Workshop: Data Science in Libraries @ LIBER 2023

Wednesday, 5th July 2023, 09:00 - 10:30 CEST

The LIBER Data Science in Libraries Working Group was founded in the Spring of 2021. The Working Group seeks to explore and promote library engagement in applying data science and analytical methods in libraries, taking into account all kinds of processes and workflows around library collections and metadata as well as digital infrastructures and service areas.

*In this pre-conference workshop, **we wish to connect practitioners of data science in research libraries, in order to share experiences and ideas on the topic.** The Working Group is also preparing a landscape analysis and survey with the aim of creating an overview of data science activities (and eventual recommendations) in research libraries. **We would like to further develop this landscape analysis and survey with the input of workshop attendees in a hackathon-style session.***

Workshop Agenda

- **09:15 - 09:30** → Introduction (15 minutes)
  - *Connecting practitioners of data science in research libraries, in order to share experiences and ideas on the topic.*
- **09:30 - 10:10** → Survey result overview (40 minutes)
- **10:10 - 10:30** → Discussion (20 minutes)
  - *Further develop landscape analysis and survey with the input of workshop attendees in a hackathon-style session.*

Welcome to the workshop collaborative document!

This is the primary collaborative document, which contains some general information as well as links to documents for each subgroup.

Links

- This Collaborative Document: https://bit.ly/liber-we-2022-dslib
- DSLib Survey: https://survey.uu.nl/jfe/form/SV_eswlDMEJuaf9nGS
- DSLib WG website: https://libereurope.eu/working-group/liber-data-science-in-libraries-working-group/
- DSLib WG open notes: https://hackmd.io/@nehamoopen/liber-dslib/
- Slides https://bit.ly/dslib-2023

Social Media

- If you come across something worth sharing with the LIBER community, tweet it with @LIBEREurope & @LIBERConference tagged!
- Before tweeting, please make sure your workshop attendees are fine with their names/photos being on social media (if they happen to be featured in the tweet).

Full List Of Attendees

1. Péter Király / software developer & researcher / GWDG / Göttingen, Germany
2. Birgit Schmidt, Head of Knowledge Commons, University of Göttingen

Collaborative Notes

DSLib WG, curr 15 members - get in touch if you are interested to join
Defn of Data Science - 3 components: CS, Maths & Stats, Domain Expertise
Computational methods - e.g. text mining, ML

DS in libraries = use of these methods in the delivery or improvement of library services…
DSLib Activities:

Collections as Data
term coined by T Padilla
usually the data is used for our own services

Library Intelligence
Geared towards the improvement of traditional library services and support for decision-making by library management
E.g. data-driven item suggestions

Research Support
Support researchers through the research lifecycle
E.g. data management planning, research software engineering, ensuring data FAIRness

Research Intelligence
E.g. data: metadata of pubs, other reserach outputs; activities: analysis workflows

Survey results

n = 50 - survey still open
The survey is part of an ongoing landscape analysis

Categories - Research Support most prominent (18 out 50), Collections as Data (11), all have been covered
Describe the current DS activities in your library - hard to categorize

Library intelligence
Catalogue related: e.g. automated cataloguing, enrichments of bibliographic records, automated subject indexing, automate the acq of metadata, data quality, visualization of authority file record
Annif tool mentioned several times (Finnish Nat Libr), can extract from full text
Usage analysis: analyzing user behaviour (biptip tool), data on space occupancy, journal renewal, semi-automated de-selection, evidence based acq, analyzing query logs, COUNTER stats, loan stats

Cmts
Usage stats
Analysis on OA
Tapir project, TIB, automatically connect PIDs for pubs, projects, etc.

Research intelligence

Bibliometric analysis, campus output breaks down by publisher share, analyse research impact

Research support
Geo-referencing (crowdsourcing), hand-written text recognition (HTR) w Transkribus, publishing metadata, attribute PIDs to elements of metadata, RDM advisory services
E.g. Project Time machine, maps, trying to make the maps more precise, use of crowd-sourcing

Other
Automating the process for systematic lit reviews, ChatBot, text mining, analyzing watermarks, automated interactive dashboards, trend analysis

Q10 departments, teams
Single person, existing unit w new tasks, several units, cross-cutting WG, …
Research support, data mgt, schol comms, research software eng, acq and cat

Q: any use case not covered
A: U Manchester, bibliometrics team, experimenting with ChatGPT to write Python code by team members that do not code

Describe staff responsible
Background - LIS, CS, DH, other
Qualifications - self-taught, workshops, field-specific training
Roles - librarians, data specialists, data scientists, research data engineers & consultants, metadata managers

Who does the library collaborate with?
Vendors, suppliers; IT departments; univ WGs, individual reserachers, uni grants office
Data stewards, other academic libraries, publishers, research consortia

Challenges
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 data, researchers lack a god understanding of techniques

Copyright issues - e.g. at UGOE we have lots of data, but cannot use the resources in different contexts

Frequency of challenges - lack of staff

Legal or ethical challenges - how do libaries address them?
Considering rights and terms in data collecting phase, copyright info point
CARE principles, CC0

No personal data are collected
Ethical aspects of bibliometric analyses

How are the DS activities typically funded?
Internal funding, univ funding
Government, national, EU

What activities are planned?
publishing/author profiles, cat w AI, usage analytics, name disambiguation and clustering, duplicate detection, dev data science skills (for staff, researchers, students), …, knowledge graph/hub

Discussion
Anything missing?
Was there a q on tools used? No - would be interested
XX: sometimes have the data but do not have the questions for them - what are the main / most important questions that we can answer w our data?
Peter: usage, user behaviour; libr metadata far from perfect - improve the data
Exploratory data analysis, checking what I have, distribution etc.
András: library data - other types of data, e.g. repositories metadata
XX: usage of DS for RDM, concrete examples?
Peter: enrichment of data, adding identifiers, providing feedback for the institution, what research data
B: data quality, relevant for data curation
Nora: shape of the data you are working with, OCR - researchers assume that it is perfect
Peter: Dataverse, API to connect w Jupyter notebooks, Raven tool for R, dataviewers
YY: compile a list of tools and describe what to do with them
ZZ: software engineer, but no experience w academic world, understanding of the environment
Peter: hard to answer, skills and knowledge not in one person, …, would like to merge these areas
Nora: partnered w XX Univ, cultural heritage course, 1 year to trial, can you teach programming skills to librarians, successful, several left the library sector and went to the commercial sector, some stayed and got promotions, digital scholarship groups, code reviews, v friendly, you can ask/share anything, breaking down the silos
Pablo: in order to capture what is going on in libraries, track the use of software solutions, PowerBI etc.
Peter: teaching comp methods, Carpentries
Nora: training prgm, some case studies on how the training has grown, people w different roles