Prospects for a Scientific Software Innovation Institute in Biological Collections Digitization: Interim White Paper, September 2011

Executive Summary

Biological collections voucher and document the identity and distribution of species on our planet; they enable us to reconstruct the past, understand the present, and predict the future of Earth's biological diversity and man's impacts on it. The potential now exists to capitalize on the massive historical and ongoing investment in biodiversity collection and curation through the use of research software to acquire, mobilize and publish the data associated with species voucher specimens.

Significant cyberinfrastructure (CI) development has already advanced in the biological collections community, and the research enterprise is positioned to be transformed by software investment and organization at a national level. The mobilization and engagement of the ecological and evolutionary data associated with specimen collections is a grand challenge for the 21st Century.

Through its recently announced Advancing Digitization of Biological Collections (ADBC) program, NSF has created a national coordinating Hub and Thematic Collections Networks (TCNs) to begin the process of undertaking a national digitization effort. One of the most significant challenges facing this initiative is the design and support of innovative software to streamline the process of capturing and mobilizing collections data.

Such tools need to be capable of dealing with complex workflows. They must be adaptable to collection management logistics unique to each institution. They need to be scalable systems that can support high-throughput, production-scale capture, annotation and mobilization of collections data on a national scale. Coordination is needed to pull together the heterogeneous efforts and software products generated by the collections community in order to achieve the requirements analysis, software engineering and technical training and support needed to congeal a scalable processing environment.

While new technologies for data capture and mobilization may emerge from the ADBC program, the extent of the available funding for ADBC, and the considerable demands on these funds, mean that large scale technology development is unlikely to be viable under the auspices of this program. This "technology gap" is a critical barrier to large-scale collections digitization. Realistically, many collections will not receive direct funding from ADBC to support digitization, so the availability of robust, scalable, and user-friendly technologies to support and accelerate data capture will be a critical element for incentivizing participation in the national effort.

Development of industrial scale technologies is a role that could be filled by an S2I2 in Biological Collections Digitization. The role of such an institute would be three-fold; to pick up and develop new technologies emerging from TCNs; to monitor technological developments if fields beyond the collections

community (e.g. engineering, computer science, library/information science) and adapt promising applications for use in collections digitization; and to work with the ADBC Hub to disseminate these tools to the wider collections community.

However, the S2I2 would also have a role that goes far beyond collections digitization. Just as the S2I2 would draw information from such diverse fields as software engineering, image processing, robotics, industrial engineering, and management science, the outputs of the S2I2 in terms of novel software tools may have applications in communities far beyond biocollections, such as libraries, archives, arts and humanities research, and both formal and informal education.

Scientific Software Institute in Biological Collections Digitization

Biological collections voucher and document the identity and distribution of species on our planet; they enable us to reconstruct the past, understand the present, and predict the future of Earth's biological diversity and man's impacts on it (Suarez and Tsutsui, 2004; Elith et al., 2006). Collection-holding institutions, including museums and herbaria, contain the results of over 300 years of biological and paleobiological inventory and sampling. The data associated with the specimens in these collections represent a baseline not only for species identity and occurrence, but also related ecological, climate, niche, environment and biological community information.

The potential now exists to capitalize on this massive investment in biodiversity collection and curation through the use of research software to acquire, mobilize and publish the data associated with species voucher specimens. Significant cyberinfrastructure (CI) development has already advanced in the biological collections community, and the research enterprise is positioned to be transformed by software investment and organization at a national level. The mobilization and engagement of the ecological and evolutionary data associated with specimen collections is a grand challenge for the 21st Century.

Process. This report presents the findings of two recent workshops: one held at the Field Museum of Natural History, Chicago, on March 22-24, 2011, which explored the potential for an S2I2 in collections digitization, and the other held March 4-5 at the Sam Noble Museum of Natural History, Norman, Oklahoma, under the auspices of the CollectionsWeb RCN, which looked at existing technologies and workflows for collections digitization. Programs and lists of participants for both meetings are attached as an appendix to this report; emphasis was placed on including participants from fields other than those traditionally associated with biological collections, including information science, software design, computer science, mechanical engineering, robotics, image recognition, cloud computing and workflow engineering.

Background. A series of NSF-supported meetings in 2010 led to the development of a comprehensive strategic plan for the digitization of the nation's biological research collections[1]. The prime objective of this strategy was to create a publicly available, sustainable and comprehensive national collections information resource[2]. The strategic plan conceived this as a unified effort involving federal funding for data acquisition and the development of cyberinfrastructure to promote efficient and standardized capture and mobilization. In August 2011, as a first stage towards implementation of the plan, NSF announced a

[1] Biological Collections Digitization Focus Group: NESCent, Durham NC, 5-7 February 2010; United States Virtual Herbarium Workshop: Missouri Botanical Garden, St. Louis, February 23-25, 2010; Research Coordination Network Data Integration Workshop: Tulane University, New Orleans, 25-27 March 2010 (www.collectionsweb.org); Biological Collections Digitization Workshop: NESCent, Durham NC, 27-29 April 2010

[2] http://tinyurl.com/2v379hl

new Advancing Digitization of Biological Collections (ADBC) program. This program envisaged support of a national coordinating Hub for collections digitization and a series of Thematic Collections Networks (TCNs) that would digitize collections across multiple institutions to address specific research questions. The first three TCNs were announced in July 2011, along with the award of the ADBC Hub to the University of Florida.

One of the most significant challenges facing the national digitization effort is the design and support of innovative software to streamline the process of capturing and mobilizing collections data. Such tools need to be capable of dealing with complex workflows, which can vary significantly across different taxonomic types and sizes of collections and which need to adapt to collection management logistics unique to each institution. The biocollections community has several interacting but autonomous software development groups focused on different parts of the challenge, but they have not been adequately coordinated or capitalized to produce the sort of scalable systems needed for the high-throughput, production-scale capture, annotation and mobilization of collections data on a national scale. Coordination is needed to pull these heterogeneous efforts and software products together to take on the requirements analysis, software engineering and technical training and support needed to congeal a scalable processing environment.

Conclusions from the Workshops. Based on the discussions held at these workshops, participants identified specific software needs and ways in which a Scientific Software Institute could facilitate the national digitization effort. Both workshops demonstrated the advantages of interdisciplinary cross-talk involving information scientists and systems engineers for projects like this. At the time of the workshops, the structure and location of the organizing center for the national digitization effort was not yet known, but based upon the program guidelines, we anticipated that the digitization HUB would focus on coordinating specimen digitization efforts across the country. Although the coordinating process would certainly allow the HUB to identify technological challenges to digitization, it is unlikely that the HUB will have sufficient resources to implement these new software solutions. Thus, we concluded that a Scientific Software Institute would be a valuable partner in the digitization effort, working closely with, but independent from the HUB.

The following is a summary of the discussions that took place in the breakout sessions at the workshops, which explored specimen digitization as a scientific, sociological and technological activity.

The Challenges of Collections Digitization

Collections digitization is the process by which information about specimens held in museums or other collections (e.g. identifications, descriptions, geographical locality information, collecting information, images, etc.) is converted from analog (e.g. ledger books, catalog cards, specimen labels) to digital form and made available to users via the web. Much of the value of natural history specimens lies with their

associated data – “unlocking” these data has the potential for massive transformative effects across a wide range of academic disciplines, as recognized in a series of recent reports[3][4][5].

Historically, the U.S. biocollections community has not been successful at digitizing collections. Less than 10% of the more than 1 billion specimens held in U.S. collections are available on-line. The bulk of the specimens that have been digitized are those that are relatively standard in size and shape (i.e., herbarium specimens) whereas the challenges of creating an efficient digitization workflow for other types of specimens (e.g., fluid collections, large animal bones) have not yet been successfully met. Digitization has to compete with other collections activities (e.g. loans, visitor support) for funds and staff time. As a result, despite more than 20 years of collections digitization, the rate of data capture has not substantially increased; at the current rate, digitization of the entire national holdings of natural history collections is a very distant goal.

Not only is there a massive backlog of specimens in need of digitization, but collections continue to grow at a rate of 5-15% per annum, and for some collections as high as 25%. New techniques such as CT scanning and DNA sequencing have increased the scope of collections data, as has the demand for accurate georeferencing of locality information. Focusing on new specimens and new data alone is not an effective long-term strategy, because it does not capitalize on the considerable value of the historical data present in museum collections.

The challenges of digitization are not limited to data capture. Converting data from analog to digital format is a critical first step, but for the value of these data to be fully realized there needs to be a stable infrastructure for the storage and mobilization of information and tools to aggregate and query data that reflect user needs. Projects like BiSciCol[6] are working on the semantic tools and linguistic infrastructure needed for resource discovery, but there will be a major need for front-end software that builds off these tools and utilizes them for the benefit of user communities.

Despite the creation of the ADBC program and a growing emphasis on funding digitization activities as part of NSF’s Collections in Support of Biological Research program, the U.S. collections community still lacks by several orders of magnitude the resources necessary for a national digitization effort, especially if the aim is to do this in the time frame necessary to address urgent environmental issues. If large-scale mobilization of collections is to be achieved then the cost of digitization will need to be reduced by orders

[3] http://www.whitehouse.gov/sites/default/files/sci-collections-report-2009-rev2.pdf

[4] http://digbiocol.files.wordpress.com/2010/06/digistratplanfinalv1.pdf

[5] http://www.nsf.gov/pubs/2009/nsf09044/nsf09044.pdf

[6] http://biscicol.blogspot.com/p/home.html

of magnitude. For instance, there have been substantial improvements in some technologies that would facilitate large-scale digitization (e.g. high resolution cameras), but the collections community lacks the interfacing software and funding to implement them. Other technologies, like industrial robotics, are still too expensive; this is not so much an issue of hardware costs, but because they require trained support staff on-site to deal with issues and their relatively inflexible programming does not work well with hyper-variable museum specimens.

While the attention of ADBC is rightly focused on tackling the massive task of capturing information from the 1 billion-plus specimens in U.S. collections, there is also the issue of exactly how this data will be mobilized and made accessible to users. This is another area where there is a compelling need for new technologies in order to maximize the user base for collections data and to promote the usage of these data. It is not sufficient to assume that if you build it, they will come. Tools are needed to facilitate and manage the interface between user and dataset. These tools will need to reflect the demands of the collections userset – people with limited time and heavy pre-existing workloads.

Technology and innovation are not barriers to collections digitization; the biggest barriers are labor and logistics. Technology, in the form of well-constructed, well-targeted software, has the potential to overcome these barriers, but to do so will require scalable solutions applied in a large-scale, coordinated fashion. Historically, the collections community has not been able to achieve this sort of coordinated response.

What innovations in software engineering and software support are needed to digitize and mobilize the massive backlog of data associated with specimens held in the nation's biological collection institutions?

Community Organization and Systems Thinking

Computerizing the data from all biological specimens in U.S. collections will require sustained intellectual and financial investment in order to realize the full scientific potential of the information they contain. A campaign to mobilize specimen data into online systems will be built upon a foundation of research practice over 300 years in the making, which has developed standard field and laboratory methods and emerged as a global, collaborative enterprise.

Thousands of geographically-distributed museum and herbarium collections are physically and administratively discrete but independent only in the sense that their specimen holdings emphasize the geographic and taxonomic interests of their past and present researchers. Myriad points of governmental, university, and private investment in biodiversity inventory, species description and curation have resulted in thousands of centers of biodiversity research and training interacting in a self-healing, virtual network.

Orphaned (decommissioned) collections are absorbed by more active centers, and protocols for specimen exchange, loans and gifts ensure that researchers around the world have access to specimens upon request. Although specimens in the U.S. are partitioned among thousands of museums and herbaria, scientists everywhere implicitly understand that all collections belong to the same collective resource—that together they comprise a single, partially-replicated sample of the plant and animal species of one planet.

Generating successive waves of disruptive, digitization-driven innovation will be done within the matrix of the distributed yet unified biocollections research enterprise. The transformation of the digitization process from one-collection-at-a-time, to the component optimization of a national, multi-institutional computerization campaign will advance if it reflects the collaborative and collective values of the biodiversity collections community.

When analyzed at a global level, collections digitization will be guided by priorities and efficiencies that only become apparent when considering the computerization of all specimens as the ultimate goal. Assessing the taxonomic and geographic coverage as well as the extent of redundancy in holdings within and among biological repositories on a national scale will suggest new ways to partition and optimize computerization strategies and deploy efficient staging and scaling of digitization efforts. Only in a global (or at least U.S.) systems context, will transformative new hyper-efficient methods be identified.

In molecular biology highly-innovative and disruptive “shotgun cloning” transformed full genome sequencing of humans and “next-generation sequencing” techniques are now completely revolutionizing molecular genetics and related fields a second time, all within a few years. Do technology analogs for transforming the process of acquiring data from biological specimens exist? At a national scale, can collection resources be partitioned or selectively sampled and then computationally reassembled? Do all specimens need to be computerized or can we model the effective contribution computerizing additional specimens of a particular taxon or locality will make—and then set digitization goals for efficiency of return on investment? Could the digitization of certain institutional collections be a large enough sample to eliminate unneeded redundant data from other collections? Could collections be sub-sampled taxonomically across institutions increase digitization efficiencies by one or more orders of magnitude? These kinds of optimization strategies and the software technologies and protocols that would be needed to support will only emerge from a systems level analysis of collections holding and science requirements done on a national scale.

Architecture

A modular internet-based software architecture designed for extensibility and change must be developed based on enterprise-scale workflows and optimizations which integrate curatorial, computational, social, and sustainability models. Web and grid services-based workflows will need to connect participants and processes in ways that reflect the distributed nature of data sources and which interconnect projects as

nodes in a systems-level design. These community workflows with highly usable thick and thin clients will require smart systems integration in order to maximize overall data acquisition rates.

The types of innovation that might be expected to emerge from an S2I2 center would include novel specimen digitization technologies and workflow designs; web-services based integration at a national and international level, new server and storage deployment options using cloud technologies; innovative 'business models' for providing incentives and sustainability options; refinement of current coarse-grained specimen metadata standards (Darwin Core), new methods for data quality enhancement and access.

Software Tools

Scalability and robustness are critical challenges to developing tools for efficient digitization of the wide range of biological collections held in U.S. repositories. Development must be based on gathering of baseline data on scale requirements, including the amount of information, number of people involved, and number of systems; this is an area where a future S2I2 center could work in collaboration with the collections and research communities. It will also be necessary to look at the economics of developing open source versus proprietary applications, and to root these discussions in sound business models and realistic cost and sustainability projections.

Software innovation is a key activity for an S2I2 center and it must be grounded on a robust understanding of local skill levels and constraints of digitization projects and researchers. An initial requirements analysis process could initiate the documentation of existing best practices, identification of bottlenecks, industrial logistics engineering analysis, and optimization.

Another opportunity for a S2I2 software development center that would serve the national digitization effort is producing software that integrates and makes interoperable independently developed, but complementary software solutions to the same digitization challenge. There are a number of instances in which software tools addressing the same digitization need has been developed in isolation. Because of the independent development, the different software solutions often offer unique features or approaches to the same problem. The Oklahoma and Chicago workshops featured demonstrations of some of these complementary software tools. For example, the Apiary Project[7] and SALIX[8] take slightly different but highly complementary approaches to the common challenge of imaging plant specimens and capturing label data from herbarium sheets. Software designed to make these tools interoperable would allow users to take advantage of the unique features of both of these solutions. Similarly, software could be developed for integrating competing, but complementary, approaches to digitizing glass slide collections or drawers of insect specimens. An interoperable georeferencing platform, uniting the unique capabilities

[7] www.apiaryproject.org

[8] http://daryllafferty.com/salix/

of GEOLocate and the Biogeomancer workbench was a proposed deliverable of one of the unsuccessful CU-led HUB project. Across-platform software collaboration could even be extended to integrating software designed to meet very different digitization challenges (e.g., integrating solutions for digitizing herbarium specimens with those developed for digitizing insects or glass slides). This "handshaking" approach to software development – integrating existing, complementary software solutions - addresses the overall digitization challenge, eliminates the need for developing entirely new software solutions, and allows competing software systems to coevolve and coexist.

Human Resources

No process of software innovation can succeed without strong stakeholder (researcher) engagement in the process. Digitization has the potential to harness the energy and expertise of a wide range of participants, including researchers, domain experts, technologists, and citizen scientists. There is the potential for software that facilitates crowd-sourcing of work, leveraging massive human resources to attack large-scale digitization challenges. Weighed against this are the human resource issues, including impact on current staff, professional development, and policy and organizational issues. Developing efficient workflows begins as a sociological process that drives technology development. It requires a social setting; contributors who have the needed experience and feel empowered to share their knowledge.

What incentives do collections institutions and scientists require to participate in the coordinated national scale digitization effort?

There are a few existing consortia that can serve as a model for this larger scale effort. VertNet is an example from the museum community; in the botanical community there are regional consortia such as the California Consortium of Herbaria, the Pacific Northwest Herbarium network and the Southwestern Herbarium Consortium. These efforts have leveraged the technological and organizational skills of a few institutions to create a community whose members see specific advantages to their own collections and to the scientific community as well. The existence of successful smaller-scale consortia may also be a disincentive, in that the community is used to the idea of home-grown, small scale projects and may be daunted by the large-scale, industrial models that will need to be applied to achieve large-scale digitization (e.g. Google). There is a fear of a loss of control at the level of the individual institution/collection.

The biggest incentive for most curators and collection managers is the potential for digitization to make better use of scarce time, by streamlining the process of dealing with basic collections queries and duties, allowing staff members to spend more time utilizing their specialist expertise and knowledge. Services

provided by an S2I2 could accelerate the rate of acquisition and assimilation of new collections and their associated data by providing institutions with readymade solutions to issues of data management and ingestion. Ultimately this would lead to better curated collections and better service to the research community. It's also important to recognize that curators are both resource managers for digitization and collection users. As a result digitization technologies that benefit data users can also incentivize data providers.

Creating a unified mechanism for querying all collection data in the country that are participating will is compelling because of the breadth of scientific questions that could be addressed using such a resource. Scientific discoveries based on the use of such a resource will bring prestige for smaller collections that participate. Data usage tracking will allow administrators to make effective arguments for resources. Such metrics are a very important incentive, but the need to be able to record not only institution-specific queries, but also the follow through; how many queries lead to loans/visits/publications. There is a potential role for new technologies to mediate this process. Finally a large, cloud-based dataset will have emergent properties of the sort seen with GBIF, creating further incentives for participation.

A major incentive would be the development of a mechanism by which museums could receive funding support to help with digitizing what they have or accepting new material because they're sharing data/metadata. That kind of compensation mechanism does not currently exist.

Software tools applied to the dataset will provide new metrics for managers of collections data that will help to strengthen the argument for further resources. Examples of these include the number of new species discovered and documented after creation of the online resource (compared to prior to its creation); the decreased cost of research per new species published; the increase in the number of collaborative research efforts; the downstream use of digital resources cited in publications; and the number of annotations provided on data aggregation sites using S2I2 tools.

The application of social networking technologies to collections digitization would have a number of benefits that might incentivize participants. These include the ability to harness collective wisdom in order to tackle problems such as specimen identification[9] [10] and georeferencing of localities[11] at a global level. This also allows for improved public exposure of, and even participation in, the research process as citizen scientists, providing valuable STEM education opportunities which would be a further incentive for many institutions and funding bodies.

[9] http://www.galaxyzoo.org/

[10] http://nmnh.typepad.com/100years/2011/03/crowdsourcing-via-social-media-allows-rapid-remote-taxonomic-identification-.html

[11] http://www.museum.tulane.edu/geolocate/

Modular workflows have potential advantages in a number of areas. They allow better definition of entry points to the digitization process, which is a significant incentive to digitize - difficulties in knowing where to start can act as a significant barrier to beginning a digitization project. Better workflow documentation and technology integration also allows more efficient management of grant-funded projects; pre-defined workflows and tools minimize the time needed for project planning and also allow for more compelling grant applications, by producing more accurate timelines and achievable deliverables. Finally, there is the issue of responding quickly and efficiently to crisis situations - for example, a request for regional marine life data in response to an oil spill - where this would involve mobilizing data that was not yet digitized. Knowing how to assemble an appropriate workflow and the amount of time/money needed would allow institutions to obtain sufficient funding and get started quicker.

If the S2I2 were to act as a software repository, researchers writing grants would be able to tap this as a resource, reducing the need to spend limited grant funds on developing specialized applications to answer specific research questions. However, it would also be necessary to address the question of sustainability. Users are reluctant to commit to new technology, only to have the project evaporate and never produce anything substantial. By ensuring sustainable services and creating useful data products, an S2I2 could provide cautious uses with a huge incentive to commit to the digitization effort.

What research, education and outreach impacts should be targeted by a software innovation institute for biological collections?

Specimens are records of our human-natural history but have received less and less study in recent years. Digitization will allow these specimens and the story they tell to return to its former glory as a research tool and provide a resource for citizen science to add value to actual research.

Tools generated by the S2I2 would allow researchers to use the aggregated information set represented by collections to define the unknown, by discovering taxonomic/geographic gaps in our knowledge base. By measuring the rate of specimens being used it will be possible to understand biodiversity for assessment and monitoring Software tools developed by the S2I2 will allow for improved access to images and data for type specimens and access to species not recognized as new. This new citable publication-quality imagery available online will speed discovery and publication.

An S2I2 would positively and reciprocally impact both bioscience and the computer/ library-info/visualization science fields by increasing crosscutting research activities and new discoveries. Collaboration with the LIS community, particularly digital archivists, could facilitate adaptation of pre-existing digital preservation practices and strategies for use with biocollections, while tools developed by

the S2I2 would allow synchronization of communal knowledge, enabling prioritization and triage for digitization across a wide range of collections.

Digitization of biocollections and their auxiliary data (field notes, particularly) could provide digital humanists and historians of science with completely new research tools and methods. The field of Digital Humanities is growing every year; working with these scholars could create innovative new research into the history of science.

In the field of Education, software tools would enable the creation of online biology courses based on access to digitized collections, with the ability to access infinitely larger study sets for labs and lectures. Biodiversity informatics (eco-informatics) degree programs would be made possible by the entire enterprise of building this research environment. Innovative access would not be limited to courses in the biological sciences; with a flexible set of online tools for course construction, collections data and images could be applied to such diverse fields as computer/info/library science classes, history, and art.

Social networking tools have the potential to involving physically distant groups of undergraduates in the digitization process. This will not only allow for the injection of organismal biology into curricula, but will also allow students to be educated, through participation, in the true nature of biology, including the uncertainties (e.g. species boundaries/identities). By helping to dispel the illusion of the infallible, inscrutable scientist this will make science more human and accessible to students. At the same time, students would be educated in information science and gain a better understanding of the importance of good data management – a critical skill for 21st Century science.

New interactive tools can increase student matriculation, research, and publication in fields benefited by biocollection mobilization-- especially organismal biology-- because data gathering activities and synthesis will become relatively effortless and fun. Students will have the tools to connect to the idea of individual organisms, populations, species, and biogeography. In doing so, an S2I2 would be helping to create a workforce for a knowledge-enabled century.

Beyond the classroom, there is the potential to develop software tools that will impact a much wider cross-section of the general public. For example, online data can be used to create on-demand field guides based on current location of user, possibly in the form of smartphone apps, and to provide metrics for the number of new mobile device apps that use S2I2 products to justify funding. Such user friendly applications will allow us to bring the public, from enthusiast communities through to anyone with a general interest, into direct contact with biocollections. The tools developed by an S2I2 would facilitate delivery of bioscience knowledge to more citizens resulting in greater nature awareness. It would also provide access to the factual evidence needed by the Government to make decisions on issues of vital importance to human health, environmental protection, and national security.

The Potential for an S2I2 in Biological Collections Digitization

The ADBC program has defined roles for networks to capture data (TCNs) and a coordinating Hub that will work with the TCNs to mobilize these data. What is not clear from the ADBC solicitation is the extent to which either the Hub or the TCNs will be developing technologies for data capture, or for mobilization. Given the extent of the available funding for ADBC, and the considerable demands on these funds, large scale technology development under the auspices of this program seems unlikely.

This “technology gap” is a critical barrier to large-scale collections digitization. Realistically, many collections will not receive direct funding from ADBC to support digitization, so the availability of robust, scalable, and user-friendly technologies to support and accelerate data capture will be a critical element for incentivizing participation in the national effort.

Development of industrial scale technologies is a role that could be filled by an S2I2 in Biological Collections Digitization. The role of such an institute would be three-fold; to pick up and develop new technologies emerging from TCNs; to monitor technological developments if fields beyond the collections community (e.g. engineering, computer science, library/information science) and adapt promising applications for use in collections digitization; and to work with the ADBC Hub to disseminate these tools to the wider collections community.

However, an S2I2 could also have a role that goes far beyond collections digitization. Just as the S2I2 would draw information from such diverse fields as software engineering, image processing, robotics, industrial engineering, and management science, the outputs of the S2I2 in terms of novel software tools may have applications in communities far beyond biocollections, such as libraries, archives, arts and humanities research, and education.

Appendix: Workshop Participants & Programs

CollectionsWeb Workshop IV: The Digitization Challenge

4-5 March 2011

Sam Noble Museum of Natural History, Norman, Oklahoma

Anne Barber, Arizona State University
Hank Bart, Tulane University
Jim Beach, University of Kansas
Janet Braun, Sam Noble Museum
Joe Cook, University of New Mexico
Wayne Elisens, University of Oklahoma
Andrew H. Fagg, University of Oklahoma
Colin Favret, AphidNet
Robert Gropp, AIBS, Washington DC
Robert Guralnick, University of Colorado
Dean F. Hougen, University of Oklahoma
Michael Mares, University of Oklahoma
Andrea Matsunaga, University of Florida
David P. Miller, University of Oklahoma
Paul Morris, Harvard University
Amanda Neill, Botanical Research Institute of Texas
Larry Page, University of Florida
Alan Prather, Michigan State University
Sridhar Radhakrishnan, University of Oklahoma
Nelson Rios, Tulane University
Gareth Russell, New Jersey Institute of Technology
Katja Seltmann, North Carolina State University
Steve Westrop, University of Oklahoma
Jim Woolley, Texas A & M University

S212 Workshop in Biological Collections Digitization

22-24 March 2011

Field Museum of Natural History, Chicago, Illinois

Ben Anhalt, University of Kansas
Brian Anthony, MIT
Hank Bart, Tulane University
Jim Beach, University of Kansas
Jason Best, Botanical Research Institute of Texas
Ann Chervenak, University of Southern California
Donald Eisenstein, University of Chicago
Linda Ford, Harvard University
Robert Guralnick, University of Colorado
Seth Kaufman, Whirl-i-Gig
Mark Leggott, University of Prince Edward Island
Bertram Ludascher, University of California, Davis
Bill Moen, University of North Texas
Amanda Neill, Botanical Research Institute of Texas
Chris Norris, Yale University
Dean Pentcheff, Natural History Museum of Los Angeles County
William Piel, Yale University
Marc Pignal, Musée Nationale d'Histoire Naturelle, Paris
Alan Prather, Michigan State University
Patrick Sweeney, Yale University
Barbara Thiers, New York Botanical Gardens
Andrea Thomer, University of Illinois
Paul Tinerella, University of Minnesota
Louis Zachos, University of Mississippi

RCN Workshop

Sam Noble Museum, Norman, OK

4-5 March 2011

The Digitization Challenge: Biology and Engineering

Thursday, 3 march

Arrive in Norman

Friday, 4 March

07:35 Vans will leave Embassy Suites for the Sam Noble Museum

08:15 Welcome, Michael A. Mares

08:30 Previous Workshops, Hank Bart

09:00 Goals for this Workshop, Alan Prather

09:30 Steps Required to Digitize Museum Specimens: Preservation Factors, Janet Braun

10:00 Break

10:30 Tour of Museum Collections

12:00 Lunch (Classroom)

01:30 Strategies for Digitizing Infradispersed Collections, Michael Mares

02:00 Intelligent Systems Applications, Dean Hougen

02:20 Industrial Engineering Applications, Shivakumar Raman

02:40 The Challenging Taxa , Jim Woolley

03:00 Break

03:30 The Scope of the Mission and Identifying Bottlenecks, Rob Guralnick

04:00 Discussion

04:30 Gala Opening Reception for Black Mesa Exhibit, Natural Wonders Gallery

06:30 Vans depart Sam Noble Museum for Campus Corner restaurants

08:30 Vans depart Campus Corner for the Embassy Suites

Saturday, 5 March

08:00 Vans will leave Embassy Suites for the Sam Noble Museum

08:30 Convene and Discussion

09:00 Breakout sessions: Biologists plus non-biologists: Goals are to draft proposals for solving critical issues/bottlenecks with current or on-the-horizon technologies

10:00 Break

10:30 Reports from Breakouts

11:00 Group Discussion of Integrating Ideas and Technologies

12:00 Lunch

01:30 Reconvene. The afternoon session will be reserved for short presentations from participating groups and integration of ideas

05:00 Vans depart Sam Nobel Museum for Embassy Suites

Dinner on your own or small groups. The Embassy Suites shuttle will take you to nearby hotels on the hour or you may dine in the hotel restaurant. Please contact the desk to let them know you would like to use the hotel shuttle and to make arrangements for return service.

Scientific Software Innovation Institutes (S2I2) Workshop

Biological Collections Digitization

March 22-24, 2011

Biodiversity Synthesis Center

Field Museum of Natural History, Chicago

Tuesday, March 22

2:00pm Introductions, background, workshop goals

1. What innovations in software engineering and software support are needed?

2. What incentives do collections institutions and scientists require to participate?

3. What research, education and outreach impacts should be targeted?

2:30pm Challenges of digitizing biological collections

Overview of collections, collections data

Workflow issues

Existing technologies

3:30pm Coffee break

3:45pm Summary of Oklahoma RCN Meeting

4:05pm Case studies (2 X 20 minutes)

- Marc Pignal, Muséum national d'Histoire naturelle, Paris
- Jim Beach & Ben Anhalt, University of Kansas

4:45pm Plenary discussion

5:45pm Break for day

Wednesday, March 23

8:00am Continental breakfast

8:30am Introduction: New Perspectives

- Brian Anthony, MIT
- Mark Leggott, University of Prince Edward Island
- Dean Pentcheff, Natural History Museum of Los Angeles County

& Ann Chervenak, USC Information Sciences Institute

9:30am Break out session

- Group 1: Innovations
- Group 2: Incentives
- Group 3: Impacts

10:45am Coffee

11:00am Plenary discussion

12:00pm Lunch

1:15pm Analysis and Synthesis: workflows, software, and community coordination

- Louis Zachos, University of Mississippi
- Bertram Ludäscher, UC Davis
- Jason Best, BRIT
- Hank Bart, Tulane University

2:15pm Discussion session

3:00pm Automate 2011 tradeshow visit

Travel by taxi from outside the Museum

5:00pm Break for day

7:00pm Group dinner

Details to follow

Thursday, March 24

8:00am Continental breakfast (?)

8:30am Plenary on technology and tradeshows

9:30am How can we optimize the rate of digital capture?

- software for digitization?
- software for workflow optimization?
- other options – crowd sourcing, etc?
- identify software gaps, data pipelines
- delivery mechanisms

9:45am Breakout session

- Group 1
- Group 2
- Group 3

10:45am Coffee

11:00am Plenary

12:00pm Ongoing discussion over lunch – roadmap and next steps

1:00pm Meeting ends