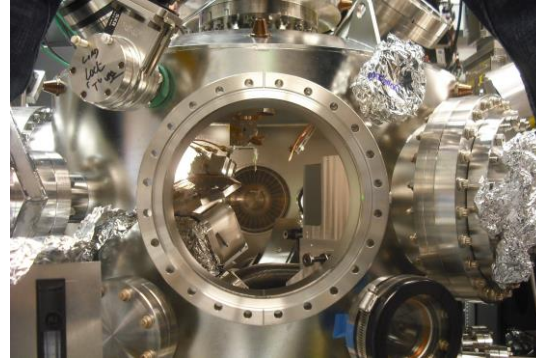# Open scholarly metadata

Dr. Teresa Kubacka
ETH Library
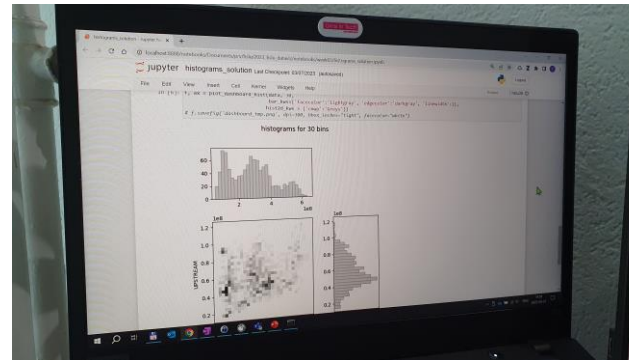Open Data Forum, 16.06.2023

Data scientist at a library?

I belong to a group of people who studied physics…
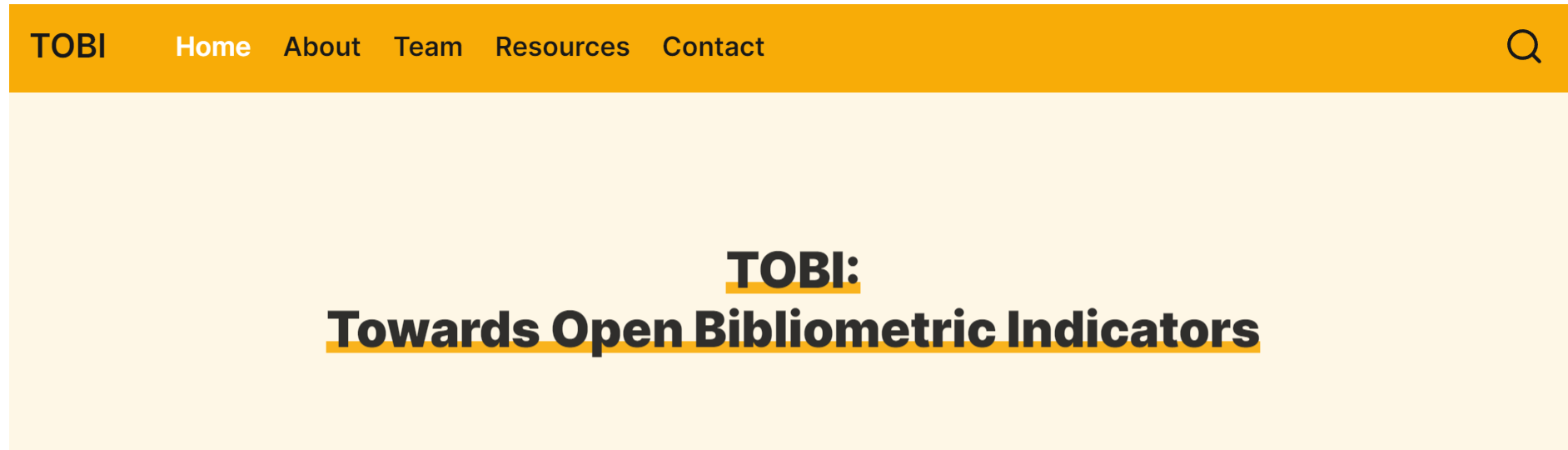
…but now do something else.

Project TOBI

TOBI   Home   About   Team   Resources   Contact

TOBI:
Towards Open Bibliometric Indicators

ETH zürich   ETH Library   swissuniversities

TOBI is a project lead by the ETH Library and co-funded by swissuniversities. It aims to evaluate the quality of open bibliometric data sources regarding research affiliated with Swiss Higher Education Institutions. The project has started in Q1 2023 and will wrap up in Q4 2024.

I will glimpse over many topics, so reach out to me afterwards!

ETH zürich

ETH Library

- ETH Zurich: ~11 000 employees, ETH Library: ~300 employees
- ETH Library provides services for the whole ETHZ community
- 3 main areas of competence:

| Books, collections and archives	Research support services	Trainings
- Books, journals, other media - Archives (old periodicals, books; historical documents of ETHZ) - Special collections (e.g. entomological, mycological, materials, history of science, maps, photos, literature…) - Digitalization - Libraries as physical places (e.g. place of study)	- Scientific communication, contracts with publishers, APCs - Open Access repository - Research data management and long-term archiving **Scientometrics** - Academic librarianship, literature search - Scientific writing and knowledge management	- Trainings and workshops for students and researchers - Consultations ("Book a librarian")

Scientometrics team at the ETH Library

ETH Library → Research Support Services → Knowledge Management (KOM)

Left to right:
Annette Guignard
Dr. Roland Suri
Mahmoud Hemila
Dr. David Johann
Dr. Teresa Kubacka
Dr. Julian Dederke
Simon Willemin
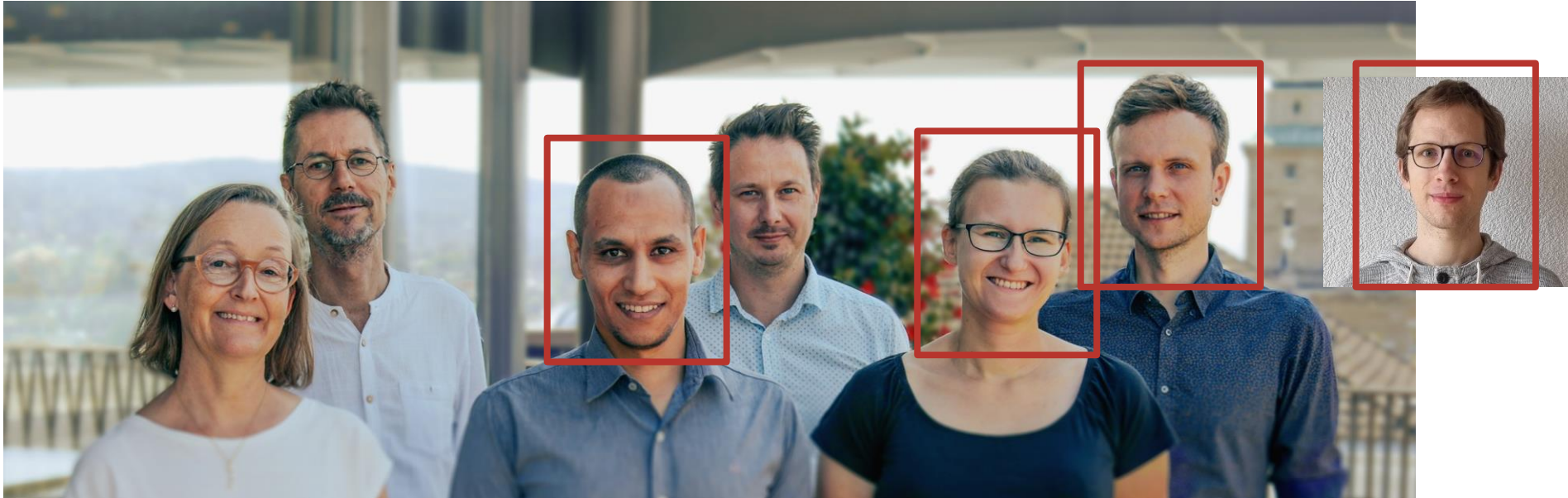
- Services in the field of scientometrics (Bibliometrics/Altmetrics)
- Information management / Tools that help you stay up to date

We started in 2020 and officially provide services since July 2021.

What is scholarly metadata?

Full-text search (incl. title, abstract)

Journal, publisher

Authors

Date

Grant ID

Keywords (given or extracted)

Metadata: Research field, topic …

Citations (ingoing and outgoing)

Affiliations

DOI and reference

ETH zürich   ETH Library   DOI: 10.1139/cgj-2016-0338   03.07.2023   6
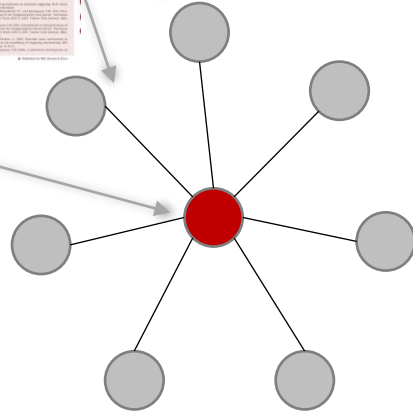
Source paper

Reference (1st degree)
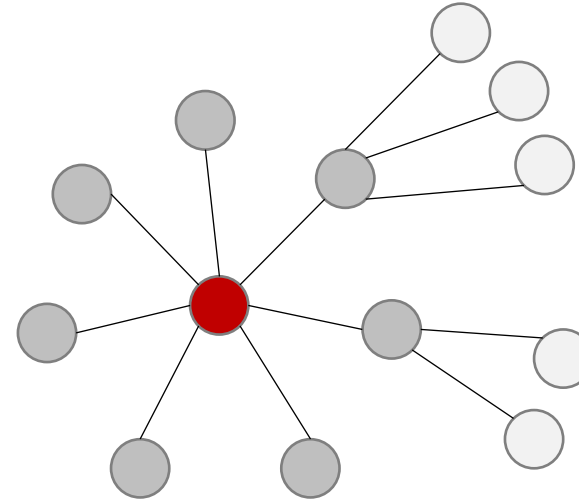
Source paper

Reference (1st degree)

Reference (2nd degree)

Because the scientists cite each other's work, we can build a **network of references and citations**

What do we need the metadata for?

- Knowledge discovery and literature search

- Research assessment, strategic research management

- Science on science, digital humanities

ETH *zürich*

Knowledge discovery

Combine keywords with analytics to drill down
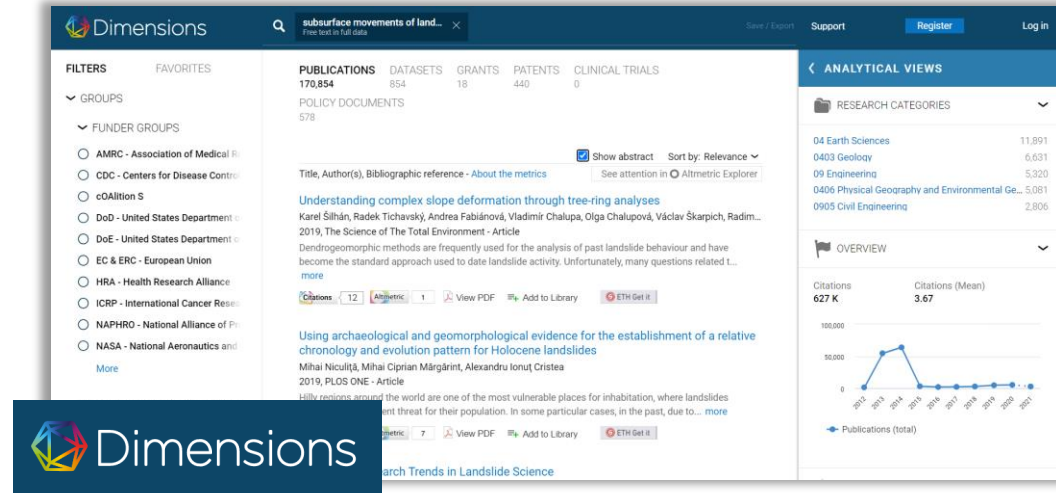
Networks based on text similarity
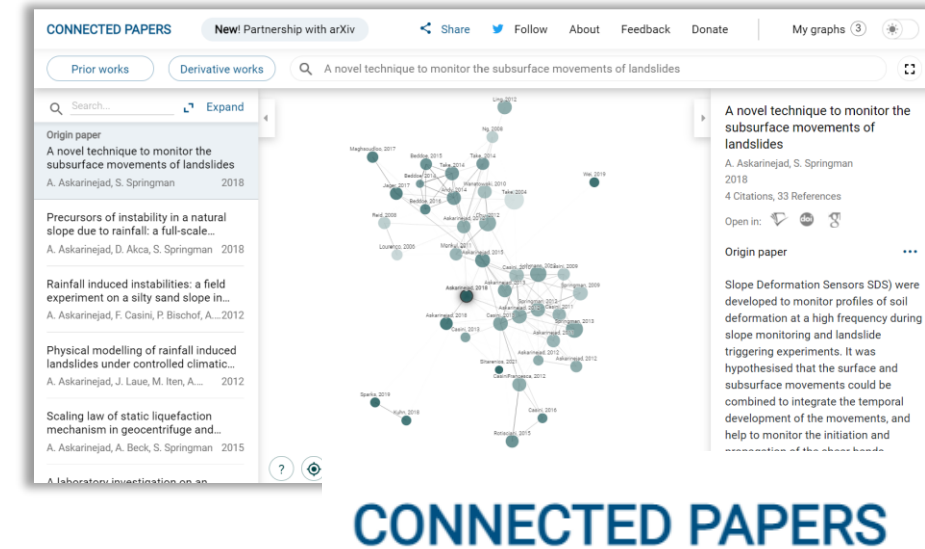
Which citations confirm given study or provide contrasting evidence

ETHzürich
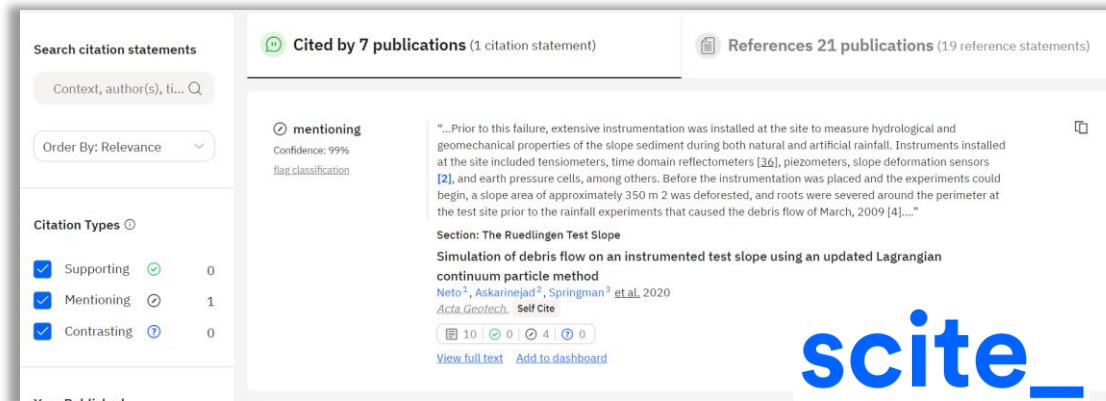
Scientometrics for research assessment

Research process

Scientific communication

Quantitative research assessment
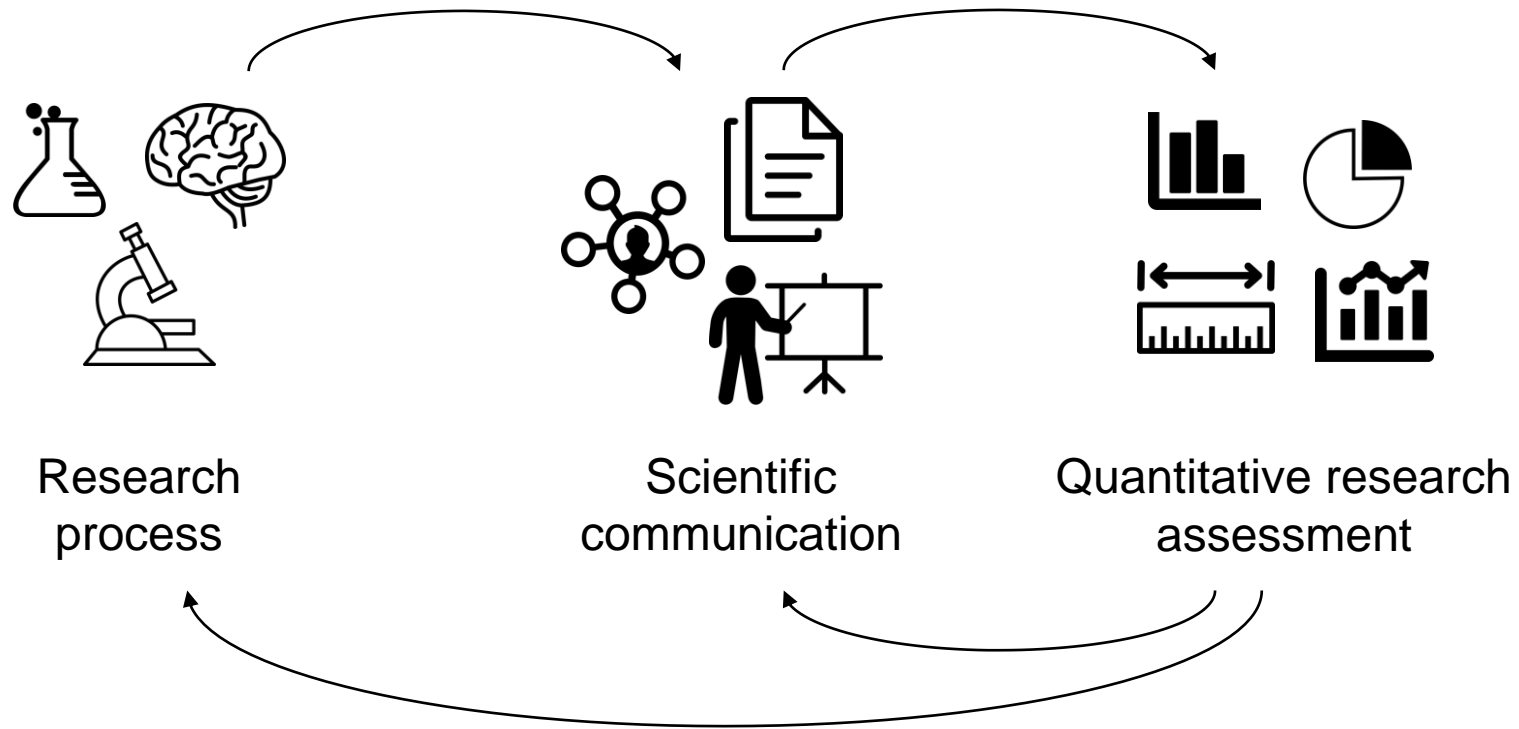
Funders

Publishers

Government

Industry

ETH zürich
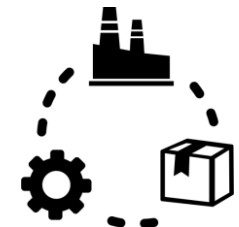
Science on science and digital humanities

ETH zürich

https://static.nomic.ai/pubmed.html
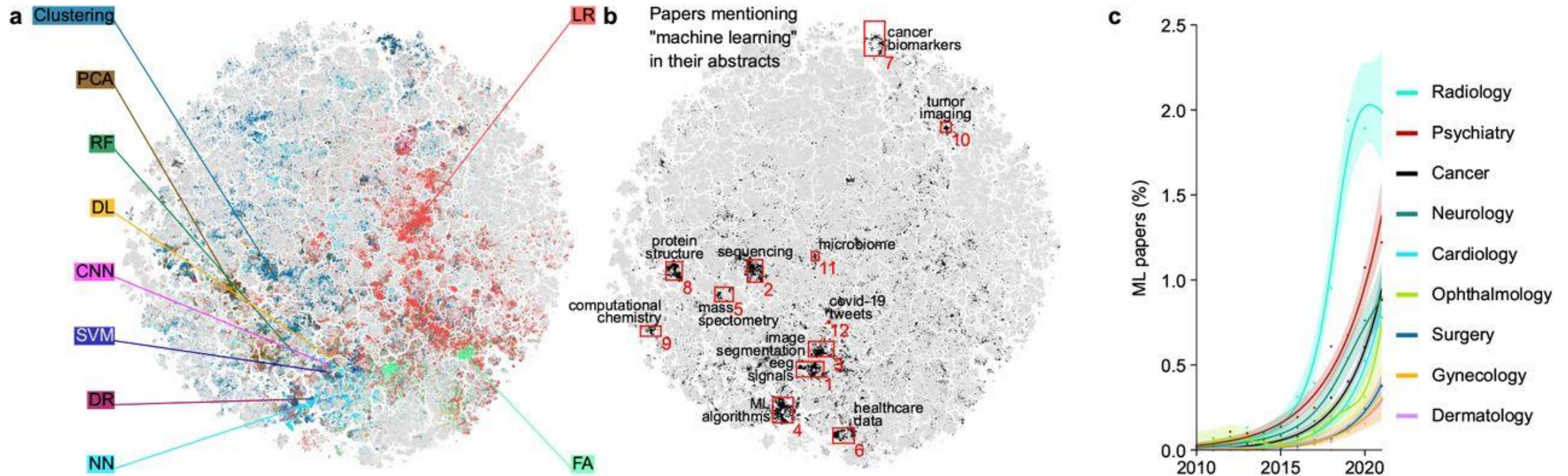
Problem: Scientific research is paywalled!

Paywalled science: single articles

General cross-coupling reactions with adaptive dynamic homogeneous catalysis

Indrajit Ghosh ✉, Nikita Shlapakov, Tobias A. Karl, Jonas Düker, Maksim Nikitin, Julia V. Burykina, Valentine P. Ananikov & Burkhard König ✉

Nature (2023) | Cite this article

66 Altmetric | Metrics

Abstract

Cross-coupling reactions are among the most important transformations in modern organic synthesis[1,2,3]. Although the range of reported (het)aryl halides and nucleophile coupling partners is very large considering various protocols, the reaction conditions vary considerably between compound classes, necessitating renewed case-by-case optimization of the reaction conditions[4]. Here we introduce adaptive dynamic homogeneous catalysis (AD-HoC) with nickel under visible-light-driven redox reaction conditions for general C(sp^2)–(hetero)atom coupling reactions. The self-adjustive nature of the catalytic system allowed the simple classification of dozens of various classes of nucleophiles in cross-coupling reactions. This is synthetically demonstrated in nine different bond-forming reactions (in this case, C(sp^2)–S, Se, N, P, B, O, C(sp^3, sp^2, sp), Si, Cl) with hundreds of synthetic examples under predictable reaction conditions. The catalytic reaction centre(s) and conditions differ from one another by the added nucleophile, or if required, a commercially available inexpensive amine base.

This is a preview of subscription content, access via your institution

Access options

Access through your institution

Access Nature and 54 other Nature Portfolio journals

Get Nature+, our best-value online-access subscription

24,99 € / 30 days

cancel any time

Learn more

Rent or buy this article

Get just this article for as long as you need it

$39.95

Learn more

Subscribe to this journal

Receive 51 print issues and online access

We are sorry, but there is no personal subscription option available for your country.

More Information

Prices may be subject to local taxes which are calculated during checkout

Paywalled science: price we pay for monopoly

- Scientists need to read other people's work to do their job

- Scientists need to publish their work so that others can re-use them

- Gold standard is the peer-review ("fact-checking", integrity check)

- Peer-review process has been captured by the journal publishers

Scientist	**Editor** $$$
Paid by the taxpayer	Paid by the publisher

Scientist	Editor
- Produce the content	- Get rid of obvious fakes
- Typeset the manuscript	- Language check
- Review work of others	- Find the reviewer
- Pay the "processing fee"	- Put the text online / in print
- Pay the Open Access fee	
- Pay to read the article	

Robert Maxwell: *"I set up a perpetual financing machine through advance subscriptions as well as the profits on the sales themselves. It is a cash generator twice over"*

Learned Publishing Vol 3 (2) pp 73-77

ETH *zürich*

OpenAPC: how much does it cost to publish?

OpenAPC: Crowdsourced data from the libraries
on the fees they paid for publishing articles
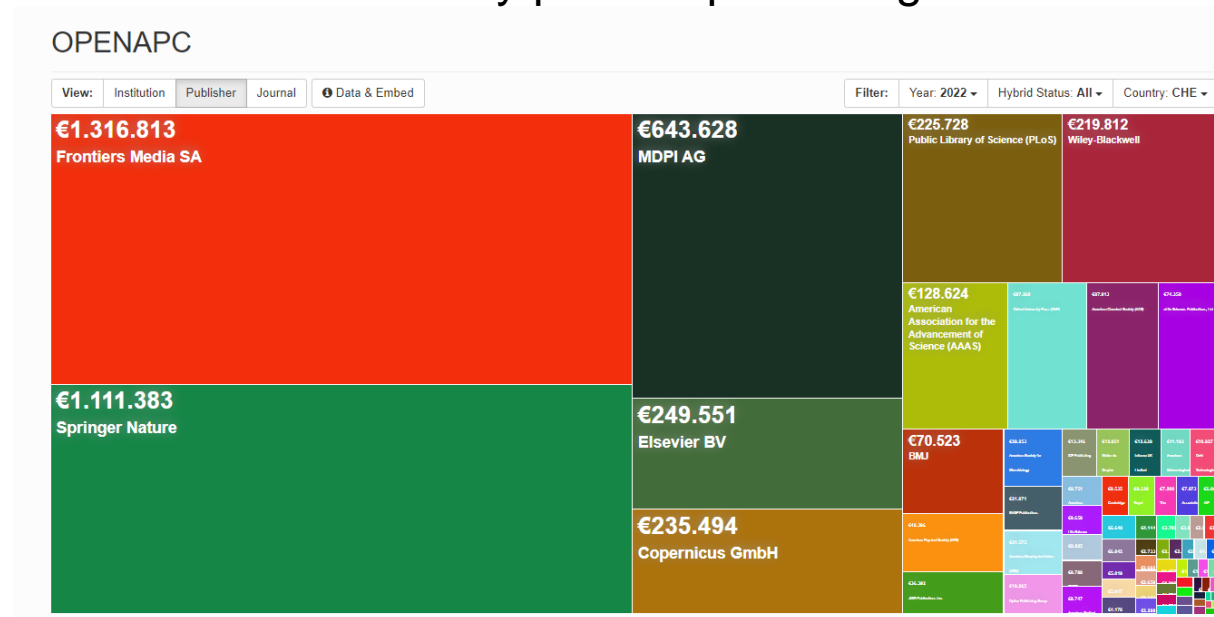
Total CH: ca. **50 000 publications in 2022**

ETH zürich

The big business of academic publishing

Big publishers have profit margins higher than big tech!

The core of Elsevier's operation is in scientific journals, the weekly or monthly publications in which scientists share their results. Despite the narrow audience, scientific publishing is a remarkably big business. With total global revenues of more than £19bn, it weighs in somewhere between the recording and the film industries in size, but it is far more profitable. In 2010, Elsevier's scientific publishing arm reported profits of £724m on just over £2bn in revenue. It was a 36% margin - higher than Apple, Google, or Amazon posted that year.

https://www.theguardian.com/science/2017/jun/27/profitable
-business-scientific-publishing-bad-for-science

Apple FY22 Income Statement

https://www.sankeyart.com/sankeys/public/all/81/
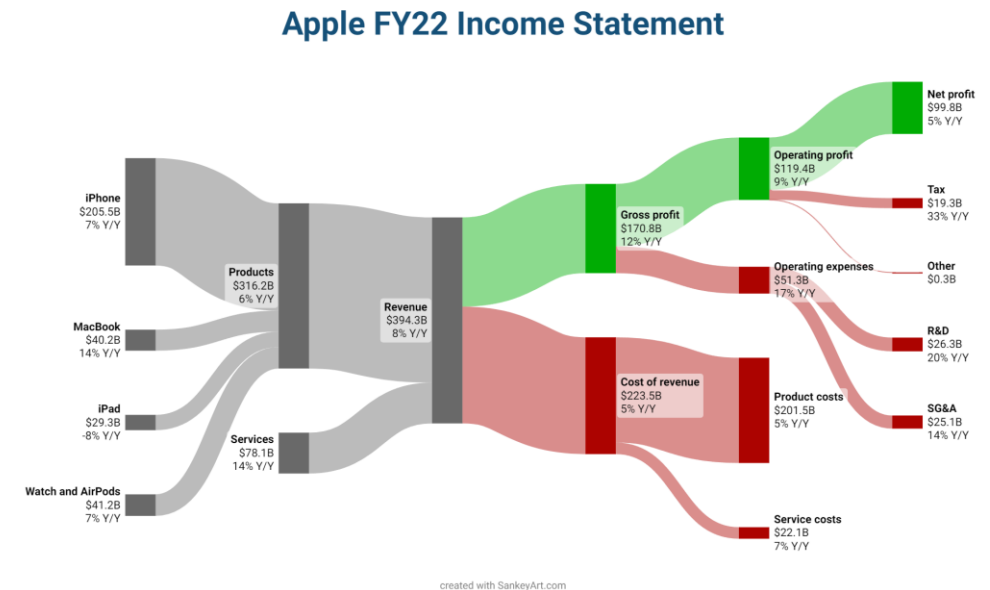
They try to secure additional income sources: selling usage data, licenses for metadata databases…

Scrape and pirate is not an option

One of the world's largest science publishers, Elsevier, won a default legal judgement on 21 June against websites that provide illicit access to tens of millions of research papers and books. A New York district court awarded Elsevier US$15 million in damages for copyright infringement by Sci-Hub, the Library of Genesis (LibGen) project and related sites.

"The Court has not mistaken illegal activity for a public good," said Maria A. Pallante, president and CEO of the Association of American Publishers – a trade group that Elsevier belongs to – in a statement released on 22 June. "On the contrary, it has recognized the defendants' operation for the flagrant and sweeping infringement that it really is and affirmed the critical role of copyright law in furthering scientific research and the public interest."

https://www.nature.com/articles/nature.2017.22196

sci-hub

enter your reference

🔑 open

common ownership of the means of production,
free access to articles of consumption

database   about   Elbakyan   stats   donate
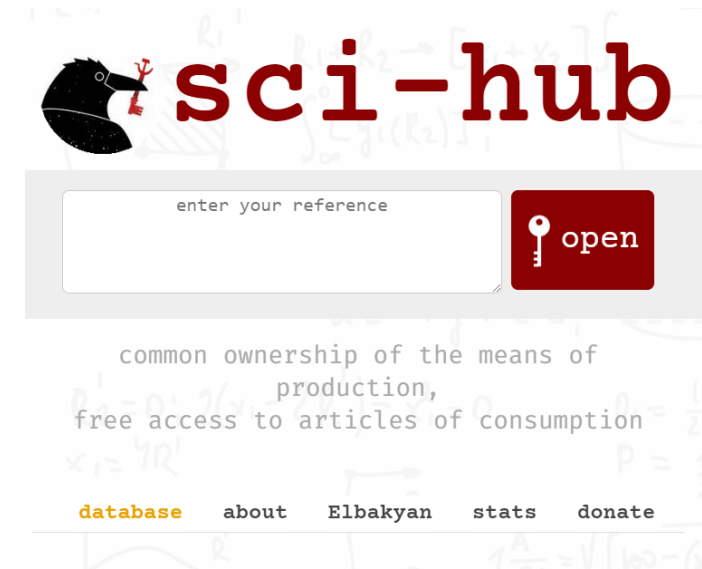
Courts on the side of publishers!
You can get in big trouble

Paywalled science: databases

Metadata databases are crucial for research assessment and
literature search – they got paywalled too!

https://www.scopus.com/home.uri
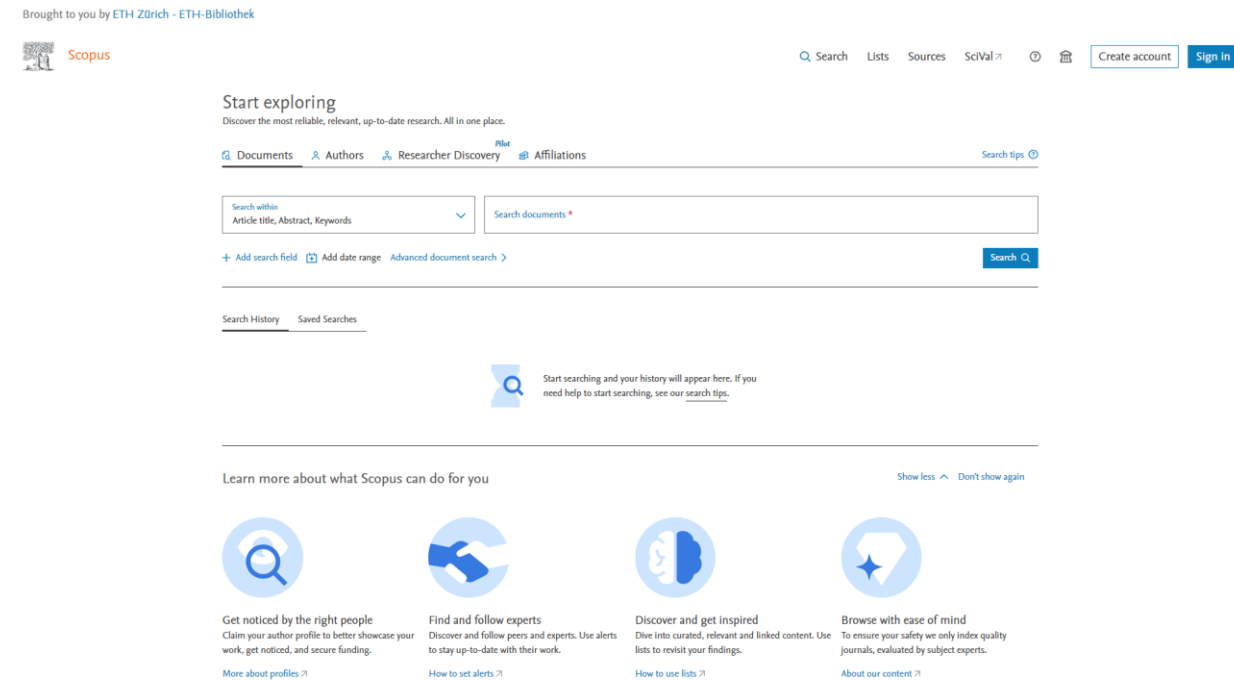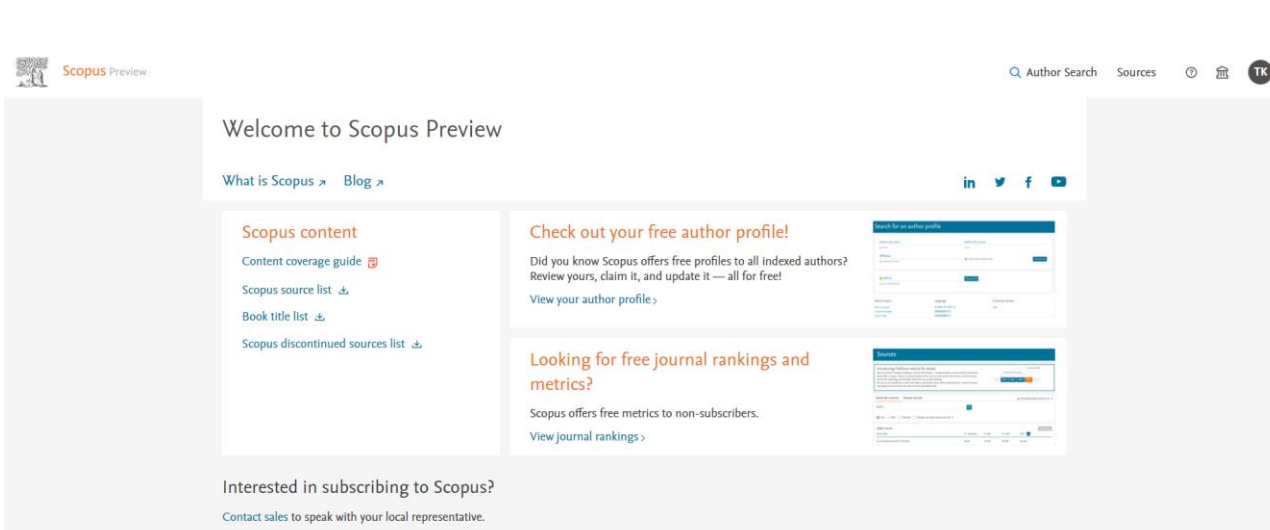
The biggest database providers

Clarivate

Web of Science™

Scopus

Dimensions

Publisher owned

(Google Scholar)

no API, no raw data

How expensive are the metadata databases?

Confidential, but if you google it…

COST

The exact pricing is confidential, but **the annual cost of USU's subscription to Web of Science is enough to purchase a very comfortable home** in Logan.

https://libguides.usu.edu/scopus/rationale

Web of Science has an annual subscription cost equal to the 2012 median price of a single-family home in the United States. Web of Science isn't an aggregator of content, like JSTOR. It's primarily an index: it aggregates citations, but not the full text of the publications themselves. It does some other important things, too, like providing cited reference searching and article-level metrics

https://libguides.library.arizona.edu/wos/rationale

ETHzürich

There used to be an open alternative…

Microsoft Academic

Article   Talk

From Wikipedia, the free encyclopedia

Microsoft Academic was a free internet-based academic search engines for academic publications and literature, developed by Microsoft Research, shut down in 2022. At the same time, **OpenAlex** launched and claimed to be a successor to Microsoft Academic.[1][2]

https://en.wikipedia.org/wiki/Microsoft_Academic

So what to do now?

ETH *zürich*

TOBI: Towards Open Bibliometric Indicators

Project TOBI

- Rising body of open bibliometric databases

- It would be great to use them instead of the commercial ones, but are they any good?
  - Missing affiliations and author attributions
  - Kinds of available metadata
  - Systematic biases (language, discipline etc.)
  - Scraped "fake" articles?
  - …

- Important because they will be used for strategic decision making!

- Scattered coverage: how to merge them sustainably?

San Francisco Declaration on Research Assessment

DORA

- Response to relying on proxy indicators in research assessment and to overrelying on metrics

For organizations that supply metrics

11. Be open and transparent by providing data and methods used to calculate all metrics.

12. Provide the data under a licence that allows unrestricted reuse, and provide computational access to data, where possible.

13. Be clear that inappropriate manipulation of metrics will not be tolerated; be explicit about what constitutes inappropriate manipulation and what measures will be taken to combat this.

14. Account for the variation in article types (e.g., reviews versus research articles), and in different subject areas when metrics are used, aggregated, or compared.

To be fully open and transparent, we cannot use paywalled data which is curated as a black box!

https://sfdora.org/read/

Project TOBI

- Not quite research, not quite service – external funding from Swissuniversities

- Started in March 2023: 3 months old!
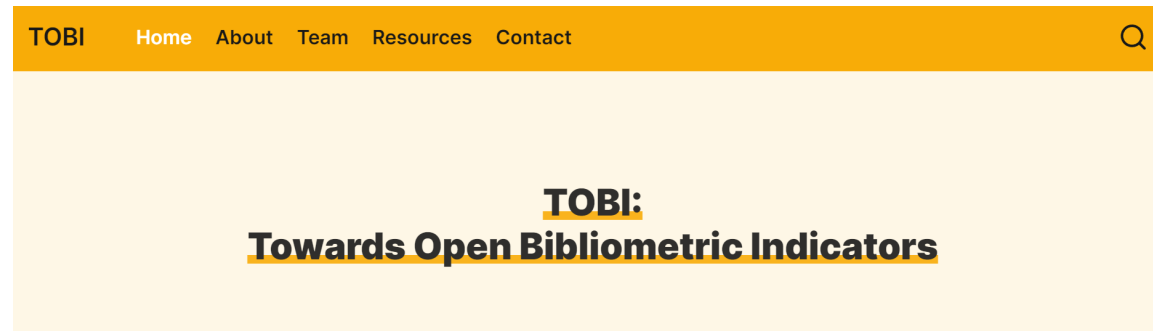
https://eth-library.github.io/tobi/

TOBI: results as of now

- Identified ~100 datasets
  - Categorized and annotated them
  - Compiled a db of metadata of datasets for easy comparison
  - Writing a report / whitepaper

- Literature review

- Investigated APIs

- Setting up infrastructure to analyze the raw data – database dumps (**we are looking for advice**!)

- Metadata tracking app

TOBI: Track your open scholarly metadata

https://tobi-open-metadata-tracker.streamlit.app/

TOBI: impact and synergies

- Swiss landscape
    - Big initial investment into using such data – not very common in CH
    - Provide starter code and data for an evidence-based decision making
    - Prime new users to use open data as the first priority
    - Open question: infrastructure with a live database for Swiss HEIs? **Looking for partners**!

- How to measure impact of science?
    - What data may be missing from our «academic view»?
    - Potential of merging with OGD! **Looking for ideas**!

- Non-academic impact and applications
    - **Looking for your input**!

Thank you for the attention!

teresa.kubacka@library.ethz.ch
@paniterka_ch
www.linkedin.com/in/tkubacka
www.teresa-kubacka.com

ETH *zürich*

Dr. Teresa Kubacka
Senior Data Scientist
teresa.kubacka@library.ethz.ch

ETH Library
Research Support Services
Knowledge Management

HG H 11.1
Rämistrasse 101
8092 Zürich

www.library.ethz.ch