Analyzing open data sources with R

Visualizing Open Access Summer School 2021

Dorothea Strecker

Goals

get to know some open data sources for Open Access

→ learn how to use open data sources

→ learn about the importance of persistent identifiers

get to know some R basics

→ learn how to prepare and analyse data

→ learn how to visualize data

What if I have a question?

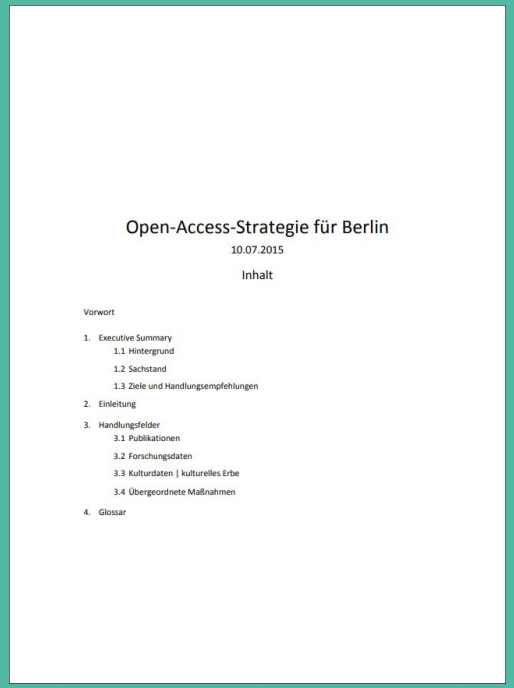
- If you have a question, feel free to ask. Please use the “raise hand” function or the chat.
- After the workshop, I will share Notebooks that explain all the steps in detail, so don’t worry if you miss a few details along the way!
- Please keep your microphone muted to reduce background noise.
- Feel free to turn your camera on or off, depending on your preference.

Outline

1. Open Access monitoring
2. Open data sources for Open Access
3. About R
4. **15 minute break** (+ time for installing software)
5. Interactive live demo: Setup
6. Querying open data sources
7. Analysing and visualising data

Open Access monitoring

The Open Access transformation requires monitoring for better decision making.

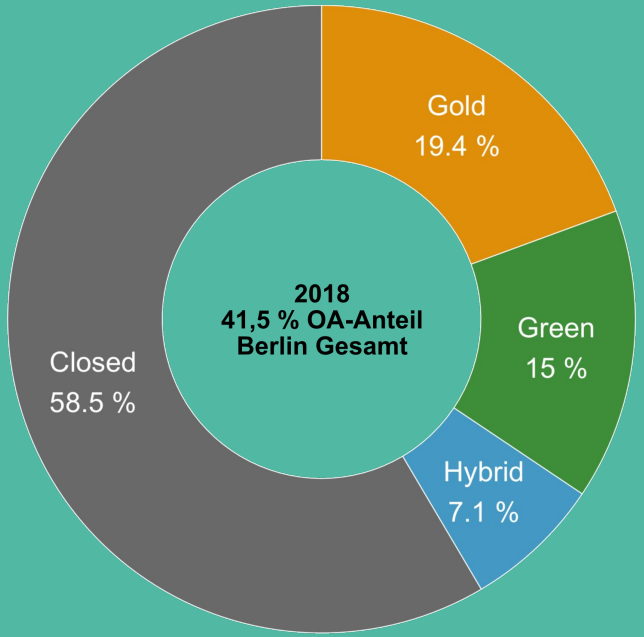


Open-Access-Strategie für Berlin

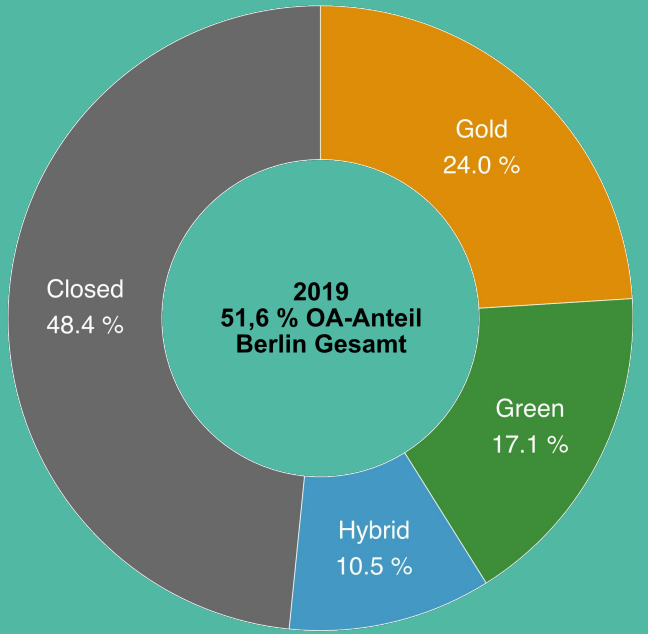
10.07.2015

Inhalt

Open Access strategy for Berlin
(Senat von Berlin, 2015)



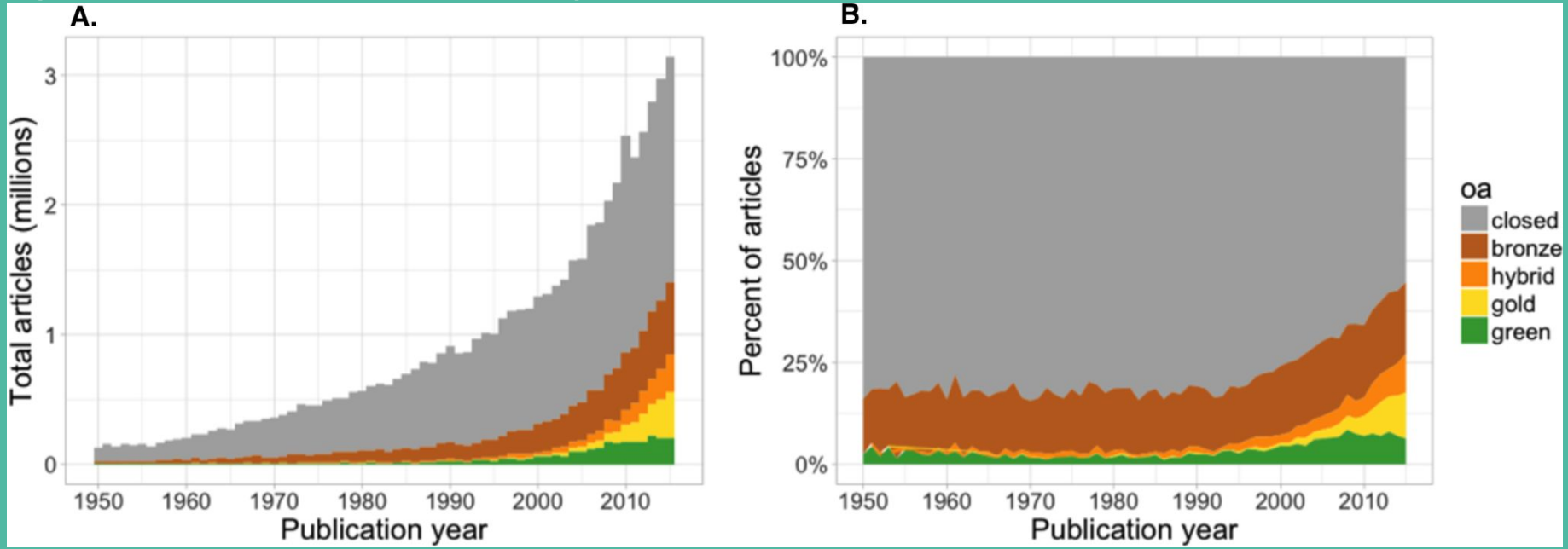
Proportion of Open Access, 2018
(Kindling et al., 2020)



Proportion of Open Access, 2019
(Kindling et al., 2021)

Open Access monitoring

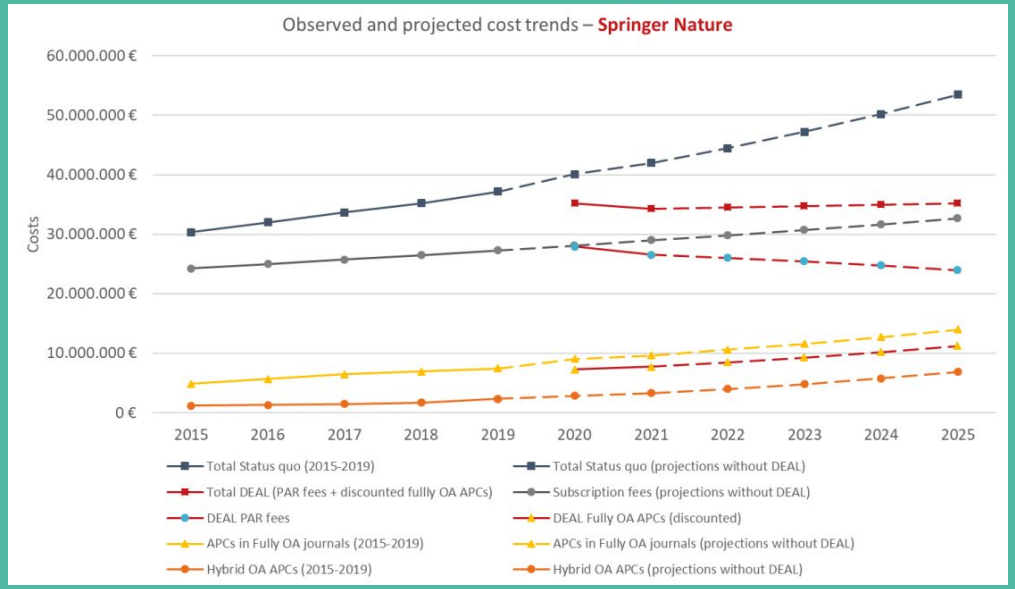
Depending on the specific use case, Open Access monitoring can be focused on publications, costs, or other aspects.



Number of articles (A) and proportion of articles (B) with OA copies, estimated based on a random sample of 100,000 articles with Crossref DOIs. (Piwowar et al., 2018)

Open Access monitoring

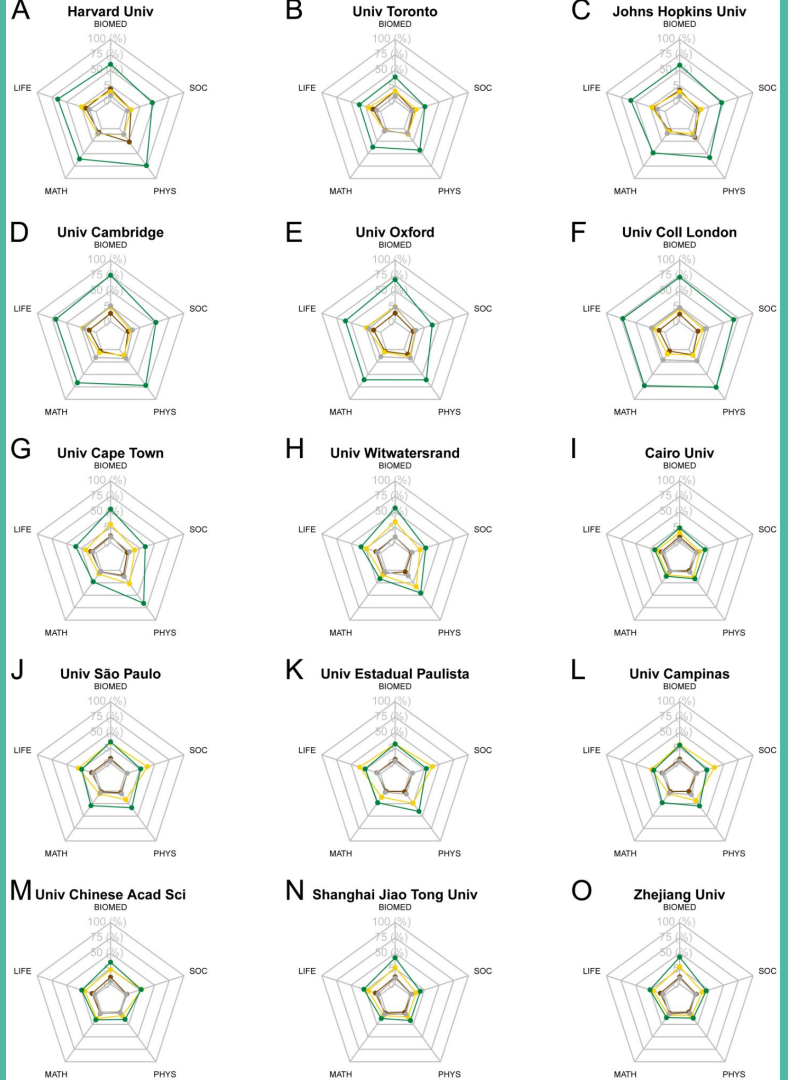
Depending on the specific use case, Open Access monitoring can be focused on publications, costs, or other aspects.



development of costs under different scenarios under the DEAL agreement with Springer Nature (Schimmer, Dér & Campbell, 2021)

Open Access monitoring

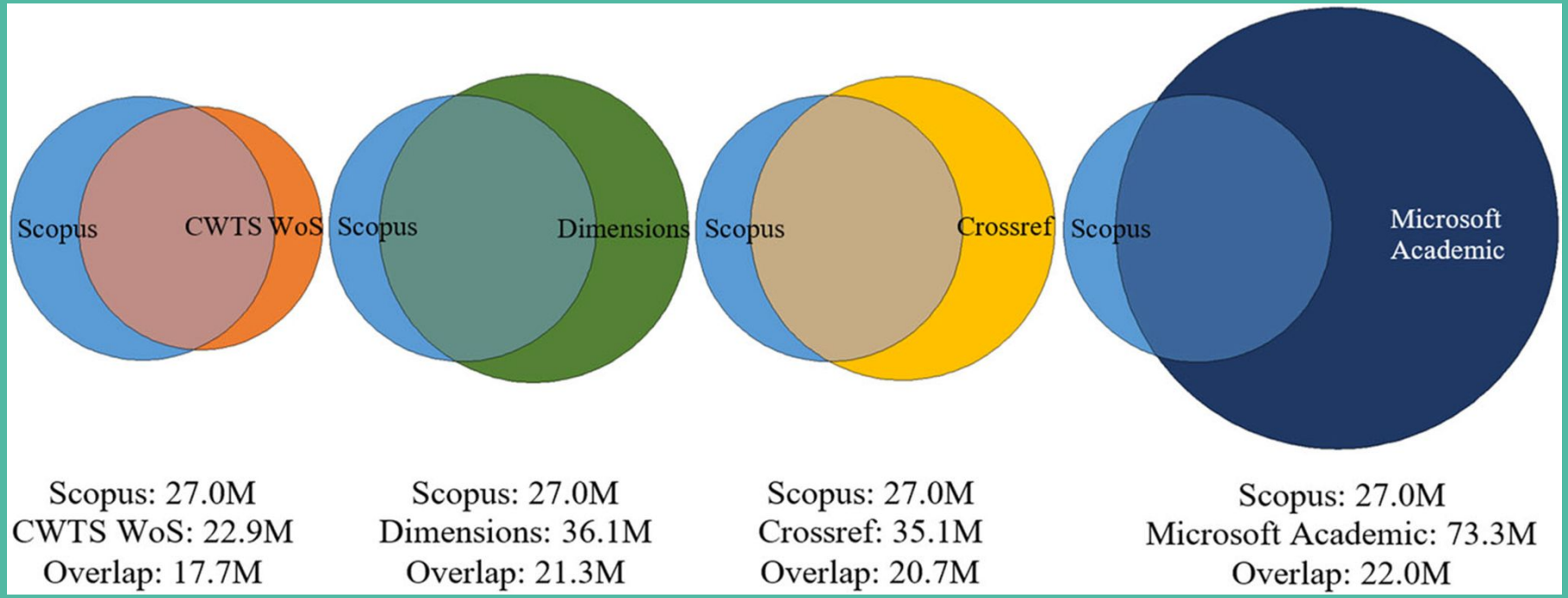
Monitoring can be conducted at different levels, for example at the institutional or national level.



OA disciplinary profiles for top three universities with the largest output for North America (A–C), Europe (D–F), Africa (G–I), South America (J–L) and Asia (M–O).
Colors refer to OA types. Brown: bronze OA; yellow: gold OA; green: green OA; grey: hybrid OA.
(Robinson-Garcia, Costas & van Leeuwen, 2020)

Open data sources for Open Access

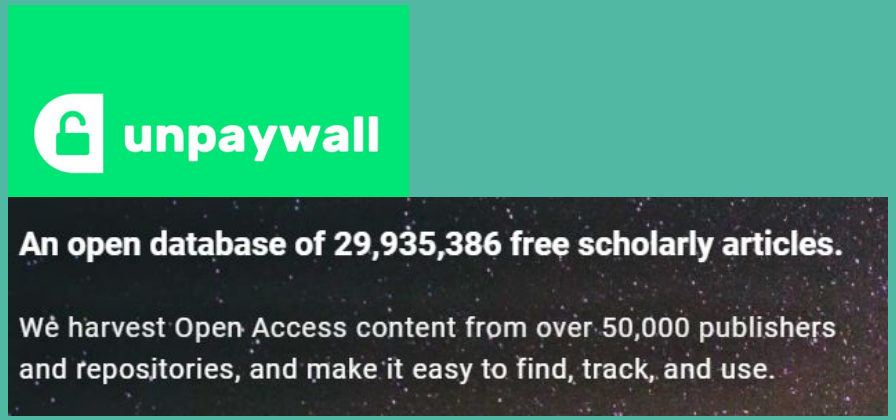
bibliometric data



Overlap of documents between Scopus and other data sources.
(Visser, van Eyck & Waltman, 2021)

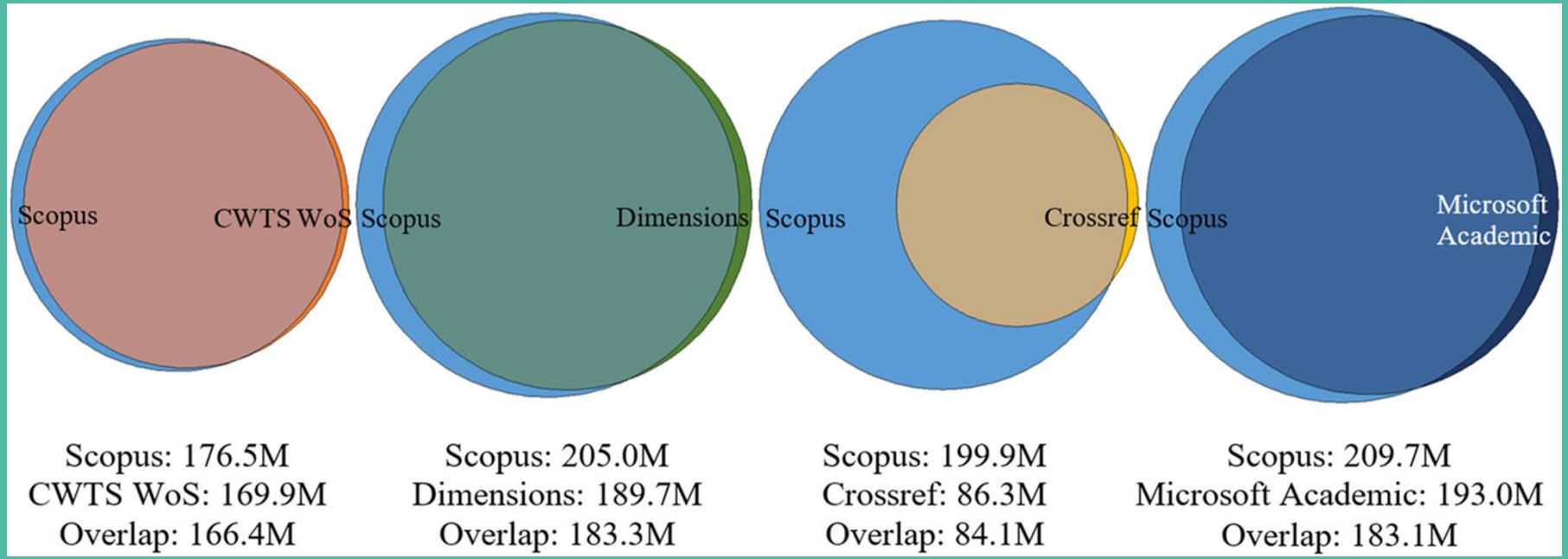
Open data sources for Open Access

Open Access status



Open data sources for Open Access

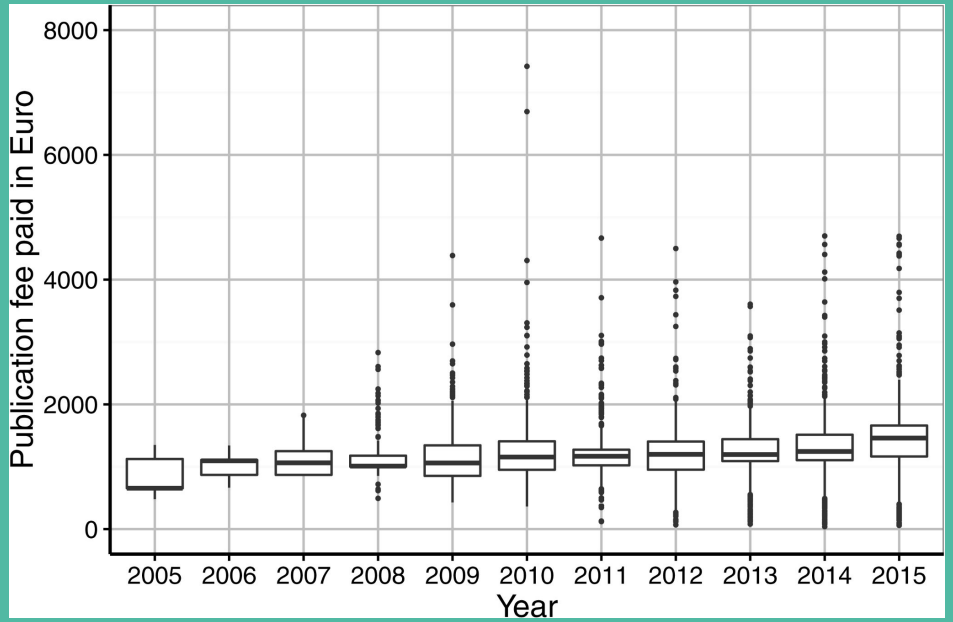
citation information



Overlap of citation links between Scopus and the other data sources.
(Visser, van Eyck & Waltman, 2021)

Open data sources for Open Access

costs



Institutional spending on publication fees by German research organisations per year (in €).
(Jahn & Tullney, 2016)

About R

free software

“an environment within which statistical techniques are implemented” (What is R?)

extendable, for example via packages (currently: 18094 on CRAN)

Do you have any prior experience with R?

Do you need assistance installing R & RStudio?

Interactive live demo: Setup

You find everything you need in the GitHub repository:

https://github.com/dorothearrr/visOA

You have **two options**:

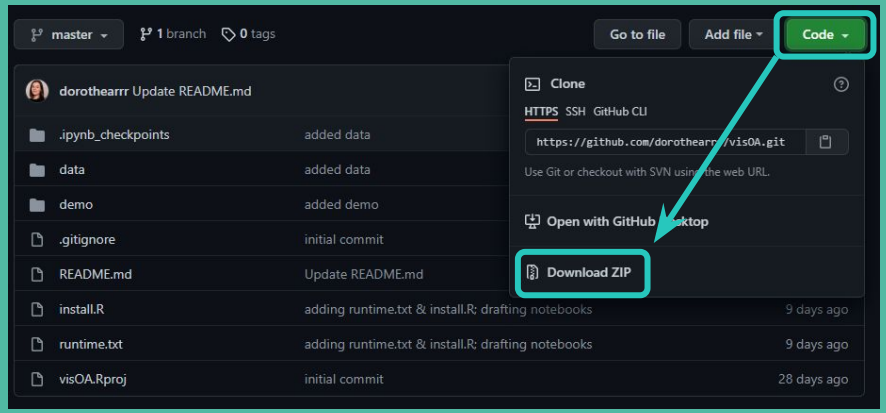
- using **RStudio** (recommended)
- using **Jupyter Notebooks in binder** (less stable, slower)

During the workshop, you can either follow along live, or copy / paste prepared code chunks (you find them in **visOA/demo/**).

Interactive live demo: Setup

using **RStudio** (recommended)

Step 1: clone the GitHub repository



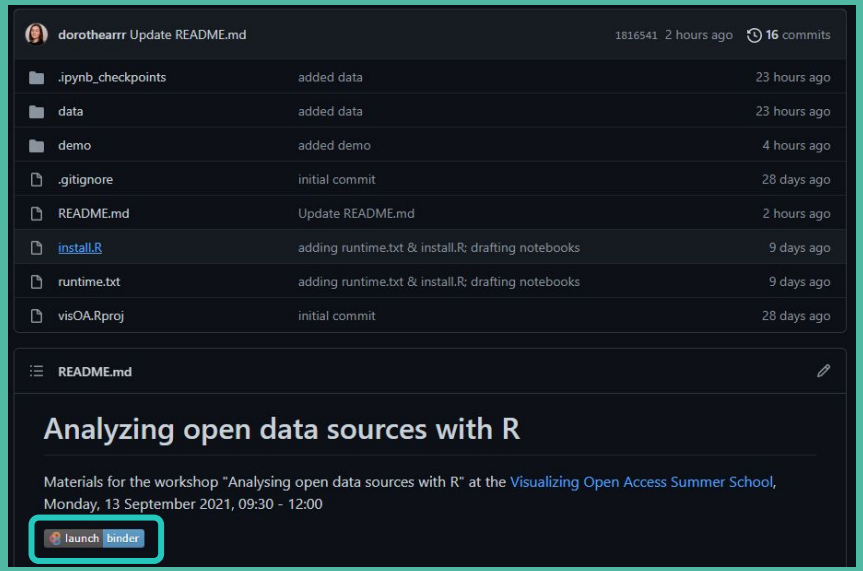
Step 2: unzip the folder and open the file **visOA.Rproj** with RStudio

Interactive live demo: Setup

using **Jupyter Notebooks in binder** (less stable, slower)

Step 1: click on the binder link in the GitHub repository

Step 2: wait for binder to load (this will take a few moments)



Bonus: Start / keep learning R

useful & free resources for learning R

courses:

- Software Carpentry
- Exercism

books:

- R for Data Science
- Advanced R
- Fundamentals of Data Visualization
- ggplot2: Elegant Graphics for Data Analysis

cheat sheets

References

Jahn, N., & Tullney, M. (2016). A study of institutional spending on open access publication fees in Germany. *PeerJ*, *4*, e2323. https://doi.org/10.7717/peerj.2323

Kindling, M., Delasalle, J., Finke, P., Hampl, M., Neufend, M., & Voigt, M. (2021). *Open-Access-Anteil bei Zeitschriftenartikeln von Wissenschaftlerinnen und Wissenschaftlern an Einrichtungen des Landes Berlin: Datenauswertung für das Jahr 2019*. https://doi.org/10.14279/depositonce-11774

Kindling, M., Hampl, M., Finke, P., Voigt, M., & Hübner, A. (2020). *Open-Access-Anteil bei Zeitschriftenartikeln von Wissenschaftlerinnen und Wissenschaftlern an Einrichtungen des Landes Berlin: Datenauswertung für das Jahr 2018*. https://doi.org/10.14279/depositonce-9606

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., & Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, *6*, e4375. https://doi.org/10.7717/peerj.4375

Robinson-Garcia, N., Costas, R., & Leeuwen, T. N. van. (2020). Open Access uptake by universities worldwide. *PeerJ*, *8*, e9410. https://doi.org/10.7717/peerj.9410

Schimmer, R., Dér, Á., & Campbell, C. (2021). *The DEAL Cost Modeling Tool*. https://doi.org/10.17617/2.3331716

Senat Von Berlin. (2015). *Open-Access-Strategie für Berlin*. https://doi.org/10.17169/REFUBIUM-26319

Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, *2*(1), 20–41. https://doi.org/10.1162/qss_a_00112