Symbiota Integrations: Exploration of Historical and Current Methods of Data Sharing Across a Decentralized Portal Network and Goals of Extending Interoperability Globally

Edward Gilbert[‡], Beckett Sterner[‡], Mark Aaron Fisher[‡], K. Samanta Orellana[‡], Katie Pearson[‡], Gregory Post[‡], Lindsay J. Walker[‡], Logan Wilt[‡], Jenn M. Yost[§], Nico Franz[‡]

‡ Arizona State University, Tempe, United States of America
§ California Polytechnic University, San Luis Obispo, United States of America

Corresponding author: Edward Gilbert (egbot@asu.edu)

Abstract

Over the last decade, the Symbiota open-source software has been readily available to establish occurrence-based data portals that represent the taxonomic and geographic expertise of a specific community of researchers. Reasons for establishing a data portal vary, but often focus on:

1. data mobilization via the creation of public data access points (e.g., in-house search and export tools, Application Programming Interface (API) access, publication tools pushing data up to aggregators);
2. tools and workflows that support active specimen digitization projects
3. a method for staging and preparing datasets for analysis to answer specific research questions (e.g., data assessment, correction, augmentation).

The software functions as a Content Management System (CMS) allowing any dataset to be collaboratively augmented, modified, and managed online. Currently, the software provides support for over 1000 collection datasets to manage their specimen data directly within a Symbiota portal as a live managed dataset. Portals often include “snapshot” data imported from externally managed systems, which are updated on a regular schedule. Depending on the goals of a project, portals will vary in the composition of live to snapshot collections, though most contain a mixture of both. In this respect, data portals serve as intermediate aggregators, integrating multiple specimen datasets that collectively represent a community-based research perspective.

Symbiota portals typically function as mid-level data aggregators that are community driven by a group of researchers with expertise within a specific taxonomic domain. This decentralized approach has been shown to promote the emergence of multiple

regionally, taxonomically, or institutionally localized, self-identifying communities of practice. Each community is 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.

In contrast to the vision of pushing data from the source to the global aggregators and ultimately out to the research community, Symbiota records are distributed across a growing array of sub-aggregators. For instance, Arizona State University Vascular Plant Herbarium's specimen data consist of a live managed dataset within SEINet with subsets of their data pushed out to the Portal de Biodiversidad de Guatemala and the Co operative Taxonomic Resource for American Myrtaceae Symbiota portals as snapshot record sets. Not only does this support research associated with each of the portal communities, it exposes the records to researchers with local and taxonomic expertise to review, correct, and comment on the occurrence data. While the Symbiota portals provide tools for these communities to annotate the distributed snapshot records, the annotations need to be directed back to the source collection. Aside from the technical challenges, there are social negotiations that need to be considered. Collection managers might not want to integrate external edits, or the collection might be understaffed without anyone to approve the information transfer. Issues associated with “round-tripping” back to the source are complicated. Nevertheless, global coordination is feasible through automatable data sharing agreements that enable efficient propagation and translation of biodiversity data across communities.

Within this presentation, we will explore ways specimen and annotation data have been shared across the Symbiota portal network, as well as the associated technical and social challenges we have encountered. We will also present recent enhancements in tracking project metadata, data provenance, record annotations, and the establishment of a public API architecture. These developments are leveraged to regulate machine-to-machine annotation propagation to enhance interoperability by providing support for real-time transmission of occurrence annotations across the distributed network of Symbiota portals. By demonstrating methods and challenges associated with data sharing across the Symbiota portal network, we strive to contribute to the global discussion of data sharing, but more importantly, solicit input and direction from the greater community on how we can improve data sharing beyond the Symbiota network.

Keywords

collection management system, data mobilization, biodiversity data aggregator, data provenance

Presenting author

Edward Gilbert

Presented at

TDWG 2023

Conflicts of interest

The authors have declared that no competing interests exist.