

## Using Symbiota to establish a global, decentralized model for high-quality data aggregation: Novel concepts and designs to improve the interoperability of occurrence-based biodiversity data

**Mr. Edward Gilbert<sup>1</sup>**, Samanta Orellana<sup>1</sup>, Katelin Pearson<sup>1</sup>, Gregory Post<sup>1</sup>, Dr. Laura Rocha Prado<sup>1</sup>, Dr. Jenn Yost<sup>2</sup>, Dr. Beckett Sterner<sup>1</sup>, Dr. Nico Franz<sup>1</sup>

<sup>1</sup>Arizona State University, Tempe, United States, <sup>2</sup>California Polytechnic State University, San Luis Obispo, United States

G6: General, Lecture Theatre 1, Appleton Tower, 11 Crichton Street, EH8 9LE, June 8, 2022, 9:30 AM - 11:00 AM

The Symbiota software platform (<https://symbiota.org>) has risen to prominence as an international tool for assembling, networking, and distributing datasets associated with biological collections. The open-source software package distributed via GitHub has been used by numerous research teams to establish data portals based on specific taxonomic and geographic themes. Portals function as Content Management Systems (CMS); occurrence data are managed directly within the portal as "live datasets", though often augmented with the import of data "snapshots" originating from external systems yet otherwise aligned with the portal's community of practice and data focus. In this respect, data portals additionally serve as mini-aggregators, integrating multiple specimen datasets that collectively represent a community-based research perspective.

One could argue that Symbiota's mid-level aggregator functionality compounds the further fragmentation of occurrence data. Rather than conforming to the vision of pushing data from the source to the global aggregators and ultimately the research community, specimen data are distributed across a growing array of mini-aggregators. However, the decentralized approach has been shown to promote the emergence of multiple regionally, taxonomically, or institutionally localized, self-identifying communities of practice. Communities remain empowered to control the social and informational design and versioning of their local data infrastructures and signals. The upfront cost of decentralization is more than offset by the long-term benefit of achieving sustained expert engagement, higher-quality data products, and ultimately more societal impact for biodiversity data.

To mitigate the negative consequences of fragmented datasets across a decentralized network, the Symbiota Support Hub, a new domain of iDigBio, has implemented a number of enhancements that allow distributed portals to function as an integrated network of data aggregators. Improvements in tracking project metadata, data provenance, record annotations, and the establishment of a public Application Programming Interface (API) architecture that regulates machine-to-machine annotation propagation have enhanced interoperability by providing support for real-time transmission of occurrence annotations across the distributed network of Symbiota portals. This enables the platform to continue to be used for establishing decentralized, domain-specific knowledge communities, while also achieving the goals of the centralized paradigm in making data findable and accessible on a global taxonomic and spatial scale.

