



Humanities research based on big cultural heritage data

Eetu Mäkelä, D.Sc.

Assistant Professor in Digital Humanities / University of Helsinki
Docent (Adjunct Professor) in Computer Science / Aalto University

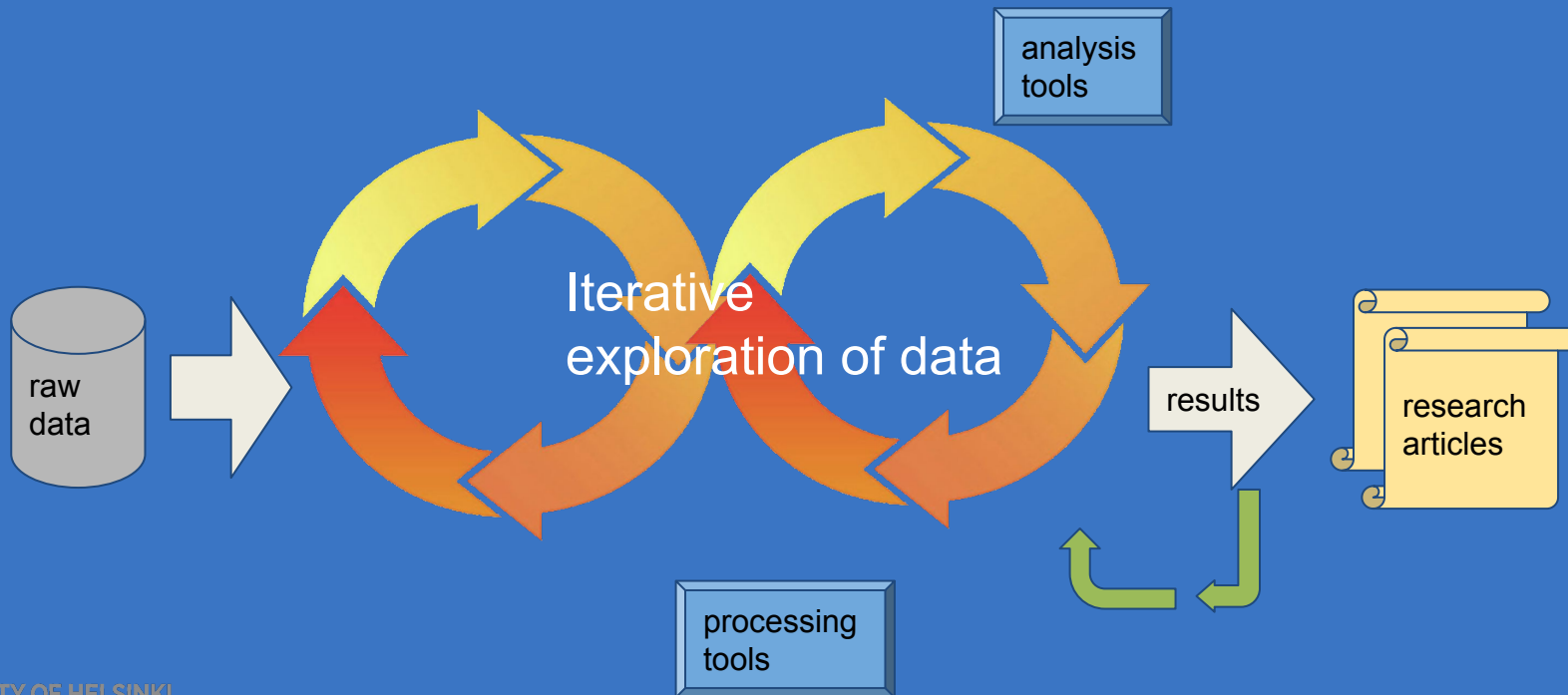


Research process

1. Have data
2. Magic (?)
3. Something interesting shows up
4. Profit!



Digital humanities research process





Open data in the digital humanities - the good

- Great aggregators pushing for CC0 licenses, publishing participating data: [Europeana](#), [Digital Public Library of America](#) & [The European Library](#)
- Influential national libraries moving to co-operative open (linked) data
 - [Library of Congress](#), [Deutsche Nationalbibliothek](#), [British Library](#), [Bibliothèque nationale de France](#)
- Museums, Galleries and Archives catching up: [British Museum](#), [Finnish National Gallery](#), ...
- Glue available: [VIAF](#), [CIDOC-CRM](#), Getty [AAT](#), [TGN](#), [ULAN](#), [CONA](#), [Pleiades](#), [DBpedia](#), [Wikidata](#), ...



Open data in the digital humanities - the bad

- Academic libraries have a long tradition of collaborating with library service companies (primarily EBSCO Information Services, ProQuest LLC and Gale Cengage Learning) to produce services
- Often, they also participate in content creation projects, and then hold the rights for that content
 - e.g. Early English Books Online (ProQuest), Nineteenth Century Collections Online (Gale), State Papers Online (Gale)
- But, this is also a wider culture **inside** humanities, e.g. Electronic Enlightenment



Research question ↔ data

- “Which places published the most French philosophy in the 18th Century?” – I know, I’ll ask the French national library database



Research question ↔ data

- “Which places published the most French philosophy in the 18th Century?” – I know, I’ll ask the French national library database
 - But is their data free of bias?
 - How is the information stored?

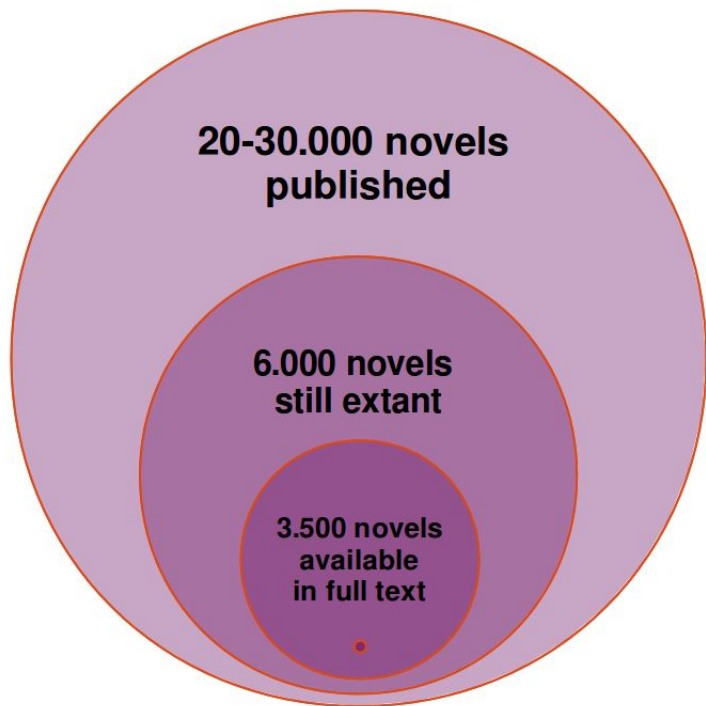


Research process

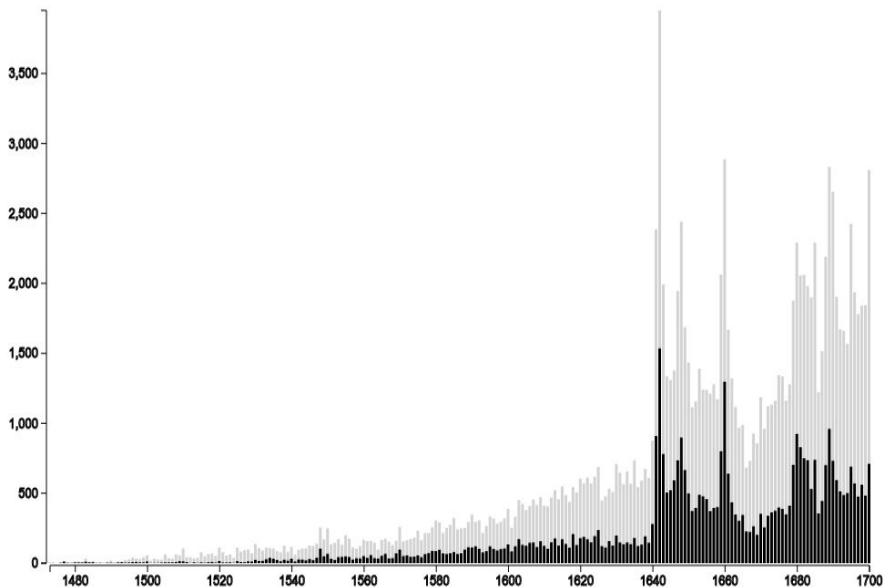
1. Have data ← 0. **Get data, understand magic that went into data**
2. Magic (?)
3. Something interesting shows up
4. Profit!



What's in there?



EEBO-TCP AND ESTC TEXT COUNTS





Library catalogue contents

Leader *****ngm 22*****1a 4500

245 04 \$a The Adventures of Safety Frog. \$p Fire
safety \$h [videorecording] /
\$c Century 21 Video, Inc.

246 30 \$a Fire safety \$h [videorecording]

260 ## \$a Van Nuys, Calif. : \$b AIMS Media, \$c 1988.

300 ## \$a 1 videocassette (10 min.) : \$b sd., col. ;
\$c 1/2 in.

500 ## \$a Cataloged from contributor's data.

538 ## \$a VHS.

521 ## \$a Elementary grades.

530 ## \$a Issued also as motion picture.

520 ## \$a Safety Frog teaches children to be fire safe,
explaining that smart kids never play with
matches. She shows how smoke detectors work
and explains why they are necessary. She also
describes how to avoid house hold accidents
that lead to fires and how to stop, drop,
and roll if clothing catches fire.

650 #0 \$a Fire prevention \$v Juvenile films.



Documentation!!!

- 81 pages of documentation on the exact annotation practices used in a digital edition of the Potage Dyvers
- Library cataloguing standards:
 - 302 pages of ISBD
 - 750 pages AACR, 1056 pages of RDA
- 1020 pages of the SPECTRUM standard for museum cataloguing
- A single page of field descriptions in the Schoenberg database



Documentation?

<https://pro.europeana.eu/data/linked-open-data-data-downloads>



The missing documentation

- “We changed our cataloguing standards once in the 80’s, and then a second time in 1998.”
- “Most of our older entries have actually been copied from the national library that has different cataloguing standards”
- “A lot of the publications from the middle of the 18th century are simply missing, as they were never indexed.”
- “This database was gathered based on the whimsies of what the participating researchers researched. It’s probably thus quite biased.”



Open data in the digital humanities - the ugly

- Different forms of encoding, typos

(Paris,) Paris [Paris,] [Paris]

(Paris) A Paris À Paris (Paris

(Paris.) [A Paris]

Amsterdam. - et Paris

Amsterdam ; et Paris

Amsterdam. - et à Paris

Amsterdam [Paris]

(Paris. - Amsterdam

A Amsterdam [i. e. Paris]. M. DCC. LXX.



Data woes: viaf.org

- Automatic conversions from “Lastname, Firstname” to “Firstname Lastname” does not always work due to bad data

100 1 _ [_ta Arlincourt, tc vicomte d' td \(Charles Victor Prévôt\), td 1789-1856](#)

100 1 _ [_ta Arlincourt, tc vicomte d' td \(Charles Victor Prévôt\), td 1789-1856](#)

100 1 _ [_ta Arlincourt, Charles Victor Prévost, tc vicomte, td 1788-1856](#)

100 0 _ [_ta Charles-Victor Prévost d'Arlincourt tc écrivain français](#)

100 1 _ [_ta Arlincourt, Charles Victor Prévôt d' td 1789-1856](#)

100 1 _ [_ta Arlincourt, Charles Victor Prevot d' td \(1789-1856\).](#)

200 _ 0 [_ta Arlincourt, tc Visconde de, tf 1789-1856](#)

100 1 _ [_ta Arlincourt, Charles Victor Prevost d' td 1788-1856 tc Vicomte](#)

100 1 _ [_ta Arlincourt, Charles Victor Prevost d', tc Vicomte, td 1788-1856](#)

200 _ | [_ta Arlincourt tb Charles-Victor Prévost d' tf 1788-1856](#)

100 1 _ [_ta Arlincourt, Charles-Victor Prévost d' td \(1788-1856\).](#)

100 1 0 [_ta Arlincourt, Charles-Victor Prévost d' td 1789-1856](#)

200 _ 1 [_ta Arlincourt tb , Charles Victor Prévôt tf <vicomte d'>](#)



<schema:name>Charles-Victor Prévost d'Arlincourt</schema:name>
<schema:name>Charles Victor Prévôt ~d'œ Arlincourt</schema:name>
<schema:name>Charles Victor Prevot d' Arlincourt</schema:name>
<schema:name>Arlincourt</schema:name>

<http://viaf.org/viaf/41896578/>



Automatic OCR

THIS not Saint *George* we Sing of here,
Nor *George*, the fatal Duke *Villier* ;
Nor *George a Green*, nor *Castriot*,
Nor *Buchanan* the learned *Scot* ;
But 'tis of *George* the Valiant *Monck*,
That made *Van-Trump* in's Blood dead-
And in the Seas his Navy funck. (drunk.
Oh! this is our brave George!

Is not- Saint George we Sing of here,
Nor George, the fatal Duke Villier ;
Nor George a Green, nor Castriot,
Nor Buchanan the learned Scot ;
But us of George the Valiant Monck,
That made Van-Trump in'S Blood deod
and in theseus his Navy snuck. (drunk,
Ok I this is our brave George !

THE

Firste volume of the Chronicles of England, Scot- lande, and Irelande.

CONTEYNING,

The description and Chronicles of England, from the
first inhabiting unto the conquest

The description and Chronicles of Scotland, from the
first originall of the Scottes nation, till the yeere
of our Lord. 1571

The description and Chronicles of Irelande, likewise
from the firste originall of that Nation, untill the
yeare. 1547.

Faithfully gathered and set forth, by
Raphaell Holinshed.

AT LONDON,
Imprinted for Iohn Harrison.

~ ~k ~

~ l I ~ li ~]J]O DmU ~ ov O ~ ii |

~ ~ 1l ~ ~ - \ O ~ Si ~ \ r <, St ~ 5, o t %, \ ~ t, \ ~ ~ ~

~ ' . - bn EIs ~ l br ~; < ~ 5n ~ 1 ~

~ 1 1 ~ t ~ 3mo 71 ~ k ~ 7noost I 3o ~ rsd
~ i ~ mlm 87il fif ~ s ~

~ ' 3, Ilmo ~ l. 6n 3 ~ nm / 17 ~ = io \ ~ ~ 7g ~ i

.... ~ -, ~. ; l l ~ 1 B ~] 8 ~ . ~ ~ ' ~

' ~ ~ @ ~ ~ ~ ~ pA til Sns t' - b ~ ~ I \ U \ ' i: ~]
~ ~ ~

I I noin ~ Hodol ~ o] bs Jni ~ qml ' ~ 1 11

1 ~ . 1 ~ ~ 1 11

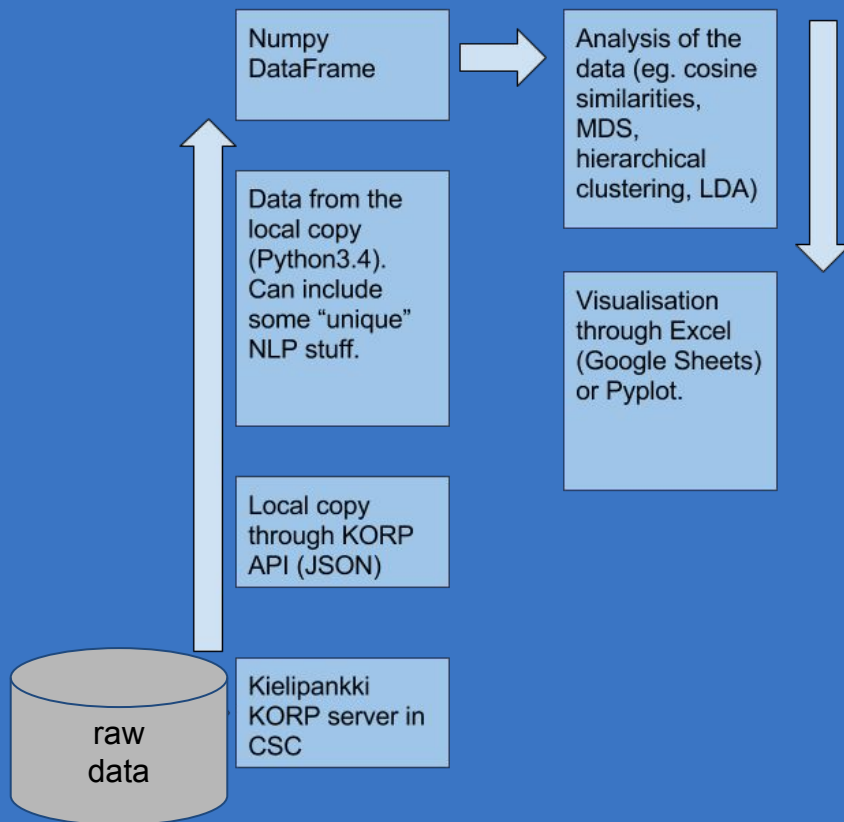
" ~ ~ [? ' ~ 9 ~ 9 ~] bo O \ ~

„ - - - . ~ 13 ~ ~ ~

- : ~ _ 1

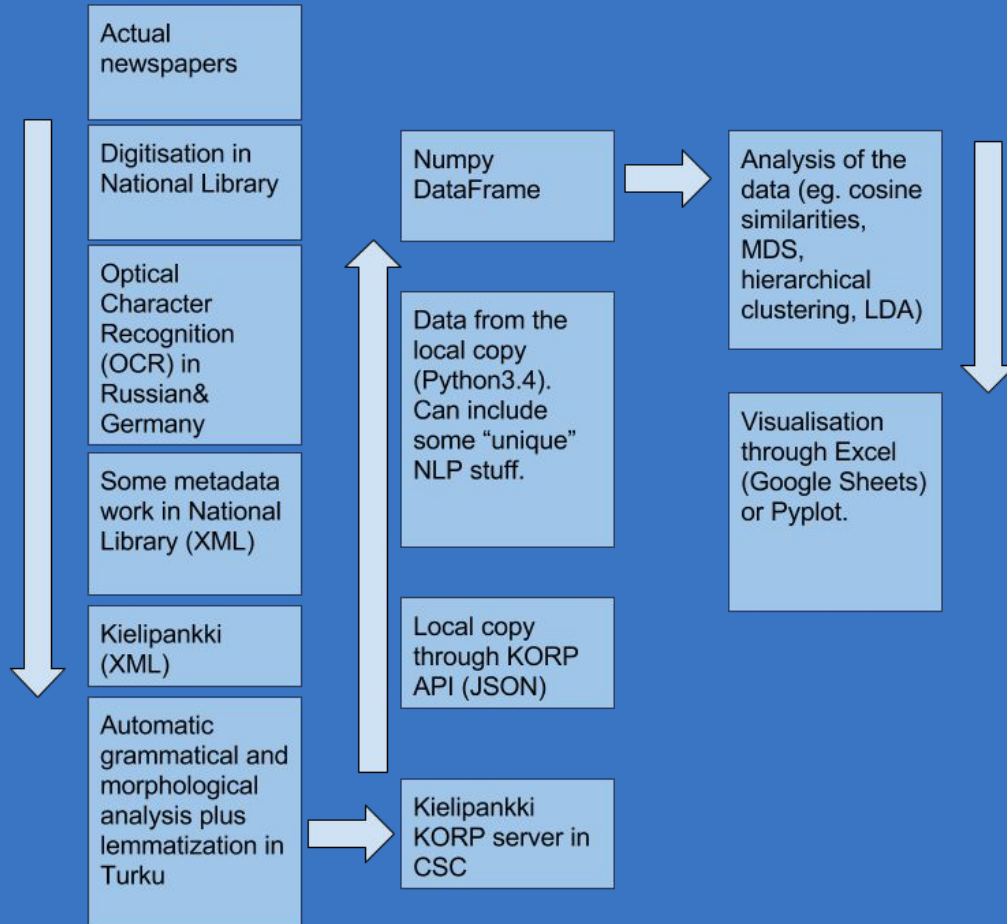


KLK Newspaper Pipeline: from archives to a researcher



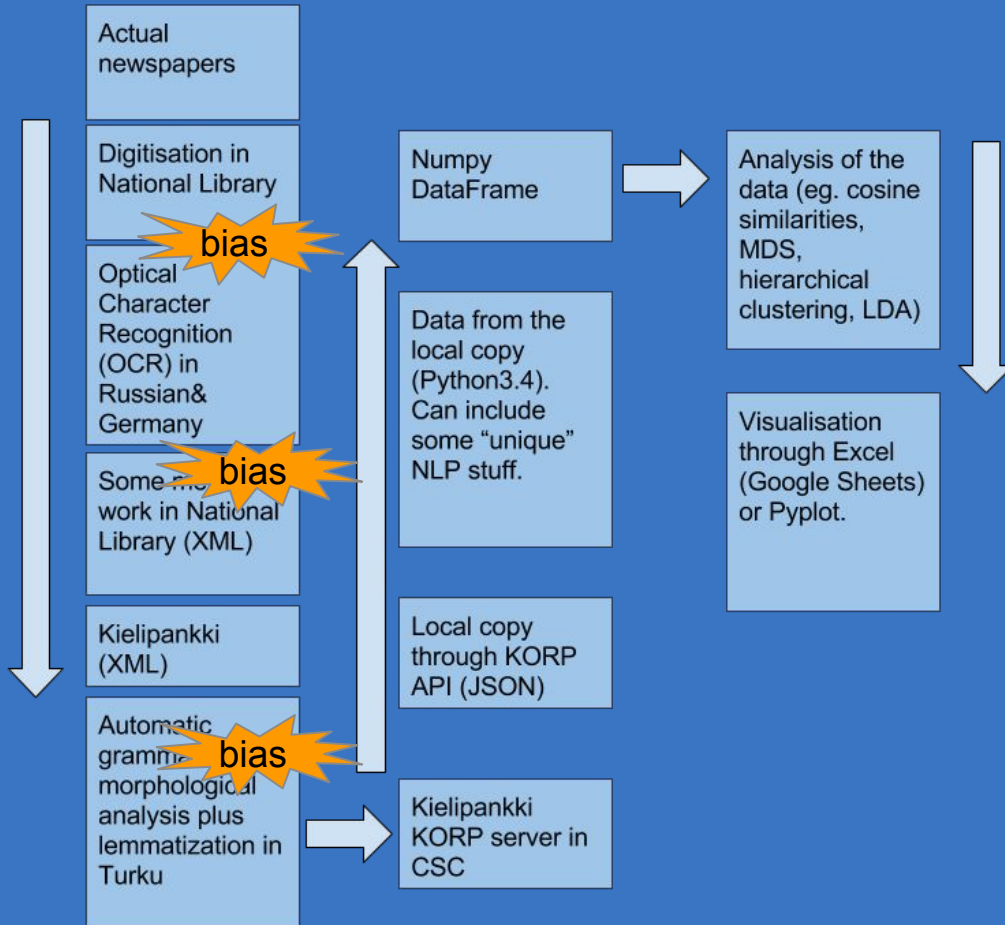


KLK Newspaper Pipeline: from archives to a researcher



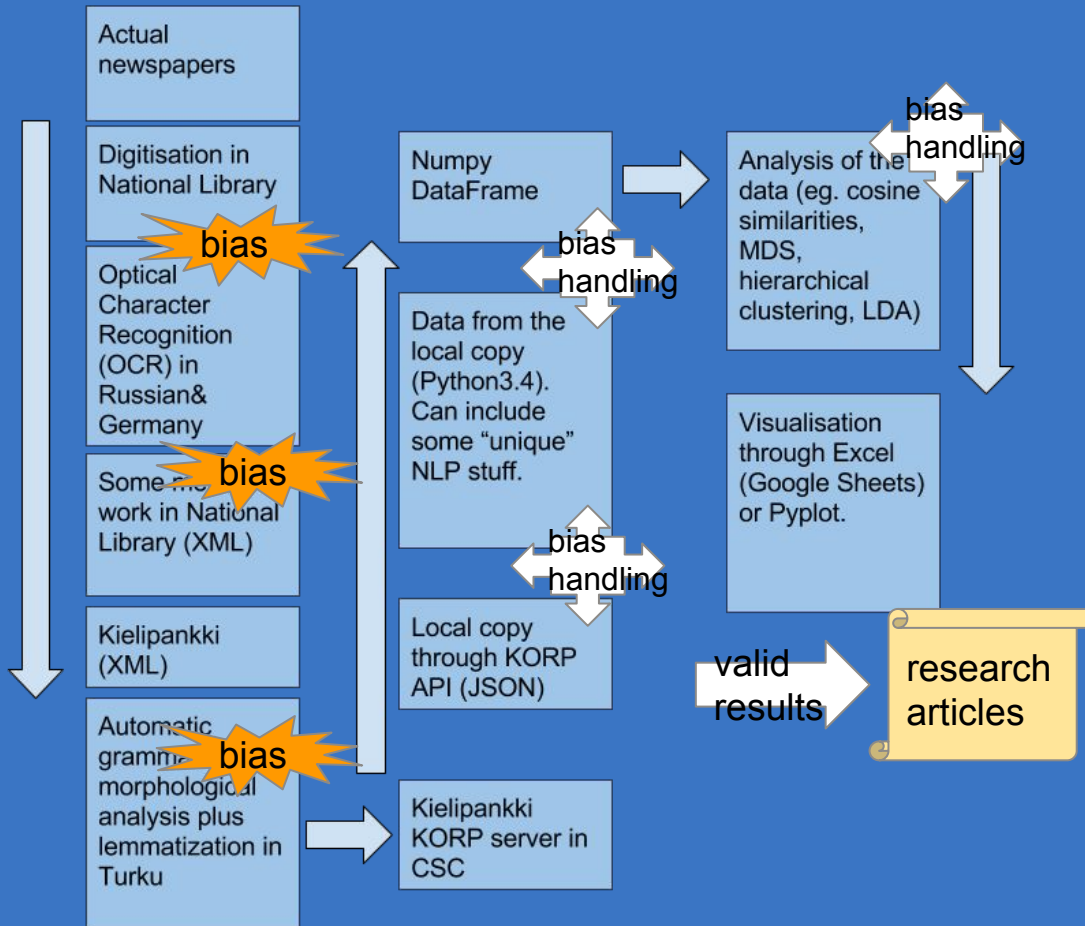


KLK Newspaper Pipeline: from archives to a researcher



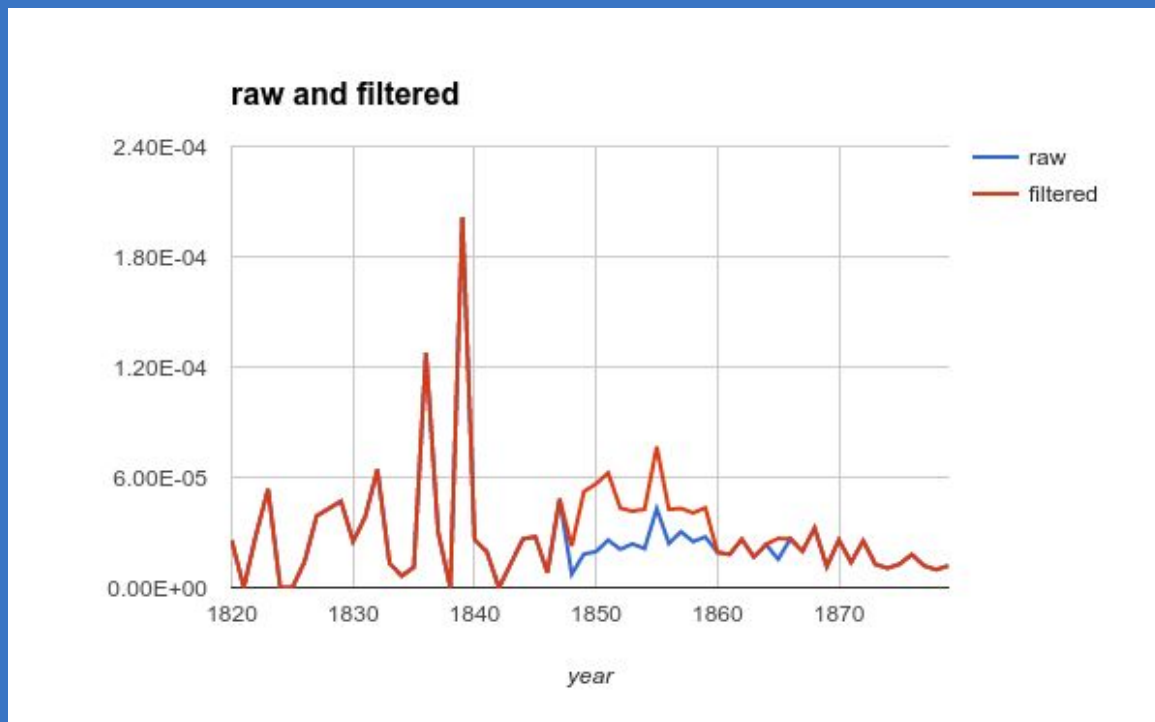


KLK Newspaper Pipeline: from archives to a researcher



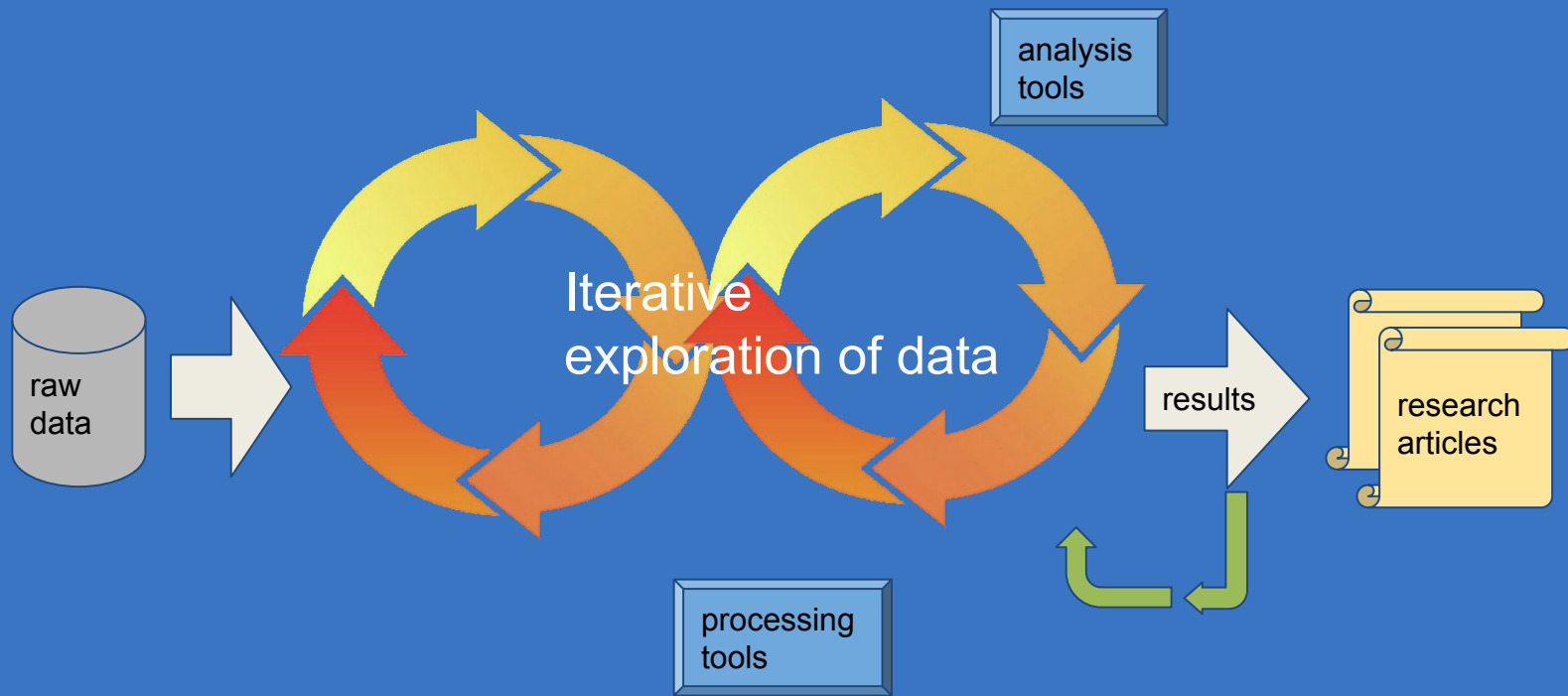


Data woes: National Newspaper Collection (KLK)



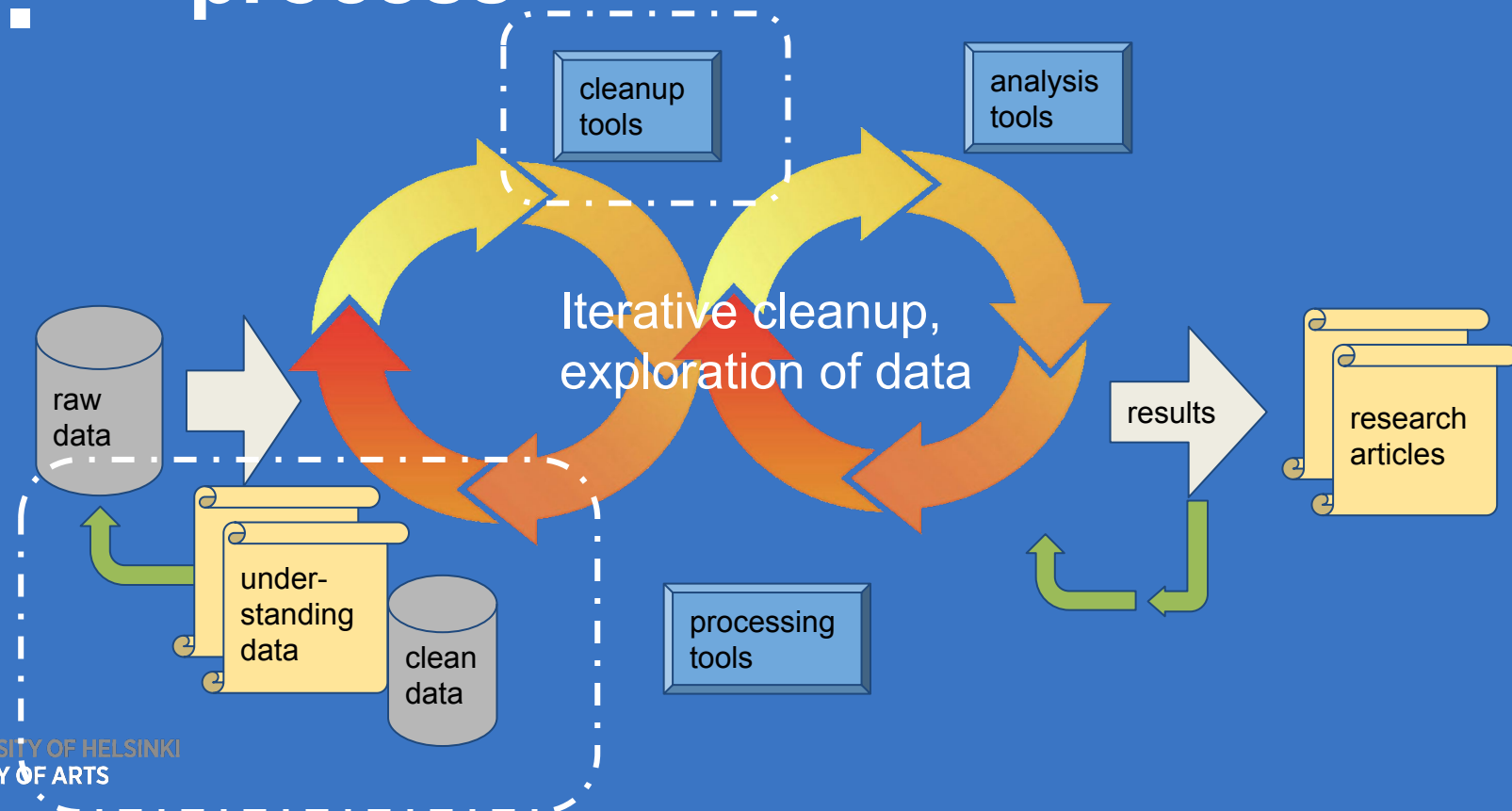


Digital humanities research process



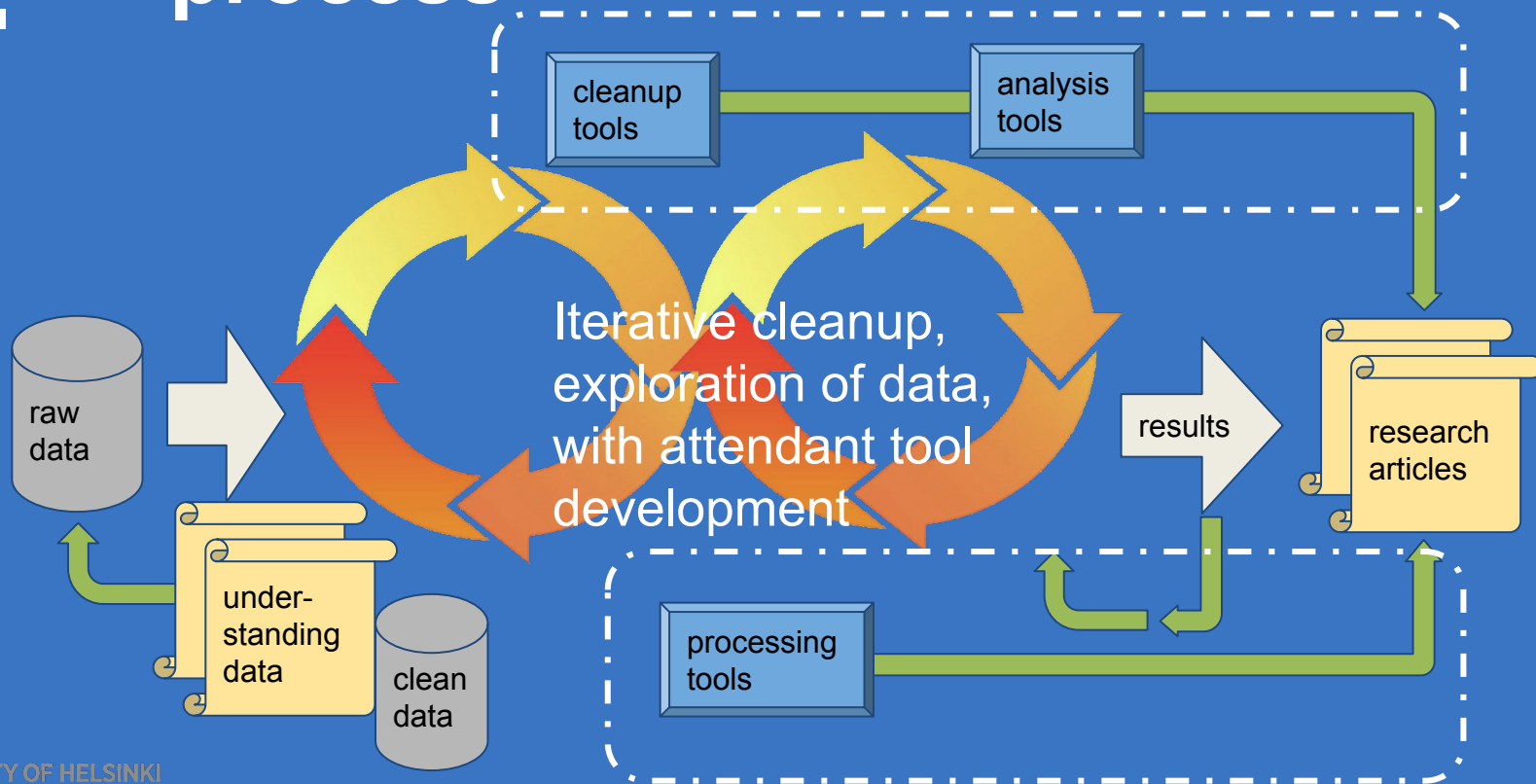


Digital humanities research process



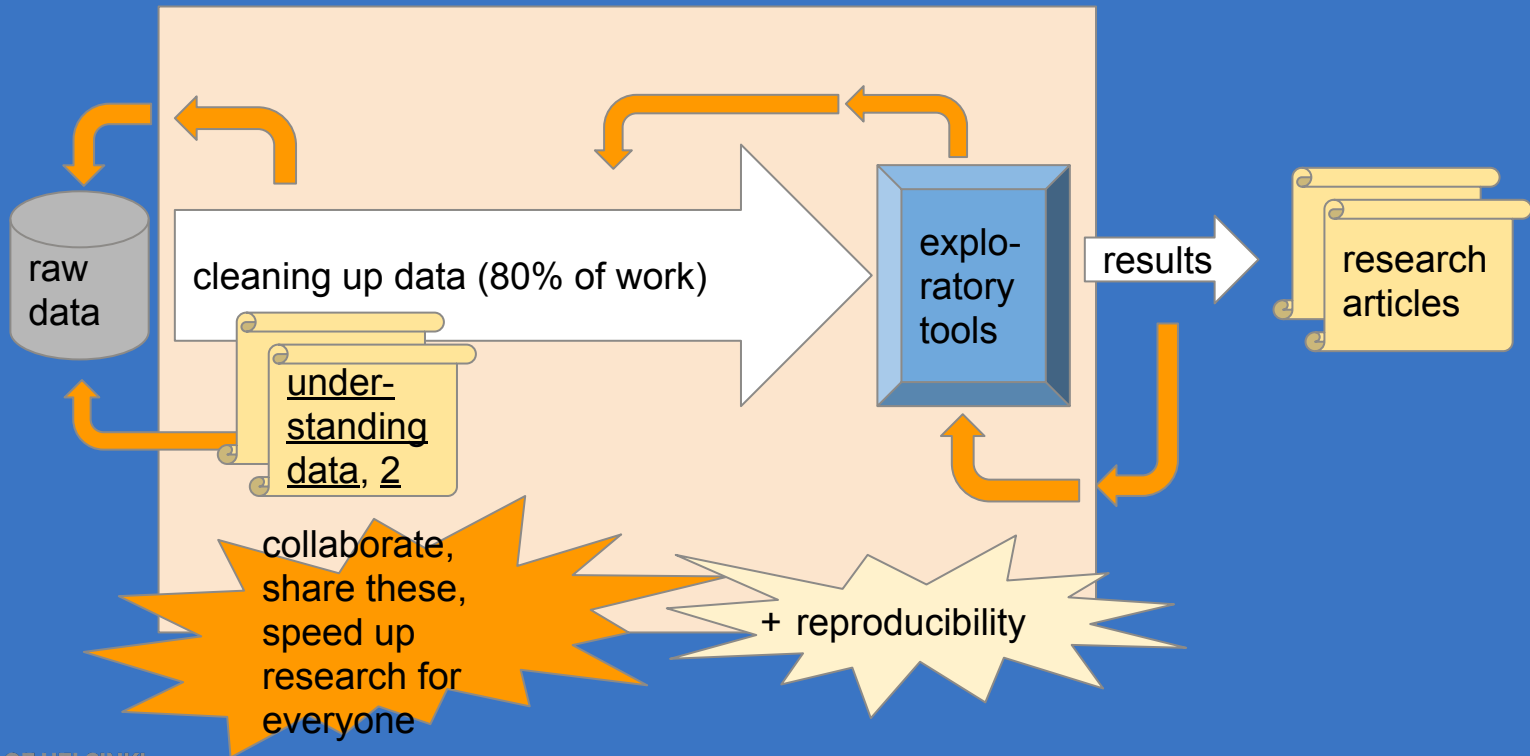


Digital humanities research process





Leverage collaboration, open science workflows to reduce individual workload





Tools to support research

Understand

[Aether](#)

[vocab.at](#)

[Voyager](#)

[Breve](#)

[CORE](#)

Import

[LAS](#)

[ARPA](#)

[Karma](#)

[OpenRefine](#)

Edit

[OpenRefine](#)

[FiCa](#)

[Wrangler](#)

[SAHA](#)

[Fibra](#)

[Snapper](#)

[nodegoat](#)

Reconcile

[Recon](#)

[Silk](#)

Organize

[SKOSJS](#)

Explore

[VISU](#)

[Palladio](#)

[Octavo](#)

[Khepri](#)

[nodegoat](#)

Publish

[LDF.fi](#)

- Chastelier, Jean
- Mersenne, Marin
- Bredeau, Claude
- Peiresc, Nicolas-Claude Fabri de
- Cornier, Robert**
- Mydorge, Claude
- Stanihurst, Henry de
- Lefebvre
- Unidentified sender
- François, Jean René
- Hoguette, Philippe Fortin de la
- Descartes, René
- Doni, Jean-Baptiste

Persons	Date range	Places
Cornier, Robert	1625-1628	Rouen

Notes

	Match	alabel	warning	plabel	link
0	None of the below				
1	Cornier, Robert, fl. 1625-1628			Rouen	[o]
2	Corker, Robert (fl. 1700)	A Cornishman			[o]
3	Ball, Robert, fl. 1634-1691	Letter-carrier for Robert Boyle			[o]
4	Bellarmino, Roberto Francesco Romolo, 1542-1621	Italian Jesuit and a Cardinal, Bellarmine, Robert; Bellarmin; Bellarmino, Roberto Francesco Romolo; Belarminus, Robertus		Rome	[o]



FiCa

roller (3)				RLR (3)	RLR (3)	
roller	rolers	yes		2	RLR	RLR
roller	roller	yes		3	RLR	RLR
roller	rollers	yes		6	RLR	RLR
ruler	ruler	unclear	er1; 1 in ed	5	RLR	RLR
rem-ber	rem-ber	no		2	RMBR	RMBR
					RMLR (2)	
					RMLR (2)	RMLR (2)
					RMLR (2)	RMLR (2)
rambler (2)						
rambler	rambler	unclear	some NP0	7	RMLR	RMLR
rambler	ramblers	no	NP0 (period	1	RMLR	RMLR
remem[[b]er	remem[[b]er	no		1	RMMB	RMMBR
					RMMR (8)	
					RMMR (7)	RMMR (7)
rememb=rs	rememb=r=ε	no		1	RMMR	RMMR
remember (2)					RMMR (2)	RMMR (2)
remember	rememb=r=	no		8	RMMR	RMMR
remember	remember	no		690	RMMR	RMMR
remember's	remember's	no		2	RMMR	RMMR
remembers	remembers	no		22	RMMR	RMMR
remembr	remembr	no		2	RMMR	RMMR
remmember	remmember	no		2	RMMR	RMMR
remembrancer	remembranc	yes	er1	1	RMMR	RMMRNSR
					RMNT (9)	
					RMNT (9)	RMNTR (9)
					RMNT (9)	RMNTR (9)
remainder	remainder	no	MC	1	RMNT	RMNTR

of small diameter which the seed is to be subjected to before it is exposed to the pressure of the great stone rollers: this he says is a late invention; but as it requires more workmanship than is easy to be had here I think cast iron rol

odel he has brought with him from England there is an apparatus for bruising the seed by making it pass between two iron rollers of small diameter which the seed is to be subjected to before it is exposed to the pressure of the great stone

<Q A 1783 FN SBENTHAM- <X SAMUEL BENTHAM- <P III,209- I [469 FROM SAMUEL BENTHAM-]] [TO JEREMY BENTHAM-] [ADDRESS-] Jere- Bentham Esq- / Lincoln's Inn / London Petersburg Sept. 13th O.S. 1783. I am at length taken into the service of this country. The rank given me is that of (Conseiller de la Cour), which is only equal to that of Lieutenant Colonel in the army. Considering that I had had no Military rank in any other country to found my pretensions on, (...) to bruise the seed in oil mills, and what objections he sees to such a substitute. In many places I imagine the rollers of stone may be cheaper; but that would not be the case here. Capper is returned and his having seen you all has attached me to him. It is probable we may be concerned together in the erection of an oil mill which gave rise to the above question. In a model he has brought with him from England there is an apparatus for bruising the seed by making it pass between two iron rollers of small diameter which the seed is to be subjected to before it is exposed to the pressure of the great stone

OED Oxford English Dictionary Quick search: Find word in dictionary GO

Lost for Words? | Advanced search | Help

Help on Search Results | Print | Email

Quick search results

Showing 1-5 of 5 results in 5 entries

Widen search? Find 'roller' in: » phrases (186) » definitions (290) » etymologies (49) » quotations (1600) » full text (751)

View as: List | Timeline Sort by: Entry | Frequency | Date

- roller, n.¹** View full entry 1295

...One of a number of (usually large) cylinders of wood or other hard material, sometimes attached to a framework, over which a heavy object can be passed...
- roller, n.²** View full entry 1678

...A jay-like bird, *Coracias garrulus* (family *Coraciidae*), having mainly greenish-blue plumage with dark blue wings and a chestnut back, noted for its characteristic tumbling display flight and found...

Your current search (entries): roller

Save Refine search

Refine your search

- Subject
- Language of Origin
- Region
- Usage
- Part of Speech
- Date of First Citation

My entries (2)

My searches (6)

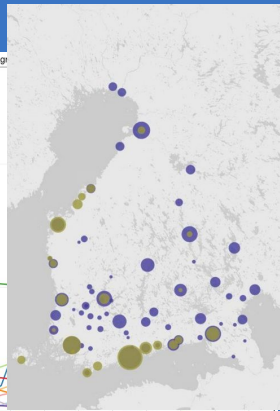
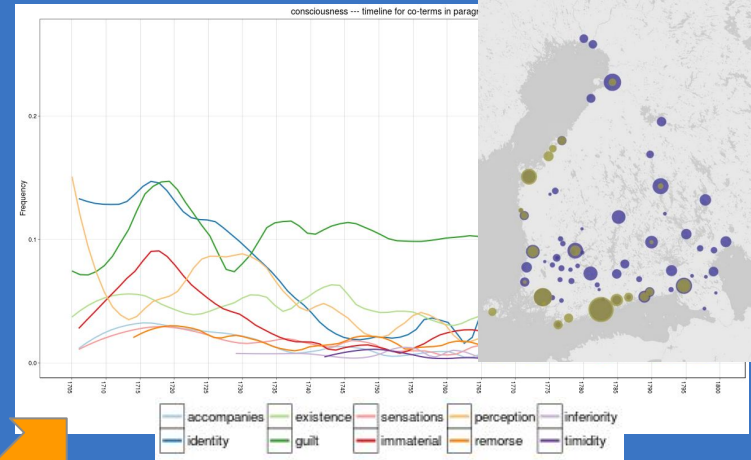
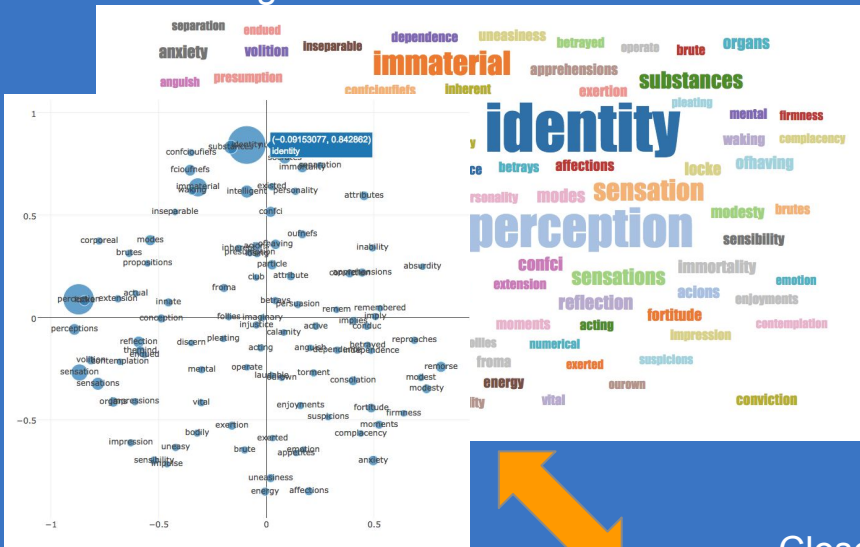
Jump to:

Entry ▾

- subduement, n.
- subduer, n.
- subduing, n.
- subduing, adj.
- subduple, adj.
- subduplicate, adj.
- subdural, adj.
- subduer, n.
- sub-echo, n.
- subedar, n.
- subedit, v.
- sub-editor, n.
- sub-official, adj.

Linguistic fingerprint, either of works or of words neighbourhood

Temporal/geographical perspective



Close reading

First 20 of 49085 Results

ESTCID	fullTitle	pubDateStart	pubDateEnd
P001934	The European magazine, and London review; containing the literature, history, politics, arts, manners and amusements of the age. By the Philological Society of London	17820000	18269999
	Thornton presented a petition from the Bank of England against the Bill, as a violation of public faith, an infringement of private, right, and establishing a dangerous precedent.		
	Fox said, he had two different objections to the Bill, the first of which was, that it was unjust to the public creditors at large, and weakening to the vital strength of public credit :- the second, that it was unjust to the Bank, as a trading company.		
	-howe'er, That the fg. of such property was the indisputable right of the Bank, and that the revenue arising therefrom was as much the property of the Bank, as the principal was the property of the public creditor.		
	He next proceeded to argue in support of his first objection to the Bill, namely, the injury it would occasion to public credit, and the injustice of it with respect to the public creditor.'		
	The contrast between the Public Creditor and Government tated particularly when, how, and where their Dividends were to be paid, and the Bank by that contrast was made a trustee; this Bill however would break that contract, and take from the Bank the trust before reposed in them; and it would be idle to say that a better security was given for even were a better security given, the contract ought not to be deviated from, unless with the consent of all parties.		
	The Banker again said was the trustee for the public , and could not, without a breach of public faith, have the trust taken away.		



OCR error handling

Configuration

Search

Max edit distance

Required common prefix length

Transposition is a single edit

Query

politeness

Keywords

- politerness (1) politeness (19035) politzeness (1)
- politenehs (4) politeess (1) politeless (3)
- politehess (1) politenesa (1) politenesj (3)
- politeneós (1) politen.ess (3) politiness (2)
- politeneps (3) politenees (16) politene.s (6)
- politeineess (2) politeness (1) politeness5 (1)
- politenejs (61) politene1s (4) politene'ss (1)
- politenes's (2) po'liteness (4) politenebs (1)
- politenesc (1) politenesl (1) pcliteness (2)
- politness (1) polivteness (1) poli:eness (2)
- politeness (49) politenss (2) positeness (2)
- politrness (2) politenegs (2) politeners (585)
- politenessa (1) politene3s (1) politenesr (1)
- politenesi (7) politcness (4) politenes3 (1)
- politenezs (1) polirteness (1) politeneds (7)
- politeneos (7) politeness (1) politeness1 (1)
- polileness (1) politeiess (1) politentess (2)
- peliteness (1) p.oliteness (6) politemess (1)
- politeneas (6) politencss (3) politen'ess (10)
- politenels (904) politenews (1) politenuss (1)
- polireness (4) politeness (1) p'oliteness (9)
- politenesk (1) politenfss (1) potiiteness (1)
- po.iteness (1) politenes.s (2) poljteness (1)
- politenerr (2) politenets (159) polite'ness (5)
- pol'teness (1) polite.ness (3) politeneis (311)
- politeness3 (1) politeniss (1) politenese (3)



HOME ADVANCED SEARCH

BASIC SEARCH > RESULTS

Search Results

Results for **Basic Search** (Entire Document Language)

Search within these results

politeneus

GO

Narrow results by subject area

[Religion and Philosophy \(1\)](#)

FI: BRA

The logo for FI: BRA features the text "FI: BRA" in a white, sans-serif font. Below the text, several thin white lines intersect to form a starburst or fiber-like pattern, with lines extending outwards from a central point.

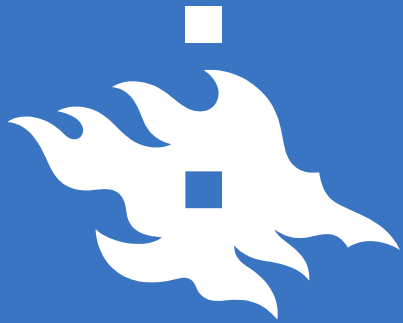
Connect digital research tools with libraries and archives around the world.

**with Dan Edelstein and Nicole Coleman,
Stanford**



Fibra – human scale tool for linked data that supports critical inquiry

1. Source information from linked datasets
2. Organize and add to data in order to build an argument
3. Capture both the data and the reasoning behind it so it will have context within the scholarly community
4. Publish the new knowledge to the community where it can be cited, re-used and built upon by others.



eetu.makela@helsinki.fi
<http://j.mp/s-makela>

This presentation:
<http://j.mp/dbhr-dhe>