Data Science in Libraries

LIBER Annual Conference
2023-07-05, Budapest

Workshop Agenda

9:15 – 9:30

About us, and what is our topic?

9:30 – 10:10

Short overview of the survey result

10:10 – 10:30

Discussions

Practical Things

Links

- Survey: https://survey.uu.nl/jfe/form/SV_eswlDMEJuaf9nGS
- Slides: https://bit.ly/dslib-2023
- Collaborative Document: https://bit.ly/dslib-2023-notes

Social Media

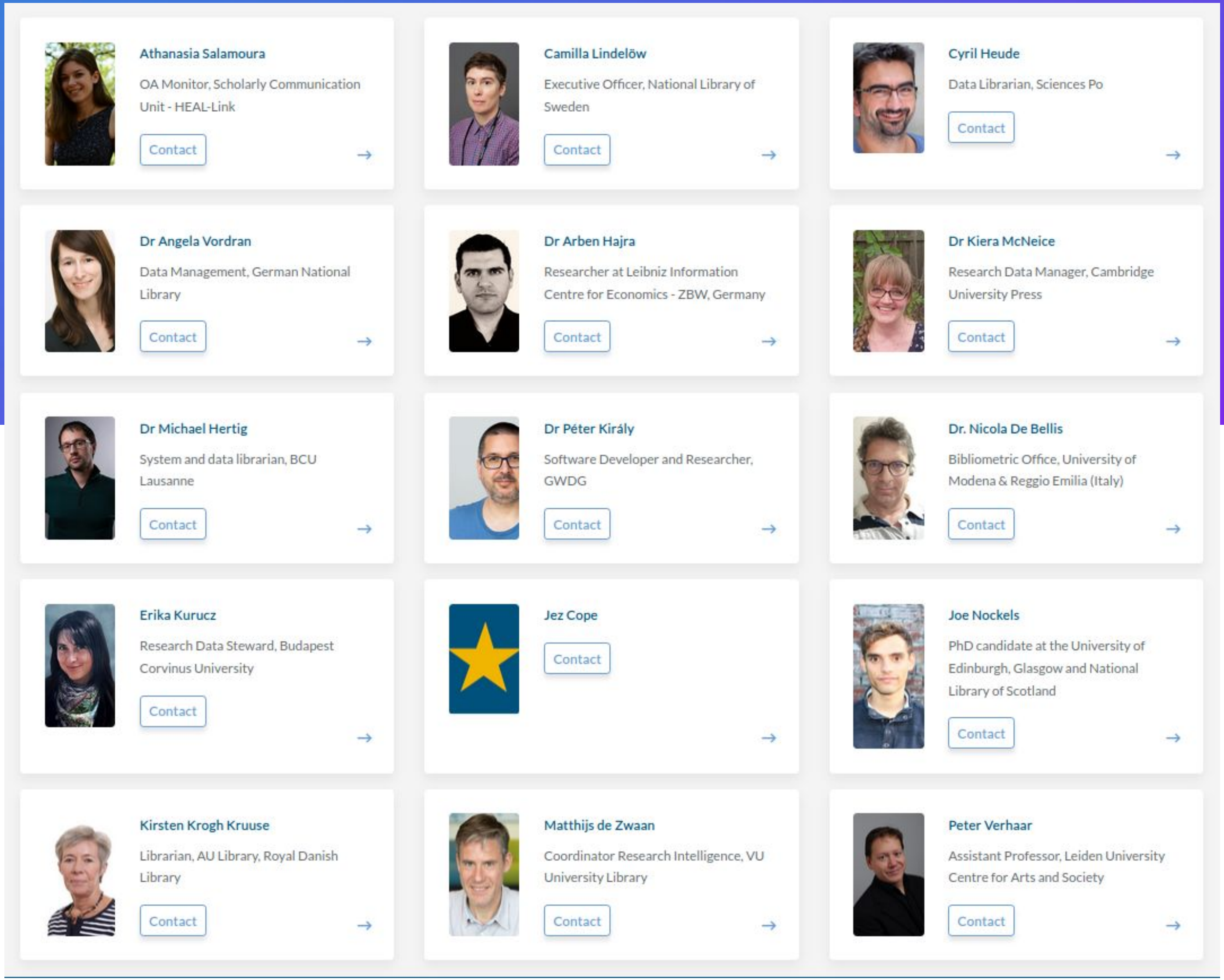
- Tweet with @LIBERConference & @LIBEREurope & #LIBER2023!
- Before tweeting, check for any privacy-related concerns in the group.
- I'm also happy to make new library friends @kiru@openbiblio.social

Have fun!

https://bit.ly/dslib-2023

PART 1

09:15 – 09:30



Athanasia Salamoura
OA Monitor, Scholarly Communication Unit - HEAL-Link

Camilla Lindelöw
Executive Officer, National Library of Sweden

Cyril Heude
Data Librarian, Sciences Po

Dr Angela Vordran
Data Management, German National Library

Dr Arben Hajra
Researcher at Leibniz Information Centre for Economics - ZBW, Germany

Dr Kiera McNeice
Research Data Manager, Cambridge University Press

Dr Michael Hertig
System and data librarian, BCU Lausanne

Dr Péter Király
Software Developer and Researcher, GWDG

Dr. Nicola De Bellis
Bibliometric Office, University of Modena & Reggio Emilia (Italy)

Erika Kurucz
Research Data Steward, Budapest Corvinus University

Jez Cope

Joe Nockels
PhD candidate at the University of Edinburgh, Glasgow and National Library of Scotland

Kirsten Krogh Kruuse
Librarian, AU Library, Royal Danish Library

Matthijs de Zwaan
Coordinator Research Intelligence, VU University Library

Peter Verhaar
Assistant Professor, Leiden University Centre for Arts and Society

https://bit.ly/dslib-2023

about us

- https://libereurope.eu/working-group/liber-data-science-in-libraries-working-group/
- founded: 2021 spring
- 15 members
- aiming: survey and landscape analysis
- monthly meetings
  - forming definitions
  - show cases
  - reading papers
  - information exchange
  - guests
- looking for chair(s)

https://bit.ly/dslib-2023

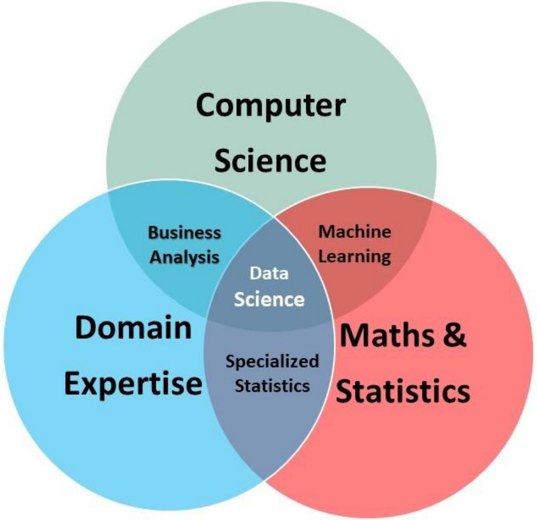
DSLib definition(s)

What is *your* definition/idea of data science in libraries?

https://bit.ly/dslib-2023

data science

is a set of computational methods for the identification of novel and actionable insights from data



https://bit.ly/dslib-2023

computational methods

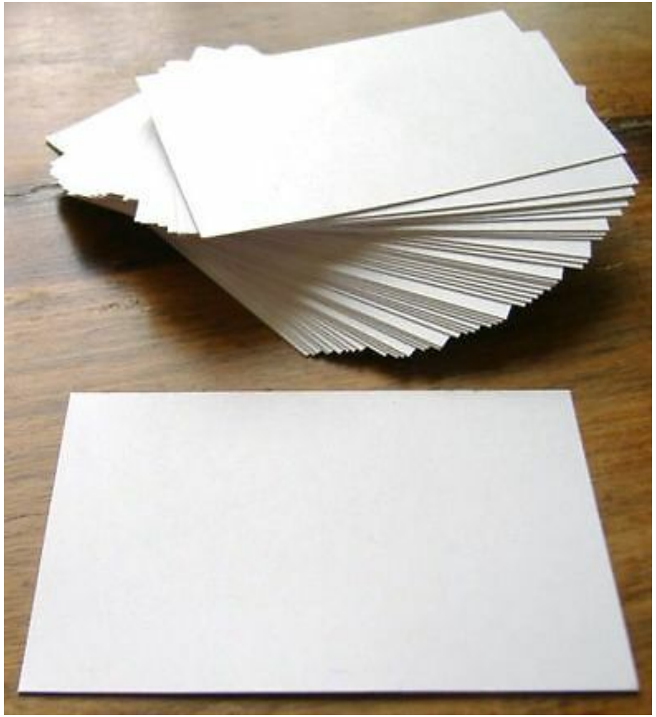
include (but are not limited to)

- ★ descriptive and inferential statistics
- ★ visualization
- ★ text mining
- ★ image processing and computer vision
- ★ machine learning
- ★ data engineering

data science in libraries

is the use of these methods in the delivery and/or improvement of library services and the delivery of data science training or services.

https://bit.ly/dslib-2023



DSLib Activities

What are some examples of data science activities you (wish to) carry out at your library?

https://bit.ly/dslib-2023

Collections as Data

Data science activities that facilitate the use of library collections in computationally-driven research and teaching. This can include activities that ensure that data from library collections are high-quality, rich with information, reliable, suitable for analysis, and easily accessible for computational interactions.

Collections as Data

- data pipelines that are created to enhance the quality of data
- machine learning and computer vision techniques used to:
  - generate data
  - discover resources
  - identify and extract rich metadata and/or full-text from documents

https://bit.ly/dslib-2023

Library Intelligence

Data science activities geared towards the improvement of traditional library services and support for decision-making by library management.

https://bit.ly/dslib-2023

Library Intelligence

- data-driven item suggestions for library patrons
- the application of machine learning techniques in the management of library material flows
- the use of library loan data analytics in collection management
- automated library analytics for day-to-day planning and annual reports

https://bit.ly/dslib-2023

Research Support

Data science activities to support researchers through the research lifecycle. This can cover areas such as research data management, research data/software engineering, digital humanities, and (digital) information skills.

Research Support

★ data management planning
★ research data/software engineering
★ ensuring data FAIRness (findability, accessibility, interoperability, reusability)
★ data curation and preservation
★ working with Linked Open Data and digital corpora
★ data science methods in (automated) systematic searches and literature reviews

Research Intelligence

Data science activities in compiling and visualizing data for decisions and benchmarking within the scientific community. Given the scale of the data available, Research Intelligence often requires the implementation of data pipelines and dashboard tools.

Research Intelligence

data:

- metadata of publications
- other research outputs
- data related to these outputs such as citations

activities:

- continuous development of analysis workflows
- combining traditional citation metrics with alternative metrics such as policy citations

https://bit.ly/dslib-2023

PART 2 – Survey results

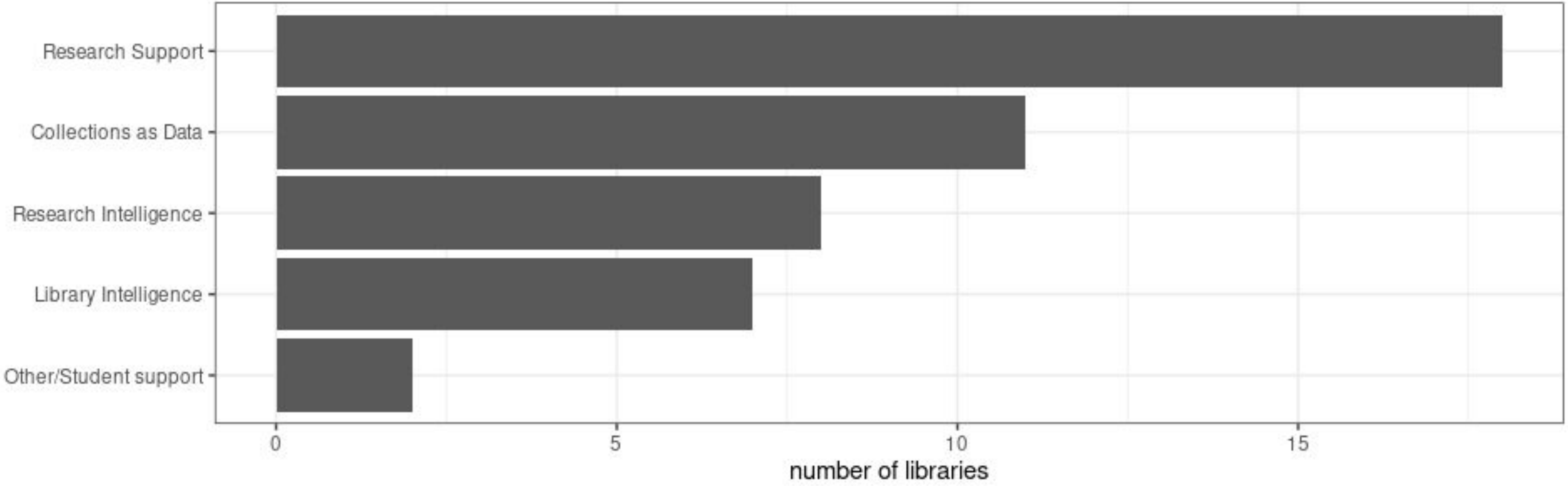
09:30 – 10:10

n = 50

https://bit.ly/dslib-2023

Q7 / Q8

Which categories apply to your data science activities?



Q9

Describe the current data science activities in your library.

Q9: library intelligence

catalogue related

- automated cataloging
- enrichments of bibliographic records
- automated subject indexing
- automate the acquisition of metadata
- data quality:
  - normalisation/disambiguation of persons and corporate bodies,
  - linking between data sets,
  - enrichment with IDs,
  - content indexing
- visualization of authority file records

https://bit.ly/dslib-2023

Q9: library intelligence

usage analysis

- ★ analyzing user behavior in the catalogue (bibtip)
- ★ data on space occupancy in study spaces
- ★ journal renewal
- ★ semi-automated de-selection
- ★ evidence based acquisition
- ★ analyzing query logs
- ★ COUNTER stats, loan statistics, Leganto stats

Q9: research inteligence

- ★ bibliometric analysis
- ★ campus output breaks down by publisher share
- ★ analyze research impact

https://bit.ly/dslib-2023

Q9: research support

- ★ geo-referencing (crowdsourcing)
- ★ hand written text recognition (HTR) with Transkribus
- ★ publishing metadata
- ★ attribute PIDs to elements of metadata
- ★ research data management advisory service

https://bit.ly/dslib-2023

Q9: other

- automating the process for systematic literature reviews
- ChatBot
- text mining
- analyzing watermarks
- automated interactive dashboards
- trend analysis

https://bit.ly/dslib-2023

Q10

Which department(s), team(s), or similar entities carry out data science activities in your library?

- ★ single person
- ★ an existing unit with new tasks
- ★ several library units
- ★ cross-cutting working group
- ★ dedicated team

- ★ research support
- ★ data management
- ★ scholarly communications
- ★ research software engineering
- ★ acquisition and cataloguing

https://bit.ly/dslib-2023

Q11

Describe the staff responsible for data science activities.

background

- LIS
- Computer Science
- Digital Humanities / Computational Literary Studies
- other

qualifications

- self-taught
- workshops
- field-specific training

roles

- librarians
- data specialists
- data scientists
- research data engineers & consultant
- metadata managers

https://bit.ly/dslib-2023

Q12

Who does the library collaborate with, internally or externally to the institution, in carrying out data science activities?

- vendors, suppliers
- IT department
- university Working Groups
- individual researchers
- university grants office
- data stewards
- other academic libraries
- publishers
- research consortia

https://bit.ly/dslib-2023

Q13

What are the challenges faced by the library in carrying out data science activities?

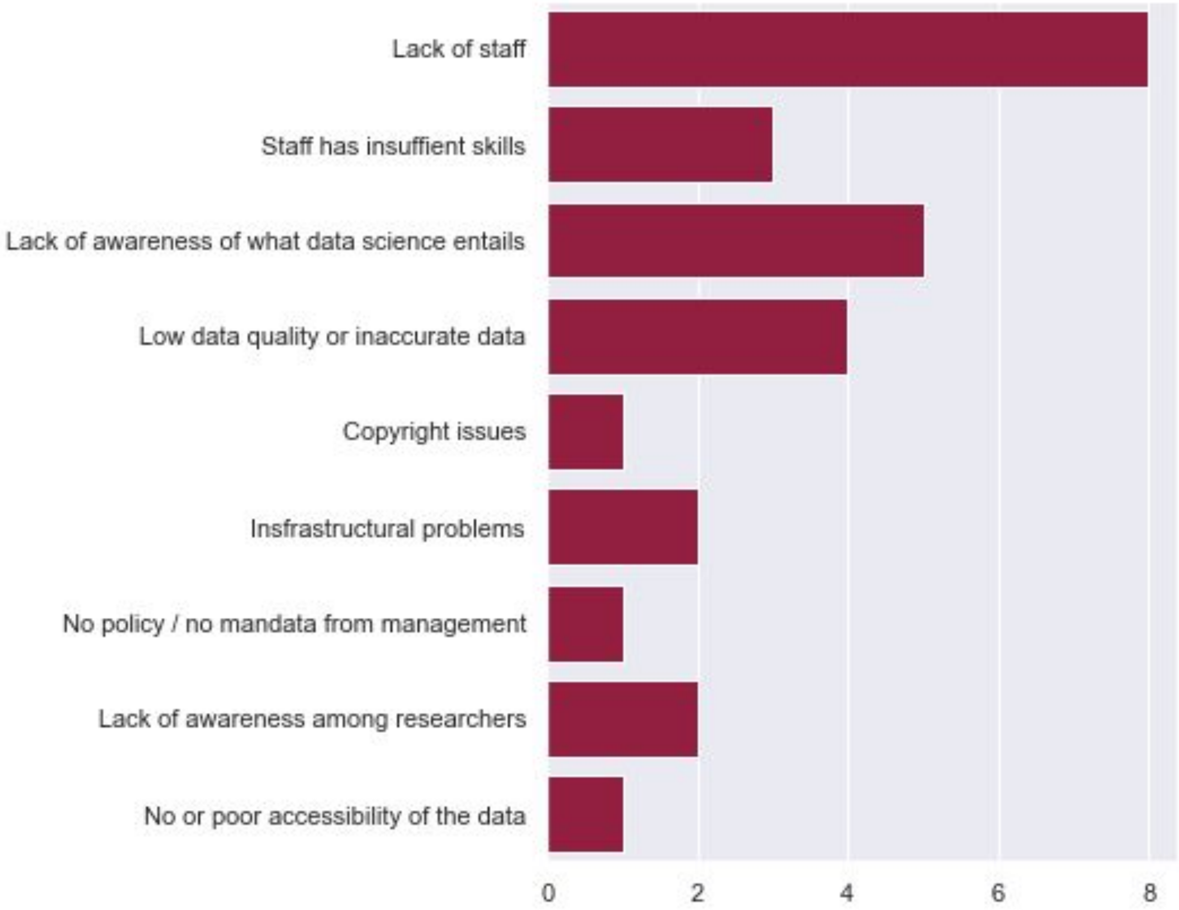
- lack of staff
- time to get the qualifications
- high demands, low resources
- OCR problems
- copyright issues
- lack of physical back-up storage
- data quality
- fixed term contracts in projects
- lack of awareness
- lack of institutional policies
- lack of budget
- researchers not sharing datasets
- researchers lack a good understanding of techniques

MVP section of this workshop

https://bit.ly/dslib-2023

Q13

What are the challenges faced by the library in carrying out data science activities?



https://bit.ly/dslib-2023

Q14

Does the library have ways of addressing legal or ethical challenges related to data science activities?

- considering rights and terms in data collecting phase
- Copyright Information Point / university legal/privacy department/officers
- CARE principles, CC0
- no personal data are collected
- ethical aspects of bibliometric analyses
- difficult questions (e.g. what can we do with abstracts...?)
- restricted materials are only accessible on the premises of the library

https://bit.ly/dslib-2023

Q15

How are the data science activities typically funded?

- internal funding
- university funding

- government
- national research funding agencies
- European research funding agencies

https://bit.ly/dslib-2023

Q16

What library data science activities are planned (if any) in the future?

- publishing/author profiles
- cataloging with AI (data enrichments, subject indexing)
- usage analytics
- name disambiguation and clustering
- duplicate detection
- developing data science skills (for staff, researchers, students)
- involvement in research projects
- collections as data
- monitoring research activities
- (research) data repository
- knowledge graph/hub

https://bit.ly/dslib-2023

Wrapping Up…

https://bit.ly/dslib-2023

Closing

★ Thanks for being here!
★ Would you be open to being contacted for the survey?
★ Would you like to join our WG or contribute in any other way?

Links:

★ Survey: https://survey.uu.nl/jfe/form/SV_eswlDMEJuaf9nGS
★ WG website: https://libereurope.eu/working-group/liber-data-science-in-libraries-working-group/
★ WG open notes: https://hackmd.io/@nehamoopen/liber-dslib/

https://bit.ly/dslib-2023

Thanks!

Email: pkiraly@gwdg.de

Mastodon: @kiru@openbiblio.social

Github: https://github.com/pkiraly

https://bit.ly/dslib-2023

Unused slides

data science activities

★ creating infrastructure for digital collections
★ clustering work bundles (FRBR/LRM)
★ links to authority records (entity recognition, entity linking)
★ automatic subject indexing (Finnish Nat. Lib.’s Annif – https://annif.org/)
★ visualization
★ statistical analysis
★ text and data mining
★ open data
★ IIIF Image API

★ language detection
★ automated retrieval of citation information
★ providing easy access to text resources
★ knowledge graph
★ Carpentries
★ book club
★ teaching in LIS courses
★ topic modelling
★ linguistic analysis
★ metadata quality assessment
★ license and right management

Landscape Analysis

The WG needs your input to develop the landscape analysis!

·

✏ Write out your thoughts and ideas in your subgroup’s collaborative document.

🗨 Discuss with your subgroup as you go.

📃 Provide a summary in the Google Doc and report back to the whole group.

Some discussion/reflection prompts:

- What kind of data do you (wish to) work with?

- What are the opportunities and challenges with respect to carrying out data science in libraries?

- If the WG works on a landscape analysis/report/advice/recommendation:
  - *What kind of information would you like to see?*
  - *What kind of information do you think all libraries would benefit from?*

- What else would you want out of our WG’s activities? Where can we provide the most valuable input?

DSLib Survey

👁️‍🗨️ Skim through the survey and imagine filling it out yourself/for your library.

What are points for improvement? What is missing?

🗨️ Discuss with your subgroup as you go.

https://survey.uu.nl/jfe/form/SV_eswlDMEJuaf9nGS