

Data and digital objects: Manual and automated analysis to improve DMPs

Heather Moulaison-Sandy and Ngoc-Minh Pham
iSchool, University of Missouri, USA
SciDataCon, June 20, 2022

What is *data*?


- Important philosophical question
 - Answered very differently by researchers in different domains, using different methodologies, pulling from different traditions, etc.
- Some kinds of data are easy to identify
 - Observations, datasets that have been formatted and published (hopefully with the requisite metadata for findability and codebooks for use)
- But arguably, there's more data, especially digital data, that is produced when doing research.
 - What about electronic lab notebooks (used in the sciences, but also potentially in the social sciences)?
 - What about publications (normally digital)?
 - Code?

Digital objects

- Digital objects can be an essential element of the research enterprise – both objects that are created and published, but also ones that are used and shared.
 - The tenets of open science require that tools and other digital objects or code be made available in support of transparency and reproducibility.
 - We acknowledge (as many others do!) that this is a form of DATA!
- Digital objects are defined in a number of ways. We can think of them as “Objects on the Web, such as YouTube videos, Facebook profiles, Flickr images etc. that are composed of data and formalized by schemes or ontologies that one can generalize as metadata” (Hui, 2012)

Digital objects and DMPs

- We define digital objects in DMPs as any digital/electronic assets that support or are a product of the research endeavor.
 - They are a kind of *data*, in the sense of data being either evidentiary or a form of record that can be collected and studied.
- Digital objects associated with research can be text files, codebooks, codes/scripts, analytical methods, videos, drafts of manuscripts for publication etc. that supplement data and allow datasets to be found, interpreted, and (re)used.
 - As a form of data, these digital objects all should ideally be subject to FAIR principles and planned for in the DMPs.



Are digital objects mentioned in DMPs?

Purpose of this analysis: To investigate consideration of digital objects in DMPs using a sample of DMPs from funded projects in the sciences, and to compare that to consideration of other elements using a text-mining approach.

Method

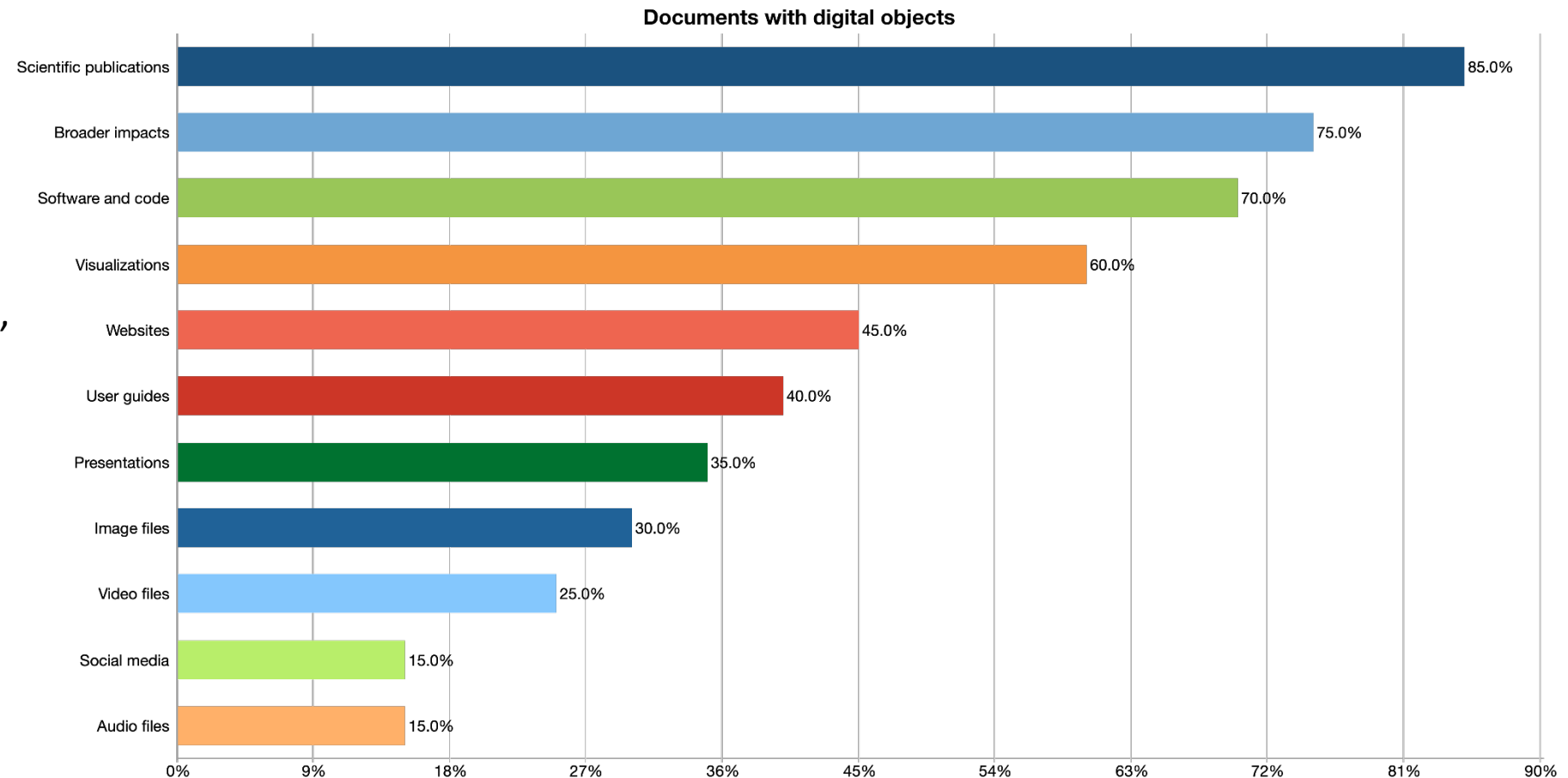
- Sample: 21 DMPs from projects funded by the Belmont Forum
- Data analysis approaches using MAXQDA software:
 - What digital objects are mentioned in particular?
 - Manually identifying and listing all the possible keyword for digital objects mentioned in the 21 DMPs
 - Assigning codes/categories to identified keywords
 - Building a dictionary on MAXQDA software with the assigned codes/categories
 - Identifying and searching keywords for digital objects throughout the corpus and assigning them to different assigned codes/categories through 'autocode with dictionary' function on MAXQDA
 - Manually reviewing results to ascertain context
 - Compiling results for analysis and visualization using 'Code frequencies' in 'Analysis' in MAXQDA
 - What is present in terms of content in the DMPs overall?
 - Automated analysis with text-mining features such as unigram frequencies and n-gram combinations offered in 'word frequencies' and 'word combinations' functions in MAXDictio in MAXQDA software
 - Manually grouping terms associated with [Belmont Forum scorecard criteria](#)

Terms/phases identified for digital objects

- **Audio files:** WAV, MP3, audio file, MP3
- **Broader impacts:** technical document, report, policy document, policy recommendation, op-ed, newsletter, narrative, brief, synthesis document
- **Image files:** photograph, image
- **Presentations:** workshop, presentation
- **Scientific publications:** publication, pre-print, paper, journal article, book
- **Social media:** Twitter, social platform, social network, Instagram, Facebook
- **Software and code:** software, simulation, model, source code, coding, script
- **User guides:** training video, restart file, ReadMe, manual, data description, data descriptor, concept note, codebook
- **Video files:** video file, Quicktime Movie
- **Visualizations:** visual, visualization, table, map, illustration, graphic, figure, diagram
- **Websites:** project website, project webpage, blog

Digital Objects Mentioned in the DMPs

- Coverage: Almost all mention digital objects (20/21 DMPs, 95.2%).
- Most common forms of DOs mentioned are **scientific publications** and **broader impacts** publications (reports, policy statements, etc.) followed by **software and code**.
- Other commonly mentioned formats include **visualizations**, **websites**, **user guides**, and **presentations**.
- **Image**, **video**, and **audio files** along with **social media** are the least common forms of DOs, which seems fitting.

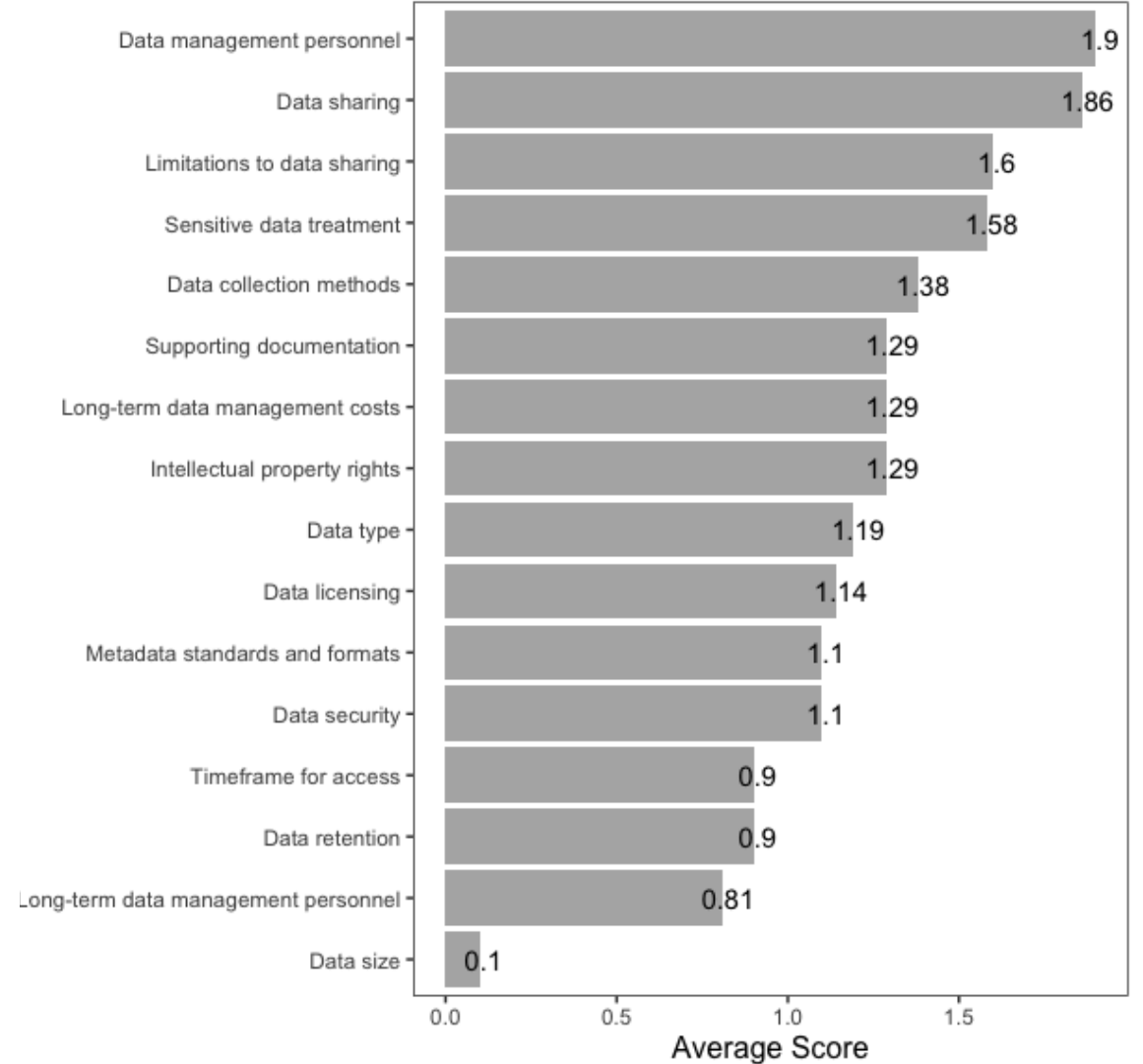


Belmont Forum Scorecard

- Scorecard criteria
 - Data
 - Data storage and use
 - Data management personnel
 - Data security
 - Data preservation concerns
 - Restrictions
 - Intellectual property
 - Supporting documentation
 - Long-term costs

Other Terms Aligning with the Scorecard

- High-performing criteria:
 - Data management personnel
 - Data sharing
- Low-performing criteria:
 - Data size
 - Long-term data management personnel
 - Data retention
 - Timeframe for access



Average scores across all DMPs, by scorecard criterion.

Discussions and Implications

- DMPs for the Belmont Forum are specifically meant to address digital objects.
- This analysis indicates at least a certain degree of compliance.
- It also finds strengths in the inclusion of personnel in the DMPs, implying that there is support for work with data and digital objects, beyond the scientists/research team.

Conclusions

- Considering digital objects is important to open science, as digital objects provide additional information that has the potential to allow for more accurate and performant reproducibility and replication of projects and results.
- Naming digital objects in the DMPs, as necessary, is the first step.
 - The Belmont Forum and other funders lead the way in this regard.
- Next steps should be investigated as to the best ways to promote buy-in on the part of researchers and data managers, and to support the publication of digital objects in alignment with FAIR Data Principles.