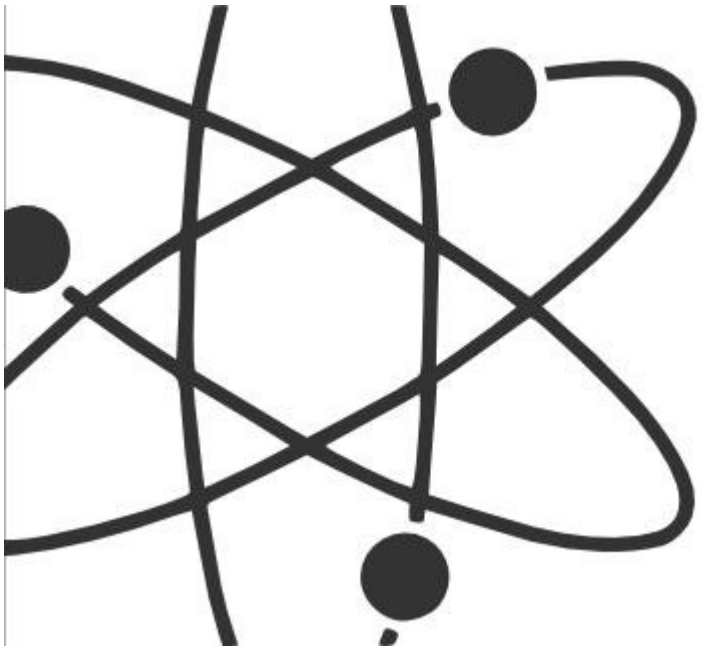
Digital Humanities, Open Science & Krakow DH 2016 Conference

Leo Lahti, University of Turku
http://www.iki.fi/Leo.Lahti

@antagomir



Open science means for instance open release of data, code, protocols, teaching material, publications, and the promotion of principles of openness, inclusivity and transparency in scientific research.

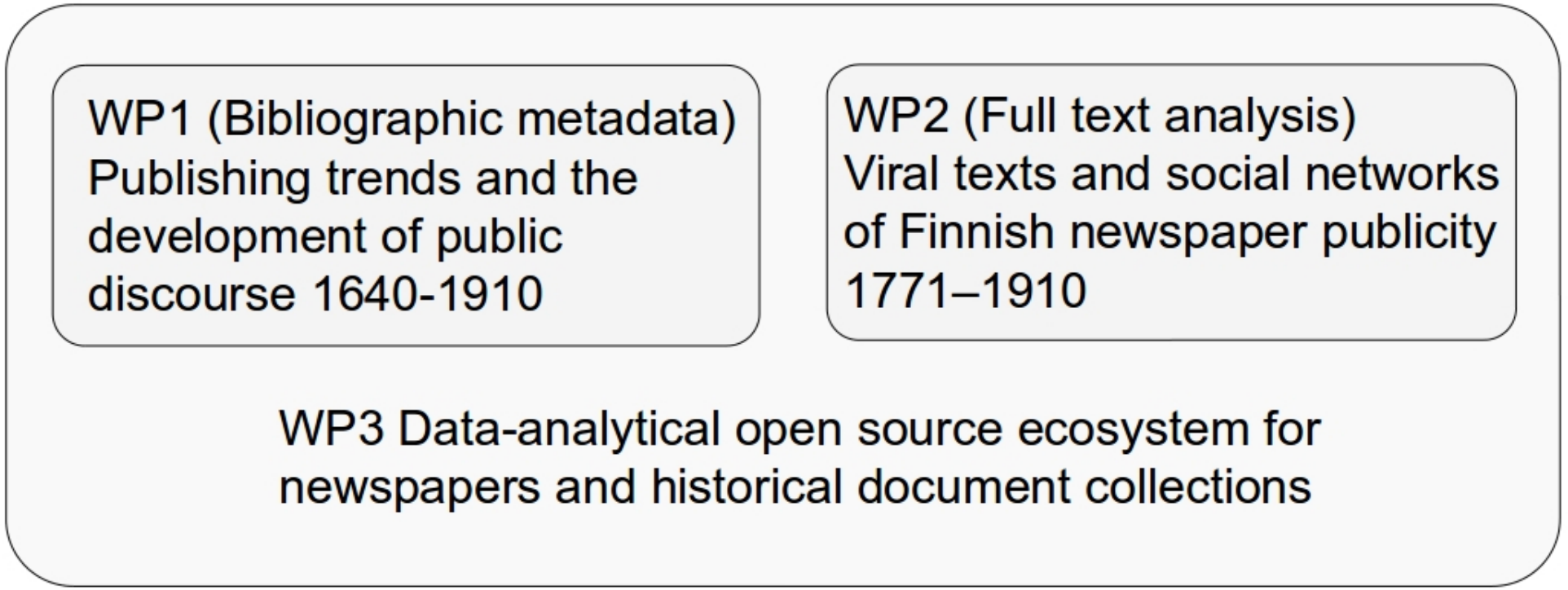
The **Open Science Finland** working group promotes openness in Finnish scientific and academic field.

Digital humanities

- Complements traditional humanities research
- Modern methods + classical questions
- Confirm earlier knowledge & discover new hypotheses
- Transparent, objective, quantitative ?
- Versatile field, many topics & approaches

Example: computational history

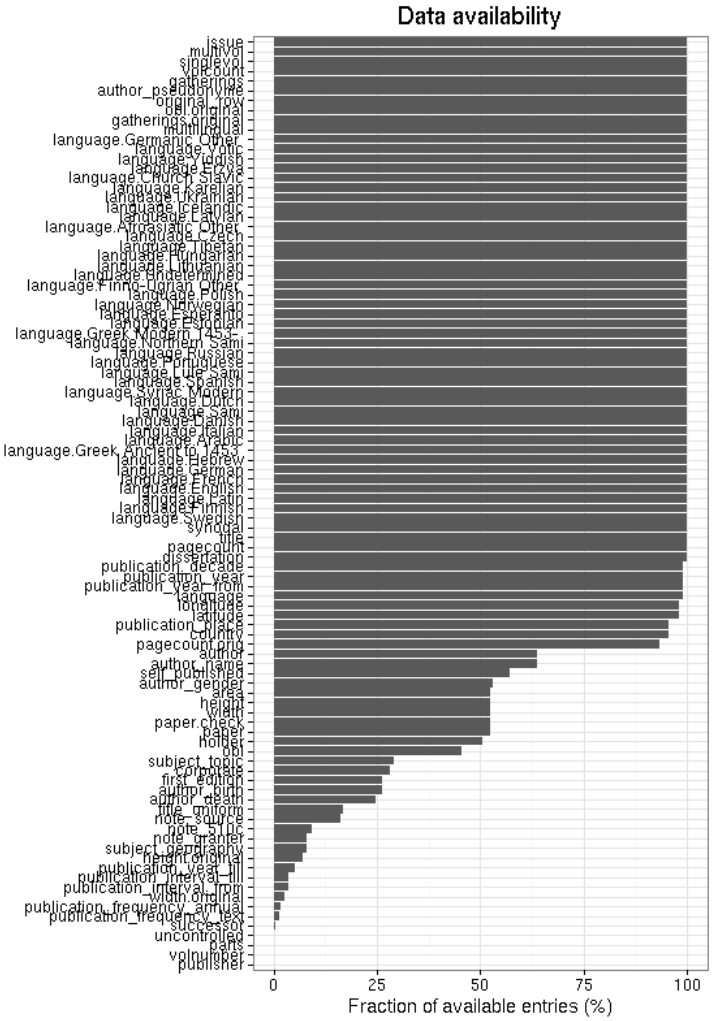
- Quantitative analysis of knowledge production in early modern era based on massive collections of bibliographic metadata & full texts
- Academy of Finland COMHIS consortium 2016-2019: University of Helsinki, University of Turku, National Library of Finland



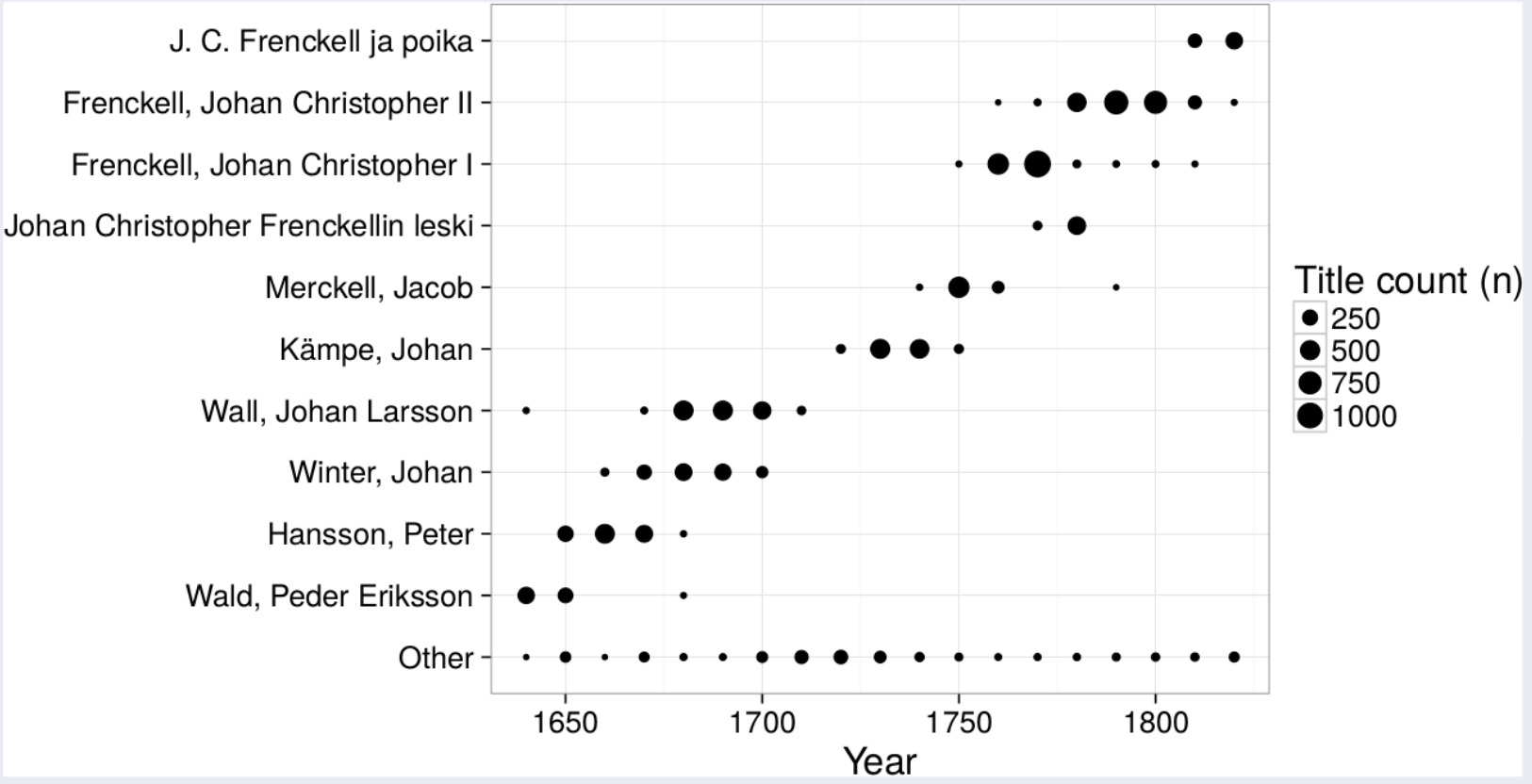
Fennica: national bibliography

Tools at http://github.com/rOpenGov/fennica

- 70,000+ documents metadata
- 1480-1910



Top publishers in Turku/Fennica



Parsed SQL data: further preprocessing needed

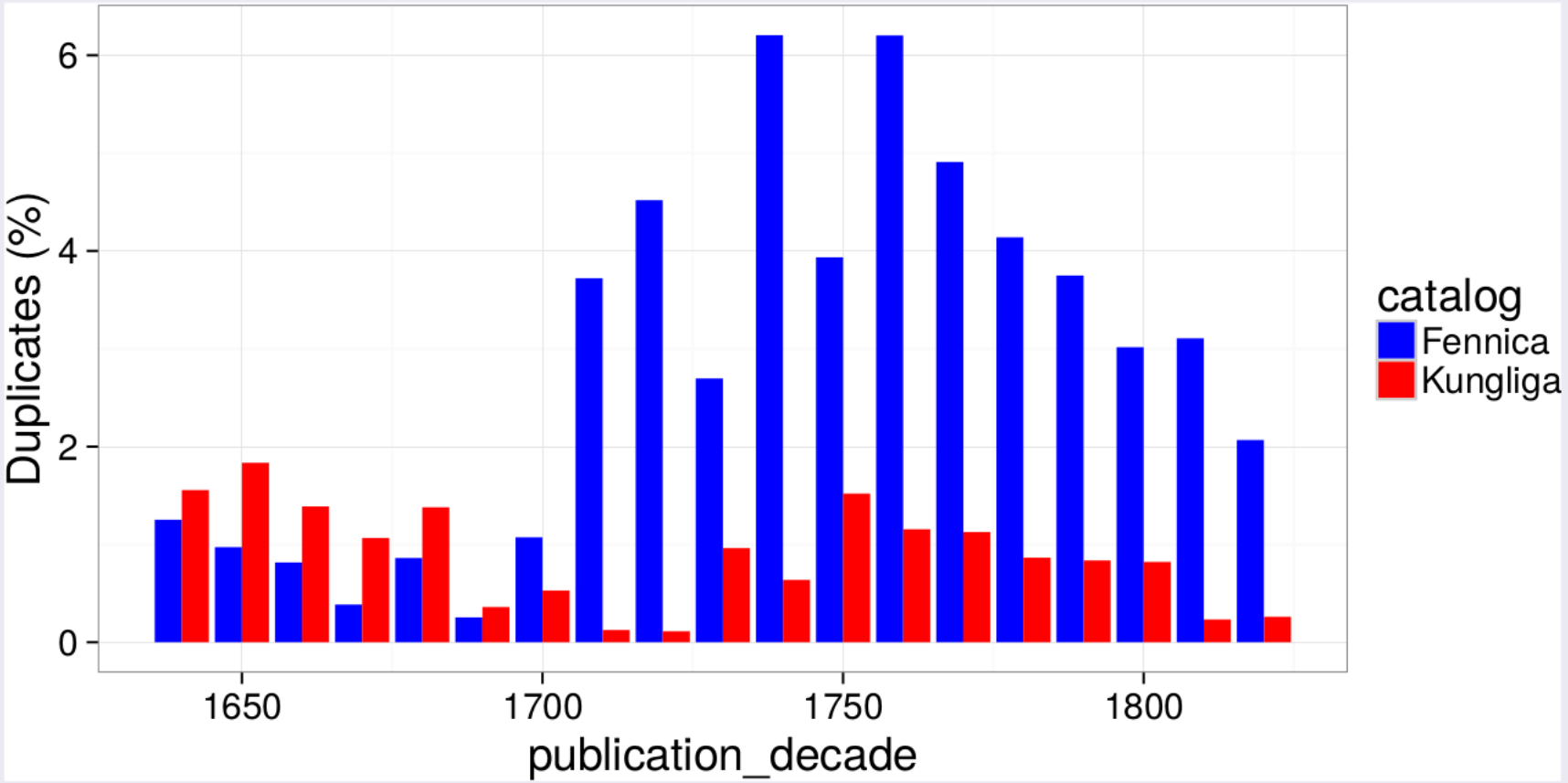
008, partial|100.a|100.d|100.d, partial|100.d, partial|240.n|245.a|260.a|260.b|260.b, partial|260.c|260.c, partial|300.a|300.c|650.a|650.y+651.y|650.z+651.a+651.z|65*-series|Language|Author|Author, dates|Author, birth|Author, death|Part/section of a work|Title|Place of publication|Publisher|Printed for|Publication date|Publication date, clean|Extent|Dimensions|Subject|Chronological subdivision|Geographic name and subdivision|Additional years

English|Bollan, William,|-1776. ||1776||Continued corruption, standing armies, and popular discontents considered;|London :|Printed, and sold by J. Almon, opposite Burlington-House, Piccadilly,||MDCCLXVIII. [1768] |1768|82p.,plate ;|4^{o}. ||Revolution, 1775-1783|United States|;;;;1775;1783

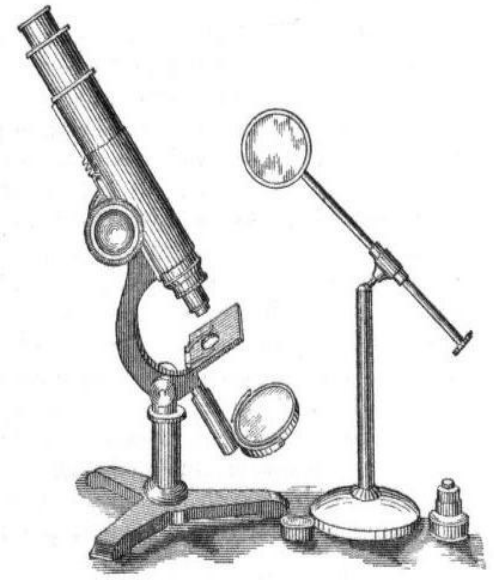
English|Trimmer,|1741-1810. |1741|1810||A description of a set of prints of English history;|London :|printed by John Marshall, at No. 4, Aldermary Church-Yard, Bow-Lane, and No. 17, Queen-Street, Cheapside,|||[between 1790 and 1800?] |1800|[2],261,[1];[2],247,[1]p. ;|24^{o}. |||Great Britain|;;;;;

English|Trimmer,|1741-1810. |1741|1810||A description of a set of prints of English history;|London :|printed and sold by John Marshall, No. 4, Aldermary Church Yard, Bow-Lane, Cheapside,|||[between 1799 and 1801?] |1801|[2],261,[1];[2],247,[1]p. ;|24^{o}. |||Great Britain|;;;;;

Duplicate publications in Fennica/Kungliga



Computation
is a new microscope
for studying massive data sets



Scaling up analysis by automatization !

Automatize everything that can be automatized:
50,000 documents - 20 minutes

[4],vii-xii,[4],222p.,plate

→ 240 pages.

Preprocess & Enrich

Clean up messy entries

```
polish_physical_extent("iii-xxiv, 118, [2] p.")$pagecount
```

```
## [1] 142
```

Enrich data (geocoordinates, gender, ..)

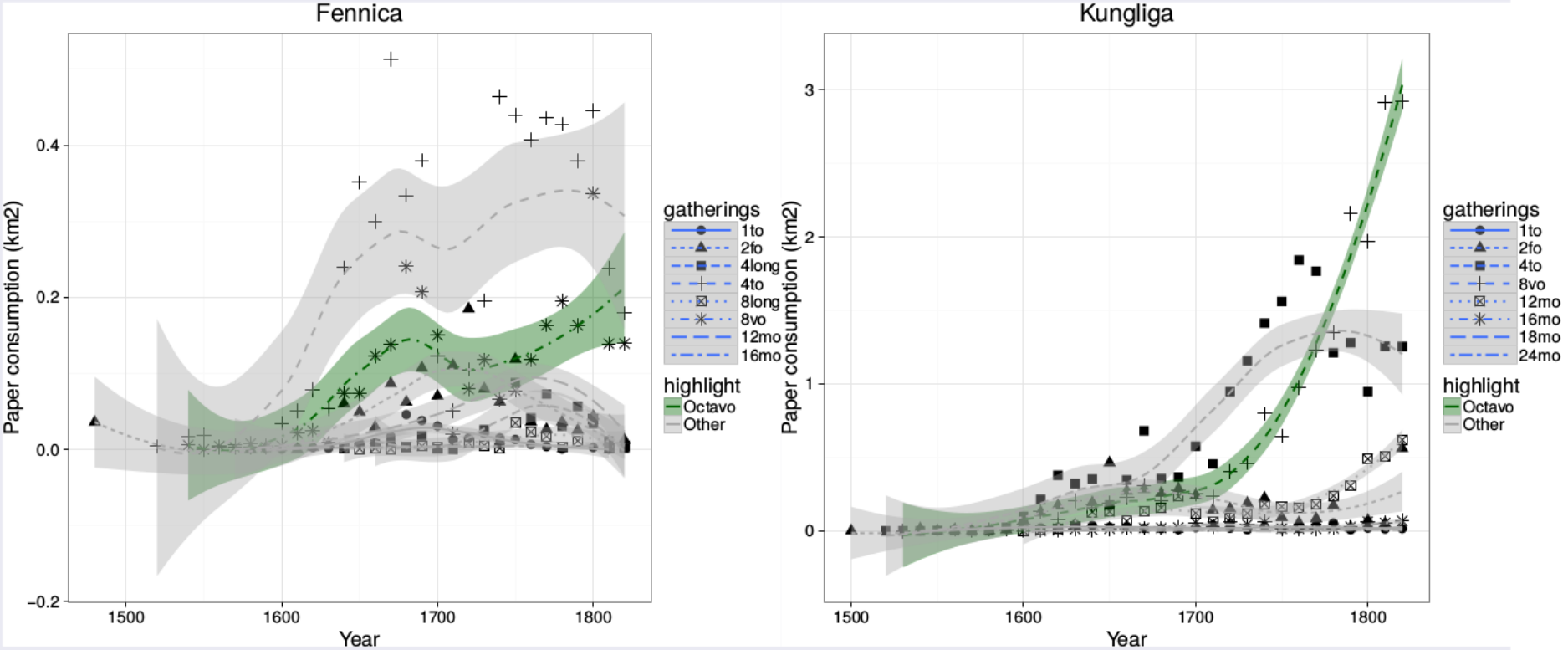
```
get_country("Porvoo")
```

```
## [1] "Finland"
```

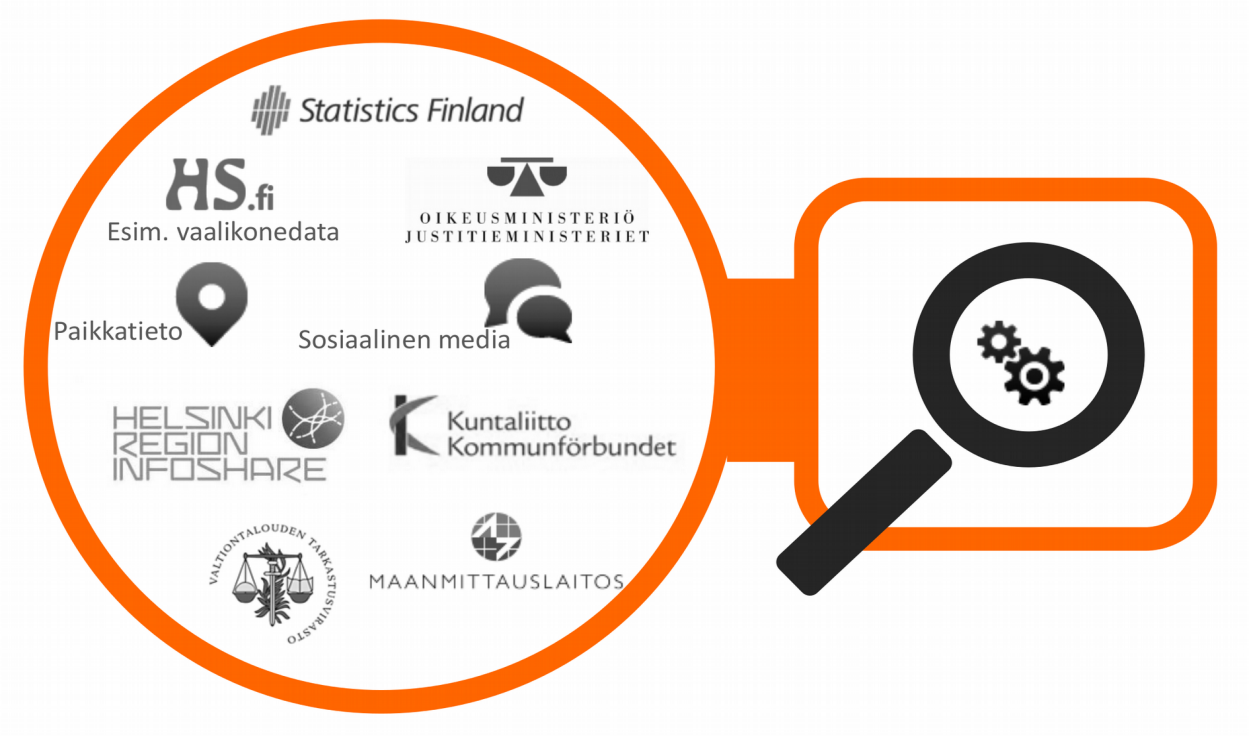
- Parse, clean up, enrich, summarise, analyze, visualize, report..

Rise of the octavo sized book as vehicle of Enlightenment

Paper consumption: different document formats over time



Fully transparent data analytics



datavaalit

OPEN KNOWLEDGE FINLAND

Apps4Finland
doing good with open data

ropenGov

Analysis tools for R growing exponentially

github
SOCIAL CODING

R

rOpenGov

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

ROpenSci

Issue with negative values (being converted to absolute values) #38

Edit New issue

Open **Hergie** opened this issue 21 days ago · 6 comments

Hergie commented 21 days ago

Hello,
Thank you so much for the very useful package! It's been a great help.

However, I encountered a problem when downloading eurostat series with negative values. In R these don't get recorded as negative but as positive (i.e. the absolute value).

This happened to me when downloading tables [nama_nace31_k] & [nama_nace21_k]. If you choose variable indic_na="B1G" & unit="PCH_PRE", i.e percentage change in gross value added, all the values are larger than 0. This is even though the volume series in levels exhibit also decreases from time to time.

What am I doing wrong or is there some kind of bug?

Many thanks!

jhuovari commented 21 days ago Collaborator

Hi,

I don't seems to be able to replicate your problem. With:

```
library("eurostat")
dat <- get_eurostat("nama_nace31_k")
subset(dat, indic_na == "B1G" & unit == "PCH_PRE" & nace_r1 == "D" & geo == "FI")
```

Labels
None yet

Milestone
No milestone

Assignee
No one—assign yourself

Notifications
Unsubscribe
You're receiving notifications because you're subscribed to this repository.

3 participants

Lock conversation

Data Science Workflow

Customized our dedicated tools for bibliographic analysis combine flexibility with automation

Automated scales well with new data and unforeseen changes; a million documents cleaned up in 30 minutes

Open source fully transparent and reproducible data analytics

Research potential greatly underestimated; world is full of metadata collections

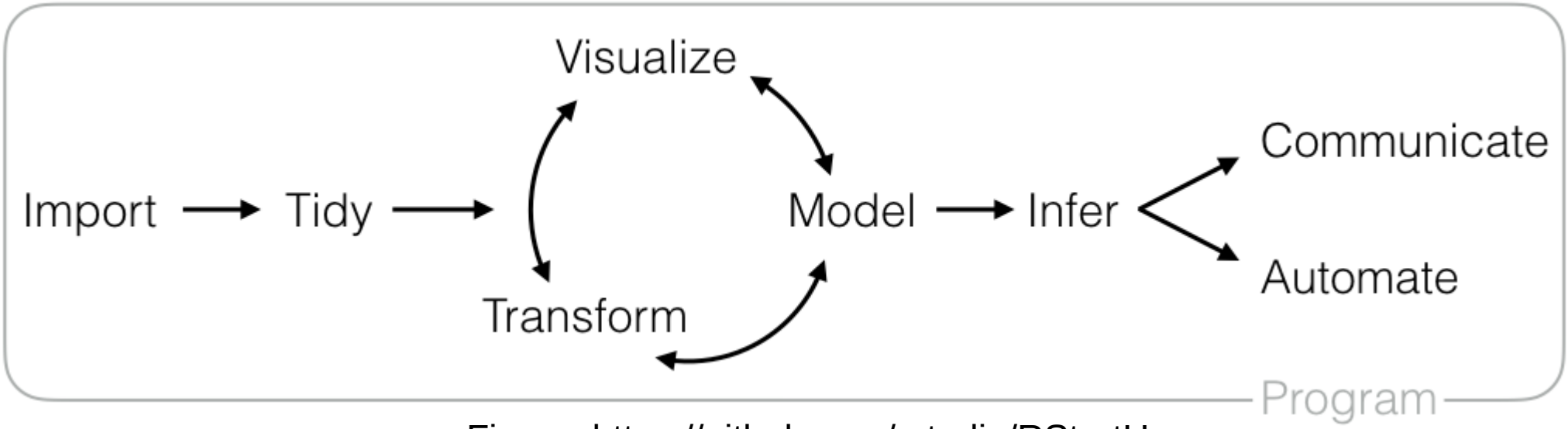
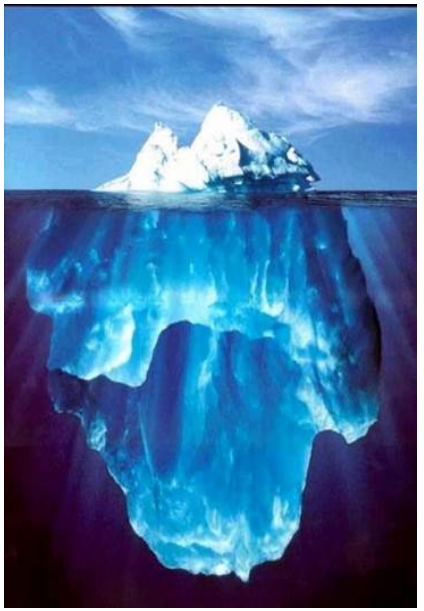


Figure: https://github.com/rstudio/RStartHere

Not everything can be automatized

→ automatize everything you can !



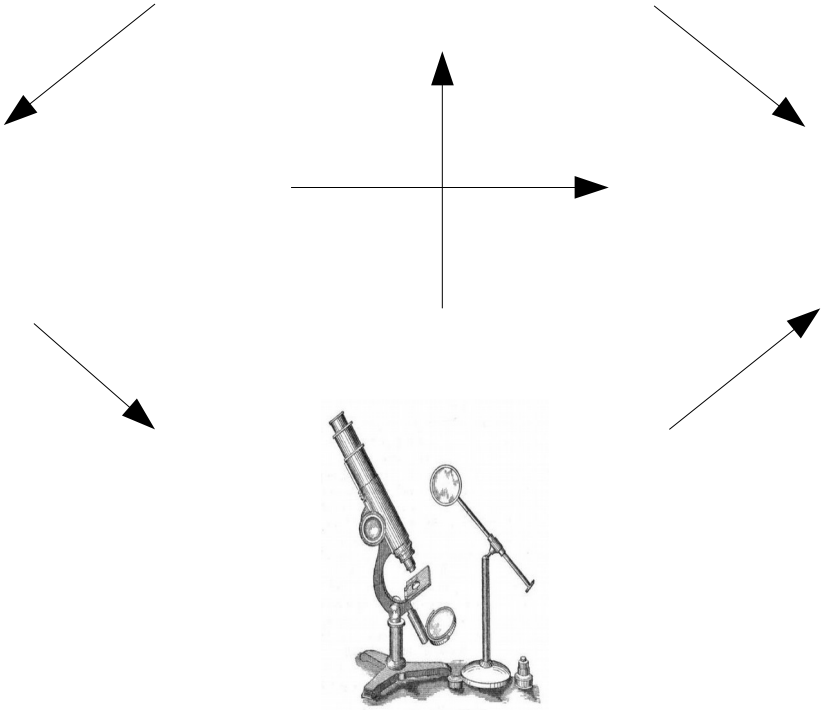
Hypothesis testing vs. hypothesis discovery?

?

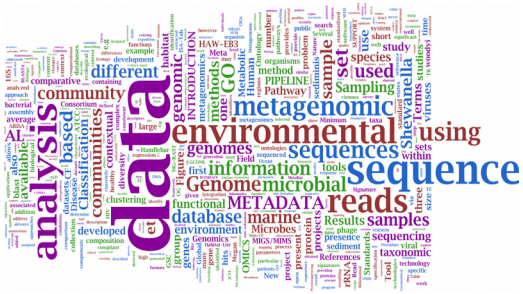
Hypothesis

Method

$$\begin{cases} X = W_x \mathbf{z} + \varepsilon_x \\ Y = W_y \mathbf{z} + \varepsilon_y \end{cases}$$



Tools



Data

Some key challenges

Data

- Quality & reliability
- Coverage & sample size
- Collection
- Source critique
- Background information

Methods

- Quality & reliability
- Feasibility of underlying assumptions
- Multiple testing
- Statistical significance
- Data integration
- Interpretation
- Visualization

Research questions

- Feasibility
- Relevance
- Clarity
- Expertise
- Privacy
- Research ethics

Data parsing Integration Significance Publication Enrichment

Preprocessing Analysis Visualization Reuse

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: Aug 30, 2005 • DOI: 10.1371/journal.pmed.0020124

PLOS | MEDICINE

Browse | For Authors

OPEN ACCESS

ESSAY

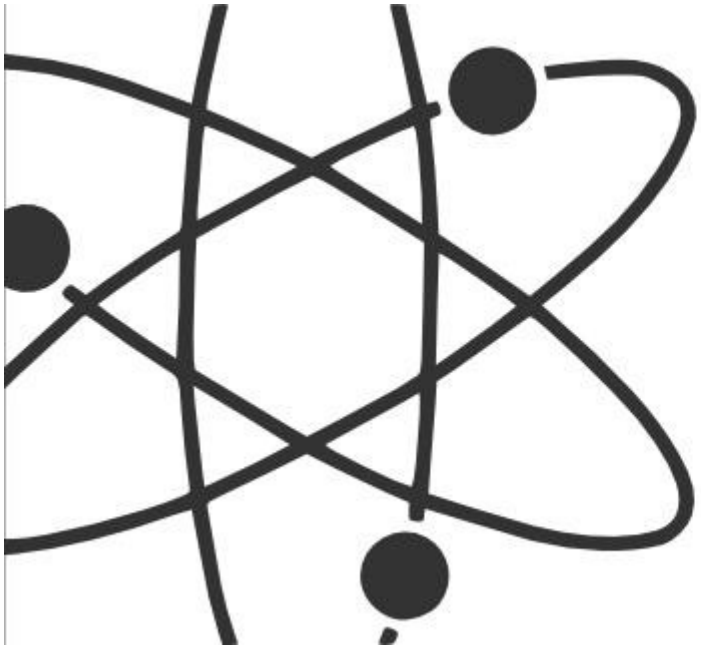
How to Make More Published Research True

John P. A. Ioannidis

Published: October 21, 2014 • DOI: 10.1371/journal.pmed.1001747

Open Science

Data – Methods – Publications - Protocols



Open science means for instance open release of data, code, protocols, teaching material, publications, and the promotion of principles of openness, inclusivity and transparency in scientific research.

The **Open Science Finland** working group promotes openness in Finnish scientific and academic field.

www: http://fi.okfn.org/wg/openscience/

FB: Open Science Finland

Other aspects of Open Science

- Open Access & Public Preprint Repositories (arXiv etc)
- Open Data
- Open Source & Methods
- Open Peer Review

All aspects of the research process, to the extent allowed by modern information technology & ethical & legal considerations !

Importance:
- Transparency, reproducibility
- Efficiency, scalability
- New collaboration modes
- New knowledge

Open Data

1) Opening up research material

2) Utilizing open data resources

Bottlenecks for data utilization

- Scattered information
- Poor accessibility
- Dirty data
- Disappearing & missing data
- Incompatible data
- Lack of analytical tools
- Privacy issues
- Commercial distribution models
- Lack of collaboration

Traditional conventions, inefficient incentives and funding mechanisms, and technical challenges form **bottlenecks for sharing of research code and data**

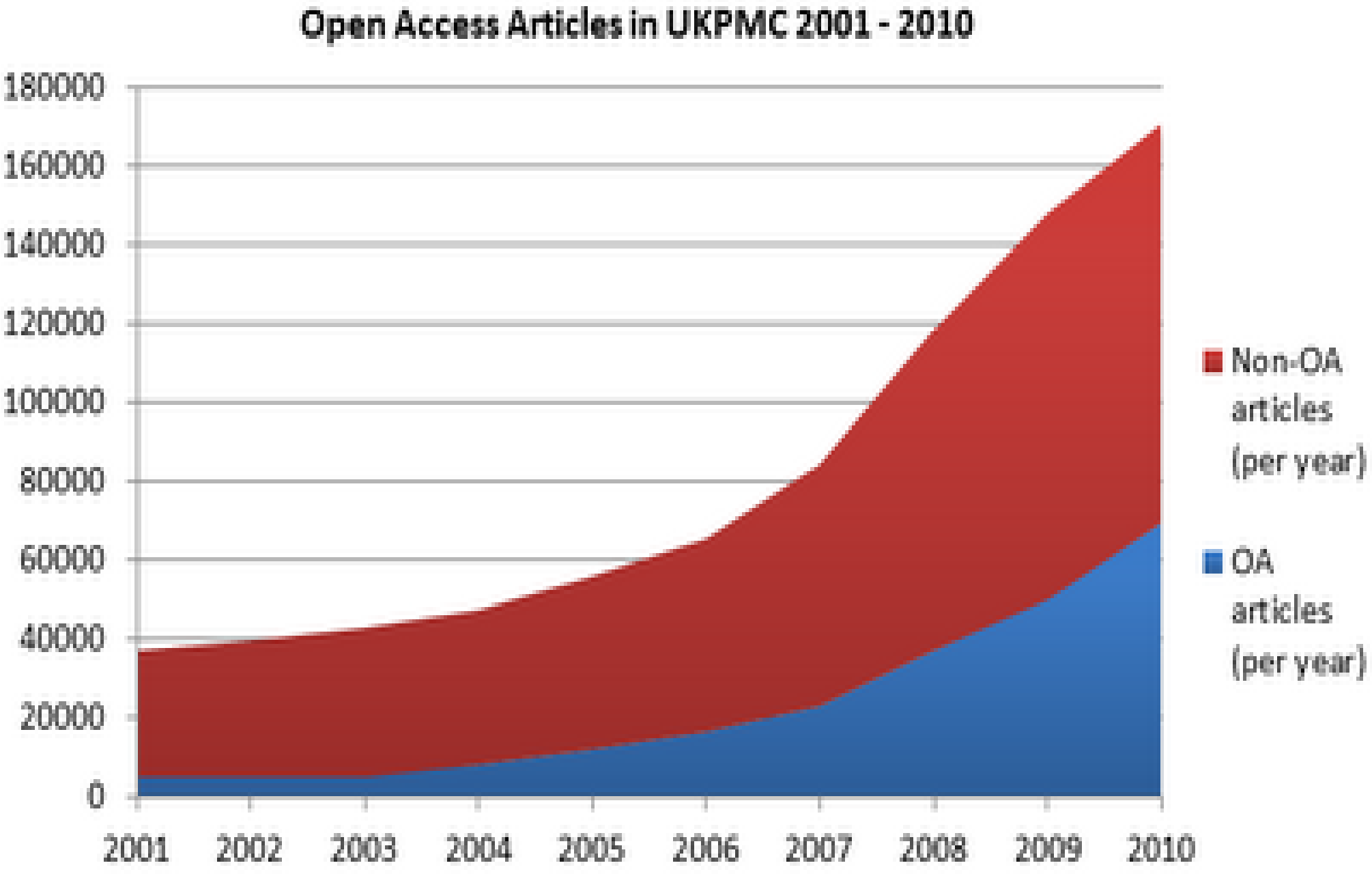
Why open data principles are not comprehended in the Humanities?

Institutions reluctant to give full access to raw data (even to researchers of the same institution). Why?

Research process not opened and research data is not shared in the Humanities. Transparency, reproduction, collaboration, new initiatives are missing. Why?

Cultural change takes time. Need for concrete & successful examples in the core field of the Humanities.

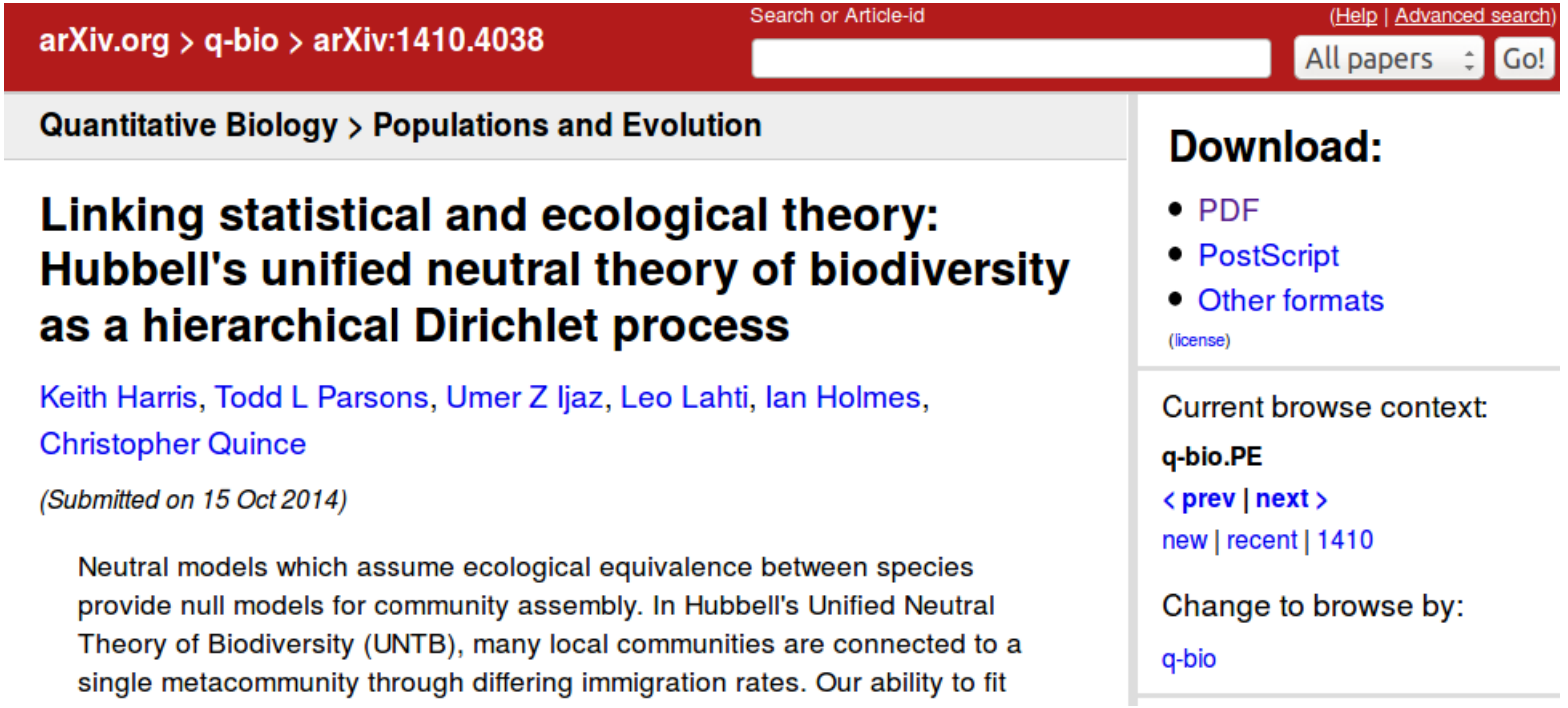
Open access publishing development 2001-2010



Public participation: preprints & comments

Preprint archives (arXiv; biorXiv; facilitates commenting during review process)

Public commenting option on published articles (experimented by Nature; PLoS..)



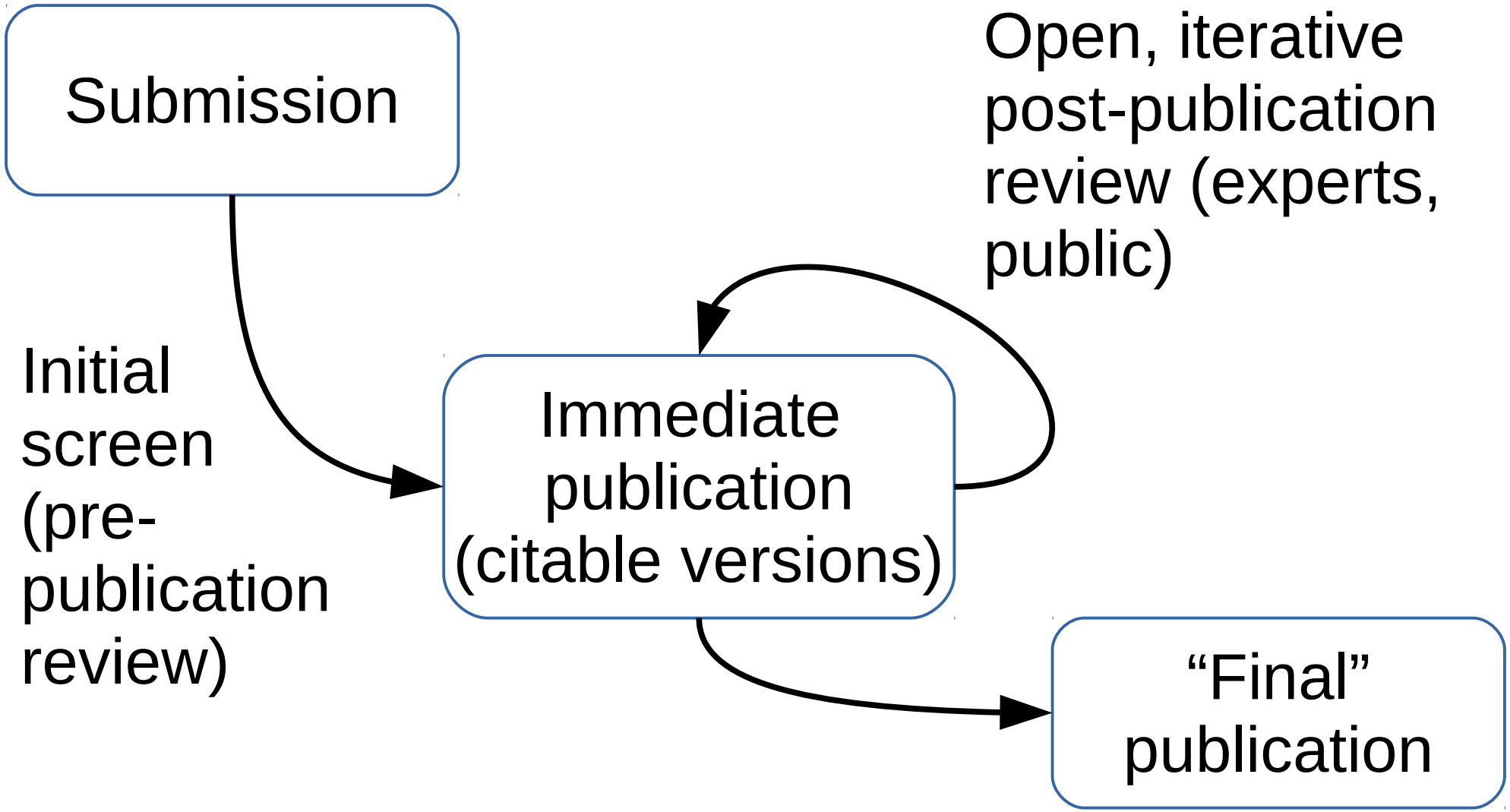
Open review is increasingly popular

PeerJ

80% of authors publish full peer review history

40% of reviewers sign their reviews

Pre- vs. post-publication review



Krakow DH 2016 conference (July 11-16)

- 650 abstracts (short & long papers)
- Posters
- Pre-conference workshops
- Panels
- Keynote presentations
- Networking..

Abstracts:
http://dh2016.adho.org/abstracts/

Our material is shared at:
https://github.com/rOpenGov/fennica

Krakow workshops

http://dh2016.adho.org/workshops/

- What Next For Crowdsourcing?
- Big Data: Complex Systems and Text Analysis
- Modeling, sharing and analyzing people’s lives
- Research data preservation and presentation
- Web Toolkit for Spatio-Temporal Network Mapping
- Music Information Retrieval Algorithms for Oral History Collections
- Complex social dynamics in a few lines of code
- HistoInformatics 2016; Historical Text Reuse Tutorial
- Introduction to Natural Language Processing

Thank You!

http://github.com/rOpenGov/fennica

Mikko Tolonen
Hege Roivainen
Jani Marjanen
Hannu Salmi
Eetu Mäkelä
Niko Välimäki
Samuel Myllykangas
Juuso Parkkinen
Joona Lehtomäki
Markus Kainu
and many others !

@antagomir