Data Registry Workshop Report

Background

A Joint Working Group on Data Sharing and Archiving (JWG), representing major professional societies that publish ecology, evolution, and organismal biology journals, was formed at a September, 2004, NSF-sponsored workshop on data sharing and archiving, hosted by ESA. Attendees adopted a consensus statement that "Our vision as members of the scientific community is to promote the advancement of science through the process of documenting, archiving, and making available the research information and supporting data of published studies." In support of that vision, the JWG made the following recommendations:

- facilitate continuing communication among professional societies on data sharing and archiving issues via a dedicated web site and periodic e-mails;
- widen participation in these activities by professional societies and international organizations; and
- support three workshops to (1) develop a strategy for creating data registries that describe data sets and provide information on how to access them, (2) identify, and develop means to reduce or eliminate, cultural and other barriers to data sharing, and (3) develop a set of requirements and recommendations for data centers in ecology, evolution, and organismal biology.

The first of these three workshops, "Data Registries for Ecology, Evolution, and Organismal Biology," was held July 11-12, 2006, in Washington, DC. Twenty-five participants representing sixteen professional societies and nine other organizations assembled to work toward three goals:

- Identify a set of common needs for, and desirable features of, data registries for ecology, evolutionary biology, and organismal biology, based on an understanding of existing resources.
- Develop recommendations, as appropriate, for shared or independent data registries for the disciplines and societies represented.
- Develop preliminary plans for implementing those recommendations.

Informational presentations

Following introductions and introductory remarks by Cliff Duke (ESA) and Sam Scheiner (NSF), Duke described the background for the workshop, and Matt Jones (NCEAS) presented an introduction to the general topic of data sharing and the relationships between data registries and data centers (repositories). Following this introduction, Jones, Jim Reichman (NCEAS), and David Baldwin (ESA) described the development of the ESA data registry as a case study for consideration by societies that may be considering launching their own registry. This presentation covered the origins of the ESA registry, discussions with the Governing Board when the registry was presented for approval, and its implementation to date.

Ease of use is a key element in getting researchers to register data sets. Laura Downey, Senior Usability Engineer for the National Science Foundation's Long Term Ecological Research Network Office, explained general concepts of usability and design of data registries, and summarized a set of usability goals and guidelines for participants to consider.

In order to provide a perspective from outside the U.S., a panel of attendees from Canada, Europe, and South America presented remarks about the interests and experiences of their organizations in data sharing and registry development. Bruce Dancik (NRC-Canada), Lindsay Haddon (BES), Allen Moore (ESEB), Hannu Saarenmaa (GBIF), and Adriana Abril (Ecological Society of Argentina) all participated.

Saarenmaa described GBIF's vision for global integration of biodiversity data, with a distributed information infrastructure that supports sharing of this data worldwide. He summarized the general operations of GBIF and the processes for contributing data to the facility and accessing the data.

Abril described Argentina's contributions to GBIF, the limitations to data sharing in Argentina and the rest of South America, and the need for a data registry in this region of the world.

Haddon stated that there is general support in the UK for data sharing, but little agreement as to how to implement or pay for it. The British Ecological Society has supplemental material posted on-line, similar to ESA's Ecological Archives, but there is a need for independent access to data, without placing it behind a subscription wall. The BES Board does not view data access as limiting research, but support the policy statement that came from the Society Summit.

Dancik noted that NRC Canada requires submission of gene sequences to GenBank, and the Canada Institute for Scientific and Technical Information (CISTI) has maintained a depository of unpublished data since 1964. Authors can deposit a wide range of material that supplements their publications in NRC journals.

Moore described the idea of data repositories as the norm in the UK for disciplines other than ecology and evolution, and the UK Research Council requires grantees to deposit data if a suitable repository exists. There is no expectation in Europe that societies themselves would initiate or host data sharing; rather, societies will facilitate data sharing by requiring it.

The last formal presentation, on data registries for evolutionary biology, was provided by Kathleen Smith and Hilmar Lapp, both of NEScent, and Don Waller, representing SSE. Smith noted that NEScent's mission is to facilitate broad syntheses in evolutionary biology, and they are committed to taking a leadership role in the establishment of cyberinfrastructure, including data registries, for the

field. Waller summarized a number of important issues in developing data registries for evolution, particularly the very heterogeneous nature of the data and the complexity of datasets, and concerns about intellectual ownership of data. Other issues that Waller raised included the question of distributed data centers vs. central ones; how much and when to require data sharing; data storage and maintenance concerns; and standardization of metadata.

Lapp described the work of the Metadata Research Center at UNC-Chapel Hill and discussed some of the technical issues involved in developing metadata standards and formats. He also pointed out the irony (or tradeoff) between making things easier for the user of data vs. the contributor of data – providing the kind of detail that makes things easy for users is difficult for contributors. Reducing the detail makes things easier for contributors but reduces the value of the data to the user.

Issues

Following the formal presentations, participants examined the preliminary list of issues presented in the agenda in the context of the information provided by presenters, to determine how to proceed toward consensus statements and recommendations at the conclusion of the workshop. The preliminary issues proposed were:

- What are the community's needs for data registries?
- What data registries already exist that may serve these needs?
- What opportunities are there to take advantage of existing registries?
- What are the advantages and disadvantages of sharing registries among societies or disciplines vs. developing society- or discipline-specific registries?
- What quality assurance procedures are needed for contributions to registries?
- Is it possible to develop a minimum specification for metadata requirements for data registries for ecology, evolutionary biology, and organismal biology, or at least for these disciplines considered separately?
- What are the usability and design requirements that should be considered to make registries as easy to use as possible?
- Can the societies agree on a minimum set of design and metadata requirements for data registries, so that contributors do not face multiple, possibly conflicting, sets of requirements in different registries?

Participants added to these the development of screening and review processes for contributions to data registries and the potential development of a common template for data registries that could be used for multiple registries. Participants also proposed a wide range of specific issues for consideration in two parallel breakout sessions planned for the second day of the workshop, broadly described as collaboration among the societies and technical implementation, respectively.

Participants also discussed potential development of a survey to distribute to their various memberships regarding data registries and sharing generally.

Breakout discussions
Participants in the "collaboration" breakout session, facilitated by Reichman, discussed a number of issues involved in determining whether collaboration among societies in developing registries would be useful or possible. Examples of issues discussed included:

- access to registries – should it be restricted to society members as a benefit, or open?
- shared vs. individual registries – what are the advantages and disadvantages to societies and to their members of shared registries?
- requirements for data registration – should authors be required to register data used in society publications?
- what are the advantages to having a common look and number of required fields for multiple data registries?

To address these concerns, Reichman proposed developing a consortium among a number of societies, which would request joint funding for one or more data registries to be housed at NCEAS or NEScent. The breakout group agreed to present this idea to the other attendees during the final plenary session.

Participants in the "implementation" session, facilitated by Downey, focused on the "how to" questions in launching data registries. These questions included:

- who hosts the registry?
- what are the technical issues and costs associated with either sharing a registry among societies or developing individual registries?
- how do we manage long-term maintenance and growth?
- should registries include both data tied to a publication and data not connected to a publication?
- should the society (or other host of a registry) have a help desk for registry users?
- what software should be used for developing the registry?

- how should one handle the tradeoff between ease of entering metadata (i.e. requiring more fields can discourage people from registering their data) vs. the ability to search for registered data (which is improved with more metadata/info for each entry)?

- can a registry itself be valuable or is it only valuable if it will eventually lead to the development of a data repository/center?

Participants distilled these questions down into guidance about what societies should consider in developing a data registry and created a list of desirable features that any data registry should have. These were presented to the other attendees in the final plenary as described below.

Outcomes

During a final plenary session, participants heard summaries of the two breakout sessions and discussed next steps for data registry development. With respect to establishing collaboration among societies on data registries, participants agreed to report back to their respective governing boards the proposal to establish a consortium of societies to develop joint data registries and to determine what societies would be interested in participating.

NCEAS and NEScent offered to lead the development of a consortium of 3 – 6 societies and to develop a joint NSF proposal to host a registry for the consortium. The consortium would review registry approaches; synonymize fields and definitions for each field; determine the need for support for a moderator or other staff, and support for professionals, such as a usability engineer, to work on design; explore the possible involvement of BioOne or another publisher; and would determine how to meet international needs, e.g. with translation of fields and/or information entered into fields.

It was suggested that registries – whether developed as part of the consortium or separately – would be seeds for data centers and for enhanced data access.

Participants suggested that we should gather data on the advantages of registries and repositories, and that a survey of societies' membership would be useful. Waller offered to develop a draft set of core questions that could be included on surveys by individual societies of their members. Because different societies are on different schedules with respect to governing board meetings, and due to concern over the ability of participants to approve a survey, it was determined that a well-designed survey would be prepared by September, 2007 and could ultimately either be distributed by the societies themselves or distributed indirectly, e.g. through "surveymonkey." We will need to determine whether to include questions only about registries or also include questions about repositories/data centers.

Participants also suggested that editorials in our respective journals would be a useful way to bring the idea of data sharing to our members.

An initial step to encourage the use of data registries, proposed by Baldwin, is to add the following to the editorial policies of society journals:

> The editors and publisher of this journal expect authors to make their data available and to deposit metadata in an approved registry.

Participants agreed to report this recommendation to their respective governing boards for consideration.

Although not requiring formal action by the societies, participants generally agreed on the list of desirable features and approaches for data registries that was developed by the implementation group. These are:

- Use of the features and metadata fields of the current ESA data registry as a good starting point
- Consider developing an adaptive user interface and adaptive search interface (For example, such that filling in top-level metadata fields like “laboratory vs. field” study or “modeling vs. review vs. original research” would modify the remainder of the form such that, for example, field researchers do not have to fill in fields that are specific to lab studies, and vice-versa.)
- Potential domain-specific additions to the metadata fields were suggested (For example, information about the structure and genetics of populations for evolutionary biology.)
- Addition of a field that states what other standards the data are compliant with (For example, the MIAME standard defines minimal metadata standards for genomics.)
- Basic search feature for specific registry
- Cross-domain search feature (but this would require some basic set of fields to be the same across various domains or some mapping among the different schemas)
- Include vouchering – to indicate the location where specimens are deposited, if applicable (EML supports this but it is not exposed in the ESA registry)
- Ability for a user to register once such that the entry will be accessible from multiple registries
- User feedback mechanism
- Feature indicating the number of times a record/entry has been viewed
- Way for non-submitters to review data and make additional annotations
- Allowing contributors to revise their metadata
- Allowing contributors to remove their metadata if not tied to a publication
- Way to browse data registry entries (alpha, categorized etc.)
- Notification to users of new entries

- Notification of new systems coming online
- Checking for broken links and possible notification of contributors about such problems

It was noted that, though sharing the administration and maintenance of a registry among societies would lower costs for each society, the long-term costs of hosting and maintaining a registry are nevertheless significant and ongoing.

Next steps

ESA staff will prepare this draft meeting report, circulate it to participants for input, and then distribute a final version to be shared with society leaders.

Cliff Duke will prepare a template article summarizing the workshop that participants can customize and include in their own member newsletters and web sites.

As part of the NSF grant, ESA's Science Office is establishing a dedicated web site, planned for implementation in September 2006, when ESA launches its own web site redesign.

Society representatives will report back to their respective governing boards about the issues discussed and will ask if they may be interested in the consortium approach to data registry development.

Society representatives will also ask their boards to consider adding to their journal editorial policies the statement that

> The editors and publisher of this journal expect authors to make their data available and to deposit metadata in an approved registry.

Society representatives will consider publishing editorials in their journals to help introduce members to data sharing concepts.

The next workshop will be on data centers, preferably in November or December 2006. ESA will solicit volunteers to develop the agenda and facilitate the workshop from among the current participants. NCEAS and NEScent are both interested in hosting the data centers workshop, and ESA will work with NSF to resolve the location and dates.