ICSR Lab:
Where research evaluation meets big data

Dr Andrew Plume

President, International Center for the Study of Research

Vice President, Research Evaluation, Elsevier

ECSS 2021: National Informatics Associations Workshop

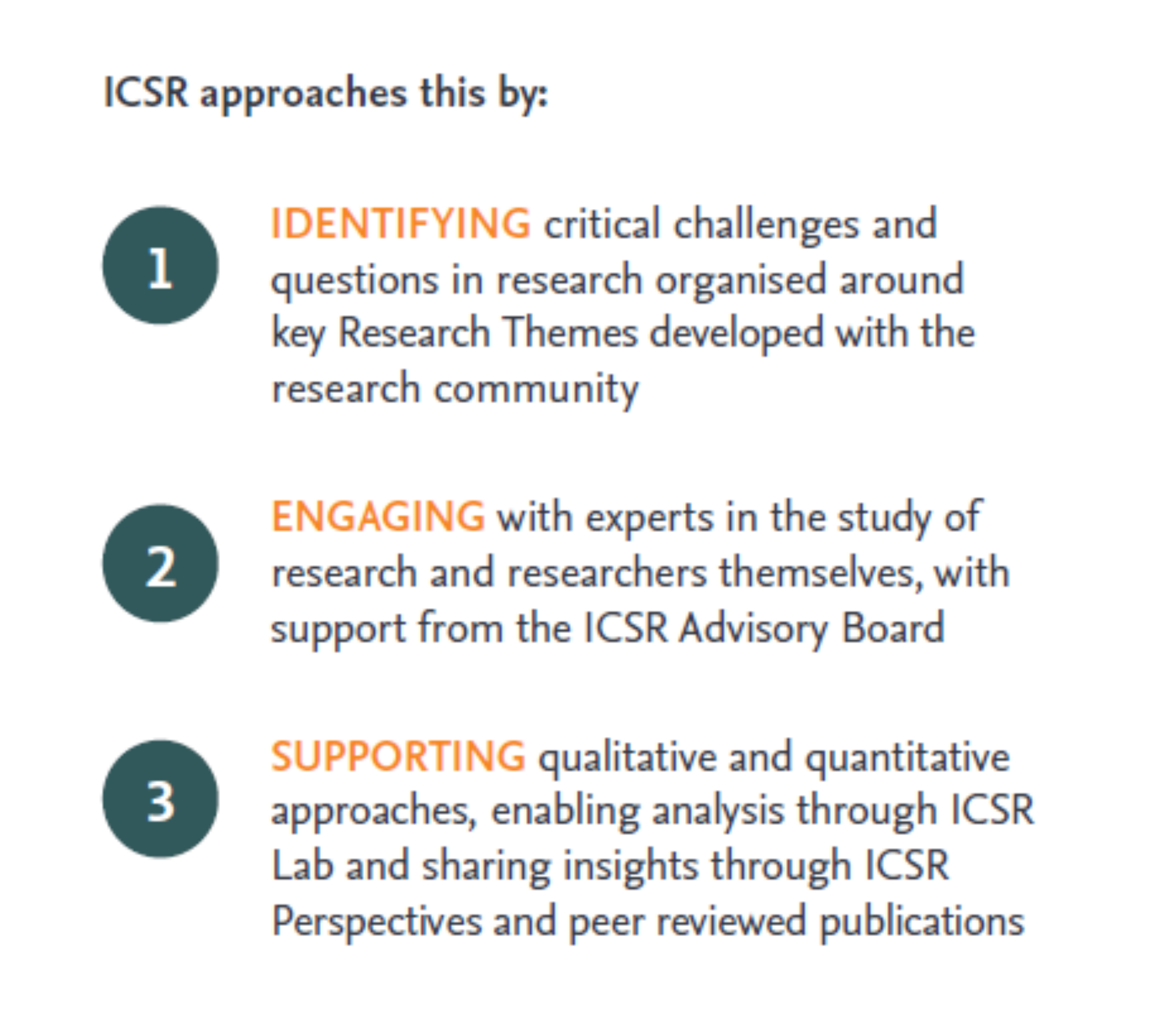
October 2021

International Center for the Study of Research

The mission of ICSR is to further the study of research and thus to contribute to the evidence base supporting the practice of research strategy, evaluation and policy.

Our vision is a world in which decisions informed by such evidence benefit research and society.

Elsevier endorsed Leiden Manifesto and signed DORA in 2020



ICSR approaches this by:

1 IDENTIFYING critical challenges and questions in research organised around key Research Themes developed with the research community

2 ENGAGING with experts in the study of research and researchers themselves, with support from the ICSR Advisory Board

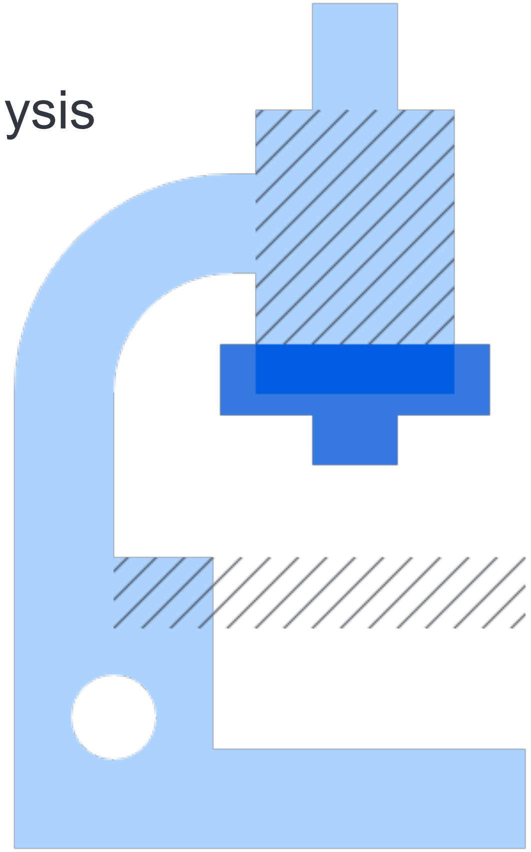
3 SUPPORTING qualitative and quantitative approaches, enabling analysis through ICSR Lab and sharing insights through ICSR Perspectives and peer reviewed publications

Introducing ICSR Lab

- A sandbox environment for **researchers** to run technical bibliometric analyses for **non-commercial**, scholarly research purposes
- Allows access to **Elsevier metadata** – including Scopus and PlumX metrics
- Free of charge – both data and computation, but researchers are expected to publish and acknowledge the data sources

Why ICSR Lab?

- Opening up the **raw data** powering Elsevier products for big data analysis
  - Extends access to volumes far beyond public APIs
  - Transparent process for access
  - Self-service with support
- Test **innovative methods** using these datasets
  - Explore new algorithms, metrics/indicators that are not yet standard
  - Sharing and iterating on code
- **Community of researchers** working alongside each other
  - Meeting collaborators and exchanging ideas
  - Collaborate across institutions with no systems access barriers
  - Reproducibility built into the platform – data snapshots and all code are retained

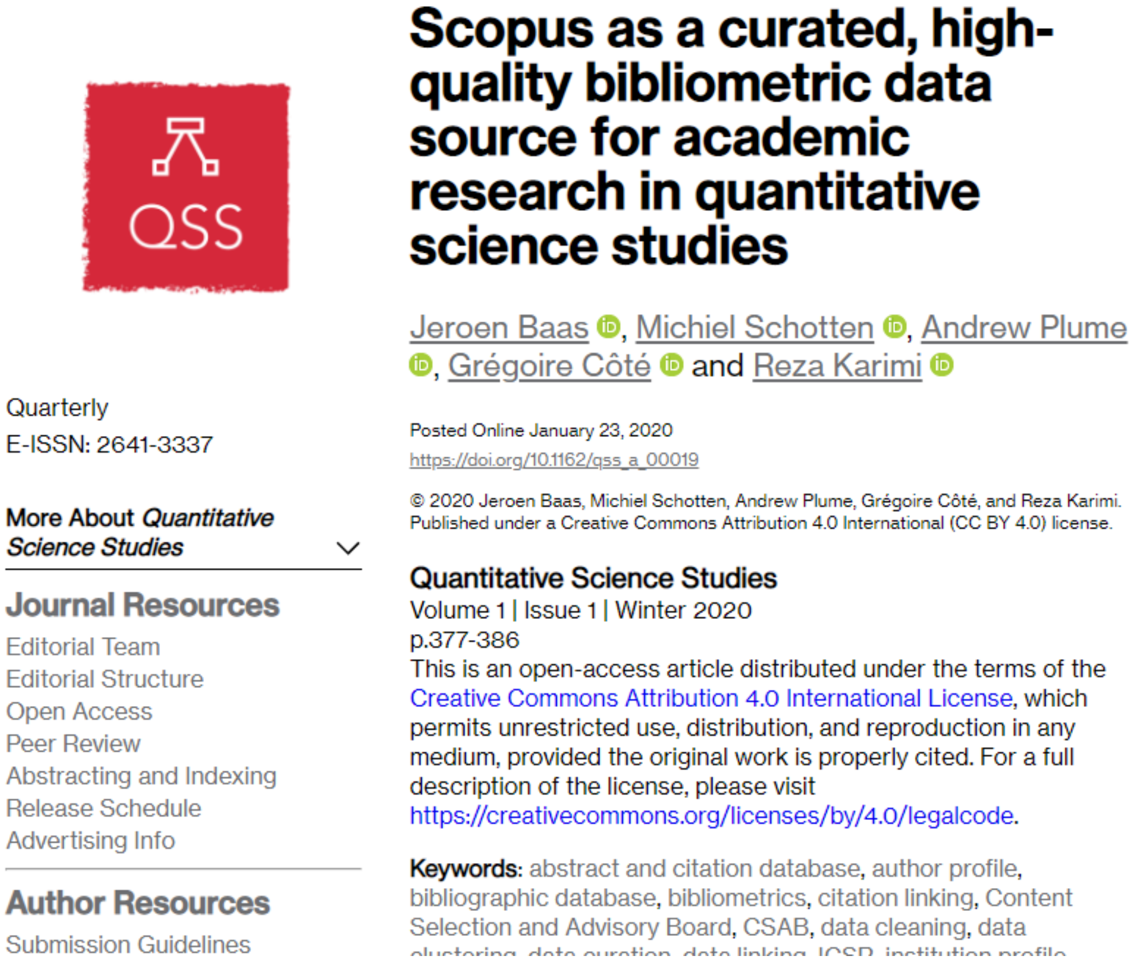


Comparison to Scopus API

APIs	ICSR Lab
Coding most analyses from scratch, or adapting sample code **CODING**	
Working with (Python) locally **ON YOUR OWN COMPUTER**	Working with Pyspark in parallel in the cloud **IN THE CLOUD**
Can view individual data points **SEE DETAIL**	Can view summary statistics (individual data points not typically exposed) **QUANTITATIVE**
Designed for small volumes – cannot process over the whole of Scopus **SMALL DATA STUDIES**	Designed for large volumes - can process over the whole of Scopus **BIG DATA STUDIES**

ICSR Lab Datasets

- Full publication metadata from Scopus:
  - **Publication metadata**, including each publication's authors and affiliations, title & **abstract (not full-text)**, language information, open access status, DOI, ASJC subject codes and more.
  - **Author names and profiles**, including affiliation details and ORCID.
  - **Institution profiles**, including name variants.
  - Ability to conduct studies of author **inferred gender**, following the methodology used in Elsevier's Gender Report 2020.



Quarterly
E-ISSN: 2641-3337

More About *Quantitative Science Studies*

Journal Resources

Editorial Team
Editorial Structure
Open Access
Peer Review
Abstracting and Indexing
Release Schedule
Advertising Info

Author Resources

Submission Guidelines

Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies

Jeroen Baas, Michiel Schotten, Andrew Plume, Grégoire Côté and Reza Karimi

Posted Online January 23, 2020
https://doi.org/10.1162/qss_a_00019

© 2020 Jeroen Baas, Michiel Schotten, Andrew Plume, Grégoire Côté, and Reza Karimi. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Quantitative Science Studies
Volume 1 | Issue 1 | Winter 2020
p.377-386

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. For a full description of the license, please visit https://creativecommons.org/licenses/by/4.0/legalcode.

Keywords: abstract and citation database, author profile, bibliographic database, bibliometrics, citation linking, Content Selection and Advisory Board, CSAB, data cleaning, data clustering, data curation, data linking, ICSR, institution profile

ICSR Lab Datasets

- PlumX metrics corresponding to the Scopus-indexed publications:
  - **Citations** encompassing traditional citations as well as for example clinical citation counts.
  - **Captures**, such as Forks and Followers on GitHub or readers on Mendeley and SSRN.
  - **Mentions** including blog mentions, comments on various platforms and Wikipedia references.
  - **Social media** including Twitter and Facebook interactions.
- SciVal Topics and various journal classifications
- NEW publication-level metadata!
  - UN SDG classification (ML approach per THE Impact Rankings)
  - Open Access flag
  - Funding acknowledgements

THE IMPACT RANKINGS

ICSR Lab Datasets

- More added on ongoing basis
  - Always keen to hear feedback from the community and users
- Can upload other (publicly available) datasets
  - Augment other datasets, such as Wikipedia, collected survey data, matching other lists of researchers via email, identifiers etc.
- No publication full text, though abstracts are available

Reproducibility in the cloud

- Data are stored indefinitely as read-only snapshots
    - Same datasets accessed by all collaborators, or when reproducing others’ studies
- Code is stored in notebooks with full version control
    - Notebooks are archived with results
    - No question of ‘silent’ editing or manipulation
- Notebooks are exportable and code sharable
- Results are aggregated data and analysis exportable under a CC-BY-NC-ND license
    - Can be shared on data repositories or alongside publications

Local vs. cloud processing

Issues in reproducible bibliometrics big data research	Can be solved by	On cloud solution
Collaborators from different institutions may have different access rights to data, resulting in different results	Hosting data on shared university server, though permissions/licensing may be difficult	Code accesses the same set of data as it is hosted in the cloud
Insight into current state of code	Use of version control	Can review versions and see editing in real-time
Different versions of software packages, source code branch	Setting up VM, software development best-practices	Cloud computing cluster is configured with up-to-date versions and libraries
Scraping volumes of data from APIs may take time due to volume limits, change of snapshot date	Ask providers directly for data	Snapshots are less frequent, but preserved indefinitely

Reproducible bibliometrics with big data

Challenges when working locally

- Data access
  - Scraping from APIs
  - Data fetched over different periods
  - Different access rights between different institutions
  - Difficult to store/share raw data to reproduce

- Collaboration
  - Keeping up to date with latest version of code
  - Ensuring same libraries, versions etc, meaning same code gives same result

In ICSR Lab

- Raw data snapshots from one date are available
- Snapshots are preserved over time
- Same access level for all project members
- Can re-run analyses later with same results, edit and adapt code

- Work in same notebooks in real-time
- Work on same system, same results
- Visibility into output/results
- Code can be downloaded

Who uses ICSR Lab?

- Anyone doing research with an academic affiliation, aiming at answering a research question
  - Users range from Masters students to senior researchers
  - Topics of broad interest to the scholarly bibliometrics/scientometrics community
  - Projects must include an intermediate+ level coder
- Working on research themes:
  - Research careers, Research practices, Research globalization
  - Open Science, Impact of research
  - Inclusivity
  - Sustainability
- New or innovative methodology, creating/testing metrics that are not yet available in other locations
- Require large volumes of data for this question
- Results are typically shared at conferences and/or peer-reviewed journal publications

What's happening in the ICSR Lab right now?

Research practices

A new conceptual framework for international collaboration and knowledge production

Indicators of research interdisciplinarity

Investigating collaboration and team science

Applying topic modelling, citation and bibliographic coupling networks to a research field

Citation prediction using textual and author characteristics

Examining impact of author self-citations on a certain citation indicator

Detecting anomalous and manipulative citation practices at scale

Impact of research

Altmetrics as an indicator of research impact

Citations from the peer-reviewed literature to Wikipedia by subject area

Inclusivity

The relationship between citations and gender

Sex- and gender-inclusivity in COVID-19 research

Inclusion & diversity in the author pool of selected journals

Research globalization

Career development and researcher mobility

Time resolution in researcher mobility studies

Brain circulation approaches applied at a national level

National political attributes and researcher mobility

Dynamic block modelling to uncover research participation of the Global South

What’s happening in the ICSR Lab right now?

Sustainability

Linking burden of disease and **UN SDG** priorities with research output

Benchmarking national contributions to research in a particular **UN SDG**

Applying machine learning approaches to create previously intractable lexical queries for **UN SDGs**

Trends in sustainable technology development in a country

Research careers

Researcher embeddedness in the citation network as a function of career stage

Open science

Openness of the published literature on critical global health threats

Social media and changes in researcher behaviour

Science of science policy

Research governance and scientific responses to the COVID-19 pandemic

Shifting research focus during the COVID-19 pandemic

Linking climate change impacts with research output

Topical clusters and their use in research evaluation and management

Is ICSR Lab for you?

Considerations

- Are you pursuing scholarly research addressable using ICSR Lab datasets?
- Is your research topic relevant to the ICSR Research Themes?
- Does at least one member of your research team have some coding experience?

We encourage…

- New & innovative methods & approaches
- Novel metrics & indicators
- Well-defined research agendas but also exploratory studies

www.icsr.net

Thank you!