Co-sponsored iDigBio-S2I2 Workshop

Developing Robust Object-to-Image-to-Data (DROID) Workflows

Workshop Dates: May 30th - May 31st (Attendees to arrive the afternoon/evening of May 29th)

S2I2 Leads: Chris Norris & Jim Beach

S2I2 Chair/Organizer: Amanda Neill

iDigBio Leads: Jose Fortes & Greg Riccardi

Location: Gainesville, FL

Description

The digitization of information about the distribution of life on earth extends scientific work and workflows which began with the discovery of species in the field and laboratory, systematic identification and description, and the deposition of specimen vouchers in biological and paleontological repositories. The goal of digitization is to add additional value to society's monumental 400-year investment in collecting and curating samples of earth's biological and paleontological diversity, by mobilizing the data associated with those specimens to the internet. The process of converting text and image information to digital formats should be the easy part of attaining this societal goal given the extraordinary level of standardization associated with specimen creation and curation across countries and centuries. Tasks associated with the creation and optimization of information acquisition from biological and paleontological specimens lend themselves to innovative and efficient technological approaches, and to efficient optimization with the variety of computational, network, and logistics tools and services available in both academia and through commercial and crowd-sourced services.

The DROID Workshop will address the documentation and analysis of digitization workflows for biological and paleontological specimens, with a primary focus on addressing those preservation types for which digital images can become the basis for further digitization steps. The goals of the meeting are to: (1) inform and train workshop participants in the use of lightweight business process modeling, which will then be used by participants to (2) create and document reference workflow models for represented disciplines and/or preservation types. The justification for describing digitization workflows by discipline and/or preparation type is that those classes define discrete sets of logistical and data-acquisition parameters based on the physical properties and legacy curatorial practice of each discipline/preparation type. Data from labels of snakes in jars, from minute labels of pinned insects, or from herbarium specimens each present unique challenges and opportunities for digitization.

The ultimate goal of producing reference workflow models for specimen digitization is to then employ them in:

1. The evaluation of newly-proposed workflows, as tests and extensions of the reference model. Made available on the web, these models would be usable by current and newly-created digitization efforts to evaluate their own approaches against a model at some level

of maturity for completeness, outputs, efficiency, cost, staffing, technology application, and throughput.

2. Identifying gaps within respective workflows for technology tools and services that would complete them and should be prioritized as required technology for development, funding, and ongoing technical support.
3. Identify options for individual workflow tasks for optimization in various economic and technical support scenarios. Crowdsourcing might be the only viable solution for certain tasks given insufficient funding for keystroke data entry. Partial records might be the only attainable outcome for disciplines constrained by massive specimen overloads or unwieldy source material.

These workshop activities should meet the goals of iDigBio by encouraging communication and collaboration among domain experts who will document technology needs and requirements via workflow modeling exercises and group exploration of innovative solutions to the object-image-data bottleneck. While workshop funds will not be used in a session to plan grants, or to develop grant proposals, Workshop deliverables will feed into the various software innovation programs being developed in the context of NSF's Cyberinfrastructure Framework for 21st Century Science and Engineering.

Workshop Outcomes:

1. A community-developed collections workflow site showing tasks that diverge among disciplines/preparation types and those that are in common across multiple workflows.
2. A call to action to discipline-based communities to identify resources and tools to complete missing areas of the workflows.
3. A model for group assessment of differing technological solutions applied to workflow steps.
4. A model for evaluating the economic efficiency of individual steps, their location and precedence in the workflow, and the economic (including logistic, HR, collaborative goal) efficiency of the overall workflow.
5. A possible, scalable, affordable, consensus technology solution for label capture technology applicable to multiple collection types (or a record of such a vision for future application).
6. Publication of these findings in a publication(s) in Collections Forum, PLoS ONE, the iDigBio website, and/or appropriate society/discipline journals.