Conference Report: Artificial Intelligence for Data Discovery and Reuse (AIDR) and Open Science Symposium (OSS) 2020

Huajin Wang, University Libraries, Carnegie Mellon University

On October 18 and 19, 2020, the Open Science & Data Collaborations program at University Libraries organized two symposia, Artificial Intelligence for Data Discovery and Reuse (AIDR) and Open Science Symposium (OSS). AIDR aims to find innovative solutions to accelerate the dissemination and reuse of scientific data in the data revolution using the power of AI. OSS focuses on building awareness and support for the adoption of open research practices and encouraging innovative ideas about data sharing. They were hosted virtually as joint events this year due to COVID-19. There were an estimated 230 unique participants, including researchers, information professionals, research computing providers, from academia, non-profit organizations, government and funding agencies, and private sectors from all over the country.

AIDR 2020

Data reuse and extracting value from data is an important and challenging issue that involves both technology and humans. In this one-day program, some recurring themes arose from talks and panel discussions at AIDR.

Metadata and data harmonization for better AI, and vice versa

A set of widely accepted guidelines for data sharing are the FAIR (Findable, Accessible, Interoperable, Reusable) principles. However, most data shared online are not FAIR; in fact, data will not be FAIR unless the metadata describing the data is FAIR. Many communities have made great effort to generate ontologies that fit the specific needs for their data. However, having too many ontologies can be problematic. Several speakers shared their work on unifying ontology for data linking and harmonization.

Melissa Haendel shared two projects that lay groundwork for AI by harmonizing clinical data. The first is the Mondo project that uses a Bayesian approach to harmonize disease definitions from multiple sources by generating an overarching ontology, which enables disease diagnosis and mechanism discovery using multiple sources. The second project is the National Covid Cohort Collaborative (N3C) that includes massive, harmonized patient level EHR data for COVID-19 on a secure enclave that enables analytics and AI applications. Both projects highlighted that having traceable, licensed & approved, and connected data is the foundation for downstream applications of ML and AI.

In contrast to Melissa's position that emphasizes good data, Mark Musen argued that good AI is needed to generate even better data. His group created CEDAR, a web-based platform that allows investigators to describe their experiments by filling a metadata template using values selected from drop-down menus. Values in later steps are generated and updated based on user-selected values in earlier steps. This easy to use platform lowers the barrier and workload for documenting metadata, while avoiding problems of imprecise metadata often seen in spreadsheets.

Expert input and human-AI interaction

To make data truly reusable and AI truly useful to extract value from data, the human factor is essential. Rayid Ghani talked about how to build AI systems that use data to improve equitable and fair outcomes in criminal justice and other social issues. He pointed out that subject-matter experts and AI-human collaboration are the key for good AI, as the decision between optimizing for algorithm performance and optimizing for equity has to be made by humans. For successful collaboration between subject matter experts and technical implementation, we have to reason at the policy level first, before technical experts are able to build an algorithm that delivers desired outcome.This assessment is echoed by many other speakers, in projects ranging from training algorithms for ontology mapping for the Dimensions database, to training AI to understand meaning and intent from image-based advertisement (Adriana Kovashka), to analyzing human generated data (Martial Hebert).

Easy-to-use tools to facilitate adoption and dissemination

A major barrier for reusing data and sharing data in a way that is FAIR, is that it takes a lot of hard work to document data and workflow using the right ontology and controlled vocabularies, and keeping this information up to date. In addition, concerns such as licensing and privacy hinder shared data from being reused. It is thus essential to build easy-to-use platforms to facilitate data sharing and analysis pipelines. This is especially important for interdisciplinary collaboration. The aforementioned CEDAR project is a prime example of technology and user-friendly platforms helps to lower the effort for metadata creation and avoid errors. Another example is using predictive models to facilitate clinical decision making based on EHR, a type of data that is usually messy and haphazard. Jeremy Weiss talked about the TL-Lite platform, a web-based temporal visualization tool that implements a temporal machine learning framework in a step-by-step way, so physicians are able to make timely decisions based on EHR data without having to write a single line of code.

Interdisciplinary collaboration and building a healthy data ecosystem

Data reuse is a common issue shared across disciplinary boundaries. A major takeaway from last year's AIDR is that we must build a healthy data ecosystem before any AI can be done. The same sentiment was reiterated throughout this year's AIDR, especially during the fireside chat and panel discussions.

Many speakers mentioned that for different disciplines to work together, it is important to create common best practices and tools to keep track of how data was collected, acquired and transformed, and have formal ways of describing it that can be shared across domains. They emphasized that to make sure that data is truly reusable, documentation needs to be distributed with data, as well as provenance and licensing information. These needs were clearly demonstrated in the CORD-19 and N3C projects.

But data and it's dissemination is not the only piece of the puzzle; a data ecosystem also includes the community that uses data, as pointed out by Ross Epstein, Chief of Staff at SafeGraph. Despite being a private organization, Ross's company has been releasing data for free during COVID-19 and has formed a Slack group involving many individuals and organizations actively working together using these data. Indeed, an indispensable part of building a healthy data ecosystem is to ensure that all stakeholders--researchers in all disciplines, developers, data and information professionals, government and funding agencies, publishers, research consortia and universities, private sectors and non-profit organizations--all need to work together towards a common goal. For higher education institutions like CMU, we need to train students to have a deeper understanding of what data means and expose them to a wide range of data, as different data leads to different technical challenges. We need to provide incentive in the reappointment and promotion process by recognizing not only published papers as research products, but also datasets and data curation efforts. We need to build a 21-century library that serves as valued partners in research to share and capture research processes and outcomes. When asked about how to leverage the power of AI to reuse data for interdisciplinary collaboration, Martial Hebert, Dean of School of Computer Science of CMU, pointed out that not only do we need to develop tools to lower the entry barrier, but also to develop models to understand how people interact with data and to learn how a model learned from one dataset will transfer to another dataset in a different discipline.

Lessons learned from COVID-19

The COVID-19 pandemic has changed many aspects of how research is done and how information is disseminated globally. It has fostered collaboration and openness at a level that was not possible before; huge amounts of publications were produced and many high value open datasets were released in a short period of time. Lucy Lu Wang and Kyle Lo talked about the CORD-19 dataset, a structured, machine readable COVID-19 research publications corpus released by Allen Institute for AI, integrating over 70,000 papers on COVID-19, with labels created by experts and community input. Many open source NLP tools quickly followed to extract patterns and trends from this corpus to develop strategies for disease control and treatment. Similarly, Imran Haque talked about RxRx-19, a morphological profiling / cell imaging dataset and metadata created and made completely open by a private company, Recursion. The data release triggered subsequent collaboration from many researchers and organizations.

In addition to the apparent importance in data sharing and reuse, one important challenge related to COVID-19 is the fast pace that data reuse is needed. Thus, technology that deals with ever-changing data is crucial. For example, in CMU's Delphi project that integrates multiple types of COVID-19 data from many sources and develops forecasting tools using these data in real-time, one key issue is data versioning and making continuous predictions, knowing that data is continuously being updated and all the data is not there at any given time. Similarly, in both the CORD-19 project and the N3C project mentioned previously, data provenance and versioning has become essential in reusing data.

OSS 2020

OSS touched on many topics in open research, open data and open infrastructure, spanning many disciplines. Below are some common themes many speakers touched upon.

Collaborative science and team science

Science has evolved to a place where it is increasingly difficult to do research without collaboration across expertise areas and disciplinary boundaries. Albert Presto shared his research on measuring air pollution concentrations in different places across the country and using the acquired data to understand how air pollution impacts various communities. Saskia de Vries talked about their work on Allen Brain Observatory, a large brain physiology dataset hosted at Allen Institute for Brain Science, in which technicians, surgeons, engineers and data scientists work together in a big team science approach. In both talks, it was emphasized that communication and sharing between teams of different expertise is essential, and that having standardized tools and high throughput pipelines not only allows for collecting a lot of data efficiently, but also building in quality control metrics that help with reproducibility and data reuse. Alexxai Kravitz, a behavioral neuroscientist who built devices that enable high throughput data collection and real-time data sharing from mouse behavior studies, proposed that we should take the reproducibility pipeline a step back from data analysis to data collection, and that having tools and infrastructure that allow high throughput data collection and sharing is especially important for disciplines in which the published results are almost always statistically underpowered using traditional approaches.

Benefit of practicing open science

Almost all speakers clearly demonstrated how open science helped to establish collaboration, accelerate discovery and improve reproducibility in real-life research. When asked how open science has affected the way science is done, Lex commented that being able to share results online in real-time has opened up possibilities that they have never thought of before, such as collaboration at multiple sites, and shorter lag time between data collection and getting results. Saskia added that having more people use the data helps to understand more about data collection and quality control, and allows the research community to improve iteratively and enhance reproducibility. The reproducibility perspective has also been emphasized by Justin Kitzes and Ciera Martinez. They argued that having a reproducible data analysis workflow is the backbone of how researchers learn to work together.

A major benefit of open science is making new discoveries using existing data, a topic that is very much tied to the data reuse theme of AIDR. To this end, Marina Sirota, a computational biologist, shared their work building a repository on preterm birth data that houses scattered multiple omics data at one place. Using publicly available data in this repository, Marina and many other researchers were able to establish new collaboration with a broader community and accelerate new discovery and disease therapeutics.

Another major development in open science is the rise of non-traditional publishing and research evaluation. Richard Sever, Co-Founder of preprint servers bioRxiv & medRxiv, made the calculation that when new discoveries are seen as a function of time, at the current pacing, science would be speeded up 5 times in 10 years, thanks to preprints at bioRxiv & medRxiv. Varsha Khodiyar, Data Curation Manager at Springer Nature, also shared the publishing and review process to publish a data paper at Springer Nature journals including Scientific Data, and recommendations on increasing FAIRness and visibility of shared data. These talks triggered many more discussions about innovations in peer-review, a topic continued from AIDR the day before.

Challenges in open science and data sharing

Despite the benefits, open science is not without challenge and one of the major issues that surfaced is choosing what data to share. Taking the Allen Brain Observatory for example, data could be shared in all forms, from raw movies, which are bulky and associated with many auxiliary files, to various stages of intermediate data, to derived metrics at the end of the pipeline that can be shred in spreadsheets. To fit the different needs of users, the data sharing effort has gone far beyond publishing a research paper or a data paper, but deployed a website to share all stages of data, along with white papers that document how data was collected and processed and tutorials of how to use the data.

The complexity and amount of effort required for data sharing also pointed to another challenge in open science--incentive and support. One one hand, many researchers are sharing their data to meet funder mandates, but without a coordinated effort as a community, the shared data becomes an increasing amount of big data files that cannot be reused. A culture shift for data reuse is needed for better and more effective data sharing. One the other hand, it is difficult for individual labs to spend huge amounts of time and resources curating data, building software and engineering tools. As a community, more thought should be given about incentives and tools that support open science that are actually useful.

Even though one apparent benefit for making data open is to increase research reproducibility, William Thompson cautioned about the other side of the coin. He pointed out that sequential hypothesis testing on the same dataset by multiple researchers inflates the rate of false positives, thereby decaying the usefulness of the data over time. Therefore, proper sequential correction procedures should be considered when using open data.

Social aspect of open science and open access to information

While open science is still at its early stage, there are social implications to consider while advocating for its wide adoption; much of this can be learned from existing open infrastructure such as the internet. Sarah Kiden pointed out that the open nature of the internet has made it a success today because it allowed people to contribute and to add new protocols and services. However, openness does not guarantee equity; with only 53% of the world population having access to the internet, we need creative ways to include underrepresented groups in the conversation and have their voices be heard. Similarly, Kari Jordan, Executive Director of The Carpentries, shared the vision of The Carpentries to build an inclusive community teaching data and coding skills, because open science is better served by having diverse people use data to address diverse questions. She said, “inclusion is more than inviting people who don't look like you to the conversation, but ensuring that when they get there, they're able to interact and contribute in ways that are meaningful to them.”

Another outstanding problem faced by the open cyberspace is the prevalence of disinformation that is being used to drive bias and influence people’s minds. Kathleen Carley talked about how disinformation campaigns use bot networks and trolls to target receptive groups and form “echo-chambers” that elicit emotional rather than logical response. According to Kathleen, there has been a surge of such anti-science, anti-minority campaigns since April not only on social media, but some have occurred in preprint servers such as arXiv.

Conclusion

After another successful year of AIDR and OSS, there is still a lot to think about and a lot of work to do. These two events addressed many overlapping topics from different angles, including interdisciplinary collaboration, data stewardship, incentives, societal impact of open science and open data, and more. Based on post-event survey response, many audiences enjoyed the topics and synergy presented in these two joint events.

In the spirit of open science, we will make slides and recordings publically available in a repository in Open Science Framework (AIDR 2020: https://osf.io/tchdq/; OSS 2020: https://osf.io/hbt7c/). Hopefully this will serve as our small contribution towards a more open, equitable, collaborative, and reproducible research landscape.