Integrating Open Science in the Humanities: the Case of Computational History

Better Science Through Better Data 2019; #scidata19

Wellcome Collection, London

6 November 2019

Mikko Tolonen (University of Helsinki)

mikko.tolonen@helsinki.fi

Comhis
Helsinki Computational History Group

UNIVERSITY OF HELSINKI
FACULTY OF ARTS

Outline of the talk

1. (Open) Science in the Humanities?
2. Challenges of Humanities Data
3. Integrating Open Science to Humanities in Computational History
4. (Experiments to go around problems of noisy data)
5. Conclusion

Helsinki Computational History Group

“Computational history” refers to an integrated mixed methods approach to study large digitized historical sources. “Integrated” means that data science is combined to specialized subject knowledge; in the case of COMHIS, intellectual history and book history.

http://helsinki.fi/computational-history

Better Science Through Better Data

(Open) Science in the Humanities?

Science and hermeneutics

Tangible objects ≠ Subjective experience

Need for mixed methods!

photo: Time Machine project

Three aspects of OA in the Humanities

Raw/primary data

- seldom access to full data
- special arrangements when access (no possibility to share further)
- data providers reluctant to share openly

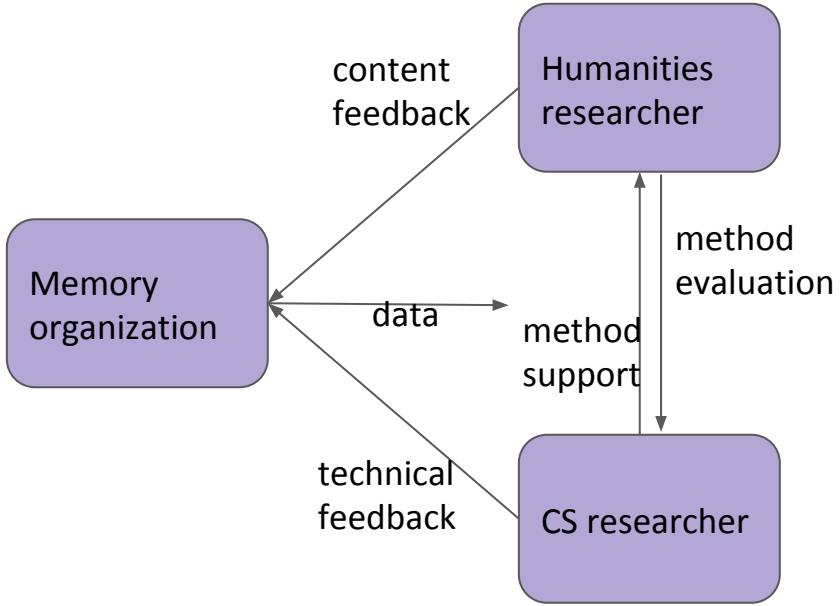
Research data

- Scholars often think they have no research data
- idea of reproducibility vague/non-existent
- little/no credit in opening research data

Open publications

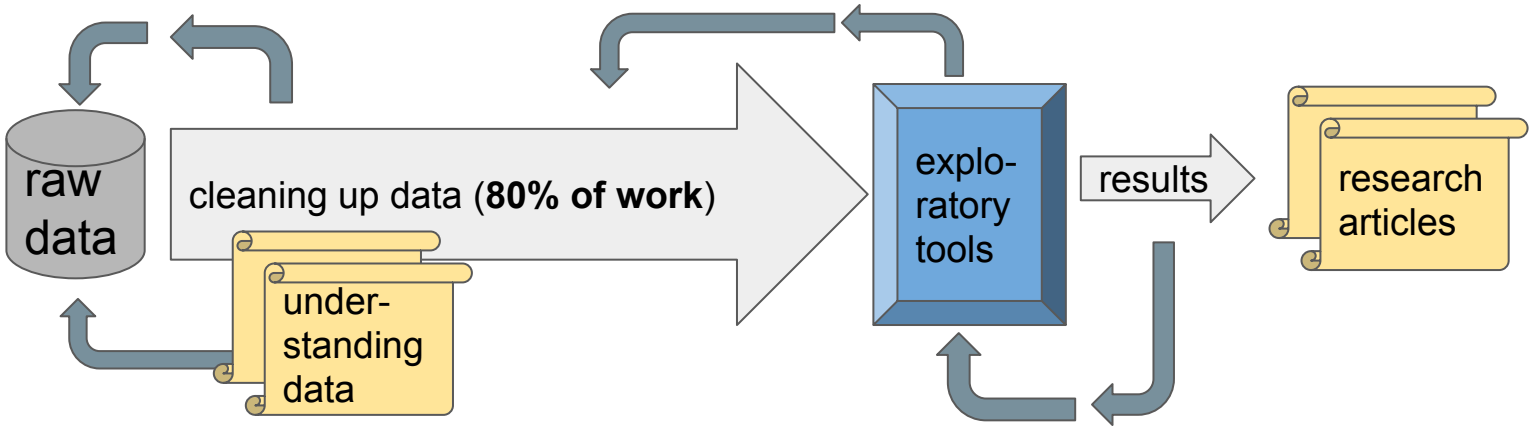
- Idea that humanities cannot afford OA
- Some think Plan-S is hurting humanities
- Scholars don’t grasp the value or relevance of OA even with respect to publications

Better data in Humanities calls for well-functioning ecosystems



UNIVERSITY OF HELSINKI

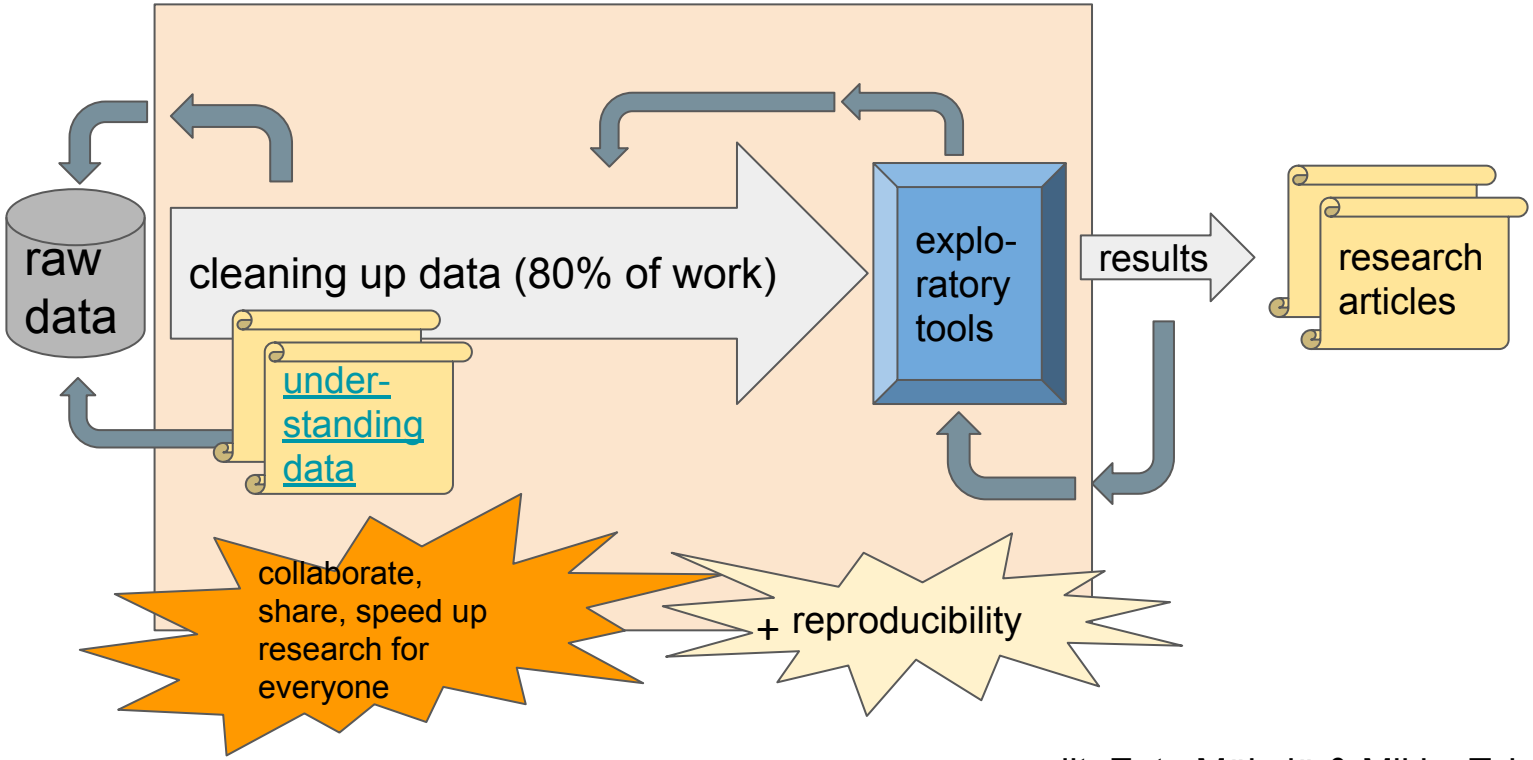
Computational history research process



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

credit: Eetu Mäkelä & Mikko Tolonen

In an ideal world: collaboration & open science workflows to reduce individual workload



credit: Eetu Mäkelä & Mikko Tolonen

Better Science Through Better Data

Challenges of Humanities Data

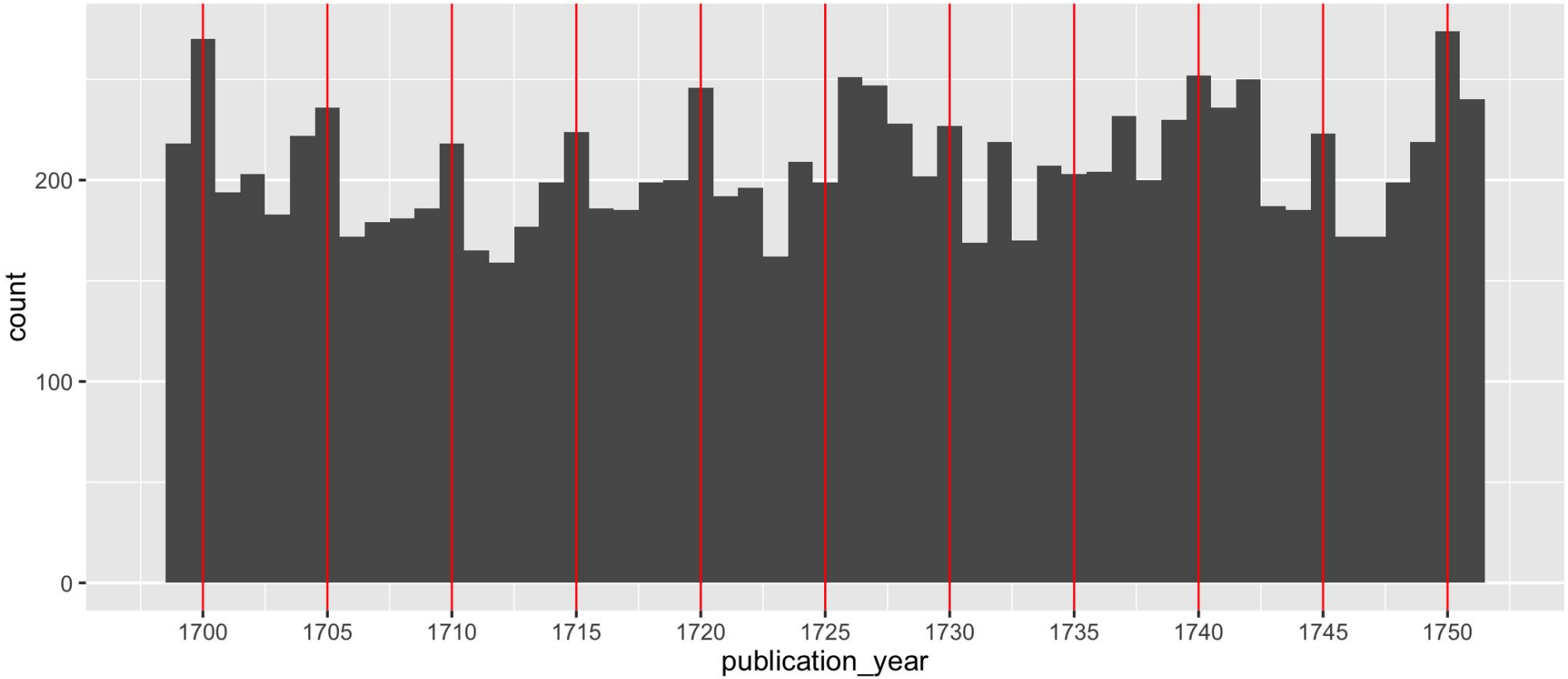
Nature of digitized humanities data

- Collections of collections
- Relevance of source criticism
- Stacking more sources does not often solve problems of bias!

Wellcome Collection

Example of bias that is particular to humanities data:
The 5-year theory with respect to ESTC data

Comhis
Helsinki Computational History Group



Burden of empty promises

Challenges with open data

- **Institutions** reluctant to give full access to data. Why?
- **Research process** is not opened and research data is not shared in the Humanities. Transparency, reproduction, collaboration, new initiatives are missing. Why?

Short answer: **Cultural change takes time**. We need concrete examples in the core field of the Humanities that actually prove OPEN DATA PRINCIPLES as useful.

Better Science Through Better Data

Integrating Open Science to Humanities in Computational History

Tradition & research culture

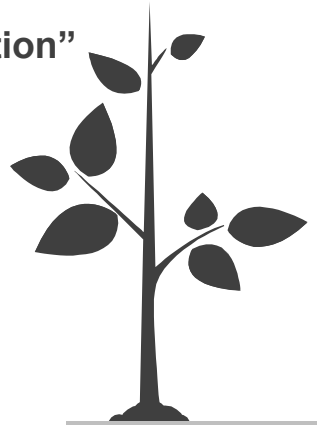
Comhis
Helsinki Computational History Group

04 **Building a research group**
A step into a different world

03 **Advancing to funded projects**
Continuing on the chosen path

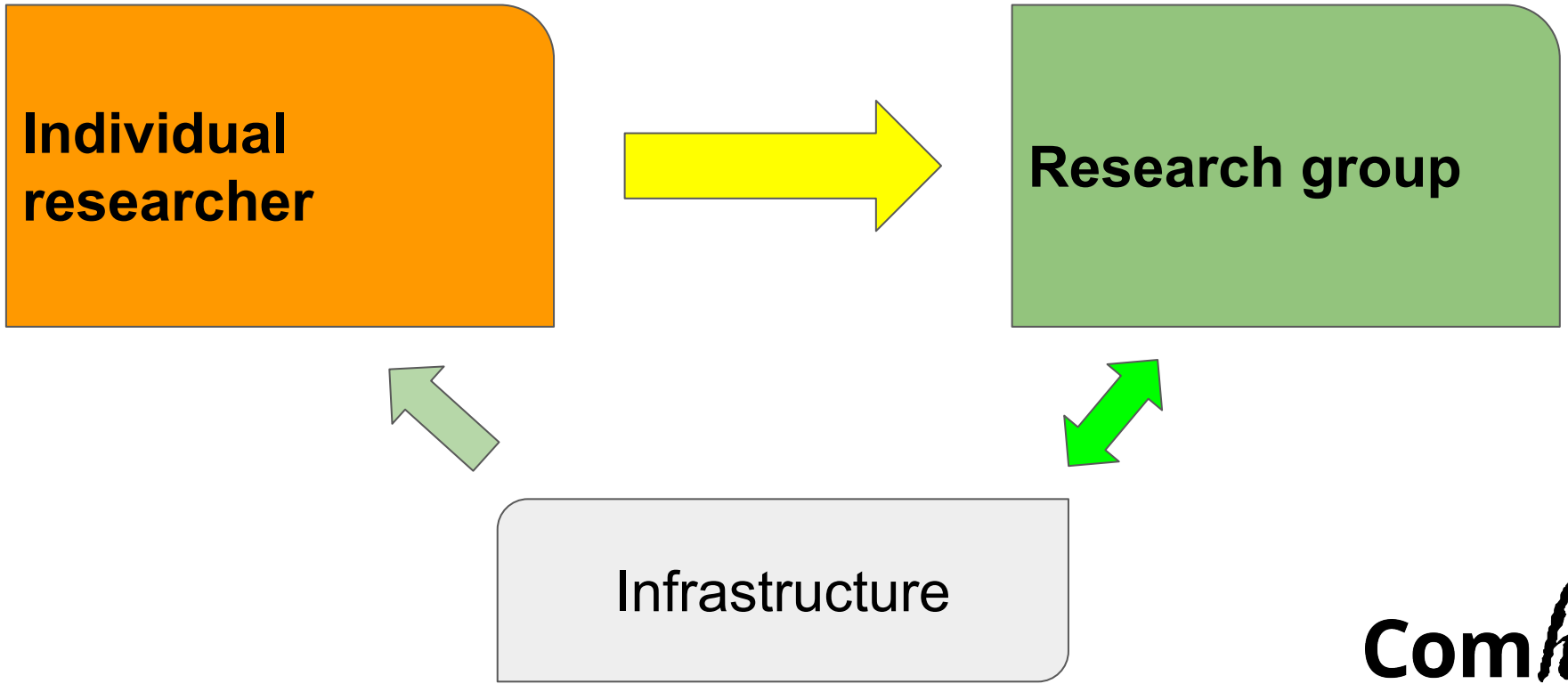
02 **Seeking “twin-collaboration”**
Realisation of what is needed as a postdoc

01 **Individual endeavours**
Dreams of expanding the method basis of intellectual history as doctoral student



[And, an overdose of infrastructural & administrative work without which the path would not be possible]

Key factor for the humanities of the future



Comhis
Helsinki Computational History Group

Helsinki Computational History Group's aim:

Understanding public communication in early modern Europe

Movement of ideas

- Metadata work based on several different library catalogues
- genres (poetry, pamphleteering); intellectual traditions (natural law tradition, ancient texts
- text reuse: genres (historical works, quoting practices)

Conceptual change

- concepts are crucial, but not directly jumping into this for various reasons
- Theoretical underpinning (historians + linguists)
- Concepts as linguistic objects (linguists + historians + CS)

Research data releases

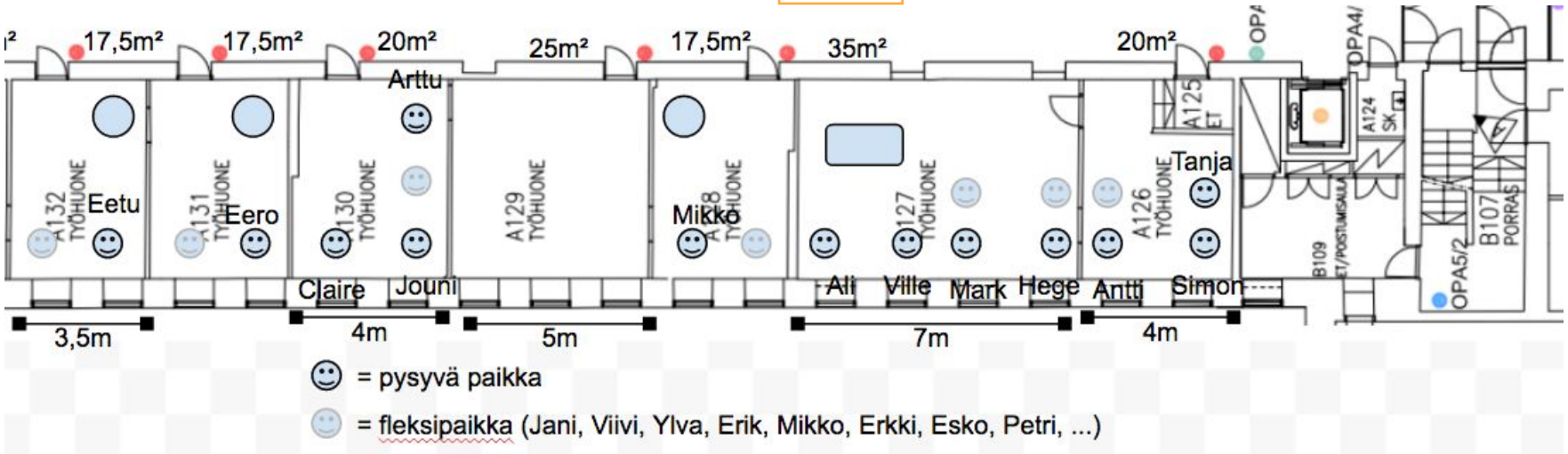
- ESTC; Fennica; Kunglica; CERL; ECCO text reuse (+ EEBO text reuse); Finnish Newspapers

Tools for others

- UIs, APIs, shiny apps etc.

2013

2017


17,5m²
17,5m²
20m²
25m²
17,5m²
35m²
20m²
A132 TYÖHUONE
Eetu
A131 TYÖHUONE
Eero
A130 TYÖHUONE
Arttu
A129 TYÖHUONE
A128 TYÖHUONE
Mikko
A127 TYÖHUONE
A125 ET
A126 TYÖHUONE
Tanja
A124 SK
B109 ET/POSTUMISALA
B107 PORRAS
OPA5/2
Claire
Jouni
Ali
Ville
Mark
Hege
Antti
Simon
3,5m
4m
5m
7m
4m
= pysyvä paikka
= fleksipaikka (Jani, Viivi, Ylva, Erik, Mikko, Erkki, Esko, Petri, ...)

Data mining in computational history

Text mining of large corpora ⟺ Metadata as a quantitative tool

- **Objective**: understanding conceptual change, uses of language
- **Sources**: full-text databases (ECCO, EEBO, Finnish Newspapers etc.)
- **Potential**: Theoretically great, the future?
- **In practice**: raw data almost never openly available; if it is, tied to limited interfaces
- **Scalability with open research data**: data-driven approach
- **Methodological perspective**: Messy to study historical sources, intellectual input not guaranteed.

- **Objective**: Quantitative study of material objects
- **Sources**: World is full of different metadata collections
- **Potential**: Greatly underestimated (even by librarians)
- **In practice**: difficulties with open access to raw data and supporting data sources, but not impossible.
- **Scalability with open research data**: fantastic
- **Methodological perspective**: excellent for borrowing best practices from other scientific fields. Quality of catalogues varies.

→ **It is the combination of harmonized (and better) metadata and full-text sources that will lead to better science in early modern intellectual history (ECCO and ESTC), for example!**

Build it yourself! Humanities questions guiding our tool building

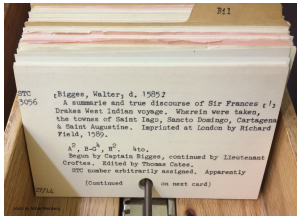
- **Bibliographic sources** as basis of networks and time series to inform the intellectual endeavours.
- **Text-reuse detection** to study influence (using BLAST to deal with OCR-mistakes).
- **Materiality explorations** of printed items based on information derived from layout, font etc.
- **Stylometry** to study particular questions of authorship.

- **Topic modelling and word embeddings** (etc.) to explore conceptual change
- **Detecting argumentative structures** based on syntax.

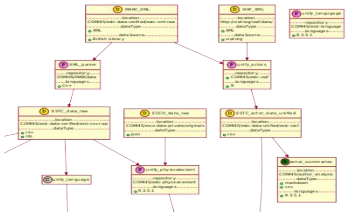
Comhis
Helsinki Computational History Group

From library catalogues to research reports

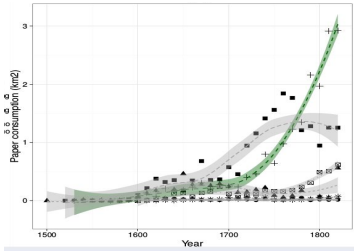
Research potential



Open data science ecosystem



Research cases



Original Articles

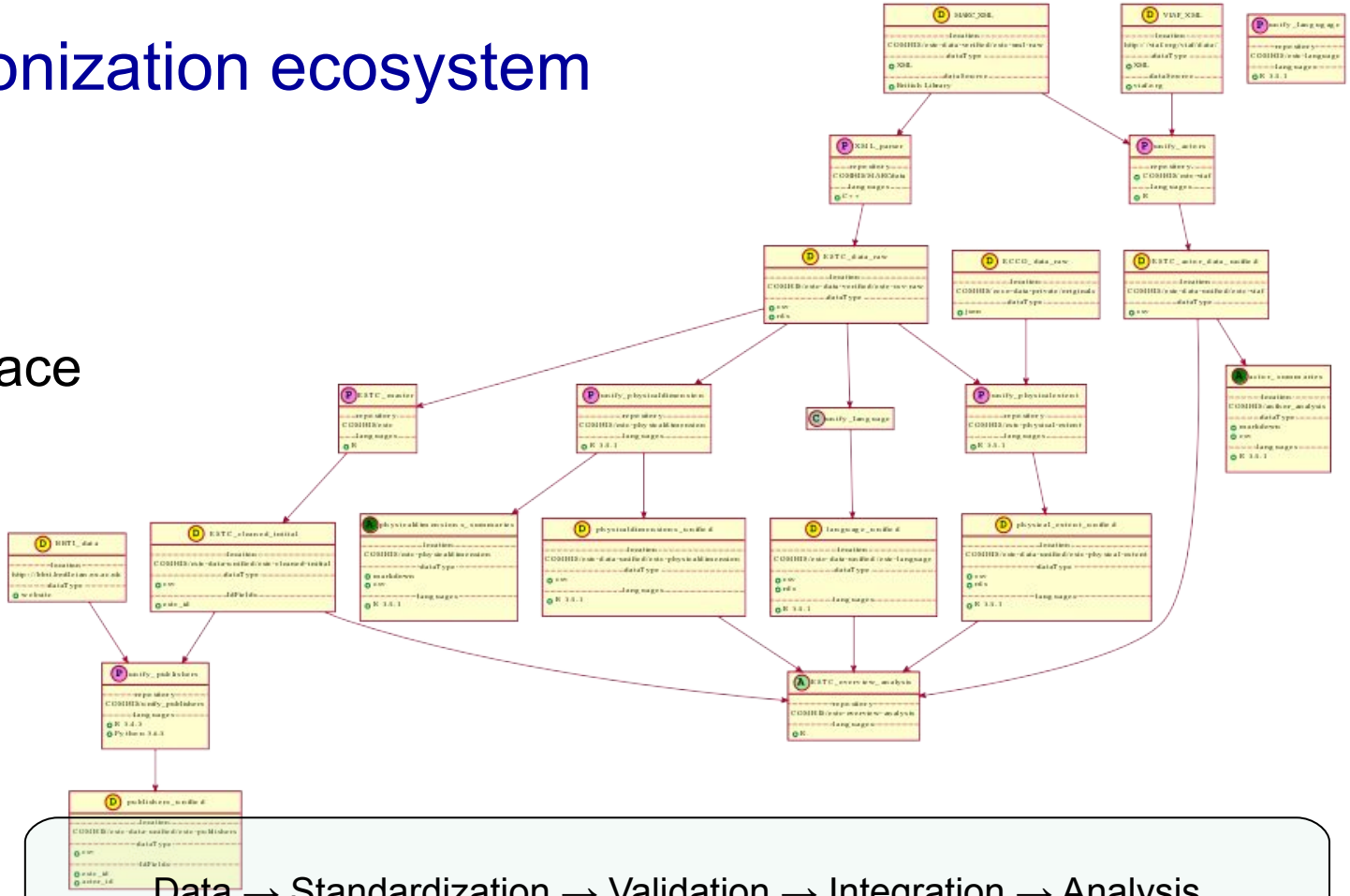
Bibliographic Data Science and the History of the Book (c. 1500–1800)

Leo Lahti, Jani Marjanen, Hege Roivainen & Mikko Tolonen

Pages 5-23 | Received 07 Jul 2018, Accepted 10 Oct 2018, Published online: 07 Jan 2019

Data harmonization ecosystem

Authors
Publishers
Editions
Publication place
Gatherings
Page count
Language
Genre...



Data → Standardization → Validation → Integration → Analysis

Our virtuous cycle of better data & better science

- Combining **harmonized metadata to full-text sources (ESTC & ECCO)** -> Enables text mining in a new way, upcoming this academic year.
- Using **full-texts to enrich metadata (ESTC & ECCO)** -> Feeding back to the loop, better quality data, detecting subject/topics for example.
- Combining **text reuse information to metadata (ESTC & ECCO)** -> feeds back to edition information.
- **Re-OCRing (ECCO)** -> Feeds back to all processes that combine ECCO and ESTC.

Comhis
Helsinki Computational History Group

Experiments to go around problems of noisy data

Text reuse detection to study influence

- BLAST -bioinformatics software
- computational analysis on the whole ECCO, that is +200,000 texts
- ~130,000 reuse fragments in Hume's *History* alone

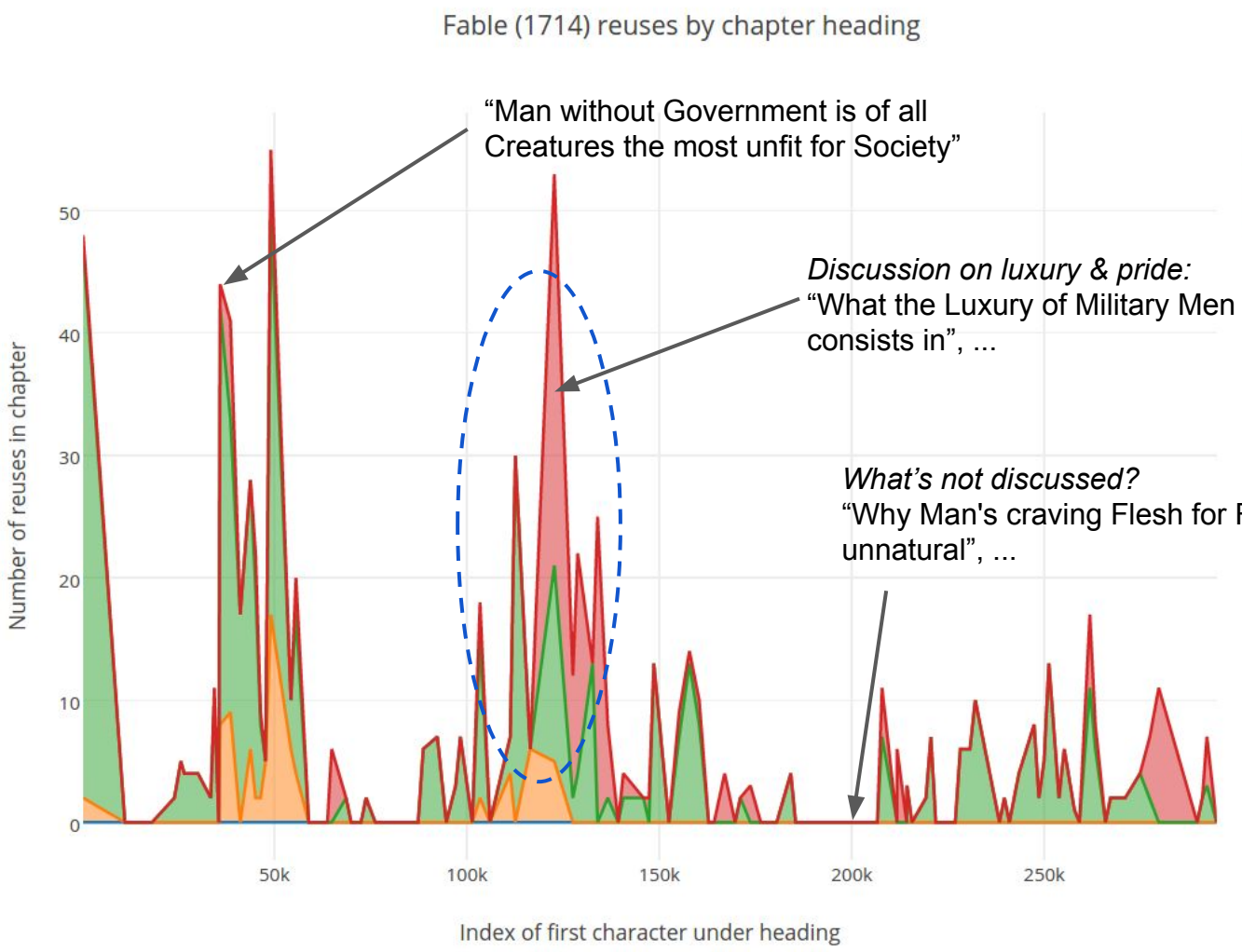
  (Whole ECCO has millions of interlinked reuse fragments)
- ~150 - 3000 characters / reuse fragment

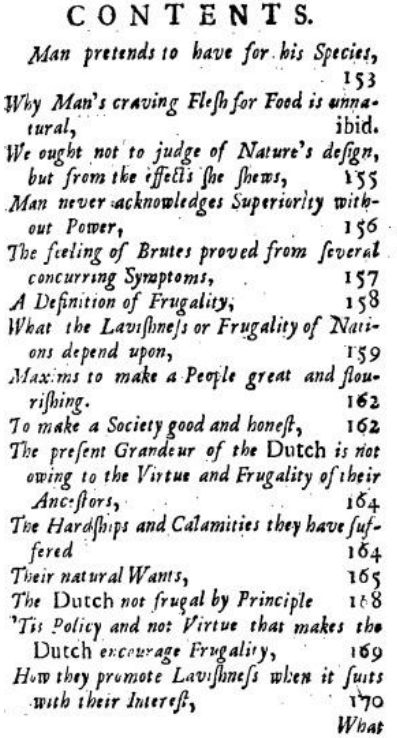
“Boscobel: or, the compleat history of His Sacred Majesty's most miraculous preservation after the battle of Worcester, which was fought Sept. 3, …” *(1660), Blount, Thomas, 1618-1679*

“broad pieces to the king, judging they would be necessary to him in his present condition; for he durfi carry no money about him in his mean garb and short cut hair, except about ten or twelve Lhillings in silver. Windham hereupon went to Lime, and spoke with Elef- don about hiring a Ihip, which he undertook; but not till he was told, it was for His Ma- jefty's transportation. During the four or five dayv\" which the King this first time staid at Windham's, where he was was known by most of the family, e heard the bells ring, and feeing a company got to- gether in the church-yard, which wa4 very near the” [...]

A general history of England. ... . Containing an Account of the first Inhabitants of the Country, and the Transactions in it, from the earliest … (1754) *Carte, Thomas, 1686-1754*

" 300 broad pieces to the king, judging they would be necefhry for him in his prefelit condition; for he durst carry no money about him in his mean garb and his short cut hair, ex- cept about ten or twelve shillings in silver. Windham hereupon went to Lyme, and 1poke to Elefi'on about hiring a Ihip, which he undertook: but not till he was told, it was for his majetfy's transportation. During the four or five days, which the king, this sirss time, flaid at Windam's (where he was known to most of the fa- mily) he heard the bells ring, and feeing a company got together in the church- yard, which was very near the” [...]





CONTENTS.

https://plot.ly/~villepvaara/7/

Better Science Through Better Data in the Humanities

Conclusion

What to learn from other fields?

- **Research support** databases
- **Collaboration & reuse** through open data
- Importance of **standards**

Nature Reviews Genetics **14**, 89-99 (February 2013) | doi:10.1038/nrg3394

Reuse of public genome-wide gene expression data

Johan Rung[1] & Alvis Brazma[1] About the authors

Data Science in Humanities

Potential

- New methods, classical research questions
- New scale of quantitative analysis
- Quality through collaboration

Pitfalls

- Data quality easily overlooked
- Existing tools drive research
- Expertise lacking

Science and hermeneutics

Tangible historical objects

Subjective historical experience

Need for new methods and clear principles for data sharing!

Humanities collaboration for better data

- Crowdsourcing experts
- Collaboration with different field of science, national libraries, infrastructures and projects
- Collaboration with companies that do digitization
- Interoperability & dealing with noise and bias

→ We need right kind of infrastructures for specific purposes that enable collaboration between researchers, companies and libraries.

NEWS EYE

A Digital Investigator for Historical Newspapers

NewsEye is a research project advancing the state of the art and introducing new concepts, methods and tools for digital humanities by providing enhanced access to historical newspapers for a wide range of users. With the tools and methods created by NewsEye, crucial user groups will be able to investigate views and perspectives on historical events and development and, as a consequence, the project will change the way European digital heritage data is (re)searched, accessed, used and analysed.

Workflow

The core concept of NewsEye is a seamlessly integrated armoury of tools and methods that will improve the users' capability to access, analyse and use the content in the digital libraries of historical newspapers.

Four Case Studies: the inner test material used by the project's humanities research groups will be from European newspaper datasets from the three partner libraries focusssing on the period 1850-1950.

4 NewsEye case study topics

1. Media & Journalism
2. Gender
3. Migration

Computer Science

Digital Humanities

La Rochelle Université

Universität Rostock

University of Helsinki

universität innsbruck

universität wien

Université Paul Valéry Montpellier 3

National Libraries

Österreichische Nationalbibliothek 600K

16 40 The National Library of Finland 500K

BnF Bibliothèque nationale de France 250K

Total number of pages used for the project (1.35M)

NewsEye consortium

Text Recognition & Article Separation

> Article Separation
> Automatic Text Recognition
> Layout Analysis

Semantic Text Enrichment

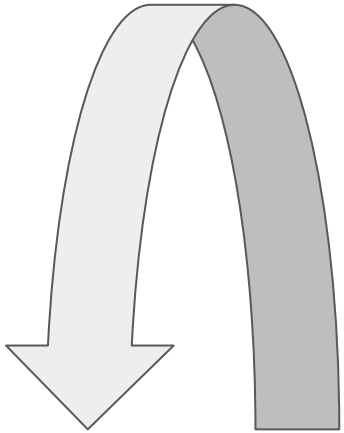
> Named entity recognition
> Stance detection
> Novelty detection
> Event detection

From projects to large-scale sharing?

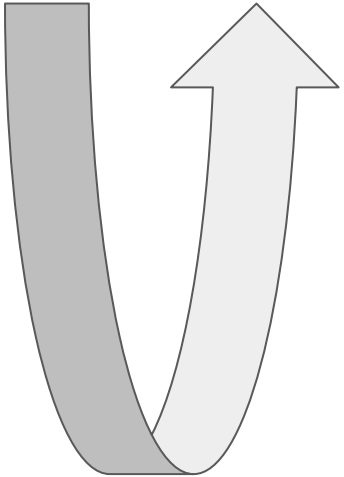
- Dariah-FI: Researcher-driven ecosystem of services for data-intensive social sciences and humanities (SSH) research

Themes	Modules	Focus areas
Data access and documentation	Social Sciences and Humanities Big Data	Digitised and born-digital data
Research methods and tool development	Analytica	Computational techniques and environments
Dissemination of best practices	Information Interaction	Researcher support

extras



Research questions



Initial data

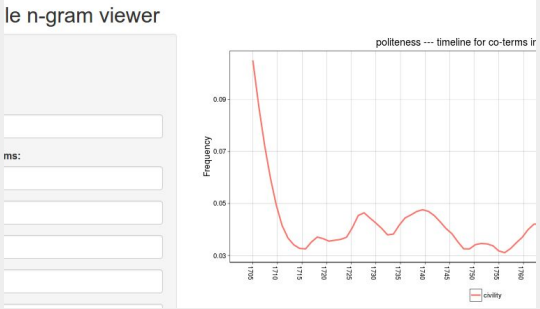
ECCO

ESTC

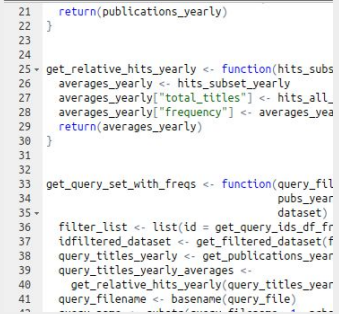
Evolving set of analysis and processing tools

APIs

Applications



Scripts



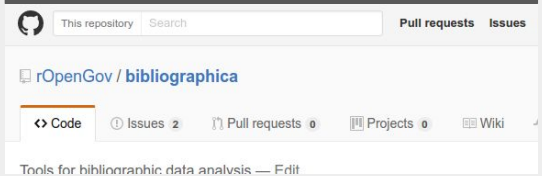
Etc...

...

Project goals

Research publications

Public tools, APIs, code



Refined data

publisher	latitude	longitude	publication_
printed for J. Cooke; and sold by the booksellers of B...	51.50853	-0.125740	London
re-printed at the Old Post Office in Fish shamble street	53.33306	-6.248890	Dublin
Sold at the Bible and Heart, in Cornhill	51.50853	-0.125740	London
printed for J. Bew	51.50853	-0.125740	London
NA	37.56596	14.282404	NA
Printed by J. Evans, Long-lane, West-smithfield, Lon...	51.50853	-0.125740	London
printed for Thomas Ewing, and William Smith	53.33306	-6.248890	Dublin
sold by J. Williams, and all the booksellers in town a...	51.50853	-0.125740	London

Challenges in reuse & sustainability

Data

- Often siloed for various different reasons
- Most humanities data not digitized currently (< 5% of relevant cultural heritage)
- data providers reluctant to share openly

Methods

- Often scattered
- Borrowed from other fields of science; do not necessarily suit the research questions and start guiding the work
- Reinventing the wheel

Expertise

- Often fragmented
- Cross-disciplinary collaboration often difficult because of the gaps in research cultures
- Ineffective collaboration between partners