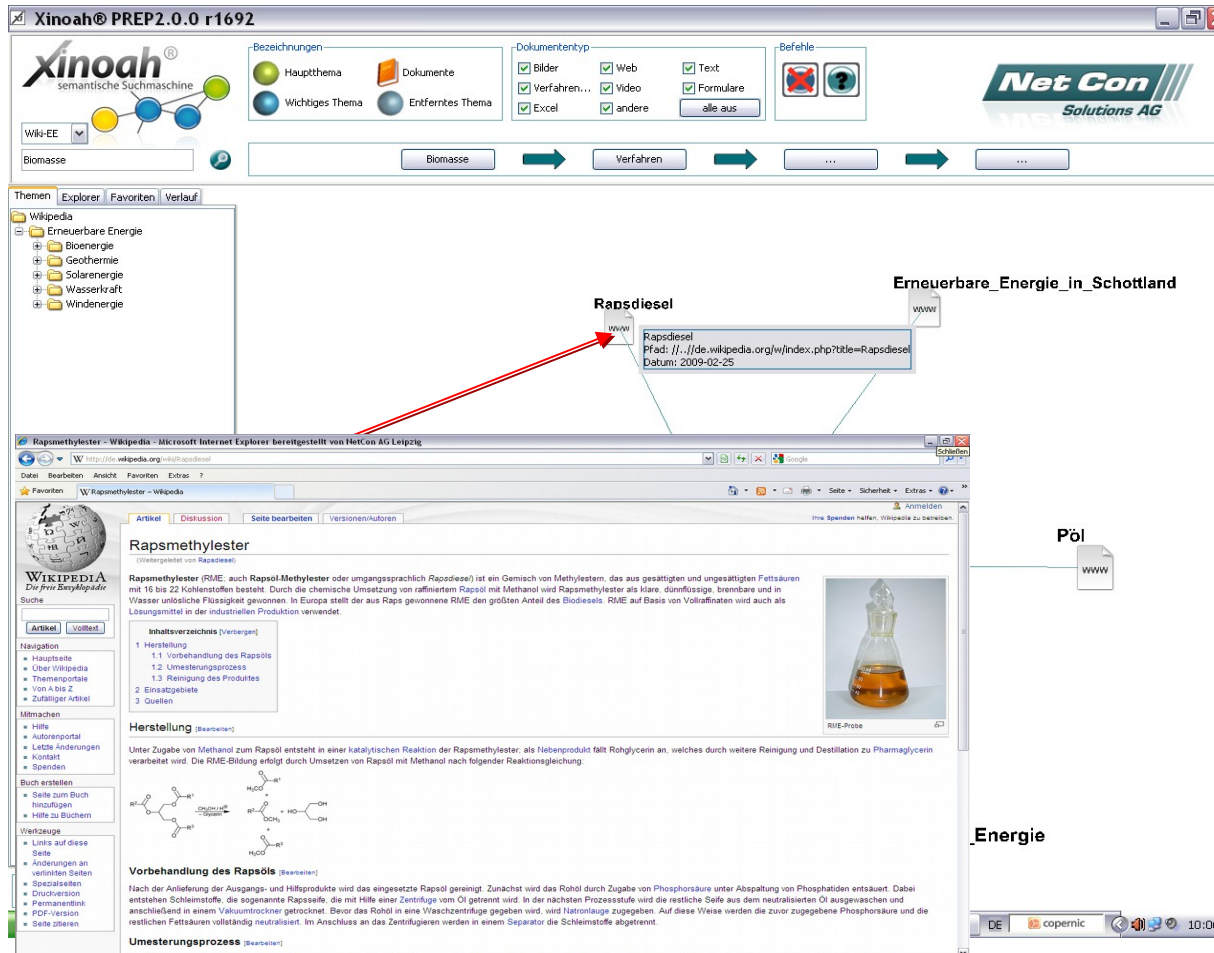


How to find relevant information in massive data?



The screenshot displays the Xinoah® PREP2.0.0 r1692 software interface. The main window shows a semantic search results for the term 'Biomasse'. The interface includes a top navigation bar with 'Themen', 'Explorer', 'Favoriten', and 'Verlauf' tabs. The 'Themen' tab is active, showing a hierarchical tree structure under 'Wikipedia' with categories like 'Erneuerbare Energie', 'Bioenergie', 'Geothermie', 'Solarenergie', 'Wasserkraft', and 'Windenergie'. The main content area displays a central node 'Verfahren' (Process) connected to six other nodes: 'Synthesega', 'Carbo', 'Einsatz', 'Umwandlung', 'BtL', and 'Herstellung'. A tooltip for the 'Verfahren' node lists actions: 'Dokumente anzeigen', 'Themen anzeigen', and 'Zu Favoriten hinzufügen'. The bottom status bar shows the Windows taskbar with various open applications and the system clock at 10:01.

How to find relevant information in massive data?



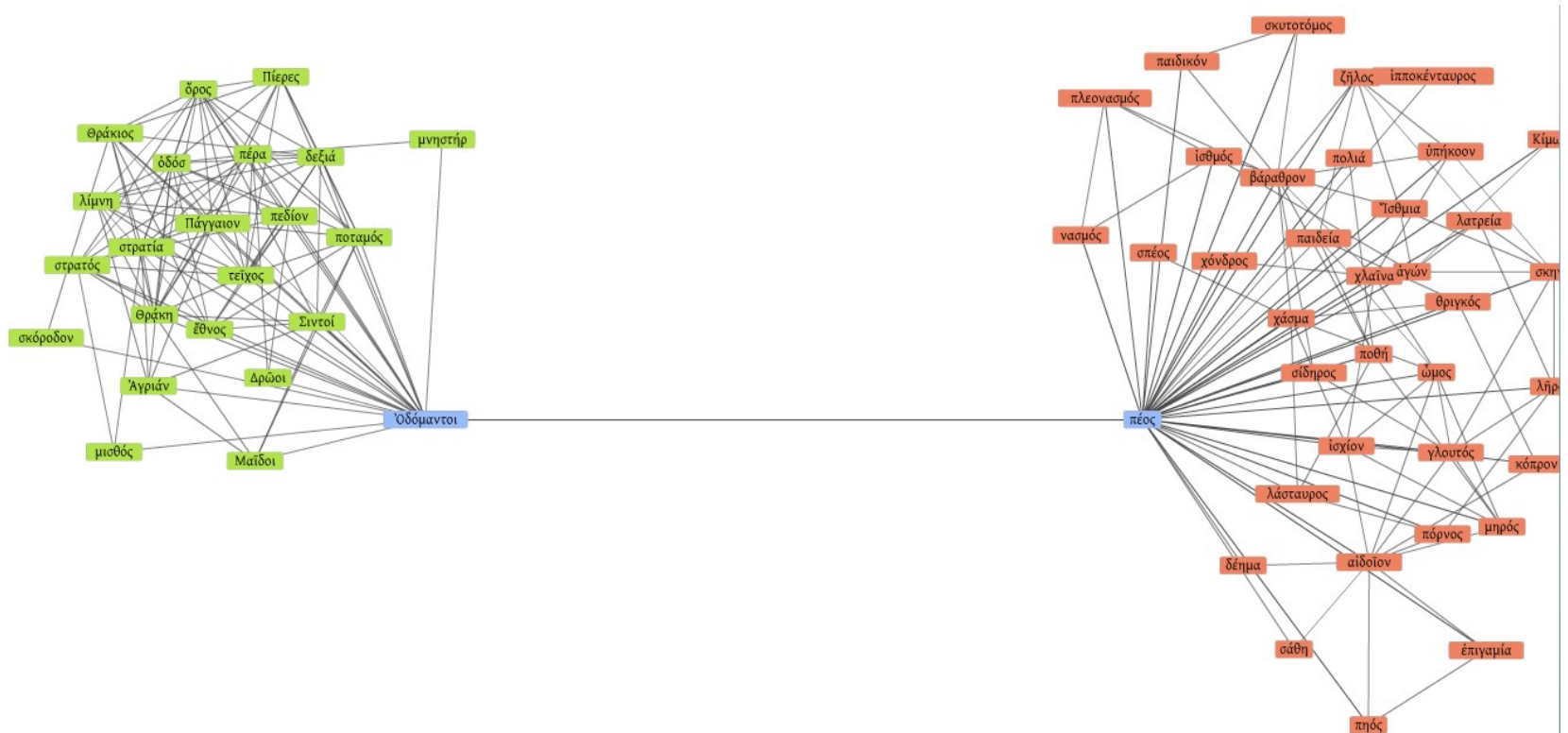
The screenshot shows the Xinoah search engine interface. The top bar includes the Xinoah logo, search filters (Bezeichnungen, Dokumententyp, Befehle), and a search bar with the input 'Biomasse'. Below the search bar, a list of themes is visible, including 'Erneuerbare Energie'. A red arrow points from the 'Rapsdiesel' entry in the search results to a Wikipedia article titled 'Rapsmethylester'. The article page is displayed below, showing the title 'Rapsmethylester', a brief description, and a chemical structure diagram. The diagram shows the reaction of a triglyceride with methanol to produce fatty acid methyl esters (FAME) and glycerol. The text describes the production process, including the use of a catalyst and the separation of the products. The article also mentions the use of rapeseed oil in the production of biodiesel.



GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Contrastive Semantics

Visualisation of Contrastive semantics



Main properties

- **Contrast:**

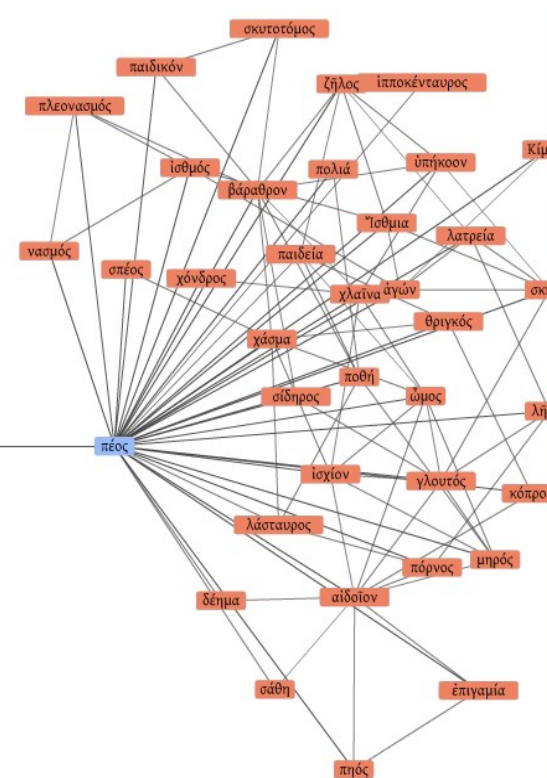
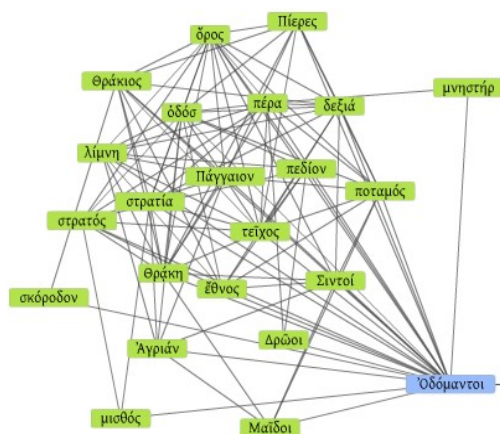
$$\text{contrast}(w_i, w_j) = \begin{cases} 1 - \text{sim}_{\text{dice}}(w_i, w_j) & \text{if } \text{sim}_{\text{dice}}(w_i, w_j) \leq \text{eps} \\ 0 & \text{if } \text{sim}_{\text{dice}}(w_i, w_j) > \text{eps} \end{cases} \quad \text{sim}_{\text{dice}}(w_i, w_j) = 2 * \frac{|K_{w_i} \cap K_{w_j}|}{|K_{w_i}| + |K_{w_j}|}$$

- **Locality:**

$$\text{dist}(w_i, w_j) \leq \text{eps}_{\text{dist}} \text{ aus } (w_i, w_j) \in C$$

- **Frequency range of contrastive semantic relations:**
 - Generally less than 10 times of common occurrences

Connectivity?



Some observations

- **Identified clusters:**
 - As shown in examples comedy
 - Sarcasm
 - Cynicism
 - Artificial ambiguity like „Michael Schumacher the red king“ (translated from a German corpus)
 - Scope to gnomology
- **Is there a relation between contrastive semantics and textual reuse?**
 - Clearly, yes.
 - „Evaluation results“: More than 90% of the contrastive semantics have a relation to text reuse



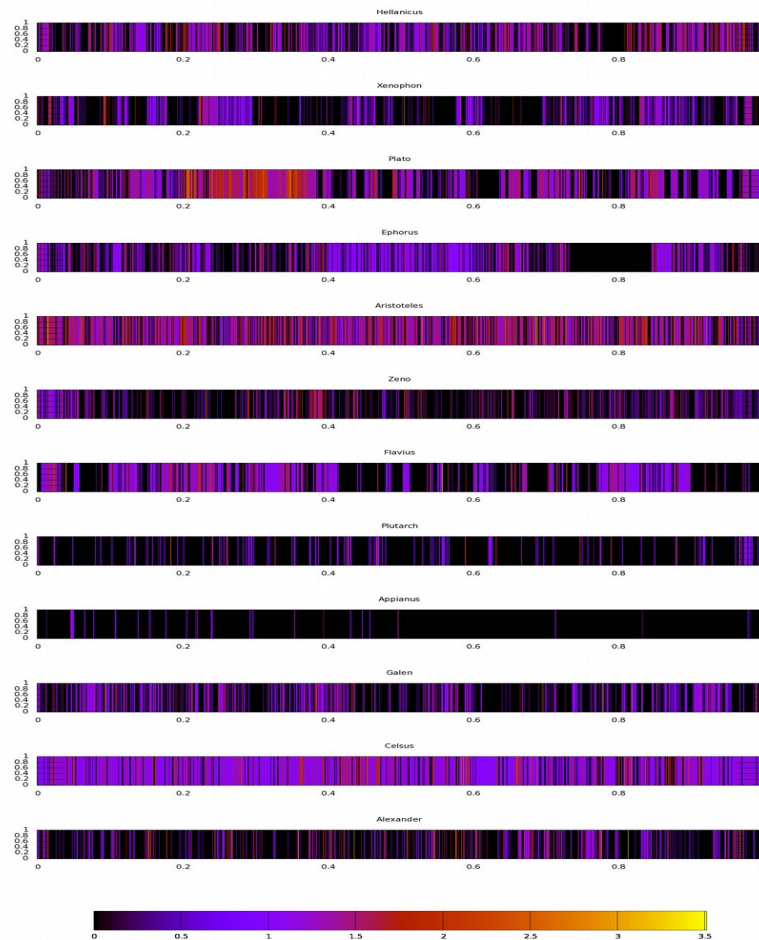
GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Historical Text Reuse

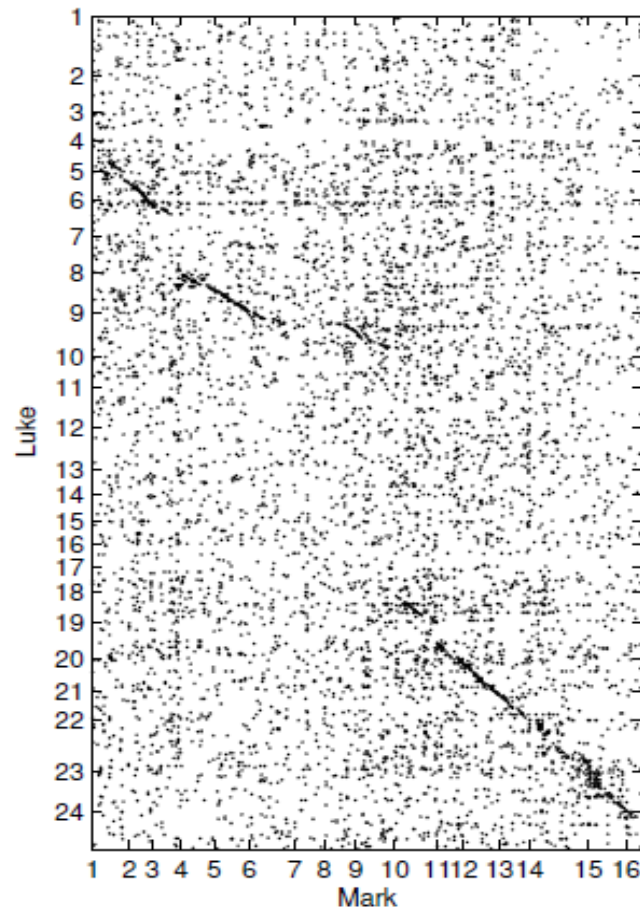
Text Reuse for Humanities and Computer Science

- **Question:** Why is Text Reuse so relevant for Humanities and Computer Science?
- **Premise:** The amount of digitally available data is growing exponentially (Big Data)
- Humanities:
 - Lines of transmission and textual criticism
 - Transmissions of ideas/thoughts under different circumstances and conditions
- Computer Science:
 - Text Decontamination for stylometry and authorship attribution, dating of texts
 - gen. Text Mining, Corpus Linguistics

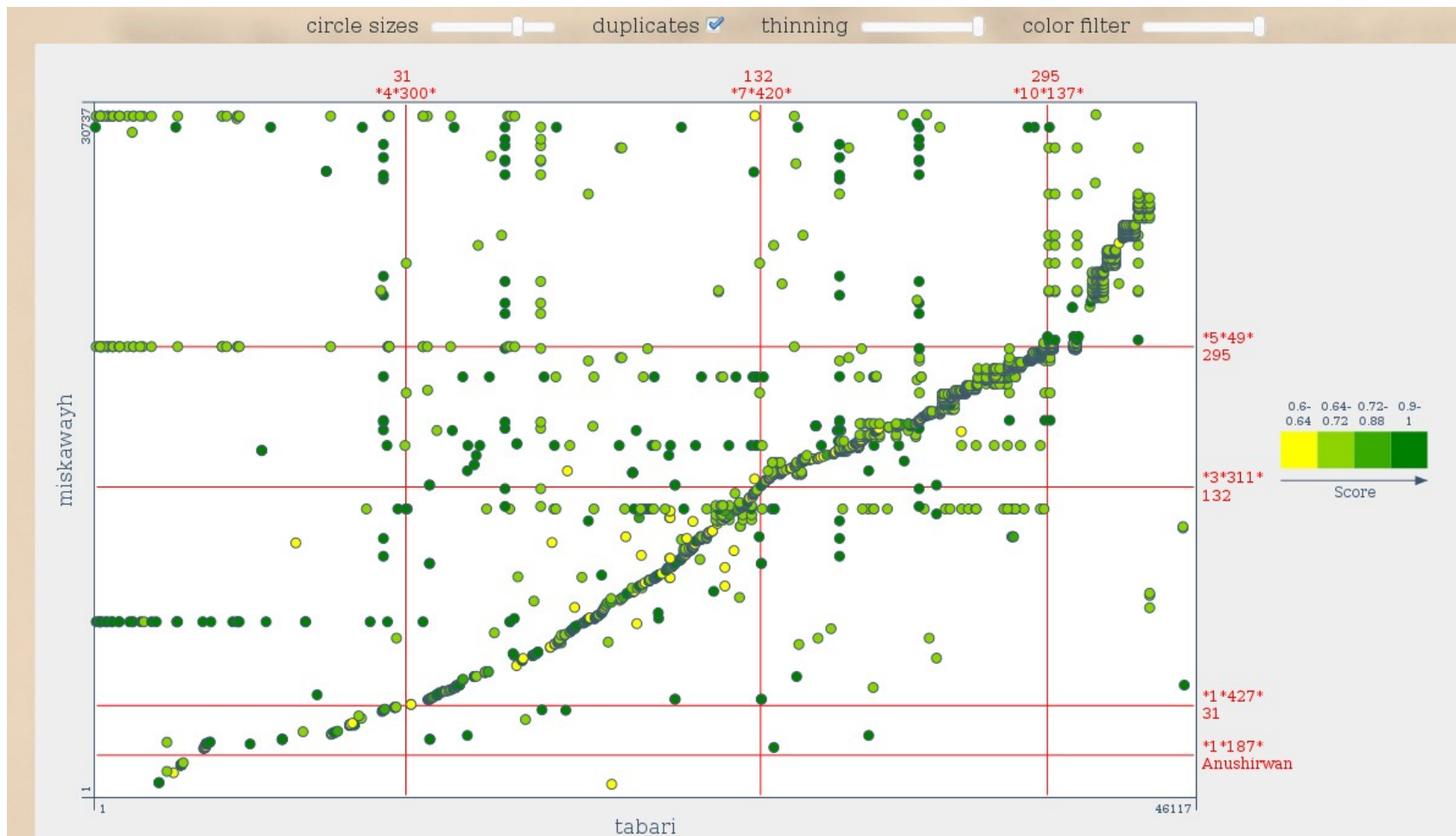
Temperature Map



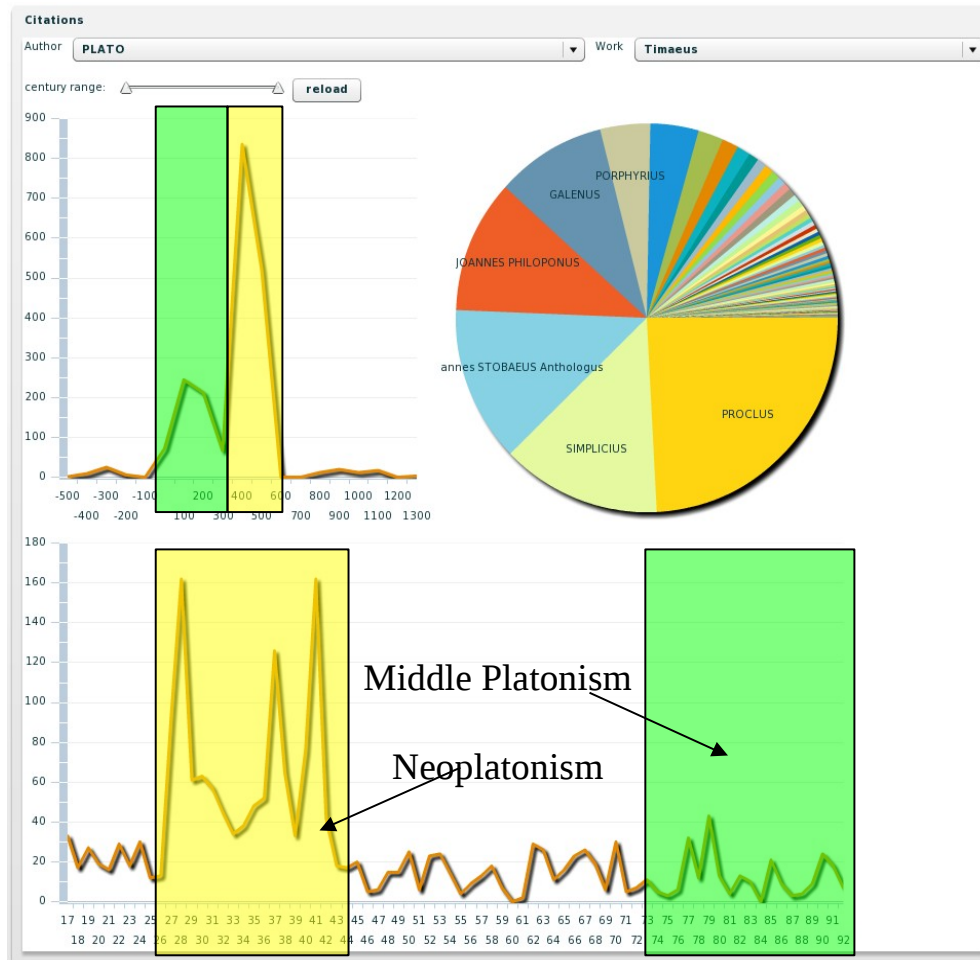
Visualisation problem: Dotplot view



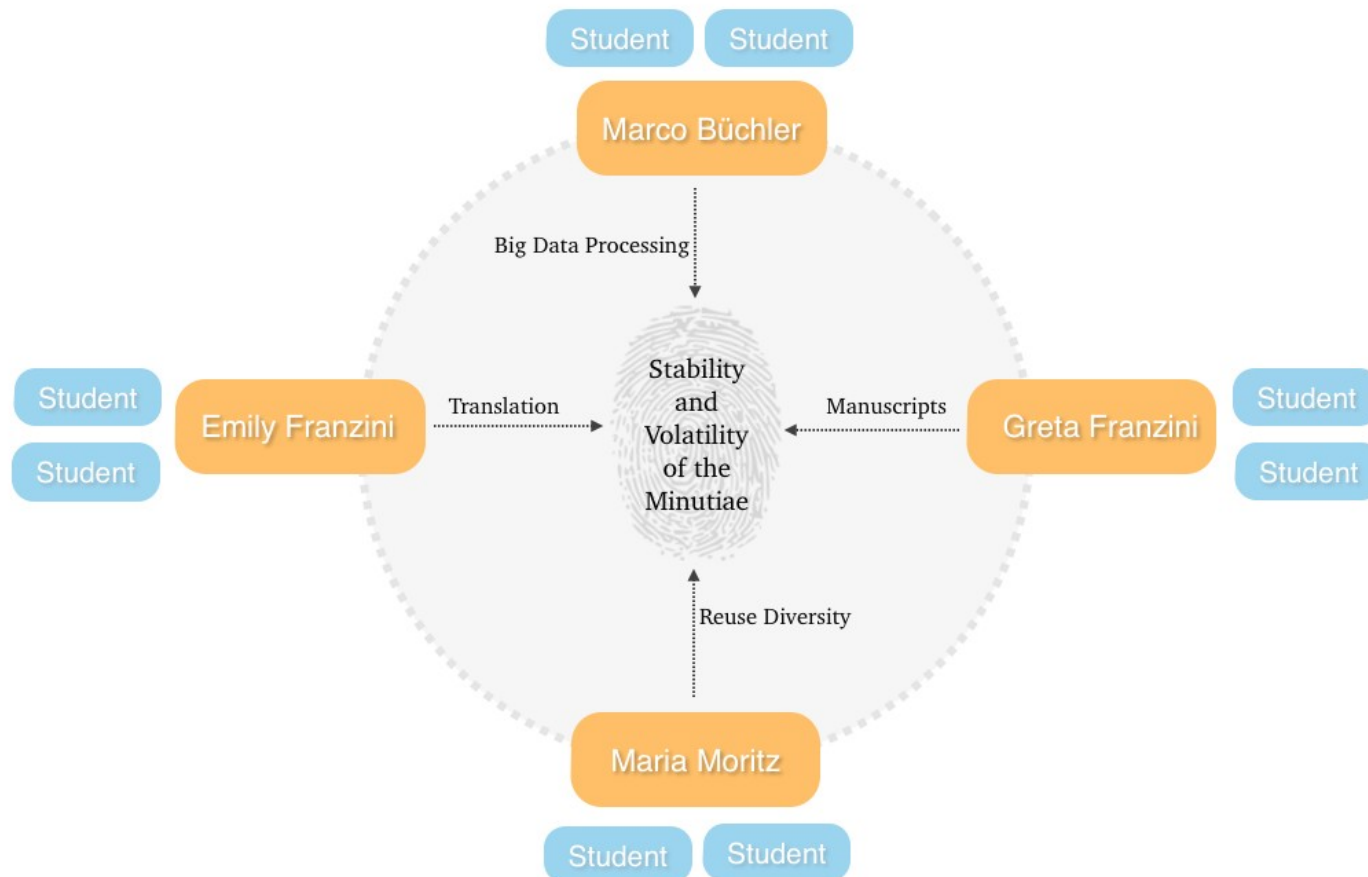
Dotplot view



Visualisation problem: Macro view



eTRAP – Electronic Text Reuse Acquisition Project





GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Thank you!

"Stealing from one is plagiarism, stealing from many is research" (Wilson Mitzner, 1876-1933)



SPONSORED BY THE



Federal Ministry
of Education
and Research

Visit us at <http://etrap.gcdh.de>