Summary Report

Federal Agency Listening Session | June 14, 2024

The Federal Agencies Listening Session of the BIOFAIR Data Network Project held on June 14, 2024 was led by BIOFAIR Data Network Steering Committee members William Moser (National Museum of Natural History), Gil Nelson (iDigBio), Nico Franz (University of Kansas Biodiversity Institute and Natural History Museum), and Brooke Long-Fox (Phoenix Bioinformatics), in collaboration with Key Domain Representative Scott Miller (National Museum of Natural History and Interagency Working Group on Scientific Collections - IWGSC).

Among the 35 session participants were 14 representatives from the Biodiversity Collections Network (BCoN) and 21 representatives from federal agencies, including the U.S. Department of Agriculture (USDA): Agricultural Research Service (ARS), Animal and Plant Health Inspection Service (APHIS); U.S. Department of Commerce: National Institute of Standards and Technology (NIST), National Oceanic and Atmospheric Administration (NOAA); U.S. Department of Defense: National Defense University (NDU); U.S. Department of Health & Human Services: National Institute of Health's (NIH) National Cancer Institute (NCI), Centers for Disease Control and Prevention (CDC); U.S. Department of Homeland Security: U.S. Customs and Border Protection (CBP); U.S. Department of Interior: Bureau of Land Management (BLM), Bureau of Ocean Energy Management (BOEM), National Park Service (NPS), U.S. Geological Survey (USGS); U.S. Department of State; National Science Foundation (NSF); and Smithsonian Institution.

Summary

Attendees discussed coalescing on a path forward to align various federal and non-federal initiatives towards data integration. Participants were encouraged to make informal contributions, sharing their perspectives in a personal capacity, not on behalf of their agencies. Discussions focused on exploring how integrating federal agency data with other sources can advance agency missions; identifying obstacles to participation of federal agencies in external networks; data aggregation through various systems like GBIF (Global Biodiversity Information Facility); the need for aggregators to retain the identity of original data sources; and the idea of “extended collections” or extending beyond individual specimens to encompass entire collections for broader data integration and accessibility. Future sessions will build on the topics discussed, aiming to develop a shared vision for an integrated data network.

Presentations

- **William Moser** provided an overview of the Biodiversity Collections Network (BCoN) and the NSF-funded BIOFAIR Data Network project (Award No. 2303588). The BCoN Steering Committee is part of a broader community representing various traditional museum groups, botanic gardens, paleontologists, culture collections, zoos, and other types of biodiversity collections. BCoN aims to promote the integration, use, and impact of biodiversity data and collections. The group has promoted the development of an Extended Specimen Network as a unifying goal for biodiversity collections over the next decade. Moser shared the goals for the Federal Agency listening session (the first of six domain-focused listening sessions) and discussed the need for a collaborative approach towards data integration. Moser highlighted the significance of the Extended Specimen Network (ESN) for linking and integrating biodiversity data and emphasized BCoN's goals of aligning data initiatives towards realizing the ESN vision.

- **Scott Miller** introduced the Interagency Working Group on Scientific Collections (IWGSC). Fifteen federal agencies that own, operate, and fund collections participate in IWGSC, which focuses on establishing collections as a necessary component of the U.S. research and development (R&D) infrastructure. Four recent reports highlighted the role of IWGSC in supporting federal collections: a 2021 White House report, "National Strategic Overview for Research and Development Infrastructure"; a 2023 White House report, "Vision, Needs, and Proposed Actions for Data for the Bioeconomy Initiative"; a 2020 IWGSC report, "Economic Analyses of Federal Scientific Collections"; and a 2023 IWGSC report, "The Unique Role of Federal Scientific Collections." Federal collections differ from other collections in unique ways:
    - They're tied to Department and Agency missions. Research is primarily mission-driven, not curiosity- or profit-driven.
    - Long-term goals and support is derived from Congressional legislation.
    - Sampling can extend over decades, creating unique longitudinal time-series for trend analyses
    - Often used for legal and regulatory enforcement.
    - Sampling and collection management may require chain-of-custody sampling and more secure storage, making them rigorous.
    - Expectations from federal collections include transparency in use and access regulations; accessibility within and across disciplines (except for cases of biothreat/defense), accountability for the services offered, long-term preservation and security, and contributing to All-Of-Government responses.

- **Dave Vieglais** presented an overview of the iSamples (Internet of Samples), an NSF-funded project to explore and implement the infrastructure (such as "server-less" technology) needed for discovery across multiple domains of environmental samples (e.g., biological, geological). A key challenge is that there are many ways to describe a physical sample-descriptions are often domain-specific. In addition, these samples are part of multiple data management systems leading to challenges for interoperability. The solution would be a common, extensible model that ensures consistent access and

enables aggregations for discovery or retrieval. The presentation highlighted the infrastructure for discovering and accessing physical sample information across domains; potential use of artificial intelligence (AI) for mapping from various systems into a common model; the challenges of diverse data descriptions; and the importance of a consistent representation and access method.

Key Topics Discussed

Need for Collaborative Data Integration

Information integration between federal and non-federal collections could lead to greater understanding to address research questions that are important to investigate from agricultural or economic perspectives (e.g. pesticide use and colony collapse). Several participants expressed enthusiasm for collaborative data-sharing and integrating efforts. A participant from USDA shared their work on creating a searchable database for white flies and the potential benefits of centralizing biological data. A comprehensive, integrated data and extended specimen network would make data easy to find and leverage for research. Integration efforts will lead to improved and easier data access for research, outreach, and education.

Potential Challenges for Integrating Federal Data with Other Data Sources

The following issues were raised during the discussions:

- A discussion about the meaning of ‘data aggregation’ revealed differences in how it is defined and understood. According to some, aggregation means storing disparate data from different sources together on a platform (GBIF, iDigBio, etc.). Others view it as a ‘collection of granular data’ (specimen level, study level, etc.). Potential differences between ‘data from a specimen’ and ‘data from a collection of specimens’ require more exploration. Clarity is also needed on the difference between ‘extended specimen’ and ‘extended collection’. A unified vision of aggregation could be based on the GBIF model.
- Challenges related to permissions and data gatekeeping in the context of aggregating collection data at federal agencies were raised. The process of obtaining appropriate permissions to share and aggregate data across different agency specimen projects can be complex. NIST, for example, are essentially simply gate-keepers for other agencies, with their biorepository serving as a kind of specimen aggregator from various projects.
- Cybersecurity requirements differ across and within agencies and change continuously. As a result of varying rules and regulations across agencies, certain software platforms are restricted. For example, at the USGS, there are concerns with access to cloud-based data storage and programs to share data (e.g. Docker). There is a lack of common data sharing and IT standards across agencies.
- There are challenges associated with giving broad access to federal data that are linked to certain restricted and proprietary information (e.g. USGS needs to integrate with oil and gas well information, international minerals data, etc.).

- The Department of the Interior has data standards across all bureaus for accessioning and cataloging non-living collections. There is a need for harmonization of data practices across the government. Sometimes it can even be challenging to identify the right taxonomic standard to use.
- An issue raised about data aggregation systems is that the stored data gets 'stale,' i.e. data is deposited and then never updated, which leads to misrepresentative and inaccurate or outdated data. Enabling "live updates" as opposed to "static dumps" – potentially through continuous live links between collections and data aggregators–would be a key issue that could potentially be addressed by data integration.

Path Forward, Areas to Make Progress

Recognizing the need for a centralized database that contains all species data in a standardized format, participants discussed some ongoing relevant initiatives and identified key areas where progress is needed.

- A framework needs to be established for identifying the types of data that need to be included or prioritized in building a comprehensive and integrated data network. First step could be starting with a platform for identifying institutions or information centers that may have certain parts of the biodiversity represented within their collections.
- Standardization of data practices is needed across the government for the way that collections information is stored and integrated. It would help to have a common set of standards that can be adopted across agencies to make the data more interoperable.
- Given that taxonomy changes regularly and taxonomic IDs vary from database to database, having an easy way to cross-reference one species across all the available databases would be a good first step to integration.
- Benefits of Darwin Core compliance, being used for the academic project TaxonWorks, were highlighted. Use of Darwin Core increases standardization and interoperability, but there are some trade-offs as it takes more work to meet that standard.
- USDA-ARS has a large specimen database available online and is currently working on becoming interoperable with genomics databases by being Breeding API (BrAPI) compliant.
- Challenges associated with access to data storage could be mitigated by having common IT practices around cybersecurity (e.g. safety of cloud storage platforms, secure methods of download, etc.).
- Importance of systems that allow for the aggregation of data from multiple sources while preserving original data attribution was highlighted.
- Participants emphasized the need for continued collaboration and discussion on how to effectively integrate and share data across different systems and platforms. Emphasis was placed on the importance of international cooperation and engagement with broader biological and environmental data communities.

Recommendations

- The community will need to address challenges related to legacy data representations and permissions for data access and integration.
- Continued discussions are needed on how the extended specimen network can integrate with federal data to expand utility.
- Practical solutions are needed for enabling data sharing and interoperability across projects and agencies.
- Harmonized data standards and IT/cybersecurity policies across agencies can enable better data sharing and integration.
- More and continued engagement is needed between the IWGSC and other agencies and data initiatives to facilitate better integration of federal data.
- Further work is needed to explore possible integration strategies and models and overcome data aggregation challenges.