Summary

The NIH Office of Data Science Strategy (ODSS)-sponsored Search Workshop explored current capacities, gaps, and opportunities for global data search across data ecosystems to enhance data discovery and reuse. Discussions during the two days emphasized the need to establish and enforce clear standards, policies, and requirements to ensure newly generated and existing data is findable, the first element of the FAIR principles, and amenable to more sophisticated search techniques. Additionally, discussions highlighted the multifaceted search landscape, with complex use cases due to the wide range of diversity and heterogeneity in science. Speakers and attendees agreed that search system design must account for these complexities in order to prevent researcher fatigue from effort required to make data searchable. There was an emphasis on the need for interoperability between data platforms, allowing enhanced data sharing for improved research, discovery, and innovation. The discussion and breakout sessions emphasized existing capabilities and opportunities, understanding and reducing bias in assembled data, incorporating emerging search technologies, and recommendations for those involved with these search systems, while aligning to three use case scenarios: data discovery, cohort building, and knowledge searches.

Introduction

The ODSS hosted a workshop from January 19-20, 2022, with more than 400 online attendees to explore current capabilities, gaps, and opportunities for global data search across the data ecosystem. Discussions included data discovery, cohort discovery, and knowledge retrieval in scientific data repository search efforts.

The meeting was co-chaired by **Dr. Stan Ahalt**, Director of the Renaissance Computing Institute (RENCI) and Professor of Computer Science at the University of North Carolina Chapel Hill and **Dr. Susanna-Assunta Sansone**, Professor of Data Readiness at the Engineering Science Department, and Associate Director of the Oxford e-Research Centre, part of the University of Oxford in the UK.

Dr. Susan Gregurick opened the meeting and Dr. Simon Twigger, Director of Data Science for BioTeam, provided an overview of the Search Listening Tour, which formed the background and precursor for the workshop, generating three focus areas:

1. **The User:** Data ecosystems must focus on the use case, including user-centric designs and effective UI/UX. Above all, they must help encourage and bolster the existing culture of data sharing.
2. **Information:** To move forward, data ecosystems must focus on data discoverability, access, and reuse, while striving to create an environment of interoperability between multiple platforms and technologies.
3. **The Data:** Ensuring data quality is key, but there should also be a bigger focus on the metadata researchers need to find information and ensure or understand the quality of datasets.

Keynotes: The Evolution, Possibilities, and Implications of Search Capabilities

The conference included two keynote addresses by **Dr. danah boyd**, Ph.D., Partner Researcher at Microsoft Research, founder and president of Data & Society, Distinguished Visiting Professor at Georgetown University, and Visiting Professor at New York University and by **Sir Nigel Shadbolt**, Ph.D., Principal of Jesus College of Oxford, Professorial Research Fellow in the Department of Computer

Science at the University of Oxford, Chairman and co-founder of the Open Data Institute, and Visiting Professor in the School of Electronics and Computer Science at the University of Southampton.

On the first day of the workshop danah boyd's keynote address focused on the issues of data discoverability, centering around key themes that provided a framework for the day's events.

1. **Query Languages:** Languages used in specialized scientific disciplines do not readily map to most query languages, creating a disconnect in search efforts. Further, personal biases and ideology are often present in a user's search query, which significantly impacts results.
2. **The Public Domain:** Most science exists behind paywalls, creating an environment where headlines do more work to convey meaning than content in the public domain.
3. **Media Manipulations:** A generally non-nefarious precedent exists for data manipulation in the public sector, where information is presented to further biases, particularly when the subject is politicized. Building alternative, search-oriented content will provide the public the ability to fact-check these manipulations more readily.

On the second day of the conference, Sir Nigel Shadbolt delivered his keynote address on the evolution of search capabilities and future possibilities, focusing on three additional themes that set the stage for the day's breakout sessions.

1. **Data as Infrastructure:** Data should be viewed as a type of infrastructure. As such, it must be reliable, safe, of high quality, maintainable, accessible, interoperable, economic, and well governed. Further, to enrich future scientific inquiry, the data ecosystem also needs to be scalable, transparent, repurposable, revisable, and must include rich metadata.
2. **FAIR Principles:** Current challenges include investigator difficulties in understanding how to support FAIR and the need for specialized training material and capacity building. Further, privately held datasets are likely less FAIR than public ones; the citability of all datasets must be addressed to move forward.
3. **Data Policy:** Moving forward, data policy will be crucial. Considerations include incentives and legal obligations, legacy data, privacy concerns, and metadata maintenance, as well as the challenge of managing secondary data and integrating concepts such as behaviors, communication, and social relationships into the management of existing and new data.

Day 1: Setting the Stage and Exploring Themes in Search

After introductions and a keynote address by danah boyd, the day continued with discussions on *Use Cases for Dataset Discovery* by Dr. Mike Huerta, *Cohort Discovery* by Dr. Anne Deslattes Mays, *Knowledge Retrieval* by Dr. Purvesh Khatri, *Ethics* by Dr. Julia Stoyanovich, and *Cultural and Social Aspects of Search* by Dr. Larry Hunter.

Panelists discussed the importance of data searchability and discovery, inherent biases and equity concerns, the importance of effective technology and its usability, and the necessity for confidence in the datasets uncovered by a search query.

Six critical points were emphasized throughout the morning by the panelists, moderators, and attendees:

1. The aggregation of data from independent sources must be improved. Creating more searchable, well-maintained, and standardized databases, using common data models, will assist scientists to appropriately utilize existing data for their needs.
    - An example shared: “I need to search for data on patients like the patient in front of me,” resonated with attendees. This is impossible without the ability to search multiple sources and requires significant coordination across communities.
2. Data repositories are key to the NIH mission of advancing scientific and medical discovery. The quality of these repositories must be carefully maintained.
3. The inherent diversity in research data must be embraced. The traditional approach of reducing heterogeneity in research data does not capture the complexity of the underlying data and creates results that are not generalizable. Variance can be a blessing, not a curse.
4. There are many sources of bias in information retrieval, including that search technologies do not correct for biases to areas of science which have received the most targeted attention by researchers and are therefore more strongly represented in repositories.
5. Data must accurately represent the real world. To understand data appropriately, researchers must consider the facets of equity, including representation, access to information, and the ability to mitigate potential downstream inequalities.
6. Discovery will only happen by moving beyond semantic relationships, accounting for biology, and rethinking allowable search query inputs to make non-obvious information accessible to scientists.

The next session explored cross-cutting data discovery themes, with input from Dr. Maryanne Martone on *Data, Metadata, and Search*, Dr. Rick Stevens on *Cutting Edge Technologies for Discovery*, Dr. Jina Huh on *Effective UI/UX*, and Dr. Matt Might on *Trust in Search*.

This session yielded four additional considerations for progress in the future of search.

1. To address a specific scientific question (use case), researchers must contend with the fluidity, heterogeneity, and possible unfamiliarity of the domains and data relevant to the specific problem. This will require a focus on the mechanisms describing data.
2. In the future, search should have the capacity to understand the intention of a query, replacing indexing with pre-trained models that are aware of the context and document-to-document relations, and shaping retrieved data sets based on previous understanding of the user.
3. Trust in search can only be achieved if the soundness, completeness, and meaningfulness of retrieved information are intact. To ensure this, there must be a clear connection between the provenance and evidence for any dataset.
4. Training will be essential for the task ahead. Moving forward with search efforts will require the participation of libraries and librarians in a training and curation capacity. It will also necessitate training students on these methods earlier in their undergraduate careers.

Workshop participants also contributed insightful comments in chat throughout Day 1, including the following ideas:

1. **Search is a powerful, essential, and lucrative tool**. While filter, sort, and ranking features are often overlooked, they have powerful implications for shaping social opinions and beliefs. How we build search becomes a feature of who is at the table.
2. **Search is, at its core, a socio-technical practice.** To manage search effectively, there must be an understanding of what the user is after, the acceptance of the fact that search is global, and its

problems are shared. Within the socio-technical context, many aspects of search have different meanings. Additionally, the roles of science influencers and social platforms are expanding.

3. **Search needs are significantly more complex than string searches.** User needs are evolving. Searches are made across multiple disciplines, evolving over repeated cycles, and based on non-traditional queries, (for example: ”find me images that look like this”). Users will need to be able to visualize and interact with results beyond simply using standard language conventions.

Day 2: Breakout Sessions on Data, Cohort, and Knowledge Discovery

Day 2 began with a keynote address by Sir Nigel Shadbolt.

Dr. Ian Fore discussed the current search landscape at the NIH, presenting three themes to serve as background for the day’s breakout sessions.

1. **Cohort Building**: The difficulties of cohort building are a consistent area of need expressed in NIH search efforts.
2. **Concerns**: Reliability and trust in search requires additional effort, without placing an overwhelming burden on researchers. There must also be consideration of socio-technical factors to ensure the uptake of promising technologies.
3. **NIH Responsibilities**: NIH has the following relevant roles:
    a. Convenor and guidance provider
    b. Provider of resources (e.g., Anvil, Biodata Catalyst, Cancer Research Data Commons, Kids First, NLM resources, NCATS Translator and others.)
    c. Funder
        - Funding research and resources
        - Ability, via conditions of funding, to encourage or require practices identified by the community that would benefit Search

Breakout sessions

Participants were sent to breakout rooms for detailed discussion of their respective topics, including multiple takeaways for the future of the data ecosystem.

Data Discovery, *moderated by Dr. Deb Agarwal:*

1. **Opportunities and Challenges:** Data providers have a heavy burden that needs to be lightened. Repositories need to provide appropriate methods to query across topics and integrate multiple datasets. This will be accomplished by emphasizing data identifiers and putting resources into high-demand topic areas.
2. **Encouraging and Supporting Citation:** Incentives and a reduction in the cost of citation will encourage better data. People will cite if there’s value to it. The future of search should look towards more auto generated DOIs and data management plans.
3. **Ease in use:** Traditionally there has been significant focus on making life easier for data providers, but additional considerations are needed for data user wants.
4. **Data Curation and Funding**: The larger scientific community cannot be expected to curate data for “free.” Human/machine hybrid curation of metadata will require serious investments of funding and time but are necessary to generate usable and reproducible data returns across multiple platforms.
5. **Integration:** Dataset search should occur within the tools and workflows data scientists currently use for analysis. Training is necessary to ensure the appropriate use of open and standardized APIs and large-scale metadata standards.

Cohort Discovery, *moderated by Dr. Adam Resnick:*

1. **Opportunities and Challenges:** A lack of strong feedback loops between data provisioning and data consumption have created a gap between those providing data and those who need it. To correct this, policies, incentives, and standards are needed.
2. **Approaches to Harmonizing Data:** Cohort creation as a centralized process is not a sustainable long-term framework. Data harmonization will require funding for human and AI-based curation and may necessitate providing deeper access into the source of data when feasible.
3. **Making Investments:** Investments should be made to establish exemplar datasets, standardize tools, and incentivize the searchability of data.
4. **Forward Thinking Capabilities:** Expanded queries are the future where users can search for like images, data streams can incorporate the dynamic state of patient-level data and new models are created for collaborative frameworks across existing repositories, commons, and healthcare ecosystems.

Knowledge Discovery, *moderated by Dr. Anita Marie Caywood Crescenzi:*

1. **Opportunities and Challenges:** To retrieve useful information, repositories should focus more on user-friendly, text-based, single-search boxes. This will allow researchers to find and access information without the need for excessive analysis.
2. **The Structure of Information:** Better standards for metadata and annotation are needed to structure information in a way that is more tailored to the user.
3. **The Threat of False Information:** Data quality should be clarified, increasing the role of curation as data rapidly changes.
4. **The NIH as Leaders:** The NIH can assist these efforts by acting for the common good, showing leadership, and providing sustainable funding to further develop resources.

Key Themes

In summary, the workshop identified the following key themes as important for NIH to pay attention to while addressing the search needs of biomedical researchers.

1. Scope

“Search” is not limited to a quick web search-result-action paradigm. As part of a growing understanding of the need to combine data collections, the workshop securely anchored cohort building as a part of search. Discussions also identified that, for a scientist, search extends over time, which is an activity that needs to be supported.

2. Policy

Modified data policy will be essential moving forward. This may include the need for clear policy on data standards and incentives to participate. Additional incentives may be necessary to increase data quality and bring about better trust in data. Machine-actionable data management plans are valuable tools for this purpose.

3. Curation and Data Quality

Data must be viewed and utilized as an economic resource, where the time, effort, and funding dedicated to producing and maintaining the data yields value by furthering research efforts and receiving potential scientific and financial gains resulting from its use. As such, data curation is essential to maintaining data quality as a fundamental part of our infrastructure. Moving forward, a combination of human and AI curation will be necessary to maintain existing datasets and incorporate new ones as they are added to growing repositories. To assist this effort, repositories may require standardized indicators of data quality. However, additional exploration may be necessary to find appropriate quality assessment methods, as some expressed counter-productive experiences with previous attempts at labeling data quality.

4. Data Collections and Heterogeneity

There is a growing need for data from heterogeneous disciplines to address complex scientific problems. In some cases, data collections aggregate diverse sources in a single collection. This is important, but not sufficient. The ability to combine datasets as needed from diverse sources in novel and appropriate ways is essential to support the imaginative process that scientists follow to address complex problems. The world is complex and federated, and solutions must be designed with that in mind.

5. Technologies

Currently, there is no quick and easy technical solution to search. To enable the functionality of future search efforts, it is necessary to consider alternatives to current methods of data indexing. While concepts like language models or AI-based methods can add features of automation to aspects of search, more research into these purpose-built techniques is needed to get closer to building the useful technologies we are striving for. Investments in technologies that will help accelerate the search paradigm and assist in maturing automated curation are critical.

6. Search as a Social Ecosystem

The individual components and organizations that contribute to search form an ecosystem. This ecosystem is a complex network that must reach beyond the large and familiar repositories to smaller ones, and to data consumers across the scientific landscape. Community engagement and active participation is an essential component of participating within an ecosystem. The following social considerations should be addressed:

- **The Code/Tools***:* Establishing a search marketplace, requires a collaborative search framework that supports innovative, “pluggable” search mechanisms, and search UIs to be deployed by researchers for specific search contexts and intents. This requires thoughtful design of a framework that supports open-source search libraries, permits alternative or competing mechanisms of search (served through APIs), and delivers results in a formalized data structure that flexibly supports display, prioritization, serendipity, and value. Additionally, this framework should support auditing the use of various components and aggregated search results, to highlight heavily used searches and aid in assessing data value.
- **The Data***:* Searching and accessing heterogeneous data across domains and jurisdictions is a sociotechnical challenge. It requires the active participation of data resources and catalogs across domains to expose key metadata that enables search, as well as understanding and agreement on their data governance and access procedures. In a competitive funding world where every data resource looks to ensure sustainability by delivering the right system to their user community, it is essential that any technical implementation or federated solution delivers

a data advantage to participating data resources. This works as an incentive and reward to buy-in and investment in conducting the necessary technical work.

- **A Social Contract***:* In addition to the technical frameworks for tools and data, active consideration and practice are needed regarding how individuals and teams collaborate effectively. This includes member interactions within the ecosystem, their incentives, and their roles, as well as data access, compliance, and reuse of others' work. It is important to gain community input, even while in progress, for the openness and sharing of work on standards. Understanding roles of other participants in the ecosystem, relying on their contributions, and the ability to trust them, require conscious work.
- **Science and Trust***:* Building a functional and adaptable search for researchers will require overcoming significant social problems associated with more commonly used, less specialized internet search platforms. Scientific search efforts must provide results that are valid, verifiable, open where possible, and without bias, or the system will have little value. Ethical concerns and confirmation bias in platforms like Google are brought about as the platforms account for language variances and previously gathered knowledge of users, to cater to their specific interests. While such user-specific catering may be useful in the future scientific search landscape, these accuracy and ethical concerns will only be mitigated if the data is verifiable, and the research methods are transparent. Search will only be beneficial if researchers are able to count their results as truth.

Speaker and Moderator Affiliations:

Dr. Stan Ahalt: Director of the Renaissance Computing Institute (RENCI) and Professor of Computer Science at the University of North Carolina Chapel Hill

Dr. Susanna-Assunta Sansone: Professor of Data Readiness at the Engineering Science Department and Associate Director of the Oxford e-Research Center (University of Oxford)

Dr. Susan Gregurick: Associate Director for Data Science and Director of the Office of Data Science Strategy (ODSS) at the National Institutes of Health (NIH)

Dr. Simon Twigger: Director of Data Science for BioTeam

Dr. danah boyd: Partner Research at Microsoft Research, founder and president of Data & Society, Distinguished Visiting Professor at Georgetown University, and Visiting Professor at New York University

Sir Nigel Shadbolt: Principal of Jesus College of Oxford, Professorial Research Fellow in the Department of Computer Science at the University of Oxford, Chairman and co-founder of the Open Data Institute, and Visiting Professor in the School of Electronics and Computer Science at the University of Southampton

Dr. Mike Huerta: Associate Director of the US National Library of Medicine and Director of the Office of Strategic Initiatives at National Library of Medicine at the NIH

Dr. Anne Deslattes Mays: NIH Data and Technology Advancement (DATA) Scholar

Dr. Purvesh Khatri: Associate Professor of Medicine and Biomedical Data Science in the field of Biomedical Informatics at the Research Institute for Immunity, Transplantation, and Infection

Dr. Julia Stoyanovich: Associate Professor in the Department of Computer Science and Engineering at the Tandon School of Engineering and the Center for Data Science

Prof. Larry Hunter: Professor of Computational Biology at the University of Colorado, Denver

Dr. Maryanne Martone: Professor Emerita at the University of California San Diego and founder of SciCrunch

Prof. Rick Stevens: Argonne's Associate Laboratory Director for Computing, Environment, and Life Sciences

Dr. Jina Huh: Assistant Professor of Human-Computer Interaction in the Department of Information Science at Drexel's College of Computing and Informatics

Dr. Matt Might: Director of the Hugh Kaul Precision Medicine Institute at the University of Alabama at Birmingham and faculty member in the Department of Biomedical Informatics at the Harvard Medical School

Dr. Ian Fore: Senior Biomedical Informatics Program Manager at NCI's Center for Biomedical Informatics and Information Technology

Dr. Deb Agarwal: SciData Division Director at the Lawrence Berkeley National Laboratory

Dr. Adam Resnick: Director of the Center for Data-Driven Discovery in Biomedicine at CHOP

Dr. Anita Crescenzi: Research Professor at the University of North Carolina at Chapel Hill in the Eshelman School of Pharmacy and the School of Information and Library Science