



# Introduction to methods for digital humanities

Eetu Mäkelä, D.Sc.

Professor (tenure track) in Humanities–Computing Interaction /  
University of Helsinki

Docent (Adjunct Professor) in Computer Science / Aalto University

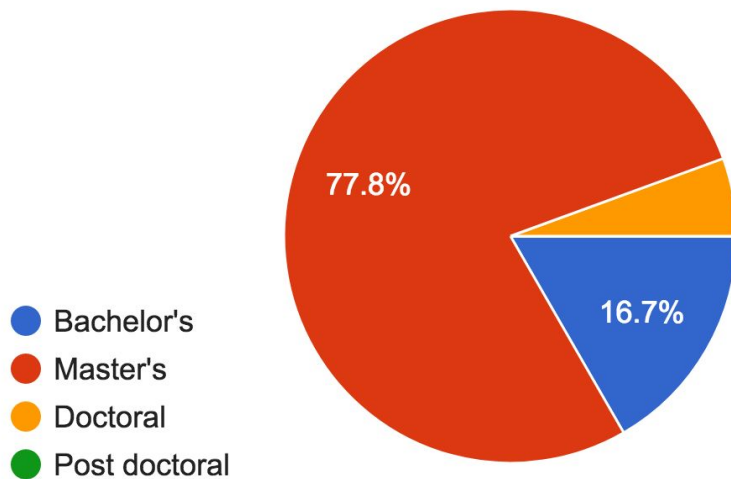


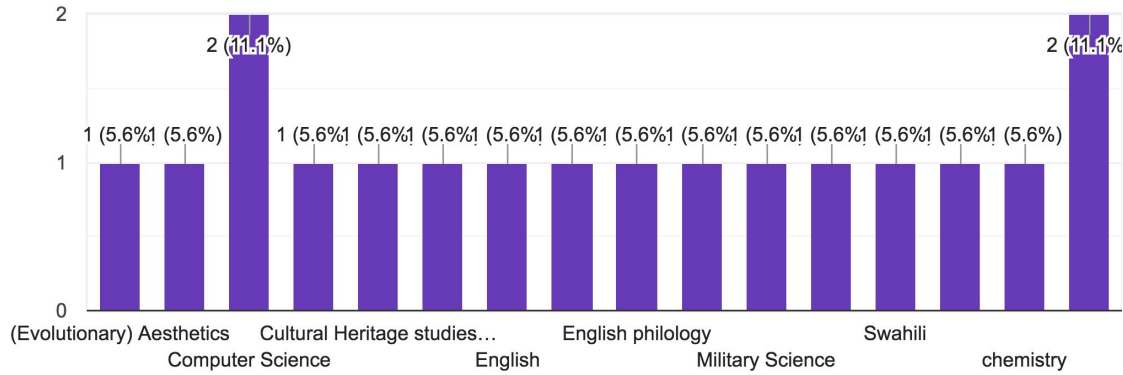
<http://presemo.helsinki.fi/meth4dh>

# Backgrounds

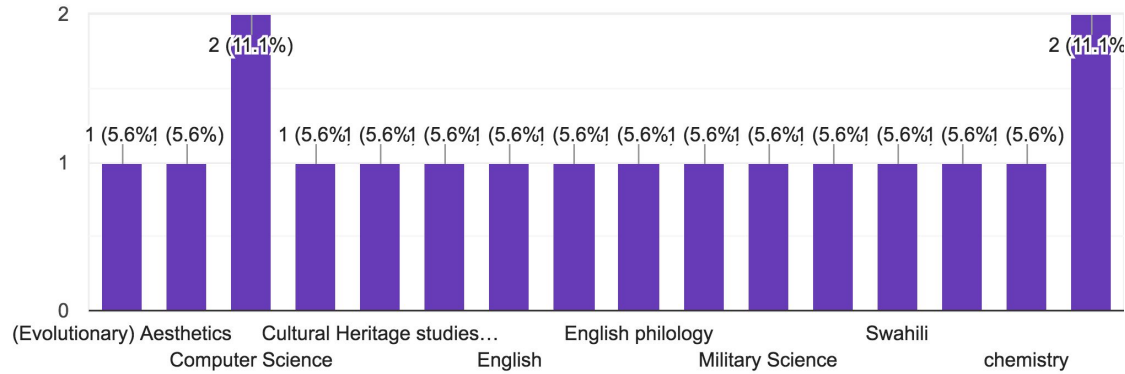
The level of studies you are currently doing

18 responses

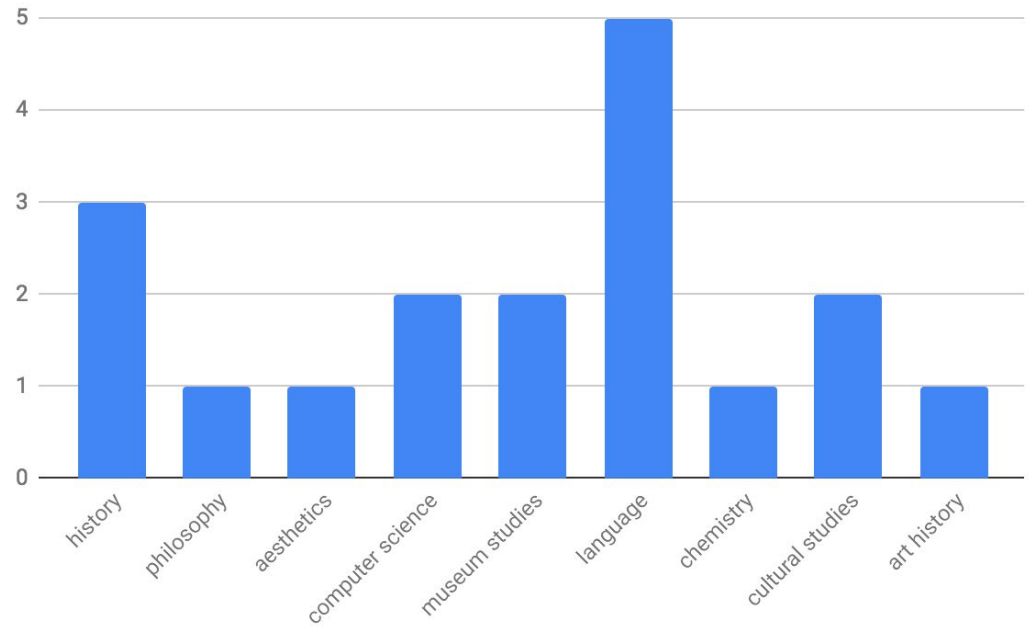




# Disciplinary backgrounds



# Disciplinary backgrounds





# Assignment: history of humanities computing

1. Did you manage to dig up any history of computational approaches applied to your own field?
2. Why didn't the 60's excitement ever turn into a true revolution?



# Recap: What to learn if you're a humanist?

1. Knowledge of easy to use end-user data processing and visualization tools
  - Easy to use for their intended purpose, but limited
2. Knowledge of the fundamentals concepts of programming
  - Frees you to process your data more efficiently
  - Allows you to more freely apply visualizations etc based on ready libraries and tutorials on the Internet
3. High-level understanding of what types of things can be accomplished with advanced CS methods
  - To be able to communicate in collaborative projects



# Recap: final projects



# Different types of data, data quality, available open datasets

Eetu Mäkelä, D.Sc.

Professor (tenure track) in Humanities–Computing Interaction /  
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!

*“Any sufficiently advanced technology is indistinguishable from magic.”*

- Arthur C. Clarke



# Research process - Magic (?)

- Hedge magic (spreadsheets, Excel graphs)
- Common ritual magic (statistics: correlation, regression, ANOVA, PCA)
  - Relatively simple, commonly understood formulae you could mostly go through with pen and paper if you wanted to
- Higher ritual magic (SVM, LSA, LDA, SnE)
  - More complex, harder to follow formulae, impossible to work through manually
- Black magic (most machine learning, neural networks)
  - You feed the machine both an input and a desired output, it derives a mostly unintelligible black box that links the two
- Flashy magic (proper visualizations)

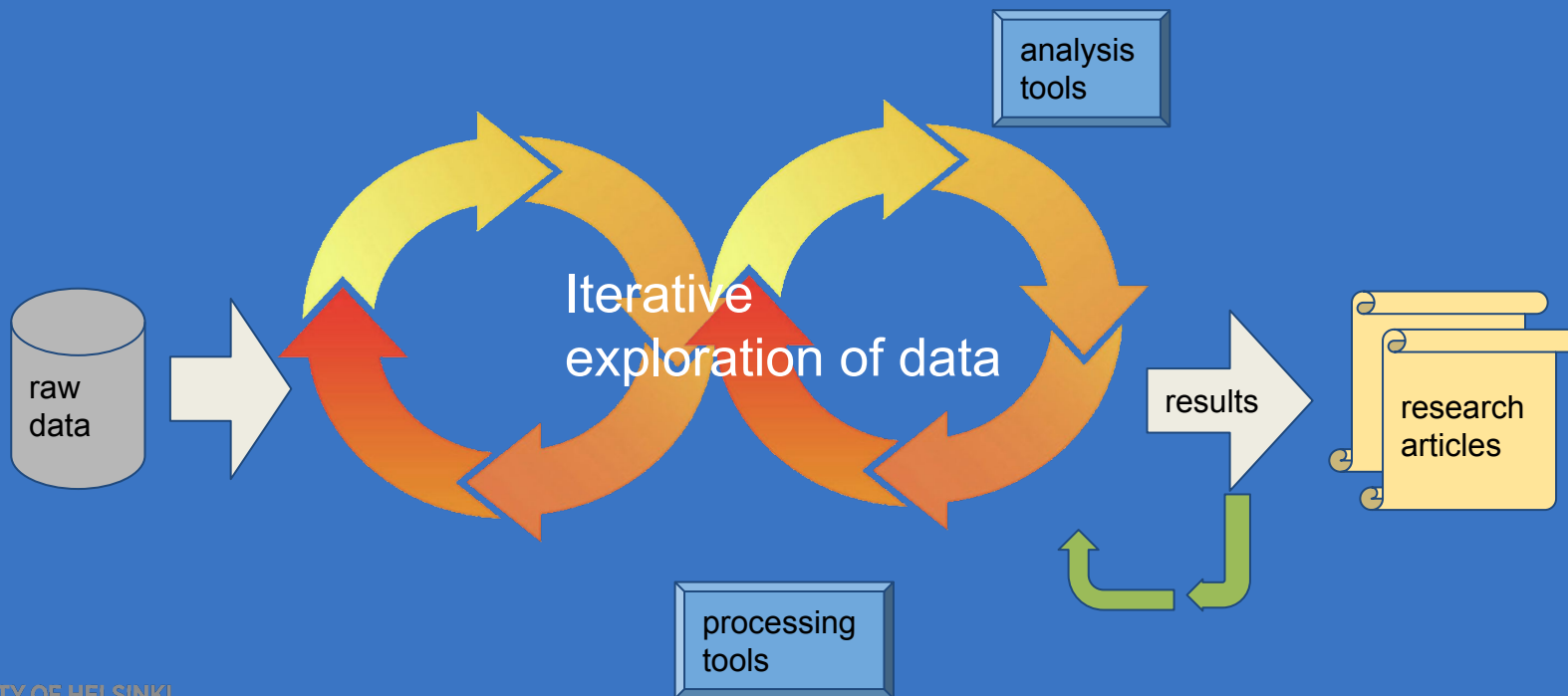


# Research process

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



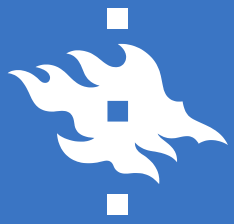
# Digital humanities research process





# 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, ...



# 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



# Archetypes of data

- Structured data - Excel sheets, databases, 2
- Unstructured data - text, 2, 3, sounds, images



## THE SONNETS

by William Shakespeare

1  
From fairest creatures we desire increase,  
That thereby beauty's rose might never die,  
But as the ripper should by time decease,  
His tender heir might bear his memory:  
But thou contracted to thine own bright eyes,  
Feed'st thy light's flame with self-substantial fuel,  
Making a famine where abundance lies,  
Thy self thy foe, to thy sweet self too cruel:  
Thou that art now the world's fresh ornament,  
And only herald to the gaudy spring,  
Within thine own buduriest thy content,  
And tender churl mak'st waste in niggarding:  
Pity the world, or else this glutton be,  
To eat the world's due, by the grave and thee.

2  
When forty winters shall besiege thy brow,  
And dig deep trenches in thy beauty's field,  
Thy youth's proud livery so gazed on now,  
Will be a tattered weed of small worth held:  
Then being asked, where all thy beauty lies,  
Where all the treasure of thy lusty days;

, praise.  
's use,  
of mine  
se'  
old,  
st it cold.

1	Work	Work_id	Author	Author_id	Date	Period	Genre	Authorial pre	Prefatory nan	Canonical pre	Canonical po	Non-canonic	Non-canonic	Roman
2	Letter to Antigone	1	Diodorus, Carystus	47.1	4.3 B.C.?	Hellenistic	medicine	yes	no					
3	Phenomena	2	Alexis	48	4.3 B.C.	Hellenistic	astronomy	yes	yes					
4	Size and Distance	3	Aristarchus	49	4.3 B.C.	Hellenistic	mathematics	no	no					
5	Letter to Herodotus	4	Epicurus	54	4.3 B.C.	Hellenistic	philosophy	yes	no					
6	Letter to Pythodorus	5	Epicurus	54	4.3 B.C.	Hellenistic	philosophy	yes	yes					
7	Letter to Menoeceus	6	Epicurus	54	4.3 B.C.	Hellenistic	philosophy	yes	yes					
8	On the Causes of the Winds	7	Theophrastus	58	4.3 B.C.	Hellenistic	philosophy	yes	no					
9	On Odors	9	Theophrastus	58	4.3 B.C.	Hellenistic	philosophy	yes	no					
10	History of Plants	10	Theophrastus	58	4.3 B.C.	Hellenistic	philosophy	yes	no					
11	On Weather Signs	11	Theophrastus	58	4.3 B.C.	Hellenistic	philosophy	yes	no					
12	Historia Plantarum	12	Aristophanes	62	3.8 B.C.	Hellenistic	philosophy	yes	no					
13	Conics 2	13	Apollonius Perg.	63	3.8 B.C.	Hellenistic	philosophy	yes	no					
14	Conics 3	14	Apollonius Perg.	63	3.8 B.C.	Hellenistic	philosophy	yes	no					
15	Conics 4	15	Apollonius Perg.	63	3.8 B.C.	Hellenistic	philosophy	yes	no					
16	Conics 5	16	Apollonius Perg.	63	3.8 B.C.	Hellenistic	philosophy	yes	no					
17	Conics 6	17	Apollonius Perg.	63	3.8 B.C.	Hellenistic	philosophy	yes	no					
18	Conics 7	18	Apollonius Perg.	63	3.8 B.C.	Hellenistic	philosophy	yes	no					
19	Two Mean Proportions	20	Eratosthenes	64.1	3.8 B.C.	Hellenistic	philosophy	yes	no					
20	On the Equilibrium of Solids	21	Archimedes	66	3.8 B.C.	Hellenistic	philosophy	yes	no					
21	On the Equilibrium of Solids	22	Archimedes	66	3.8 B.C.	Hellenistic	philosophy	yes	no					
22	On the Equilibrium of Solids	23	Archimedes	66	3.8 B.C.	Hellenistic	philosophy	yes	no					
23	Sphere and Cylinder	24	Archimedes	66	3.8 B.C.	Hellenistic	philosophy	yes	no					

### STN Online Database Archive

Browse Map Query Rank Compare Options Help & Resources Links

#### Query Books by Economic Sector

See books bought by members of Clergy

located everywhere

between 1st January 1769 and 31st December 1794

display results as table

Refresh

Export to Spreadsheet

	total number	as % for that period
1. Dictionnaire portatif de la campagne	1,400	20.74
2. Bon (le) père	1,001	14.83
3. Chrétien (le) par conviction et par sentiment	775	11.48
4. Catéchumène (le) instruit sous une forme nouvelle	612	9.07
5. Réflexions d'un homme de bon sens sur les comètes	312	4.62
6. (Abrégé de l'histoire sainte et du catéchisme)	311	4.61
7. Sermons sur les dogmes fondamentaux de la religion naturelle	222	3.29
8. Anarchie (l) médicale	202	2.99
9. (Thèses de Répécaud)	200	2.96
10. Bible	189	2.80



# Types of data

- Structured (databases) vs unstructured (text, image, video, audio)
- Clean vs messy



# Types of data

- Structured (databases) vs unstructured (text, image, video, audio)
- Clean vs messy
- **Biased?** ← **incomplete, messy, badly sampled**



# Research process

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



# Research process

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



# Data production pipelines: Early English Books Online



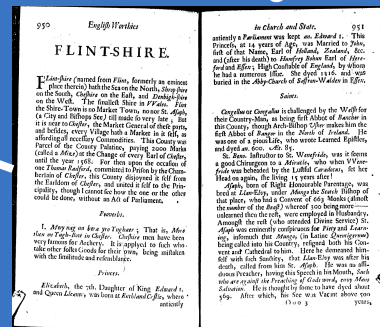
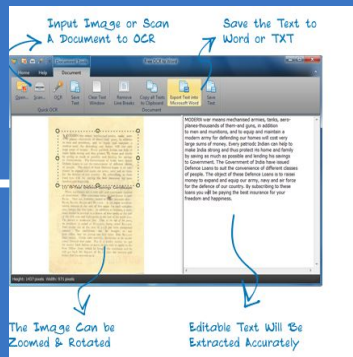
Physical books



Microfilms

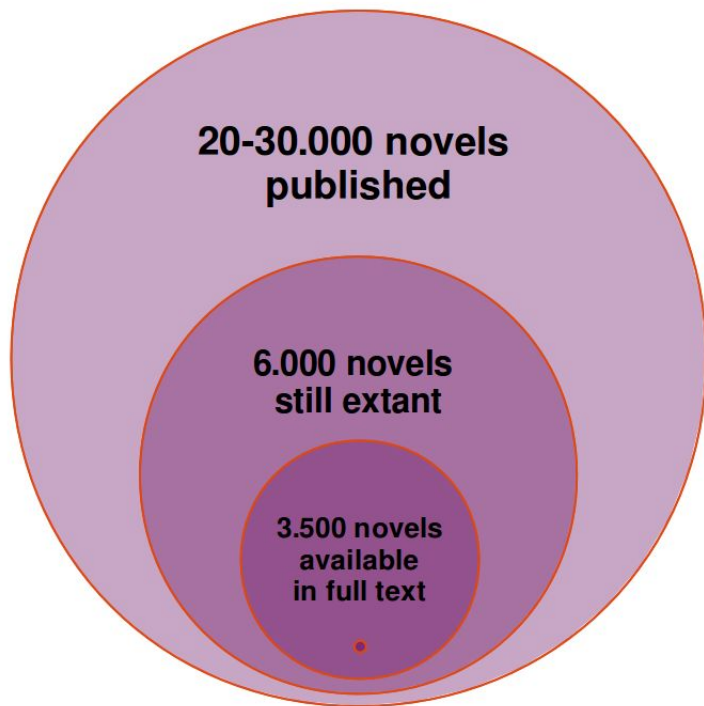


Electronic Image Scans

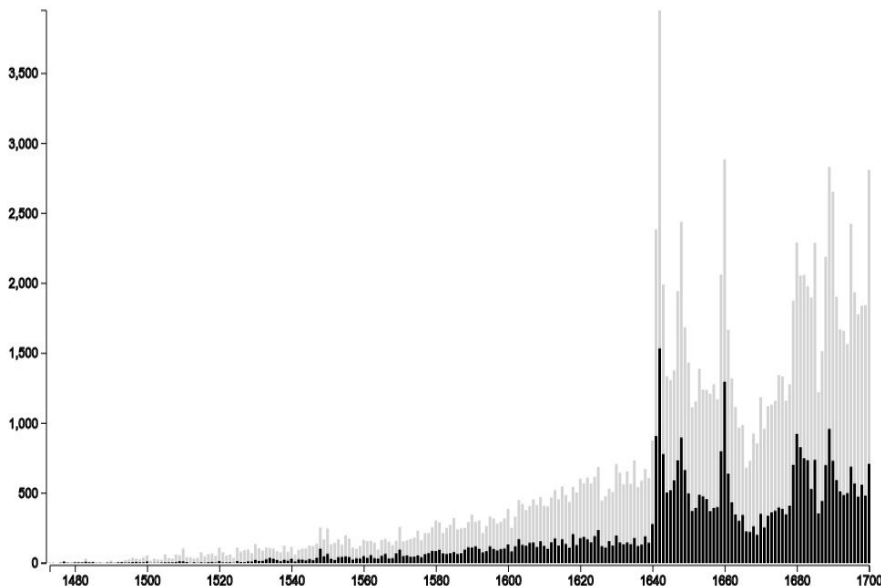




# But what's in there?



EEBO-TCP AND ESTC TEXT COUNTS





# Automatic OCR

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 '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 Redde. 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~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/17~=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 \~

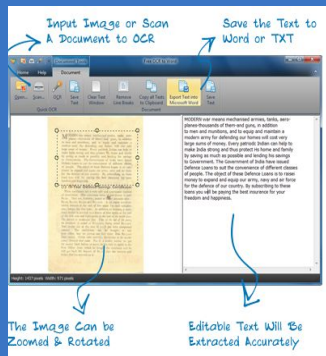
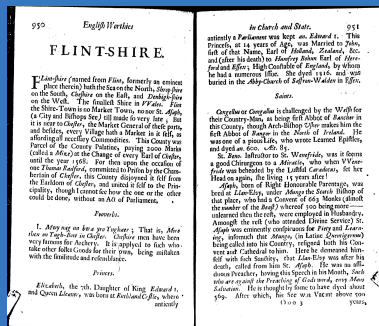
„---. ~13 ~ ~ ~

-: ~\_\_ 1

.



# Data production pipelines: EEBO-TCP (Text Creation Partnership)



Images + OCR

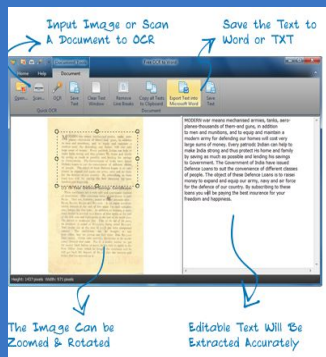
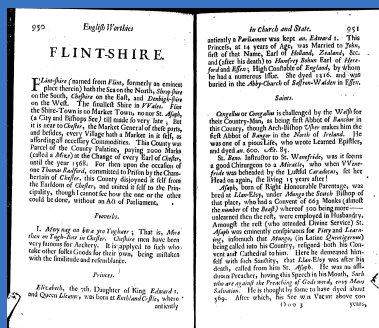
2x manual keying

# TCP

Transcribed by hand.  
Owned by libraries.  
Made for everyone.



# Data production pipelines: EEBO-TCP (Text Creation Partnership)



Images + OCR

2x manual keying

TCP I and TCP II are now available on EEBO, adding transcriptions of approximately **50%** of the texts on EEBO.



Transcribed by hand.  
Owned by libraries.  
Made for everyone.



1. All material should be recorded in the form in which it appears in the book: do not attempt to correct spelling or typographic error.
2. Illegible text, missing and damaged text, or clear but unrecognized symbols all will require some attention from us. Two extremes should be avoided as far as possible: (1) using the illegibility markers promiscuously to avoid capturing text about which there is some difficulty; and (2) "creative" capture of text that really cannot be read (from the [EBO TCP keying instructions](#))

## FLINTSHIRE.

**F**lint-shire (named from *Flint*, formerly an eminent place therein) hath the Sea on the North, *Shrop-shire* on the South, *Cheeshire* on the East, and *Denbigh-shire* on the West. The smallest Shire in *VVales*. *Flint* the Shire-Town is no Market Town, nor *St. Asaph*, (a City and Bishops See) till made so very late; But it is near to *Chester*, the Market General of these parts, and besides, every Village hath a Market in it self, as affording all necessary Commodities. This County was Parcel of the County Palatine, paying 2000 Marks (called a *Mize*) at the Change of every Earl of *Chester*, until the year 1568. For then upon the occasion of one *Thomas Radford*, committed to Prison by the Chamberlain of *Chester*, this County disjoyned it self from the Earldom of *Chester*, and united it self to the Principality, though I cannot fee how the one or the other could be done, without an Act of Parliament.

## Proverbs.

1. *Mwy nag un bwa yro Ynghaer*; That is, *More then on Yugh-Bow in Chester*. *Cheeshire* men have been very famous for Archery. It is applied to such who take other folks Goods for their own, being mistaken with the similitude and resemblance.

## Princes.

*Elizabeth*, the 7th. Daughter of King *Edward 1.* and Queen *Eleonor*, was born at *Rutland Castle*, where antiently

antiently a *Parliament* was kept *an. Edward 1.* This Princefs, at 14 years of Age, was Married to *John*, first of that Name, Earl of *Holland*, *Zealand*, &c. and (after his death) to *Humphrey Bohun* Earl of *Hereford* and *Essex*; High Constable of *England*, by whom he had a numerous Issue. She dyed 1316. and was buried in the *Abby-Church* of *Saffron-Walden* in *Essex*.

## Saints.

*Congellus* or *Comgallus* is challenged by the *Welsh* for their Country-Man, as being first Abbot of *Bancher* in this County, though Arch-Bishop *Usher* makes him the first Abbot of *Bangor* in the North of *Ireland*. He was one of a pious Life, who wrote Learned Epistles, and dyed *an. 600. Aet. 85.*

*St. Beno.* Instructor to *St. Wenefride*, was it seems a good Chirurgion to a *Miracle*, who when *VVenefride* was beheaded by the Lustful *Caradocus*, set her Head on again, she living 15 years after!

*Asaph*, born of Right Honourable Parentage, was bred at *Llan-Eloy*, under *Mungo* the Scotch Bishop of that place, who had a Convent of 663 Monks (almost the number of the *Beast*) whereof 300 being more—unlearned then the rest, were employed in Husbandry. Amongst the rest (who attended Divine Service) *St.*

*Asaph* was eminently conspicuous for Piety and Learning, inasmuch that *Mungo*, (in Latine *Quenigernus*) being called into his Country, resigned both his Convent and Cathedral to him. Here he demeaned himself with such Sanctity, that *Llan-Eloy* was after his death, called from him *St. Asaph*. He was an assiduous Preacher, having this Speech in his Mouth, *Such who are against the Preaching of Gods word, envy Mans Salvation*. He is thought by some to have dyed about 569. After which, his See was Vacant above 500

000 3 years,

## FLINTSHIRE.

**F**lint-shire (named from *Flint*, formerly an eminent place therein) hath the Sea on the North, *Shrop-shire* on the South, *Cheeshire* on the East, and *Denbigh-shire* on the West. The smallest Shire in *VVales*. *Flint* the Shire-Town is no Market Town, nor *St. Asaph*, (a City and Bishops See) till made so very late; But it is near to *Chester*, the Market General of these parts, and besides, every Village hath a Market in it self, as affording all necessary Commodities. This County was Parcel of the County Palatine, paying 2000 Marks (called a *Mize*) at the Change of every Earl of *Chester*, until the year 1568. For then upon the occasion of one *Thomas Radford*, committed to Prison by the Chamberlain of *Chester*, this County disjoyned it self from the Earldom of *Chester*, and united it self to the Principality, though I cannot fee how the one or the other could be done, without an Act of Parliament.

## Proverbs.

1. *Mwy nag un bwa yro Ynghaer*; That is, *More then on Yugh-Bow in Chester*. *Cheeshire* men have been very famous for Archery. It is applied to such who take other folks Goods for their own, being mistaken with the similitude and resemblance.

## Princes.

*Elizabeth*, the 7th. Daughter of King *Edward 1.* and Queen *Eleonor*, was born at *Rutland Castle*, where antiently

antiently a *Parliament* was kept *an. Edward 1.* This Princefs, at 14 years of Age, was Married to *John*, first of that Name, Earl of *Holland*, *Zealand*, &c. and (after his death) to *Humphrey Bohun* Earl of *Hereford* and *Essex*; High Constable of *England*, by whom he had a numerous Issue. She dyed 1316. and was buried in the *Abby-Church* of *Saffron-Walden* in *Essex*.

## Saints.

*Congellus* or *Comgallus* is challenged by the *Welsh* for their Country-Man, as being first Abbot of *Bancher* in this County, though Arch-Bishop *Usher* makes him the first Abbot of *Bangor* in the North of *Ireland*. He was one of a pious Life, who wrote Learned Epistles, and dyed *an. 600. Aet. 85.*

*St. Beno.* Instructor to *St. Wenefride*, was it seems a good Chirurgion to a *Miracle*, who when *VVenefride* was beheaded by the Lustful *Caradocus*, set her Head on again, she living 15 years after!

*Asaph*, born of Right Honourable Parentage, was bred at *Llan-Eloy*, under *Mungo* the Scotch Bishop of that place, who had a Convent of 663 Monks (almost the number of the *Beast*) whereof 300 being more—unlearned then the rest, were employed in Husbandry. Amongst the rest (who attended Divine Service) *St.*

*Asaph* was eminently conspicuous for Piety and Learning, inasmuch that *Mungo*, (in Latine *Quenigernus*) being called into his Country, resigned both his Convent and Cathedral to him. Here he demeaned himself with such Sanctity, that *Llan-Eloy* was after his death, called from him *St. Asaph*. He was an assiduous Preacher, having this Speech in his Mouth, *Such who are against the Preaching of Gods word, envy Mans Salvation*. He is thought by some to have dyed about 569. After which, his See was Vacant above 500

000 3 years,

Amongst the rest (who attended Divine Service) St. Afaph was eminently conspicuous for Piety and Learning, infomuch that Mungo, (in Latine **Quentigernus**) being called into his Country, resigned both his Convent and Cathedral to him. Here he demeaned himself with such Sanctity, that Llan-Elvy was after his death, called from him St. Afaph. He was an assiduous Preacher, having this Speech in his Mouth, Such who are against the Preaching of Gods word, envy Mans Salvation. He is thought by some to have dyed about 569. After which, his See was Vacant above 500 years

Amongst the rest (who attended Divine Service) St. *Afaph* was eminently conspicuous for *Piety* and *Learning*, infomuch that *Mungo*, (in Latine **Quentigernus**) being called into his Country, resigned both his Convent and Cathedral to him. Here he demeaned himself with such Sanctity, that *Llan-Elvy* was after his death, called from him St. *Afaph*. He was an assiduous Preacher, having this Speech in his Mouth, *Such who are against the Preaching of Gods word, envy Mans Salvation*. He is thought by some to have dyed about 569. After which, his See was Vacant above 500 years,

Amongst the rest (who attended Divine Service) St. Afaph was eminently conspicuous for Piety and Learning, infomuch that Mungo, (in Latine **Quentigermu**) being called into his Country, resigned both his Convent and Cathedral to him. Here he demeaned himself with such Sanctity, that Llan-Elvy was after his death, called from him St. Afaph. He was an assiduous Preacher, having this Speech in his Mouth, Such who are against the Preaching of Gods word, envy Mans Salvation. He is thought by some to have dyed about 569. After which, his See was Vacant above 500 years

Amongst the rest (who attended Divine Service) St. *Afaph* was eminently conspicuous for *Piety* and *Learning*, infomuch that *Mungo*, (in Latine **Quentigernus**) being called into his Country, resigned both his Convent and Cathedral to him. Here he demeaned himself with such Sanctity, that *Llan-Elvy* was after his death, called from him St. *Afaph*. He was an assiduous Preacher, having this Speech in his Mouth, *Such who are against the Preaching of Gods word, envy Mans Salvation*. He is thought by some to have dyed about 569. After which, his See was Vacant above 500 years,



honour should be done unto him whilst living, who was so solemnly disgraced after his death; his Books being then publicly prohibited by the Court of Rome. See Writers in this Shire.

*John of Monmouth*, D. D. and Canon of *Lincoln*, was chosen Bishop of *Landaff*, 1296. after that See had been 7 years vacant. He was a Learned and Pious Divine. Besides other Benefactions to his See, he procured the Rectory of *Newland*, in the Forrest of *Dean*, to be appropriated thereto; But Bishop *Kutchin* afterwards impoverished the same, more then all his Predecessors had endowed it in 400 years. This *John* dyed April 8. 1322. and was buried in *St. Maries Chapel*.

*Walter Cantilupe*, Son to *William* the elder, Lord *Cantilupe*, (whose prime residence was at *Abergavennie* in this County) was made (by *Henry 3.*) Bishop of *Worcester*. He would not yield to the Popes Legate, who complained of many Clergy men keeping their Livings against the Canons, intending to make room for the Popes Favourites, or force such irregular incumbents to a Composition. He was one of a keen nature, whose two-edged spirit did cut on both sides, against the King and Pope. Against the former, he sided with the Barons, to whom he promised Heaven for the reward of their Rebellion against their Prince, though it cost him an Excommunication from the Pope, who was the more forward in denouncing that fatal Sentence against him, because he had told *Rusland* his Legate, coming hither 1255. that he would prefer him to be hang'd on the Gallows, rather then ever consent to such expiation of the Church, as aforesaid. Lying on his death bed, he was touched with true remorse for his disloyalty, and obtained Absolution. He dyed February, 1267. whom I behold as Uncle to *Thomas*, the Sainted Bishop of *Hereford*.

Soldiers.

## Soldiers.

*Richard de Clare*, alias *Strongbow*, born (probably) at *Stringule Castle*, was Earl of *Stringule* and *Pembroke*. A person of effectual performance. It happened that *Mac Murugh*, an. 1167. being expelled his Territories, for several Tyrannies, by the Lords of *Meath* and *Conaught*, repaired to King *Henry 2.* and invited him to *Ireland*. That Politick King sent over this *R. Strongbow* (with 1200 Men) who possessed himself of the Ports of *Leinster* and *Munster*, with large Lands thereunto belonging; insomuch that the King growing jealous of his greatness remanded him home, and commanded him to surrender his Acquest into his hands, which done, he received them by regnant from the King; save that *Henry* reserved the City of *Dublin* for himself. This *Strongbow* is commonly called *Domitor Hibernia*, the Tamer of *Ireland*. Yet some of the great Lords there did still retain the Power and Title of King; Witness the Preface in the Commission, whereby King *Henry 2.* made *William Fitz. Adeline* his Lieutenant of *Ireland*, *Archiepiscopus, Episcopus, Regibus, &c. Salutem*. This Earl dyed at *Dublin*, 1177.

Sir *Roger Williams*, born of an ancient Family at *Penrofs*, was first a Soldier of Fortune under the Duke of *Alva*, and afterwards served Queen *Elizabeth*. A man extremely forward to Fight. When a Spanish Captain challenged Sir *John Norris* to fight a single Combat (which was beneath him to accept, being a General) this *Roger* undertook the *Don*. And after they had fought some time (both Armies beholding them) without any hurt, they pledged each other a deep draught of Wine, and so friendly departed. Another time at mid night, he assaulted the Camp of the Prince of *Parma*, nigh *Venloe*, slew some of the Enemies, and pierced

P p p 2

pierced

honour should be done unto him whilst living, who was so solemnly disgraced after his death; his Books being then publicly prohibited by the Court of Rome. See Writers in this Shire.

*John of Monmouth*, D. D. and Canon of *Lincoln*, was chosen Bishop of *Landaff*, 1296. after that See had been 7 years vacant. He was a Learned and Pious Divine. Besides other Benefactions to his See, he procured the Rectory of *Newland*, in the Forrest of *Dean*, to be appropriated thereto; But Bishop *Kutchin* afterwards impoverished the same, more then all his Predecessors had endowed it in 400 years. This *John* dyed April 8. 1322. and was buried in *St. Maries Chapel*.

*Walter Cantilupe*, Son to *William* the elder, Lord *Cantilupe*, (whose prime residence was at *Abergavennie* in this County) was made (by *Henry 3.*) Bishop of *Worcester*. He would not yield to the Popes Legate, who complained of many Clergy men keeping their Livings against the Canons, intending to make room for the Popes Favourites, or force such irregular incumbents to a Composition. He was one of a keen nature, whose two-edged spirit did cut on both sides, against the King and Pope. Against the former, he sided with the Barons, to whom he promised Heaven for the reward of their Rebellion against their Prince, though it cost him an Excommunication from the Pope, who was the more forward in denouncing that fatal Sentence against him, because he had told *Rusland* his Legate, coming hither 1255. that he would prefer him to be hang'd on the Gallows, rather then ever consent to such expiation of the Church, as aforesaid. Lying on his death bed, he was touched with true remorse for his disloyalty, and obtained Absolution. He dyed February, 1267. whom I behold as Uncle to *Thomas*, the Sainted Bishop of *Hereford*.

Soldiers.

## Soldiers.

*Richard de Clare*, alias *Strongbow*, born (probably) at *Stringule Castle*, was Earl of *Stringule* and *Pembroke*. A person of effectual performance. It happened that *Mac Murugh*, an. 1167. being expelled his Territories, for several Tyrannies, by the Lords of *Meath* and *Conaught*, repaired to King *Henry 2.* and invited him to *Ireland*. That Politick King sent over this *R. Strongbow* (with 1200 Men) who possessed himself of the Ports of *Leinster* and *Munster*, with large Lands thereunto belonging; insomuch that the King growing jealous of his greatness remanded him home, and commanded him to surrender his Acquest into his hands, which done, he received them by regnant from the King; save that *Henry* reserved the City of *Dublin* for himself. This *Strongbow* is commonly called *Domitor Hibernia*, the Tamer of *Ireland*. Yet some of the great Lords there did still retain the Power and Title of King; Witness the Preface in the Commission, whereby King *Henry 2.* made *William Fitz. Adeline* his Lieutenant of *Ireland*, *Archiepiscopus, Episcopus, Regibus, &c. Salutem*. This Earl dyed at *Dublin*, 1177.

Sir *Roger Williams*, born of an ancient Family at *Penrofs*, was first a Soldier of Fortune under the Duke of *Alva*, and afterwards served Queen *Elizabeth*. A man extremely forward to Fight. When a Spanish Captain challenged Sir *John Norris* to fight a single Combat (which was beneath him to accept, being a General) this *Roger* undertook the *Don*. And after they had fought some time (both Armies beholding them) without any hurt, they pledged each other a deep draught of Wine, and so friendly departed. Another time at mid night, he assaulted the Camp of the Prince of *Parma*, nigh *Venloe*, slew some of the Enemies, and

P p p 2

pierced



Walter Cantilupe, Son to William the elder, Lord  
**Cuntilupe**, (whose prime residence was at  
Abergavennie in this County) was made (by Henry  
3.) Bishop of Worcester. He would not yield to the  
Popes Legate

Walter Cantilupe, Son to William the elder, Lord  
**Cantilupe**, (whose prime residence was at  
Abergavennie in this County) was made (by Henry  
3.) Bishop of Worcester. He would not yield to the  
Popes Legate

Walter Cantilupe, Son to William the elder, Lord  
**Cuntilupe**, (whose prime residence was at *Abergavennie*  
in this County) was made (by *Henry 3.*) Bishop of  
*Worcester*. He would not yield to the Popes Legate,

Walter Cantilupe, Son to William the elder, Lord  
**Cantilupe**, (whose prime residence was at *Abergavennie*  
in this County) was made (by *Henry 3.*) Bishop of  
*Worcester*. He would not yield to the Popes Legate,



# 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 the digital edition of the Potage Dyvers
- Library cataloguing standards:
  - 302 pages of ISBD
  - 750 pages AACR, 1056 pages of RDA
    - Helmetin luettelointiohjeet
- 1020 pages of the SPECTRUM standard for museum cataloguing
- A single page of field descriptions in the Schoenberg database



# 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 Swedish language publications from the 19th 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.”



# Documentation?

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



# 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' tq](#) (Charles Victor Prévôt), [td 1789-1856](#)

 100 1 \_ [ta Arlincourt, tc vicomte d' tq](#) (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'>](#)

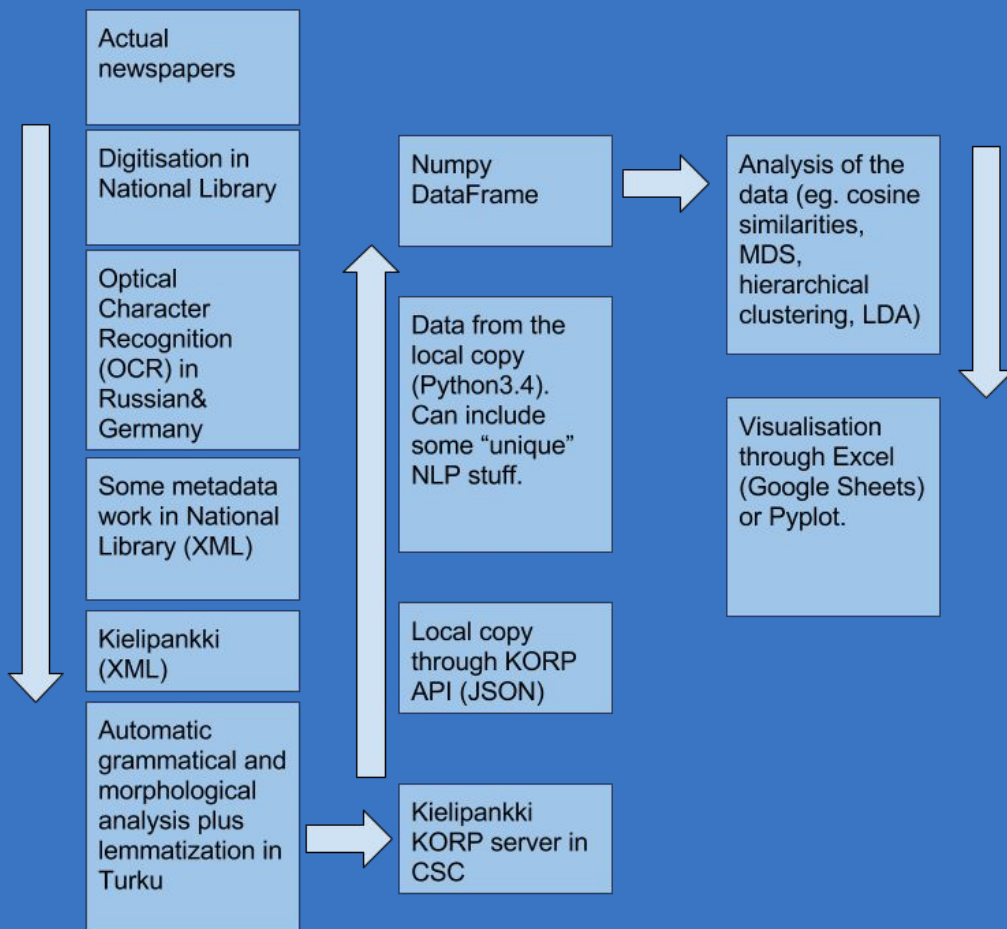


`<schema:name>Charles-Victor Prévost d'Arlincourt</schema:name>`  
`<schema:name>Charles Victor Prévôt ~d'œ Arlincourt</schema:name>`  
`<schema:name>Charles Victor Prevot d' Arlincourt</schema:name>`  
`<schema:name>Arlincourt</schema:name>`

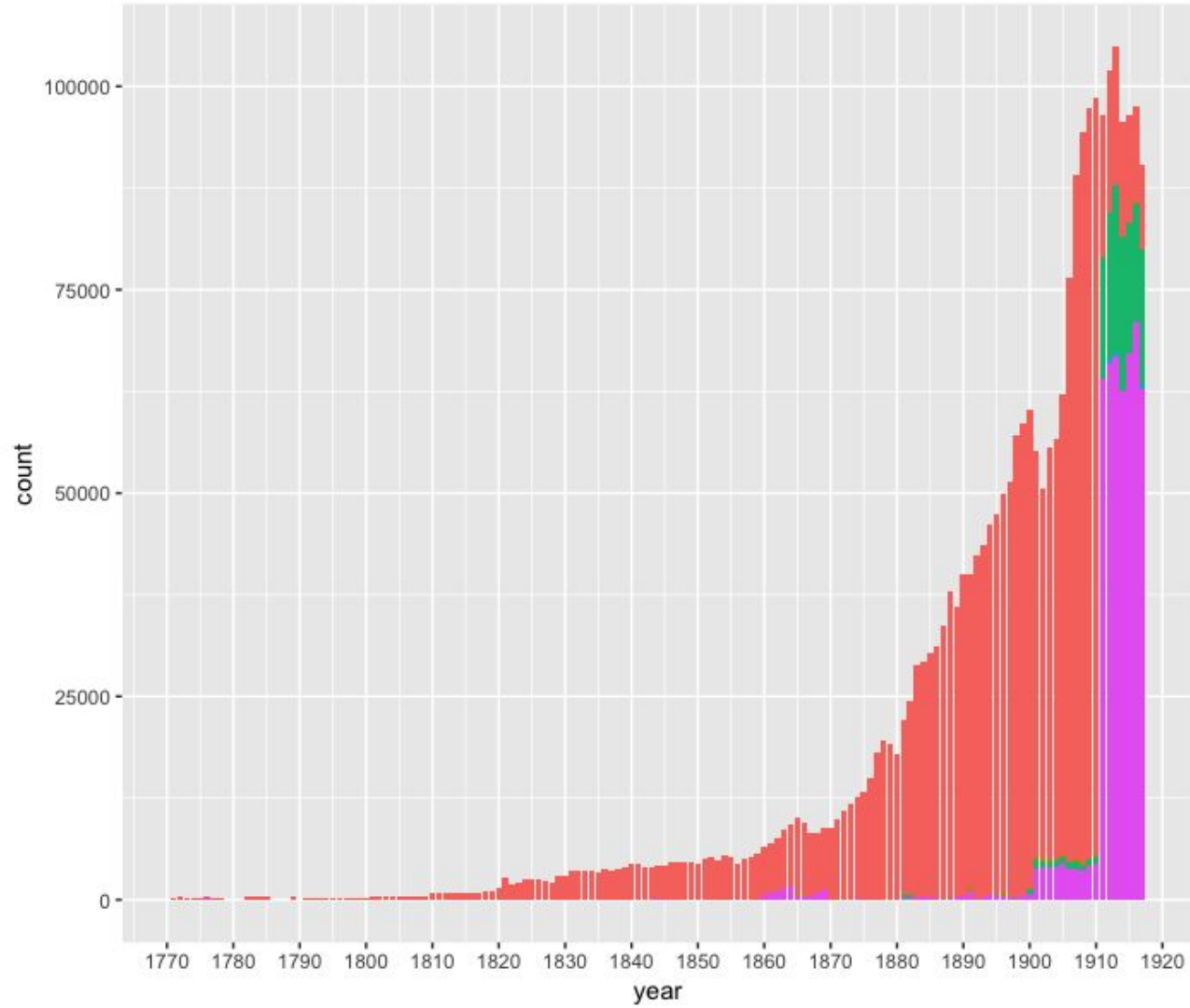
<http://viaf.org/viaf/41896578/>



# KLK Newspaper Pipeline: from archives to a researcher



22(!) different  
versions of  
pipeline





# Data woes: National Newspaper Collection (KLK)

- On the surface, the Korp API allows one to search by lemma.
- However, these lemmas have been automatically generated, and are only as good as the process that generated them
- Examples:
  - Early modern Finnish allowed words with the letter “v” to be written as “w”. These are all passed through unprocessed by the analyzer
  - Fraktur fonts, which are hard for OCR engines, appear in the early parts of the collection



# Data woes: National Newspaper Collection (KLK)

## EXAMPLE:

KLK-1800-subcorpora contains the highly frequent lemma 'niisi', because of faulty disambiguation of a morphological analysis (should be *ne : niiden* instead of *niisi : niiden*).



# Data woes: National Newspaper Collection (KLK)

- In fact, due to 1) OCR errors and 2) historical language, only a small fraction of KLK is accurately lemmatized
  - 1851-1910: 9,6% of distinct words (66,0% of tokens)
  - before 1851: 15,0% of distinct words (69,3% of tokens)
- Number of lemmas in data sets of comparable sizes:
  - KLK\_1980-2000: 201
  - KLK\_1820-1859: 528



# Data woes: National Newspaper Collection (KLK)

The data is (for obvious historical reasons) temporally imbalanced, this causes the earlier part to be more fragile to metadata problems.

EXAMPLE: “Tähdenvälejä” is located in the year 1842 instead of 1942 where it belongs, and is the only paper in the 1842 corpus.



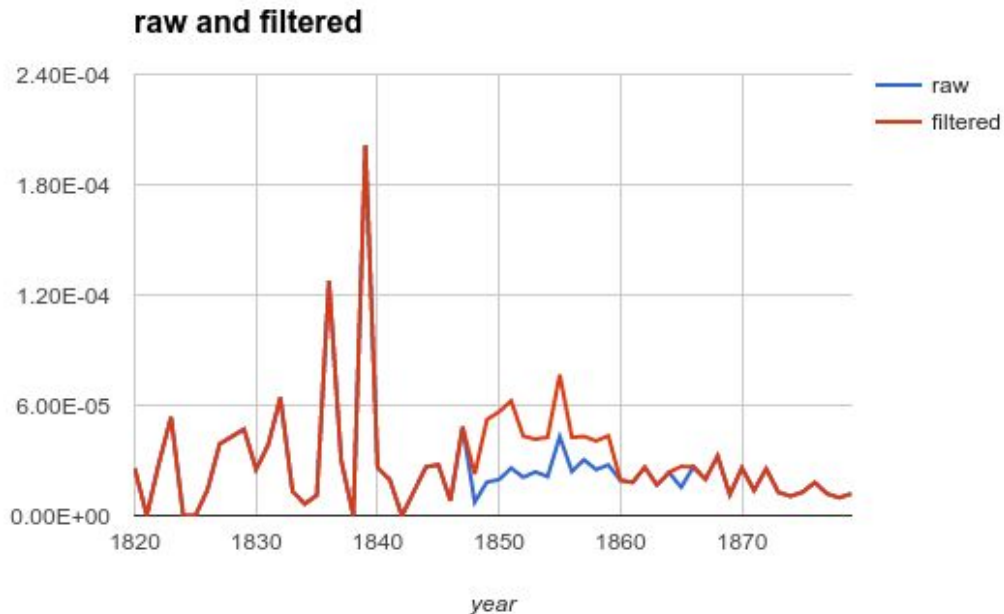
# Data woes: National Newspaper Collection (KLK)

Finnish and Swedish Newspapers are classified under separate subcorpora. Sometimes individual papers contain both languages. These have not been consistently separated.

EXAMPLE: Helsingfors Tidningar

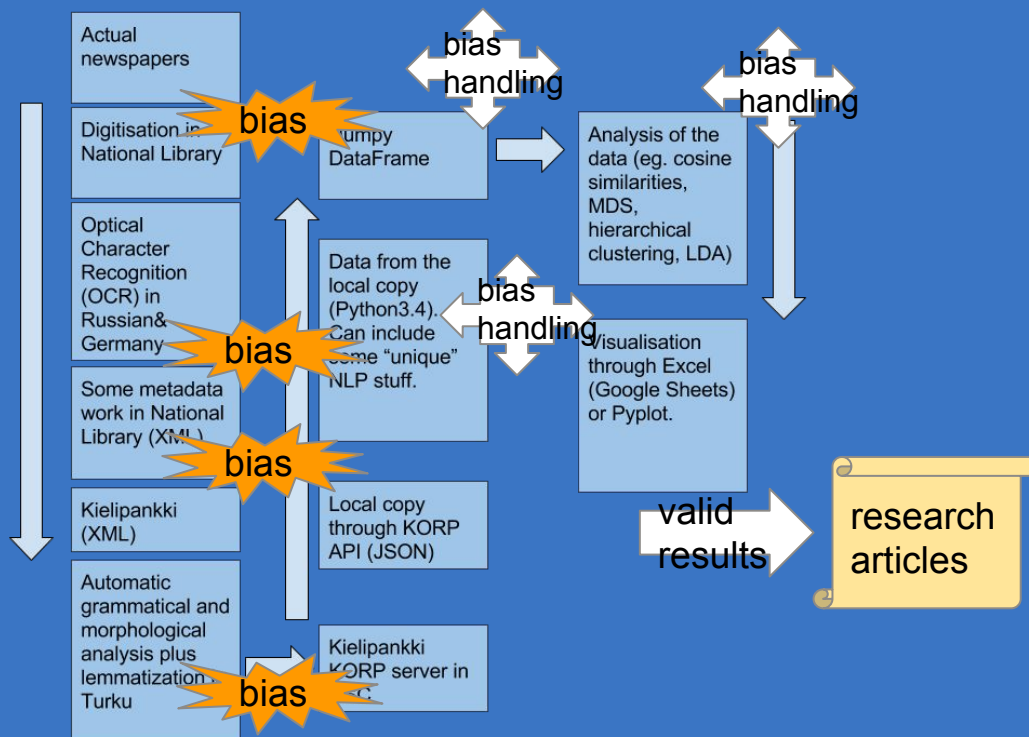


# Data woes: National Newspaper Collection (KLK)



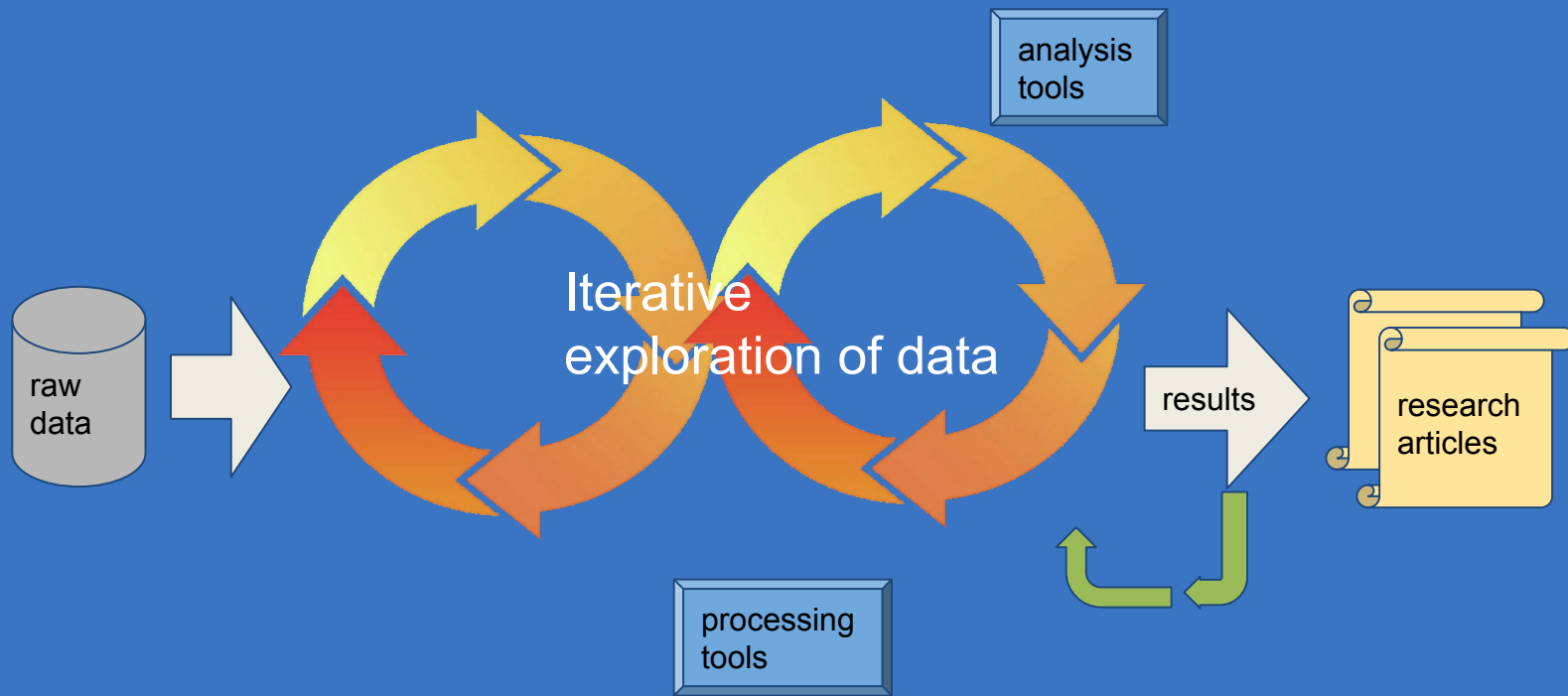


# KLK Newspaper Pipeline: from archives to a researcher



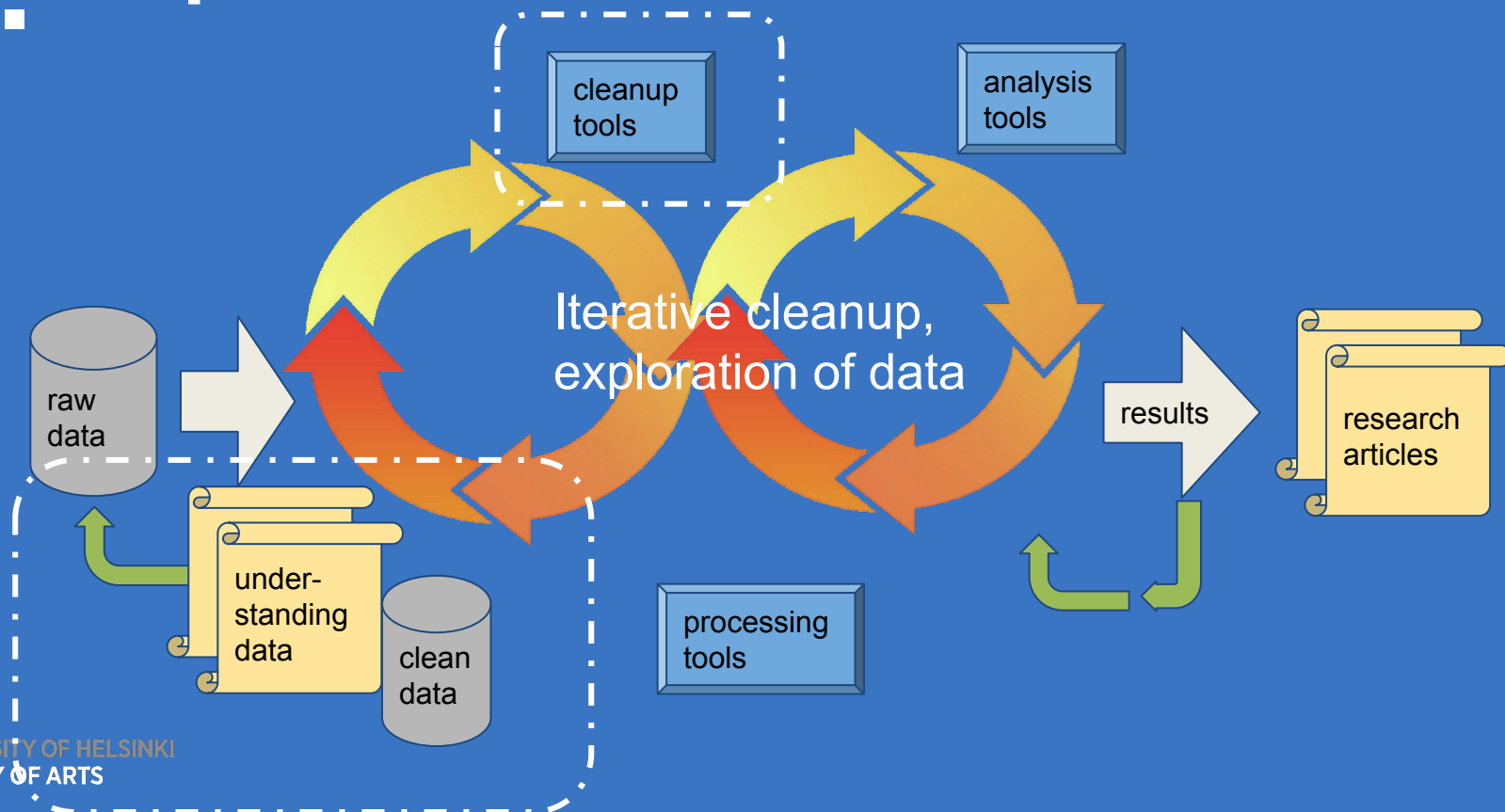


# Digital humanities research process



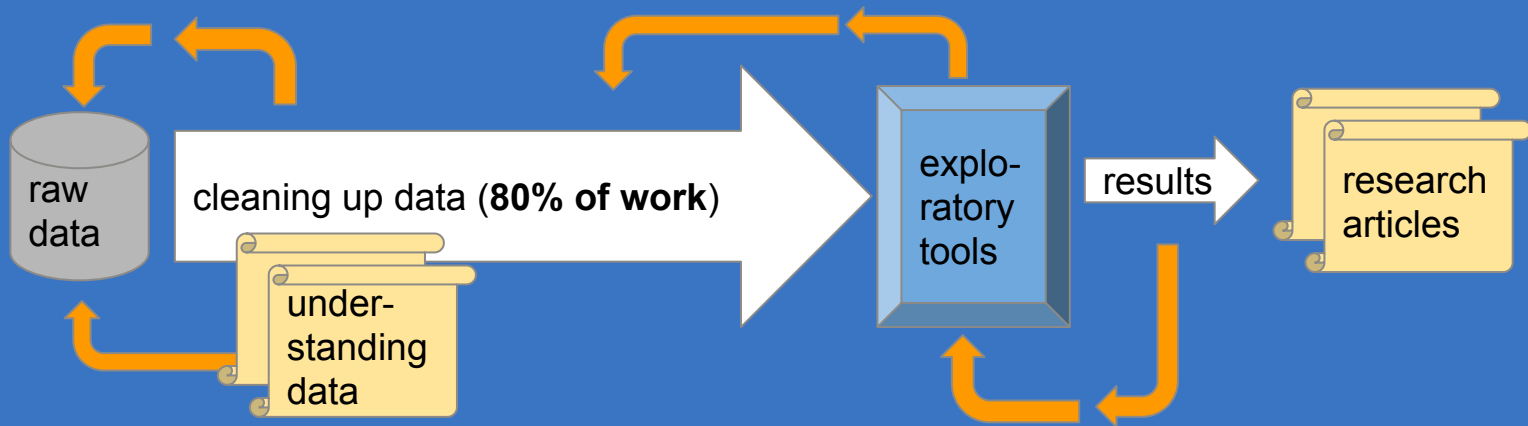


# Digital humanities research process





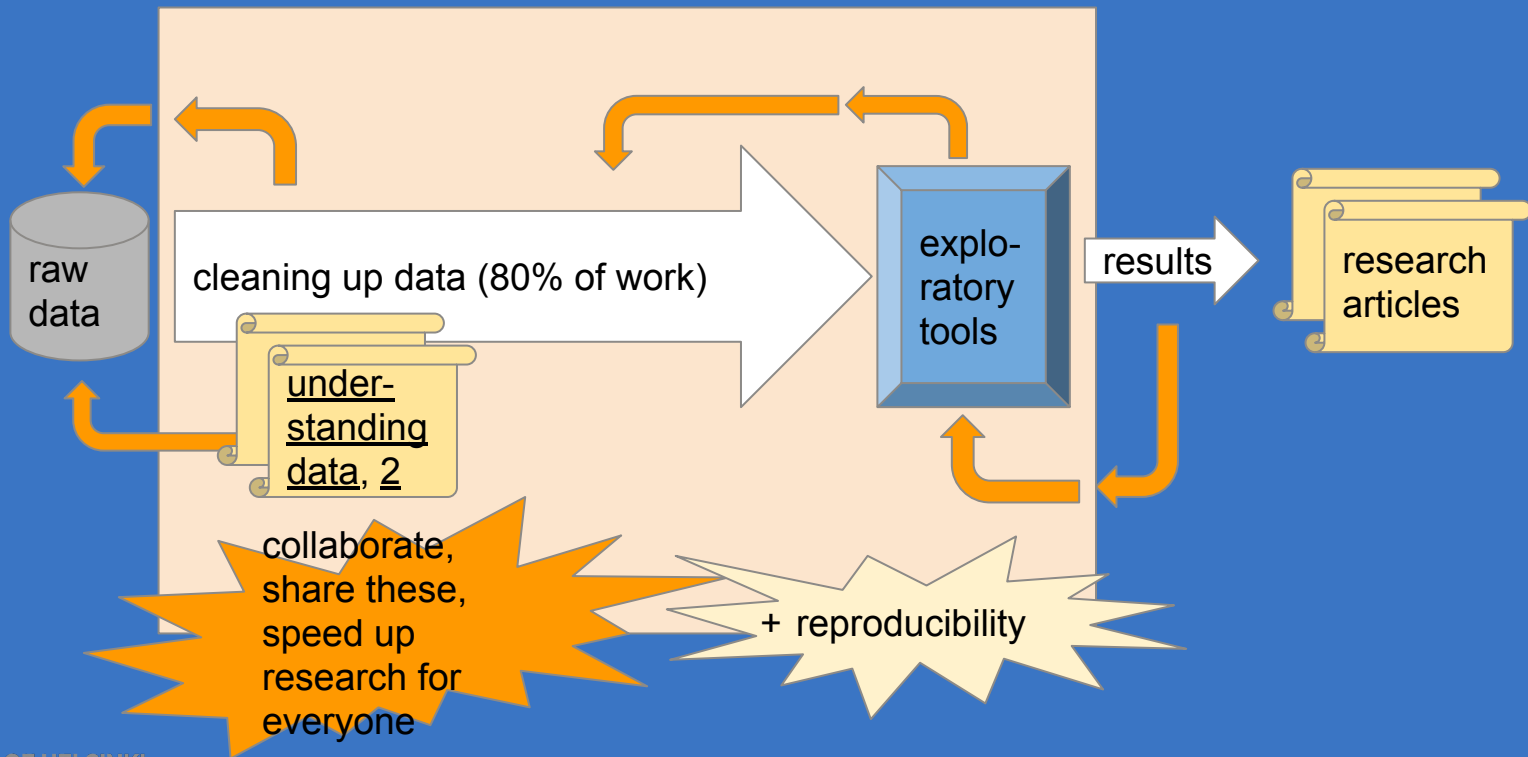
# Digital humanities research process



80% of your time for data cleanup, another  
80% for algorithms, ...



# Leverage collaboration, open science workflows to reduce individual workload





# Sample available datasets and APIs

- Korp API (hits in texts+metadata)
- Finnish national gallery API / dump (metadata)
- Schoenberg database (metadata)
- Cushman collection metadata (metadata)
- WW2 covert support networks (metadata)
- Europeana APIs (metadata)
- DPLA APIs (metadata)
- The European Library API (metadata)
- Sydney Powerhouse Museum (metadata)
- EEBO-TCP Phase I (full texts+metadata)
- ECCO-TCP (full texts+metadata)



# Assignments for next week

1. Find a dataset that could be of interest to you in your final project. Post a message on #datasets on Slack giving a link to the dataset and a note on why you selected it.
2. Read some background on visualisation for next week:
  - What is Visualization Really for? sections 2.1-2.4 for a categorisation of different uses for visualisation
  - Perception deception & Common visualization mistakes for learning to not trust visualisations blind
3. Read your first DH research article on approaches to visualising relationships between different versions of texts.



eetu.makela@helsinki.fi  
<http://iki.fi/eetu.makela>

<http://presemo.helsinki.fi/meth4dh>