DSLib group – theme exploration – August 2021 – by Neha, Linda, Pam

In the brainstorming notes, a number of elements emerge that are mentioned in common by the participants:

- **Landscape analysis:** generate an overview of developments in the field of AI and machine learning,
- **Skill development within libraries**: generate an overview of tools and skills needed for this - for example Python/R,
- **Service to external users:** establishing existing guidelines and/or generate them,
- **Service to external users:** exchange views on how we deal with meta data – for example related to CHRIS/FAIR/Grants.
- **Service to 'internal' users:**

In further specifying the follow-up steps associated with the above points, the following categorization can be considered:

- **Target groups** (e.g. management, researchers, students, colleagues),
- **Various fields of research** (e.g. potential differences and/or similarities between alpha, gamma, beta, and meds domains),
- **Policy and strategy related action points** (e.g. for example how to report on scientific & societal impact in research evaluations),
- **Pragmatic aspects** (e.g. infrastructure, tools, related costs),
- **Sharing technical knowledge** (e.g. potential differences and/or similarities within countries and libraries, exploring text and data mining services, data visualization, how to implement new techniques in research data repositories, legal aspects, training colleagues and/or researchers, guidelines),
- **Connection between theory and practice** (e.g. for example, what DS developments are short term goals and what developments are realistic in a larger time span, how do DS development tie in with overarching policy and evaluation developments, is every data set suitable for DS or should we choose when to apply it).
- **Open Science and Recognition & Rewards** (e.g. how do DS development tie in with overarching policy and evaluation developments).

Subsequently, as a working group, we can discuss the questions below, grouping them by the above categories as desired, and prioritize them.

Overarching DSLib questions:

- Is the focus on libraries only or GLAM/LAM?
- How much attention to research data (overlap with RDM wg in Liber)?
- How much attention metrics for external users?
- For upskilling: focus on resources / promotion / training on DS? (or all?)
- Resources for DS practitioners or for those who want to learn DS?
- What's the outcome: a report? Use-case descriptions (~rdm wg)? DS training / guidance for organising training / resources for training? A specific DS service created collaboratively?

Data science in libraries: a) landscape analysis

- what are libraries doing in data science/AI/machine-learning
- Text analysis of liber institutions sites for mentions of ml/ds as starting point for landscape survey
- what are the current applications of DS in LAM. Specifically I would like to get information for each use case about data sources, methods/tools, output and whether the use case is in the research or production phase.
- What are the data that the library engages with and deals with? In-house/open data? What infrastructure and tools are used? How do they publish their outcomes?

- split goals in two sections? pragmatic (infrastructure, tools) vs. overarching goals (scientific impact & societal impact)
- It would be useful to look into the ideas that are/become embedded in data practices within library and map out what happens in this regard in different libraries. From what I have seen, sometimes, there is a clash between ideas as they are discussed and as they are executed in practice. Given the increasing use of data science / data analytics methods, a discussion on this topic could bring out ideas on directions that can be taken within libraries.
- Collect use cases

Data science in libraries: b) a service to external users

- data science-related services that libraries can provide their customers (researchers): the tools/skills that are handy in research
- Guidance to researchers how to use AI techniques in their research.
- Matchmaking: identification of funding streams for researchers.
- Modelling/curating (FAIR) data as fuel for AI techniques.
- Monitoring and predicting impact of Open Science for researchers, society, citizens.
- Be able to assist researchers / provide guidance on installed or emerging topics regarding data acquisition and transformation, such as TDM ; point them to existing tools
- Position the Library as a competence center for such services (guidance on installed or emerging topics regarding data acquisition and transformation)

Data science in libraries: c) skill development within libraries

Resources

- Build a list of grant funded ml/ds initiatives to see what is successfully being funded
- Open datasets for applying ml/ds in glams
- point people to resources already developed or being worked on (e.g. AI4LAM workshops, Turing humanities and data science WG; Library Carpentry AI / machine learning lessons, project or format specific material e.g. newspapers and Living with Machines, Impresso etc, computer vision, OCR/HTR; link to other landscape analysis e.g. Europeana AI survey; Terras forthcoming etc) as reference material for their own learning and answering queries.
- cooperation on a common literature list (Zotero group)
- Existing resources for processing bibliographic data in R / Python (e.g. doi verification, isbn hyphenation, record linkage) and exchanging ideas on what and how is done in different settings (use of machine learning, dashboards, data viz streamlining and so on).

Tasks within libraries / (G)LAM

- Extraction of metadata from AV, Images.
- Explore auto-generated metadata, text summarization, etc as information discovery tools in glams
- Scaling up pilots to full coverage of collections

Skills

- how to upskill librarians in data science-related skills (this can be for use within the library too, not necessarily for researchers or students)
- Upskill the colleagues and teams in that regard through professional training
- Help people understand what DS / AI / ML is particularly good for and when/where it's harder to do well; where it fits in library data lifecycles and the issues that emerge when applying to GLAM data
- Help people understand how to choose a platform for DS work depending on their need, number of images / source files, where the work fits in a general data workflow, etc.
- Help people anticipate costs (compute, storage) and overhead for accessing DS services, associated needs for digital preservation and storage, replicability

- Help people understand how to minimise environmental impact of DS / AI methods e.g. minimising re-training, avoiding blockchain
- how to build a team of data analytics within the organisation - who will join, what are the prerequisites and how to explain to the rest of the organisation what this is about. Since the library has many different data sources, this would probably be a cross-organisational work. What kind of competence is already available, and what is needed. (https://doi.org/10.7287/peerj.preprints.3160v2)
- Joint learning/upskilling/team building. Library Carpentry is an example.