The Biodiversity Collections Index – A Proposal

Roger Hyam, Royal Botanic Garden Edinburgh, 23th November 2007

Summary

Drivers: All biodiversity research ultimately relies upon artefacts in collections. These artefacts may be dead or living specimens, illustrations, fossils, seeds or other materials. The Convention on Biological Diversity obliges governments to conserve their biodiversity. To do this researchers need to find and utilise these artefacts.

The Problem: Artefacts are widely distributed between collections and a large proportion are not stored in their countries of origin so locating them is far from straightforward. There is no central list of the collections that hold these artefacts let alone information on what these collections contain. Researchers have to rely on historical knowledge, specimens cited in extant publications and word of mouth to find research materials. They cannot know if they are missing anything significant. Governments cannot know what assets are available to help them understand and conserve their biodiversity evenat a high level.

Many institutions and networks hold a large number of collections and sub-collections, but do not expose this information to the outside world in a uniform manner. Some collections, such as those containing artefacts from a single expedition, are geographically dispersed across different physical collections often in several countries and so it is not clear where the responsibility lies for maintaining data about them. Only a tiny proportion of artefacts have been digitally catalogued. Most are invisible from the point of view of information technology and yet it is only through exploiting informatics techniques that the taxonomic impediment is likely to be overcome and our natural resources managed successfully.

Solution Envisaged

- A single internet-based, shared resource: the **Biodiversity Collections Index**.
  - Data about the collections not the artefacts.
  - Covering all taxonomic groups.
  - Created by a consortium of major institutions.
  - Contributions from all institutions globally.
  - Coordinated/hosted by Royal Botanic Garden Edinburgh.
- A website providing a single point of access for researchers and governments seeking biodiversity materials.
- Economies of scale for participating institutions through sharing of resources.
- Distributed management of data.
  - Contributing institutions have bespoke interfaces to meet their own business needs.
  - Generic interface available for all institutions.
- Web services provide machine access to data (e.g. GBIF, EoL and Atlas of Living Australia)
- Data cleaning and augmentation
  - Data mining of research databases allows expansion of data on collections - beyond what may be known by collection managers.
  - Gap analysis of collections data to stimulate further research and target artefact digitization efforts becomes possible.
- Outreach
  - Collections data becomes available to anyone with a computer.
  - "Dynamic Virtual Collections" give developing nations a single view of materials from their region - even though these may be spread across multiple institutions.
  - The central website is internationalised and localised to major languages.
  - Data is held in English and local languages where relevant.

Development of this Proposal

This proposal builds on work carried out within the *Collections Descriptions Interest Group*[1] of *Biodiversity Information Standards (TDWG)*[2] that is building a data exchange standard called *Natural Collections Descriptions* (NCD). NCD is due to be adopted as a TDWG standard in 2008. This, in turn, brings together work on collections descriptions being carried out for the European Union Framework VI programme known as SYNTHESYS[3] and the work performed by RAVNS[4] under the auspices of RLG[5].

A meeting in June 2007 at ETI[6] in Amsterdam funded through TDWG by the Gordon and Betty Moore Foundation[7] discussed the development of the NCD standard and its deployment using a software tool kit that is being developed at ETI on behalf of GBIF[8]. The assumption at the meeting was that there would be a series of national data nodes (mirroring the GBIF national nodes) that served data concerning collections held within their jurisdiction. A number of problems were identified with this model. Firstly many nations do not have national nodes and many institutions publish their data directly to GBIF. Secondly most scientists and institutions tend to work thematically. Botanists work with botanists and entomologists work with entomologists on a global basis. The science does not arrange itself around political boundaries. This becomes particularly apparent when the assignment of Globally Unique Identifiers is considered. The usefulness of GUIDs is greatly enhanced if there is only one GUID for any one real world object but it is very difficult to control the issuing of GUIDs if data on collections may be held in one or more thematic databases or national nodes or both.

It became apparent that great efficiencies would be gained if there were a single service that held the basic data for all collections and issued resolvable GUIDs for each data record. *The time to build such a service is now*. What was not apparent at the meeting was how such a service could be built, but subsequent discussions have resulted in this outline proposal.

Funding

There appears to be an important role for the BCI within the emerging biodiversity informatics infrastructure. Funding is required to build the initial system that would address:

- The primary requirements of participating institutions and data consumers.
- The identification and prioritisation of secondary requirements of participating institutions and consumers.
- Integration with domains specific, high level collections databases such as *Index Herbariorum*[9], *Insect and The Spider Collections of the World*[10], Catalog of Fishes and SYNTHESYS.
- Integration with larger biodiversity informatics projects such as GBIF, EoL[11] and Atlas of Living Australia.
- Developing a technical and financial plan for running the system within a sustainable ongoing budget.

Funding is sought for the minimum of one year and continued support at a high level for the following six months. Without commitment of sufficient resources to "prime the pump" the project will not be viable. Commitments to further support and development will be sought by the end of the first year.

Implementation

An initial development phase of one year will establish a core system that supports primary use cases. At the

1 http://www.tdwg.org/activities/ncd/
2 http://www.tdwg.org/
3 http://www.synthesys.info/
4 http://www.rlg.org/en/page.php?Page_ID=338
5 http://www.rlg.org/index.php
6 http://www.eti.uva.nl/
7 http://www.moore.org/
8 http://www.gbif.org/
9 http://sciweb.nybg.org/science2/IndexHerbariorum.asp
10 http://hbs.bishopmuseum.org/codens/codens-r-us.html
11 http://www.eol.org/

end of this phase the system will be functional and maintainable without further development. This phase will only address the most important use cases and may be followed immediately by subsequent development phases addressing new use cases. The focus will be on making the implementation of each use case robust before instigating further development. A strategy for on going maintenance of the system that permits graceful degradation of service and perpetual availability of data should funding fail is presented here.

Initial Development Phase

The initial development phase will last one year. Rather than dividing the time available into four roughly even parts: requirements gathering, design, implementation and testing, the BCI will adopt an agile software development approach[12]. After an initial period of requirements gathering and technology evaluation lasting less than three months working systems containing new functionality will be released on a monthly cycle until the end of the year. This should dramatically reduce project risk and enable the development process to adapt to changing requirements identified by user feedback.

Initial Deliverables

The most important deliverables will be prioritised during the first three months of the project in collaboration with all parties. These use cases and data import tasks will be reviewed and re-prioritised every subsequent month in the light of the extant live system. An initial list of such work items includes:

Requirement	**Functionality**
Clear legal status of data.	Legal statements on submission of data and data distribution – probably a requirement for all data fields (apart from administrative contact details) to be covered by a Creative Commons license.
A few collections are of a sensitive nature. The system must not aid extremists who may target these collections.	A policy statement will make it clear that sensitive data should not be submitted. Abuse control will mitigate against third parties posting sensitive material on the system.
Public should have access to search and browse records.	A public facing website will expose the database for general access. Thematic areas will be added to this site for different user communities as required. Common web usability techniques will be used such as: similar collections, collections near this one, maps based on public APIs such as Google Maps.
Web based tools allow creation, editing and deleting of entries.	User authentication/authorisation. Ajax enabled web forms and work flow for creation of entries whilst minimizing chances of duplication e.g. via checking of duplicate/similar names, location, zip codes as form fields are completed.
“Findability” is an issue with such databases. The system must offer better services than a simple yellow pages.	Metadata on collections will be proactively managed. At a minimum there will be taxonomic and geographic tagging of collection records. Third parties will be able to tag collections. “Owners” of records will be able to validate tagging. The administrator will be able to detect badly tagged records and target them for improvement. Metadata will be mined from

12 http://en.wikipedia.org/wiki/Agile_software_development

	external sources such as NCBI and GBIF.
Institutions must be able to maintain their own 'official' versions of records about assets they own or delegate this job to a recognised authority (e.g. Index Herbariorum). These records must be differentiated from records generated by third parties.	Mechanism to verify institutional logins will be put in place. A logo or other branding will be visible on official records for institutions.
Administrator must be able to de-duplicate and clean up entries.	Administrative tasks will be made available through web based interface.
Third parties should be able to comment on any record.	A customer-review-like interface that requires minimal login will enable notes to be added. Administrator and owners of 'official' records will be notified of postings as an abuse control measure.
System should be spam proof.	CAPTCHA tests will be included on relevant forms. Logging/notification of new posts to administrator for spam removal
System should contain enough information to be useful from the start.	Identification, prioritisation and import (or synchronisation with) other datasets e.g. NCBI, GBIF, FAO, Index Herbariorum, Insect and The Spider Collections of the World and SYNTHESYS.
Records of Institutions and Collections should be uniquely identifiable.	Life Science Identifiers will be issued for all records, maintained permanently and synonymised. An LSID to URL proxy will also be implemented for non LSID aware clients as per TDWG recommendations.
Other systems should be able to keep a synchronised internal copy of the data for their own data validation purposes.	Harvesting protocols will be implemented. At a minimum this will be OAI-PMH but could also be RSS or other protocol if required. Legal statements will make it clear that keeping and distributing a complete copy of the data is OK.
Clients need to find an Institution/Collection on the basis of partial information.	Both a RESTful web service and a web page will allow “fuzzy” searching to return a list of records ordered by relevance.
Records need to be related to entries in other lists of Institutions and Collections.	Records will bear an extensible list of foreign identifier codes.
The system must persist indefinitely.	Funding will be sought to maintain the system and curatorial dumps will allow data to persist should funding fail – see below.
The data must be backed up.	The server data and software will be backed up regularly so that it can be restored in case of catastrophic failure. The data will be distributed under a Creative Commons license on a regular basis (minimum of yearly).
The system must survive the departure of key staff.	Full administrative documentation will be prepared as part of the administrative interface

	and as a free standing document. Code will be maintained in a Subversion repository.
The system must be available internationally.	The public interfaces will be internationalised and localised as required. The administrative interface will not be internationalised in the first instance. It will be possible to enter data in multiple languages in key fields. All text will be UTF-8 encoded throughout the system.
The data should be available to other networks.	Web services, such as TAPIR, will be implemented as required in addition to the minimum LSID authority, harvesting interface and RESTful lookup service. International standards will be followed wherever possible. The TDWG standard (Natural Collections Descriptions) will inform the data model and form the basis of data exchange. Other TDWG standards will be followed for taxonomy and nomenclature.

Sustainability

Sustainability has been raised as an issue. The BCI needs to exist indefinitely but institutions are generally unable to commit resources to maintain data and services in perpetuity. The only financial mechanism likely to secure such a future for BCI would be an endowment and this is very unlikely to occur in the immediate future. BCI therefore needs a strategy to persist its data and services within a volatile funding environment – by harnessing finite term commitments. The strategy separates data persistence from service persistence.

Data Persistence: The persistence of data held in the BCI is separate from persistence of services. Long term persistence will be ensured by taking regular (minimum yearly) curatorial data snapshots and depositing these in appropriate digital archives as well as on media in physical archives. Data persistence is one of the main drivers for data deposited with BCI being governed by an open license.

Service Persistence: Services will only persist while there are resources to support them. A classification of support levels is listed in the table below. By forming a consortium of interested parties the BCI would aim to maintain sufficient funding for Technical Support Level 4 and Data Support Level 2 to be maintained on an on-going basis. Should this funding fail the system could run with reduced levels of support for some time thus permitting "graceful degradation" of the system, giving the opportunity for new funds to be found and preventing the sudden outage of a critical system. Even if the system were to fail entirely and the consortium break up it would be possible for concerned parties to create a new system based on the archived data. Once the BCI is established, clients will be able to build their own systems on the understanding that it will persist indefinitely.

Commitment by Royal Botanic Garden Edinburgh: The RBGE recognises the importance of the BCI and is willing to host the service on an on-going basis. It will commit to act as a safety net by providing Technical Support Level 2 and Data Support Level 1 following the loss of external funding but reserves the right to reduce this level of support with one year's notice. RBGE will support appropriate on-going applications for funding.

Commitment from users: The system will become an integral part of the biodiversity informatics infrastructure. Commitments will be sought from institutions and projects that depend on BCI services to provide financial and other support at least for the duration of their dependence on those services.

Continual Growth: BCI will continually adapt its services to the needs of new projects and seek funding from those projects to support development and maintenance.

Support Categories and Levels

This table describes the categories and levels of support required by the system when it is not in a

development phase. The levels within the categories are additive so that Level 3 technical support implies 22 days total (12 + 5 + 5). Likewise Level 2 data support implies 50% of someone's time (10% reactive and 40% proactive).

Category	Level	Example actions
Technical Support	1. Routine Service (approx. 5 days per year)	Backups
		Software upgrades - patching for security and other bug fixes in OS and libraries
		Generate curatorial data snapshots.
	2. Incident Response (Highly unpredictable; budget 5 days per year)	Recovery from hardware failure (power outages etc.)
		Hack attacks - restore of system and plugging security holes
		Fix/workaround critical bugs and crashes (e.g. memory leaks, log rotation)
	3. Software updates (approx. 12 days per year)	Fix non-critical bugs
		Addition of minor functionality for existing use cases
	4. Minor development (approx. 12 days per year)	Minor refactoring of functionality
		Implementation of new functionality supporting new use cases.
		New import/export tools.
Data Support	1. Reactive (approx. 10% person year)	Answering user queries.
		Abuse control on comments.
		Basic checking of new entries.
	2. Proactive (minimum 40% person year)	De-duplication entries.
		Invitation to update or confirm collection data.
		Instigate data cleaning campaigns and promotion.

Resources

The table below shows the main resources required for project start up. The two posts mentioned may each be split between more than one person and/or location *pro rata* during the course of the project. Discussions are under way for GBIF and TDWG to provide funding for a large part of these costs. The Royal Botanic Garden Edinburgh is committing significant resources and the Smithsonian Institute has expressed an interest in committing resources. Support is still needed from other parties if the project is to succeed.

Resource	Source	Value
Institutional hosting	Supplied by RBGE	US$10k

ICT, hosting and support	Supplied by RBGE	US$10k
Salaried Manager/Developer for 1 Year from 1st quarter 2008	GBIF, TDWG plus others	US$80k
Data entry and cleaning by one or more staff (equivalent to one year half time starting 3rd quarter 2008)	GBIF, TDWG plus others	US$40k
Promotion and travel budget. Training and presentation at workshops, printing etc. Small amount of graphic design for website.	GBIF, TDWG plus others	US$40k

There is now an opportunity for institutions and communities to commit resources that enable the curation of metadata directly relevant to their domains of interest and the implementation of functionality that directly supports their particular needs. Organisations that support BCI will not only benefit from the services it offers but will also be recognised as playing a leading rolein the development of the global biodiversity infrastructure.

Curators are responsible for make their collections available toresearchers. Contributing to BCI is a cost effective way of doing this.

Further Information

If you would like to become involved or have any further comments please contact Roger Hyam <roger@tdwg.org>