Humanities research based on big cultural heritage data

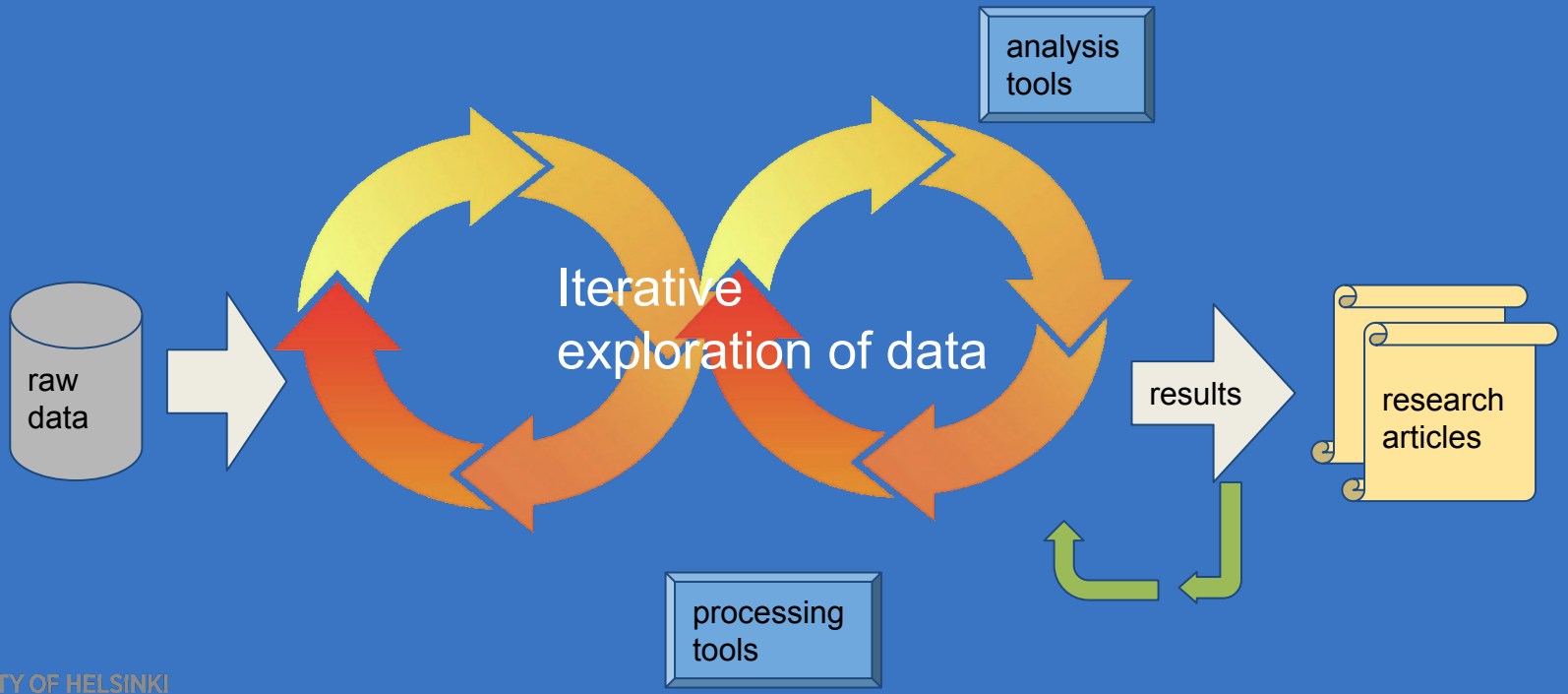
Eetu Mäkelä, D.Sc.
Assistant Professor in Digital Humanities / University of Helsinki
Docent (Adjunct Professor) in Computer Science / Aalto University

UNIVERSITY OF HELSINKI
FACULTY OF ARTS

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

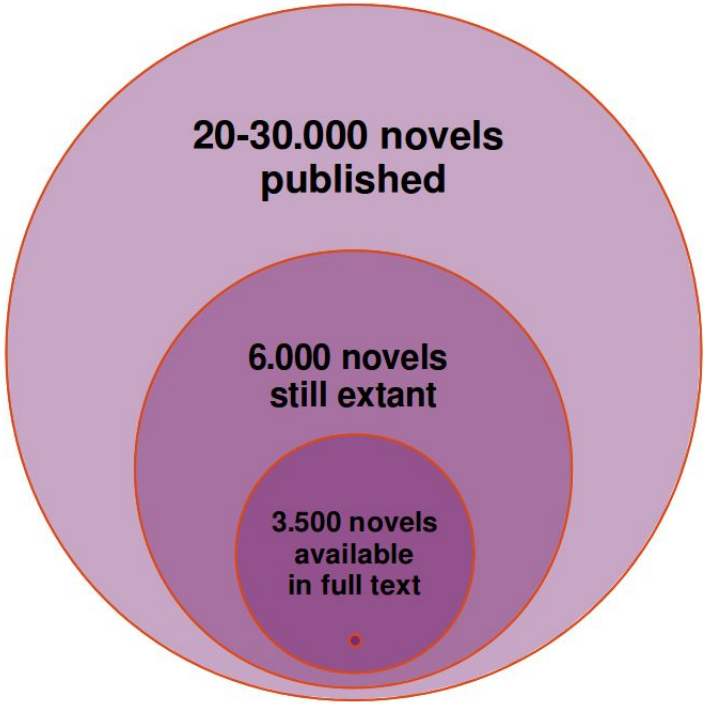
Research question ⇔ data

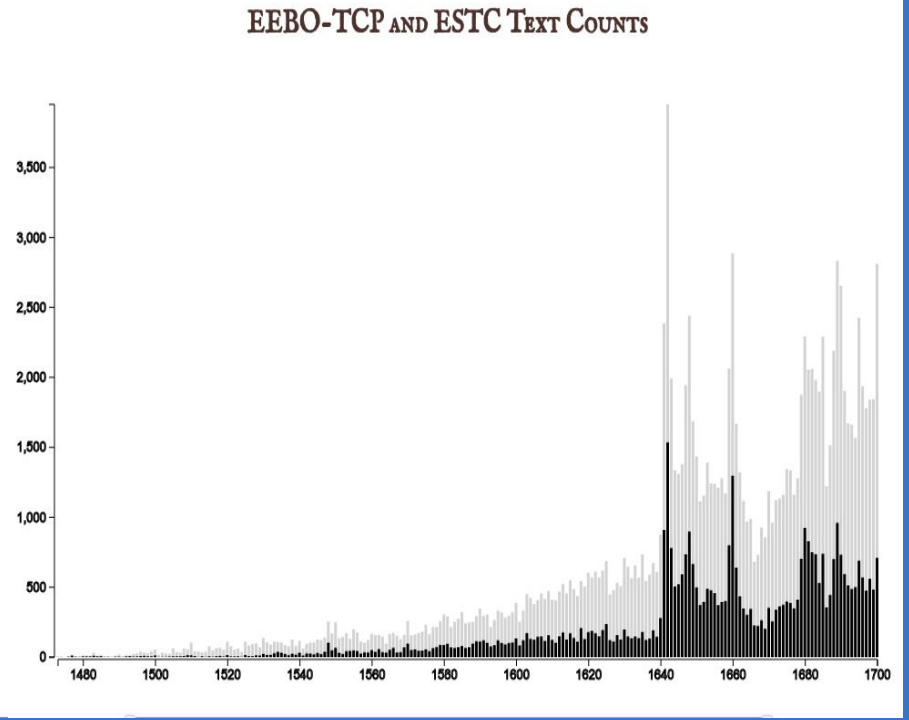
- “Which places published the most French philosophy in the 18th Century?” – I know, I’ll ask the French national library database
  - But is their data free of bias?
  - How is the information stored?

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?





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.

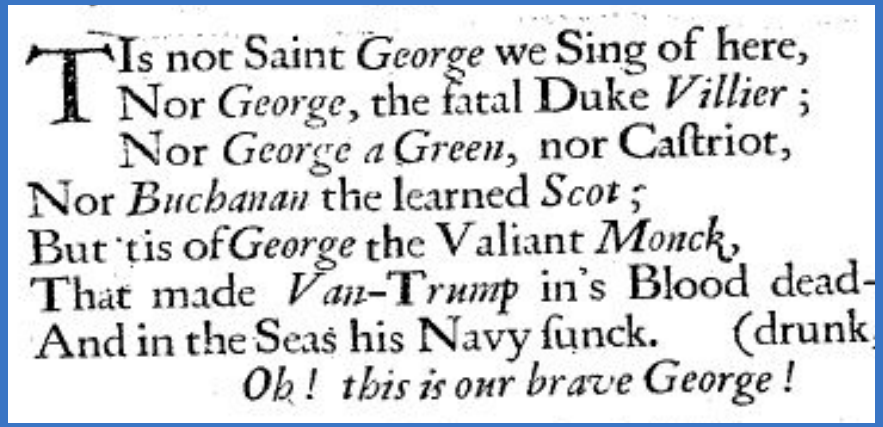
Data woes: viaf.org

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

<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



TIs 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 ſunck. (drunk
Oh ! this is our brave George !

lsw-not- Saint George we Sing of here,
Nor George, the fatal Duke Villier ;
Nor George a Green, nor Castriot,
Nor Buchanan the learned S cot q
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, Scotlande, and Irelande.*

CONTEYNING,

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

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

The deſcription and Chronicles of Irelande, likewiſe from the firſte originall of that Nation, vntill 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~1i |

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

~' .-bnEIs~l br~; <~5n~1 ~

.~1 1~t ~3mo71~l<~7noostI3o~rsd
~=i~mlm87il fif ~s ~

~' 3,Ilmo~l.6n3~nm/l7~=io\~ ~7g ~i

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

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

I I noin~Hodol~o]bsJni~qml '~1 11

1~.1 ~ ~1 11

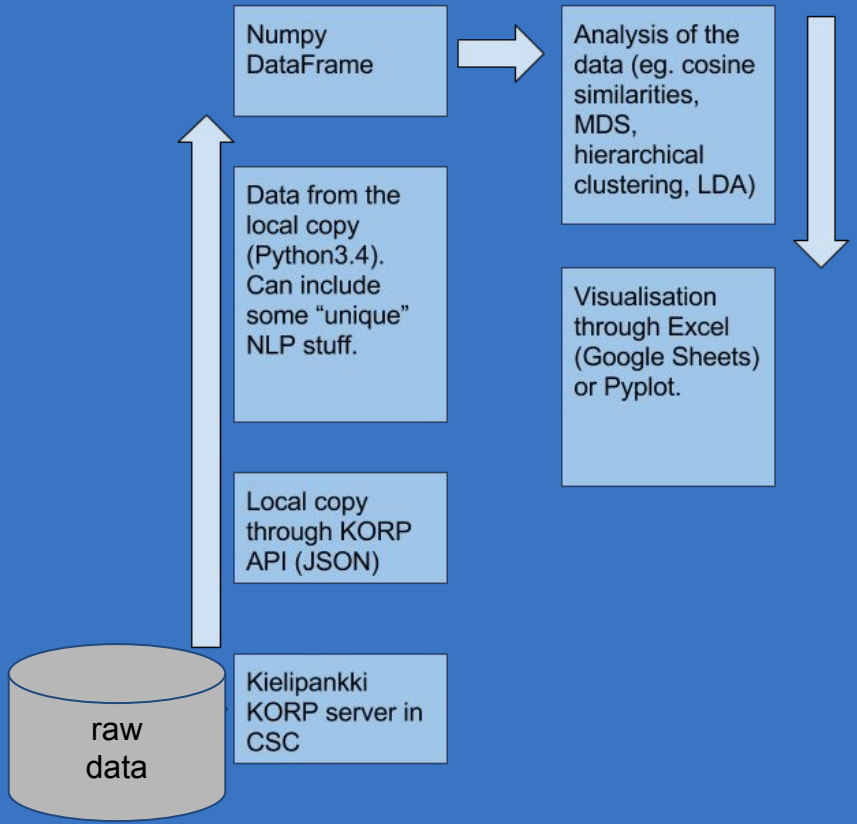
"' ~ ~ |'? ' ~ 9~ 9~] boO \~

,,.---. ~13 ~ ~ ~

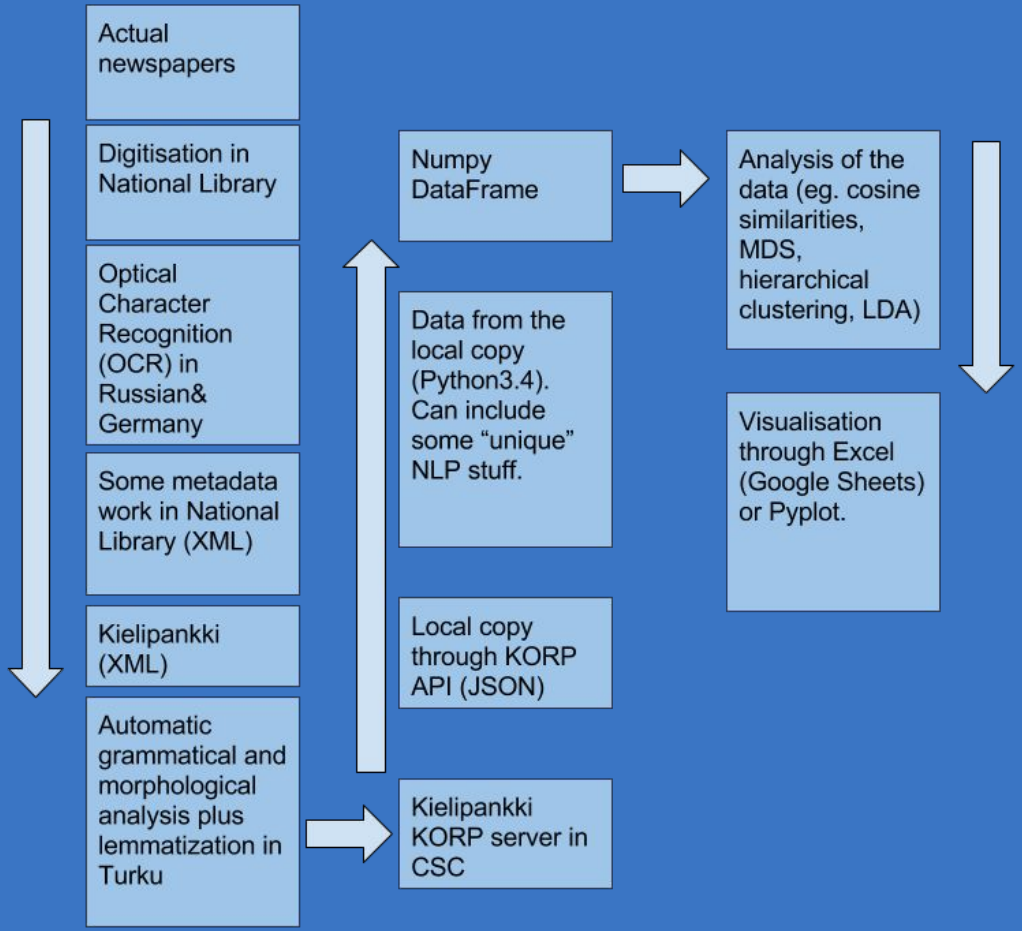
-: ~___ 1

.

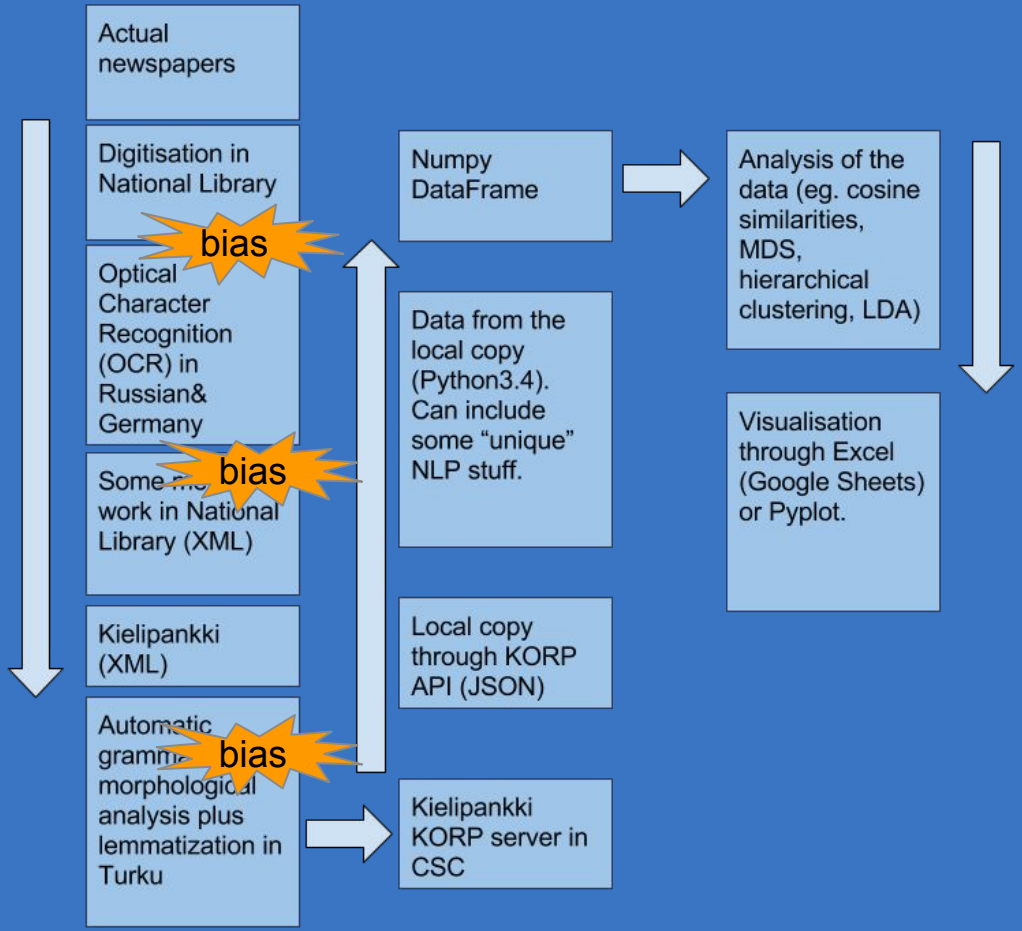
KLK Newspaper Pipeline: from archives to a researcher



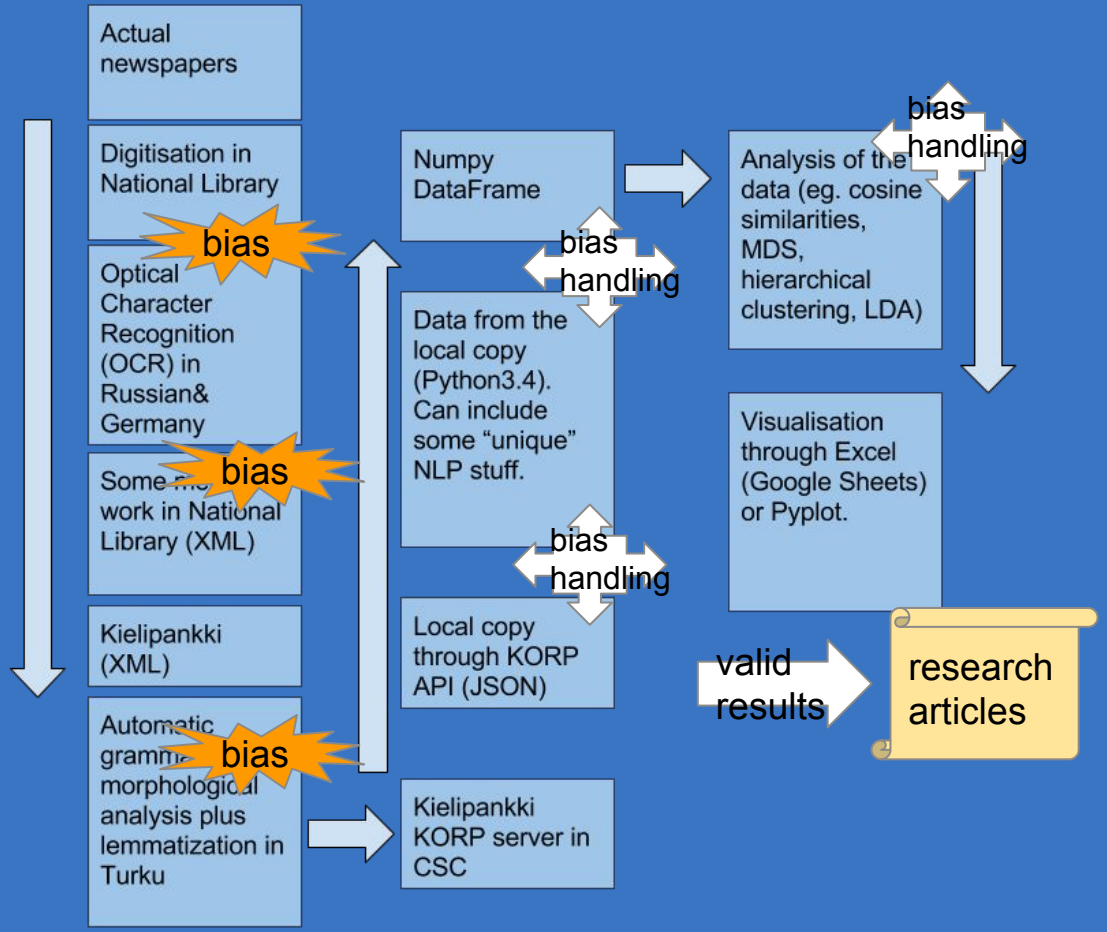
KLK Newspaper Pipeline: from archives to a researcher



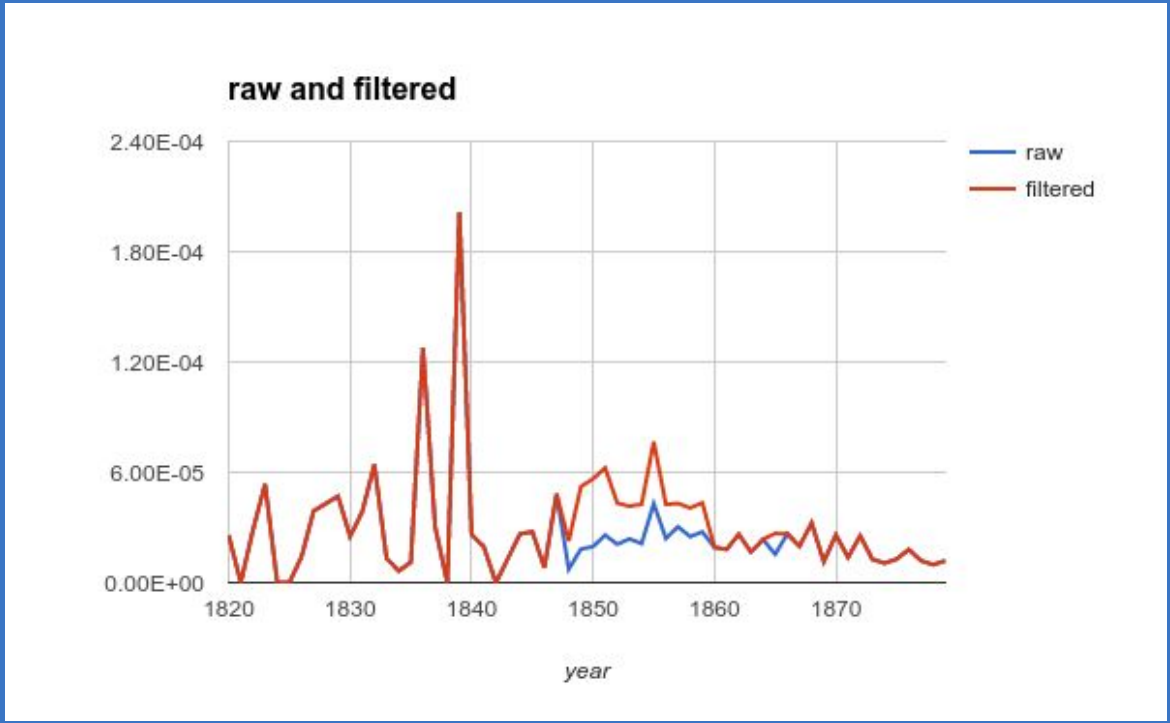
KLK Newspaper Pipeline: from archives to a researcher



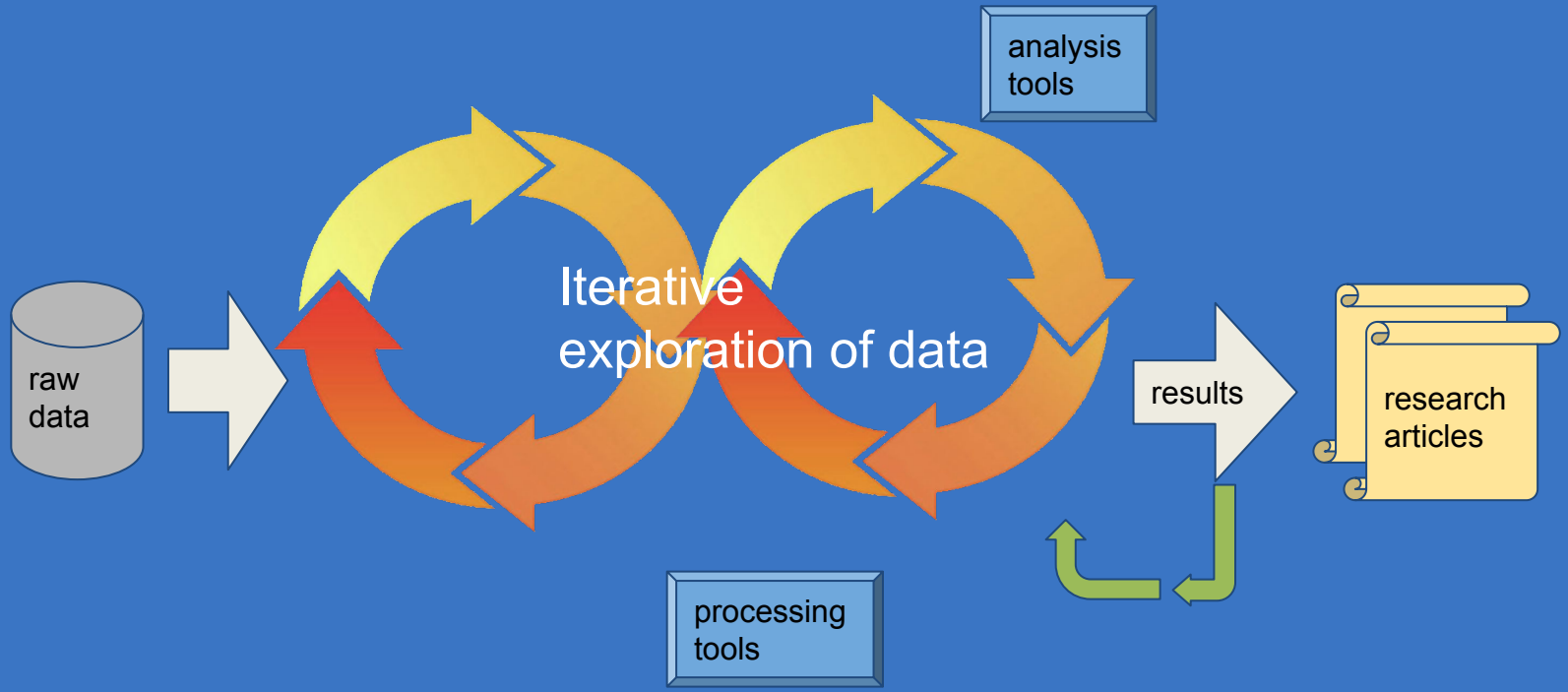
KLK Newspaper Pipeline: from archives to a researcher



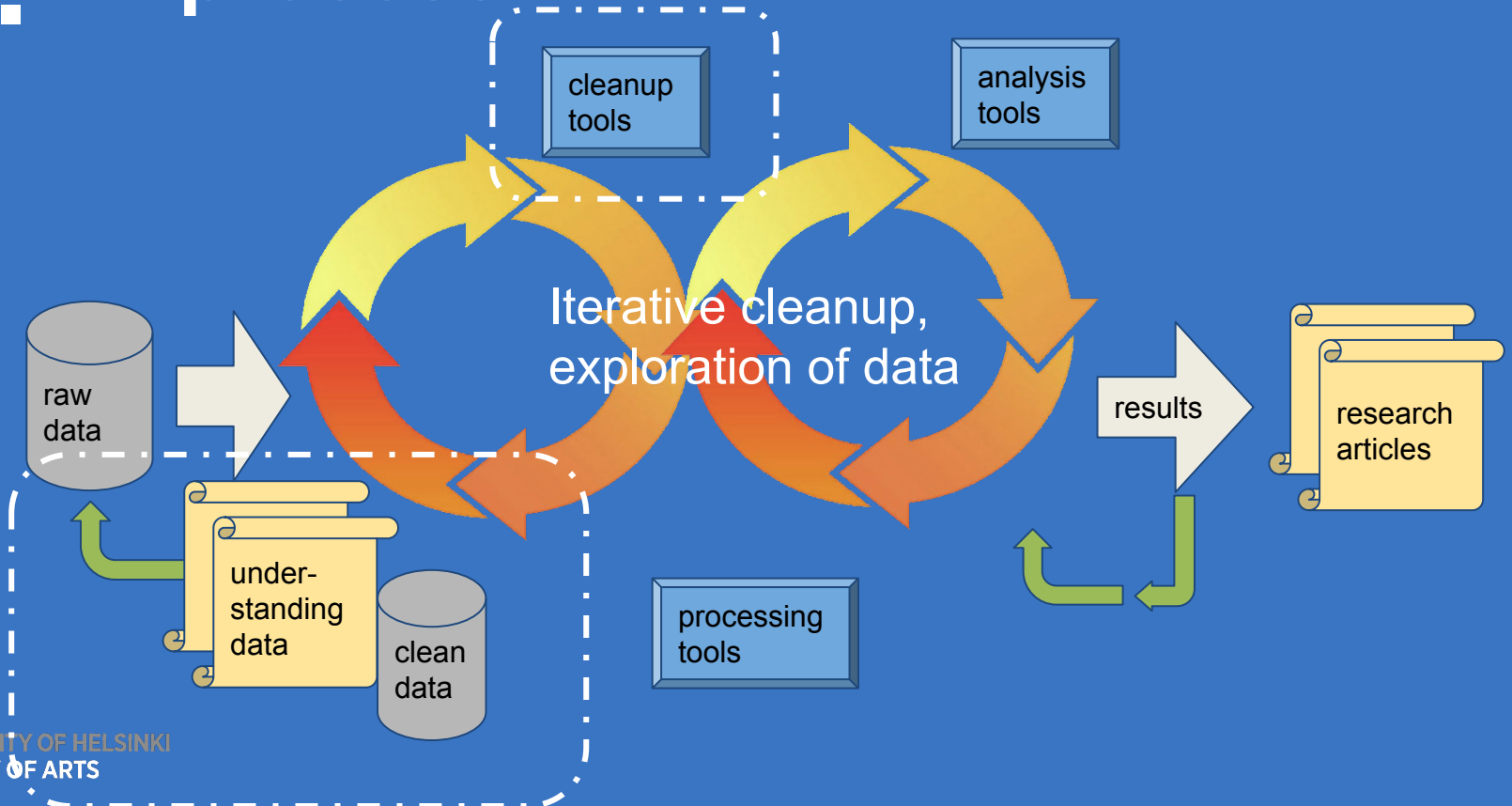
Data woes: National Newspaper Collection (KLK)



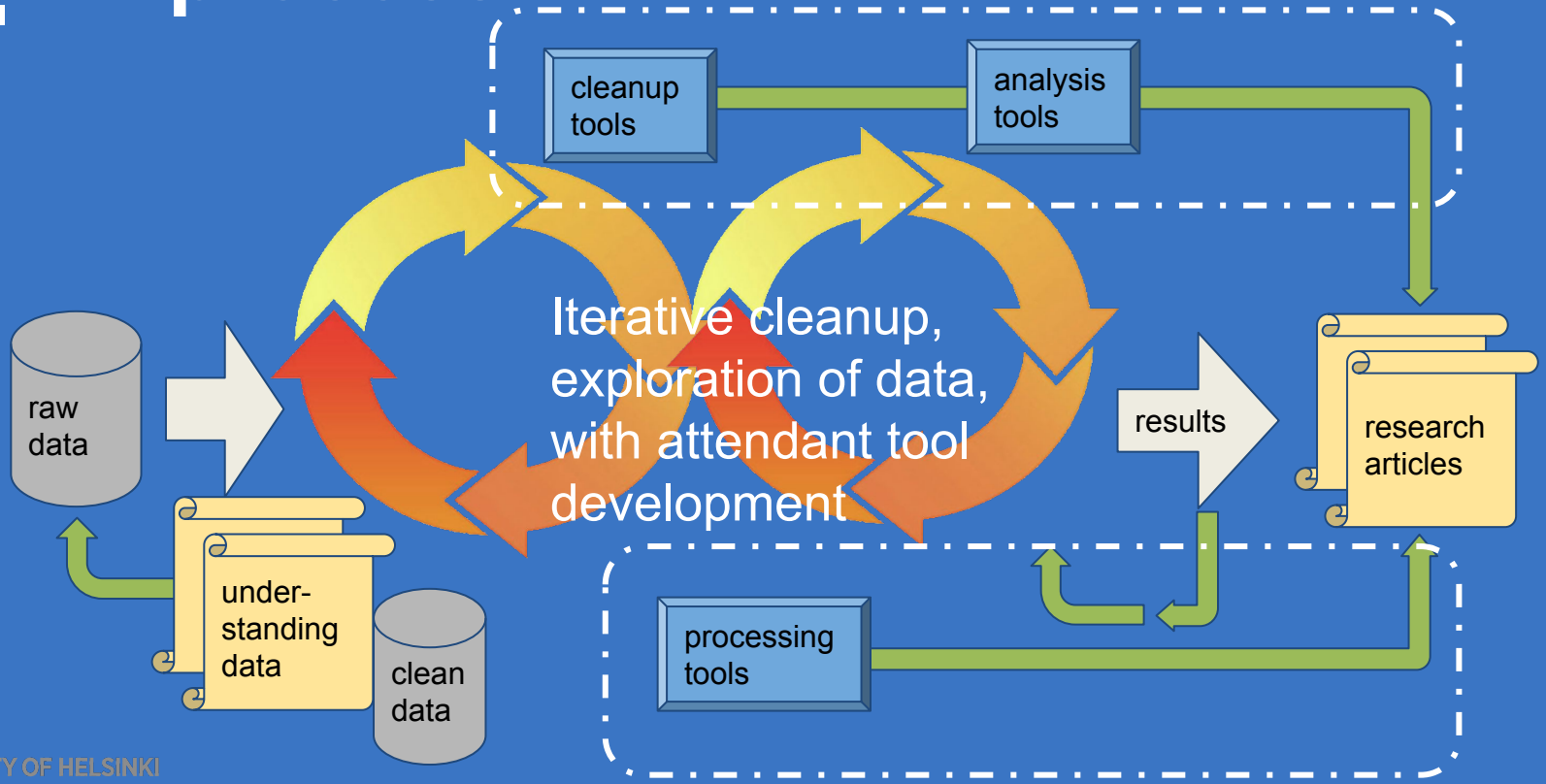
Digital humanities research process



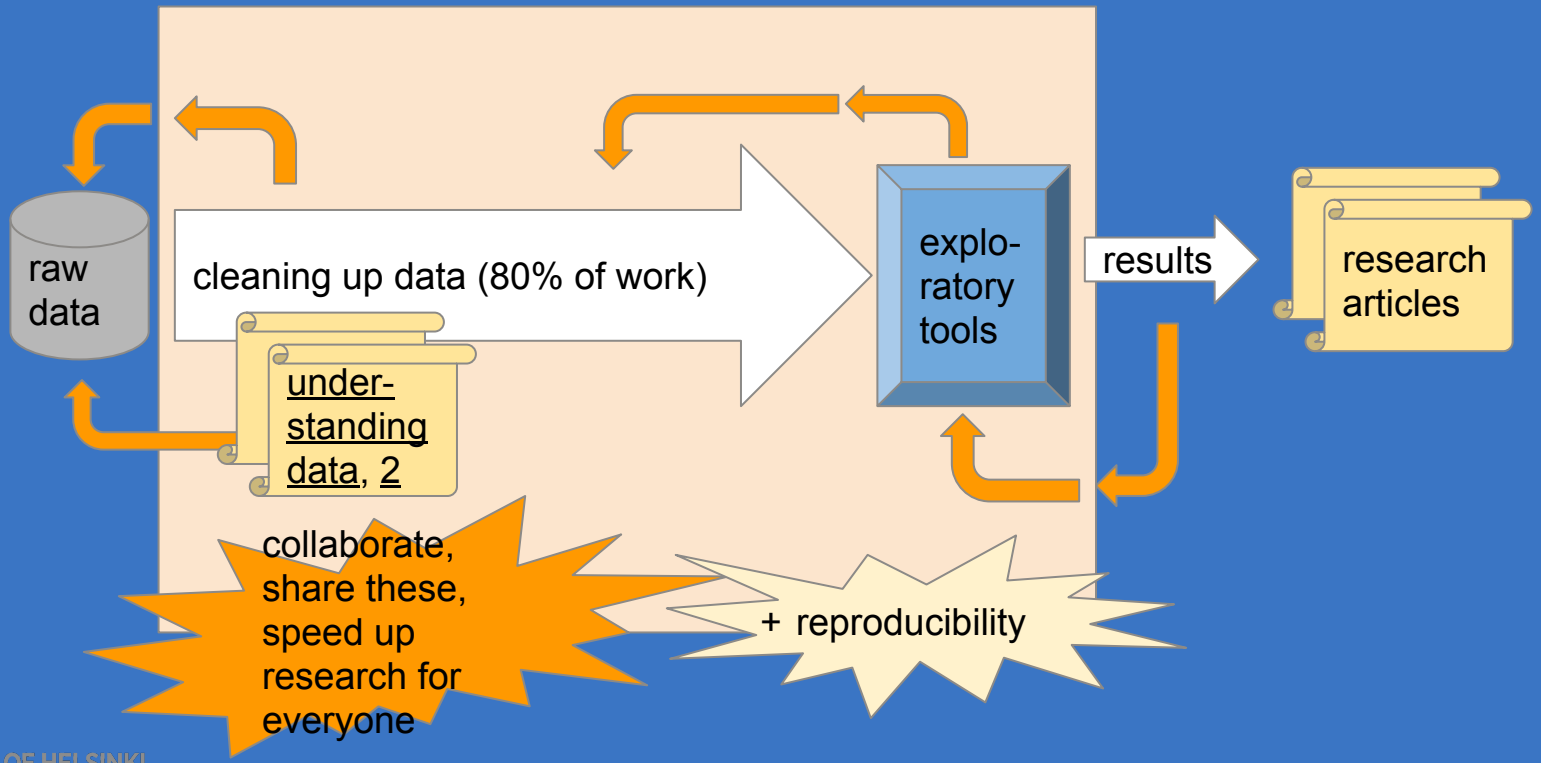
Digital humanities research process



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)**
rambler (2)					**RMLR (2)**	**RMLR (2)**
rambler	rambler	unclear	some NP0	7	RMLR	RMLR
rambler	ramblers	no	NP0 (perioc	1	RMLR	RMLR
remem{[b[]er	remem{[b[]e	no		1	**RMMB**	RMMBR
					RMMR (8)	
					RMMR (7)	**RMMR (7)**
rememb=r=s	rememb=r=s	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)**
remainder (9)					**RMNT (9)**	**RMNTR (9)**

of small diameter which the seed is to be subjected to before it is exposed to the pressure of the great stone **rolers:** 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

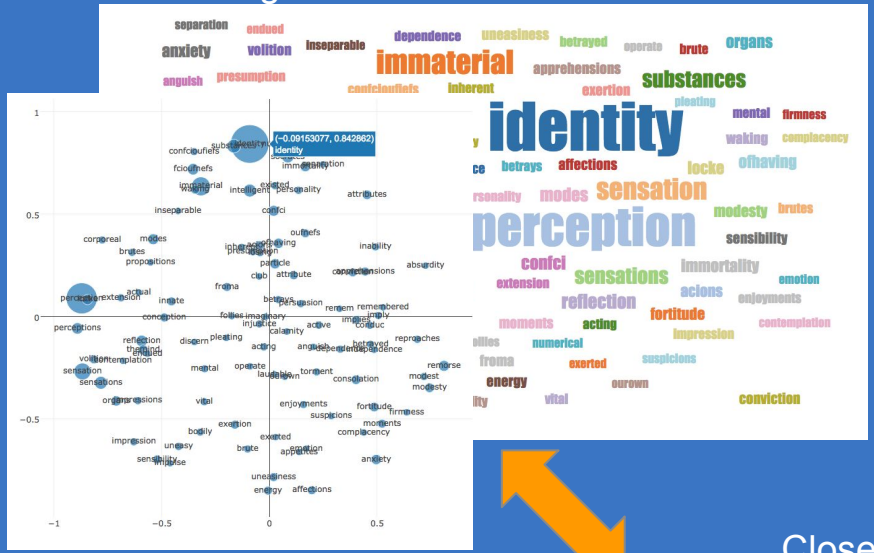
odel he has brought with him from England there is an apparatus for bruising the seed by making it pass between two iron **rolers** of small diameter which the seed is to be subjected to before it is exposed to the pressure of the great stone

<Q A 1783 FN SBENTHAM> <X SAMUEL BENTHAM> <P III,209> [] [\469 FROM SAMUEL BENTHAM\]]] [^TO JEREMY BENTHAM^] [\ADDRESSED:\] Jere=y= Bentham Esq=r= / Lincoln's Inn / London Petersbourg Septr. 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 pretentions 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 attatched 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 rolers of small diameter which the seed is to be subjected to before it is exposed to the pressure of the great stone

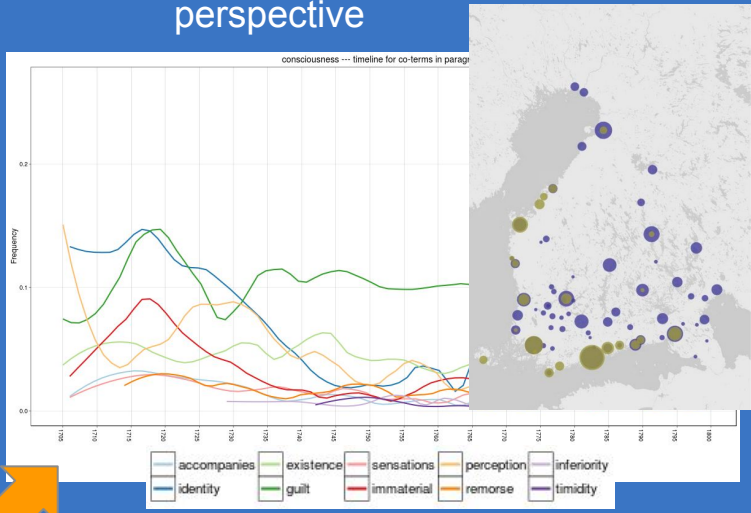
UNIVERSITY OF HELSINKI
FACULTY OF ARTS

from 6800 candidates to 1720 actual instances

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 fromthe Bank of England against the Bill, as aviolation of **public** faith, an infringement ofprivate, right, and eftablithing a dangerousprecedent.

Fox said, he hadtwo diftinEt objeations to the Bill, the firft'ofwhich was, that it was unjust to the publiccreditors at large, and weakening to thevital firength of **public** credit ;-the second,Ihat it was unjaft to tde Bank, as a tradingcompany.

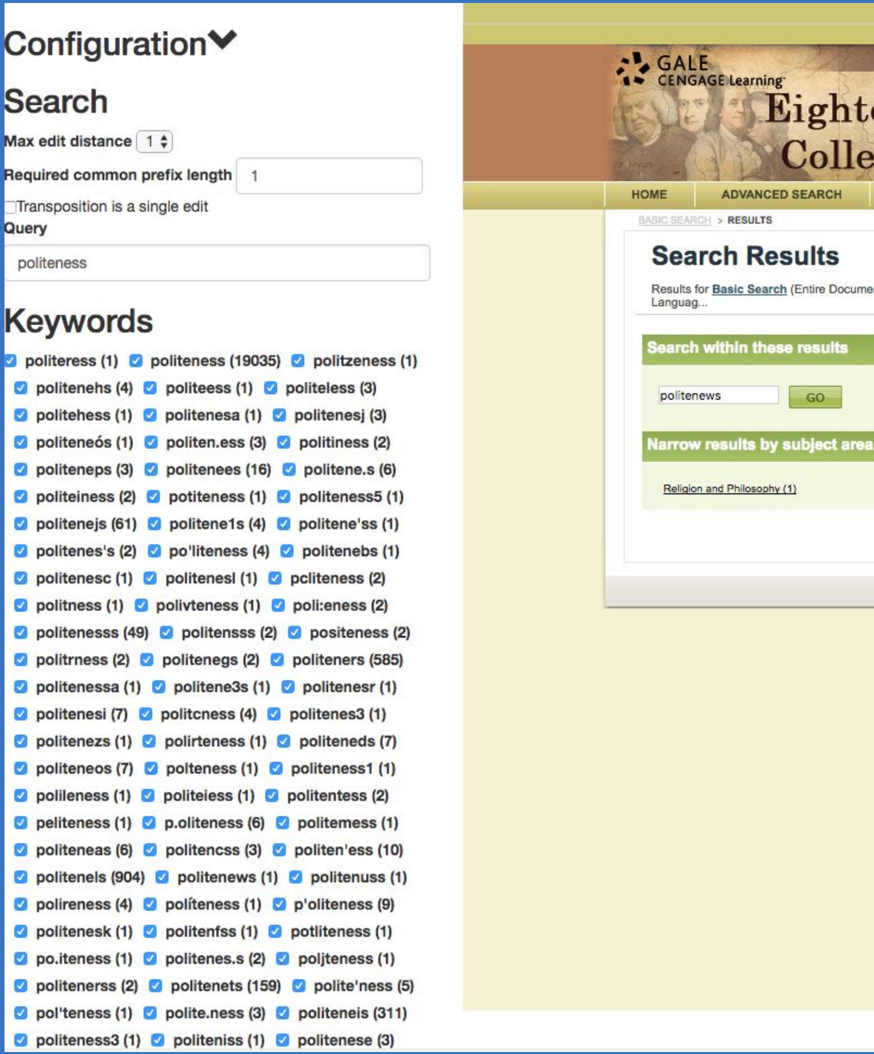
-houofe, That the fg. of such property was the indisputable tight o(the Bank, and that the revenue arising there.from was as much the property of theBank, as the principal was the property of'the **public** creditor.

He next proceeded toargue in support of his first objedion to theBill, namely, the injury it would occasion topublic credit, and the injustice of it with re-fpet to the **public** creditor.'

The contrastbetween the **Public** Creditor and Govern.ment tated partidularly when, how, andwhere their Dividends were to be paid, an4the Bank by that contrast was made liatrustee; this Bill however would break thatcontrat, and take from the Bank the trualbefore reposed in them; and it would b.idle to fay that a better security was given .for even were a better security given, thecontradt ought not to be deviated from, unlefi with the consent of all parties.

The Banuhe again said was the truflee for the **public**, andcould not, without a breach of **public** faith,have the trust taken away.

OCR error handling



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