Integrated Digitized Biocollections (iDigBio) Cloud-Based Cyberinfrastructure

José Fortes, Renato Figueiredo, Andréa Matsunaga, Alex Thompson, Matthew Collins, Jiangyan Xu
Advanced Computing and Information Systems Laboratory
Department of Electrical and Computer Engineering, University of Florida
PO Box 116200, Gainesville, FL 32611-6200, USA
{fortes, renato, ammatsun, athompson, mcollins, jxu}@acis.ufl.edu

Integrated Digitized Biocollections (iDigBio) has been funded as part of the NSF Advancing Digitization of Biological Collections (ADBC) program with the mission to organize all the information contained in vouchered biological and paleontological collections across the USA. iDigBio is developing a national resource that supports the vision of ADBC including building and deploying a customized cloud-based cyberinfrastructure for collections among other activities. The iDigBio cyberinfrastructure architecture is forward-looking and adaptable, taking advantage of (a) virtual appliances to ease the dissemination of existing tools, community standards and best practices, (b) cloud computing and storage services to aggregate and process data including not only textual information, but also media files from digitization projects, and (c) web technologies and standards to broaden data exchange and dissemination. Virtual appliances are ideal to package tools that facilitate the digitization workflows taking place at Thematic Collections Network (TCNs), which involve data capture, georeferencing, optical character recognition, data validation, and quality control, and most important of all, sharing of data with the national resource. The cloud computing and storage services based on distributed object and database stores are designed to be scalable and reliable in order to support the data management and information processing required by the estimated 1-billion biological specimens. iDigBio relies on standards, proven solutions and software reuse whenever possible, with programmatic and graphical user interfaces to interact with different stakeholders (data producers, data consumers, domain service providers, infrastructure providers and other aggregators).