Report of the 1st Canadensys meeting – January 16-17, 2009

This report was published on 2009.02.24.

For all presentations and documents related to this meeting, visit:
http://www.canadensys.net/meetings/2009_01_meeting.html

General recommendations

- Canadensys should not reinvent the wheel, but learn from other collection networks.
- Canadensys should publish or reference recommendations. Some are already available at: http://www.biodiversite.umontreal.ca/canadensys/documents.html
- We should continue the discussions started during the meeting on an online discussion group. Now available at: http://groups.google.ca/group/canadensys

Canadensys management structure

(The following reports on outcomes of both the discussion on this topic started at the meeting and subsequent consultations and drafting of terms of reference)

Steering Committee

The Steering Committee (SC) will be the main day-to-day decision-making body of Canadensys. For major issues affecting all participating institutions and collections, the SC will consult the entire membership, as well as the Scientific Advisory Board (SAB, see below) prior to making decisions that reflect the overall consensus as much as possible. Less critical decisions will be made by consensus among SC members only or with limited consultations. These members will be the Canadensys Principal Scientist (Anne Bruneau), the Biodiversity Informatics Manager (Peter Desmet) and representatives of each of the following categories of participants (striving for regional representation, as well): Botany, Entomology, Mycology, Botanical Gardens, as well as an additional Biodiversity informatics specialist. The SC will meet in person once a year and remotely three times a year. It will also appoint other committees as needed, such as the SAB.

Scientific Advisory Board

The SAB is needed to inform the discussions, consultations and decisions of the Canadensys membership and its Steering Committee with independent advice and updates on parallel initiatives and international developments. The SAB will be appointed by the Canadensys Steering Committee and be composed of a small number of key scientists, collection managers and digitization experts from organizations with parallel objectives within Canada or similar objectives outside of Canada. The SAB will also be asked to review and comment on Canadensys reports and to make recommendations to the Steering Committee regarding future directions and priorities.

Network architecture

This topic covers how the network will work technically and was introduced at the meeting by Peter Desmet with the presentation 'Data standards & protocols'.

Central or distributed network

In a central network all the data are stored (copied) on one central server, which can be accessed by the users. In a distributed network, the data remain on the servers of the different data providers, which could be accessed via a central portal. Both architectures have their advantages and disadvantages.

- It is best to work with a mixed architecture.
- Updating a central cache can take months (see GBIF). Maybe Google can provide a solution?
- Choosing an architecture for serving images (central or distributed) is another problem, and might be different than the one used for serving (text) data.

Data hosting

Some (smaller) collections might like to have their data hosted somewhere else, instead of maintaining their own server.

- This service is already provided by BOLD, CBIF and Acadia University. It has proved useful, and could be provided by Canadensys as well.
- Hosting could be done at the central server and/or regional servers (like Acadia University).
- Hosting by universities will be subject to the IT policies of that university.

IPT

IPT is the GBIF Integrated Publishing Toolkit (http://code.google.com/p/gbif-providertoolkit/), which is a tool to serve biodiversity data to the world via different services (TAPIR, zip-file, etc.) The IPT was presented at the meeting by James Macklin.

- IPT is now available in beta. A full version is expected in spring/summer 2009.
- Canadensys recommends that its participants share data via IPT, once the software is stable. Installation and use of IPT could be covered in a workshop (fall 2009?).
- There is an opportunity for Canadensys to help the development of IPT and to produce user manuals (funded by GBIF).
- The question of whether TAPIR Lite (as opposed to more complex versions) is sufficient for our network needs more examination.

DarwinCore

DarwinCore is a data exchange schema developed by TDWG (www.tdwg.org/activities/darwincore), used to share biodiversity data in a common form. It is almost always different from the local data model and thus needs to be mapped before data can be shared. The local database might hold more information than what can be shared through DarwinCore.

- DarwinCore 1.4 is the standard we will use to share data within the Canadensys network. We will use the Core, the Curatorial Extension and the Geospatial Extension. We might include the Interaction Extension as well. Using DarwinCore does not obstruct the adaptation of the TDWG LSID vocabularies (e.g. TaxonOccurence) in the future.

- We need to discuss what fields can be provided (for old and new data) per kingdom. Kingdoms need to share as many common fields as possible, to guarantee cross-kingdom search.

- In order to start these discussions, each collection should send a sample of its database to Peter Desmet. This can be an Excel file (e.g. with field names as column headers) and/or a data model of the database. Only text information is needed. If you have image information, just mention how you link your images to your database.

Duplication of data

Shared data might get duplicated in aggregators (like GBIF) if it is served via different paths (e.g. via national node and directly from the provider). GUIDs could solve this problem, but not all aggregators (if any) support GUIDs (yet).

- Canadensys needs a policy on how to (re)distribute data. If data are served directly from the data providers, this will seriously limit the use and usability of the central Canadensys portal and the promotion of the network. We would therefore suggest to set up the central portal as the only gateway to access Canadensys data.

- We need a policy on if and how we redistribute data from outside partners in the future.

- Canadensys should adopt GUIDs to limit these problems in the future.

LSIDs / GUIDs

LSIDs (Life Science Identifiers) are a kind of GUIDs (Global Unique Identifiers), which allow to uniquely reference an object (e.g. a specimen). The advantages of GUIDs are clear, but they have not been adopted by a lot of systems at this time.

- Canadensys should adopt GUIDs, but it is not clear what is the best way to implement them. More examination of this subject is necessary.

- As a start, Canadensys could send around a survey to ask how each collection deals with local unique IDs (how are they generated, does each specimen have its own unique number, etc.?). This information could be provided with the data sample to Peter Desmet.

Filtered Push

The Filtered Push (http://mantis.cs.umb.edu/wiki/index.php/Main_Page) is a bi-directional distributed network for annotations. Curators can query the network for information (authority lists, other collections and several web services) on a certain topic (e.g. a genus) and filter the information they get back (e.g. the next day) to complete their collection database. The corrected information is pushed back into the community pool. The system allows to improve the quality of collection databases and reduce duplicate effort. The Filtered Push is in development and was presented at the meeting by one of the researchers in this area, James Macklin.

- The Canadensys community would certainly benefit from using the Filtered Push. The tool will be examined more closely once it is available.

- If possible Canadensys partners should focus on the unique parts of their collections and work on the duplicated parts when the Filtered Push is available.

BCI – Biodiversity Collections Index

BCI (http://biocol.org) is a central index to collections all around the world, developed by TDWG, GBIF and the Royal Botanic Garden Edinburgh. BCI assigns an LSID for each collection and has several web services to retrieve information about these collections in a standardized form. BCI is available as a proof of concept at this stage, but will be part of GBIF's data discovery services in the future. BCI was introduced by Peter Desmet at the meeting.

- Canadensys will use the services of BCI for its collections' metadata.

- 80% of all Canadensys collections are already available on BCI (via Index Herbariorum and the Insect and Spider Collections of the World). Curators should add or review their collection information on BCI. This can be done by registering on the website or submitting the correct information to Peter Desmet. An overview of the Canadensys collections (with remarks) can be reached on http://groups.google.ca/group/canadensys/web/collections

- BCI is working on an updating strategy. The Canadensys community could review this updating strategy (via user comments) if asked by BCI.

Use of data

Only a part of our collections can be digitized with the current funding. To maximize the use of that data for research and the chances for future funding, we need to prioritize what we will digitize and share. This topic was introduced by Larry Speers in 'Data entry (digitization)' and was discussed throughout the whole meeting.

What info do we capture?

Except for taxonomists – who use the physical vouchers – label information is sufficient for all other users. "What, when and where" is the most commonly used information, but other data uses cannot be foreseen.

- It is therefore important to capture as much information as possible from the label. This might also be the most efficient way to capture information in the long run.

- The DarwinCore fields are the most commonly used elements of biodiversity information (they are the result of several use case analyses) and are probably sufficient for sharing Canadensys information. If the Canadensys community has a large amount of other information shared by several collections, we need to look for ways to share this information as well (e.g. via a custom DarwinCore extension).

What groups do we prioritize?

We should prioritize (taxonomical, geographical, etc.) groups that are most valuable for research. This can be hard to assess, especially since biodiversity uses can be extremely diverse.

- Each collection should focus on digitizing groups where it is especially strong/unique.
- We should focus on groups that receive research attention from our own Canadensys community, like ‘Carex’.
- We need to reach our “outside” users in order to know what information they are looking for.
- There is demand for good large-scale datasets by the climate community. We should keep in mind that the real strength of the Canadensys network lays in the combination of our collections (the “meta collection”), as shown by the butterfly example in Larry Speers’ ‘Data entry (digitization)’ presentation.
- Data on rare and invasive species are also often used (and are good indicators of climate change as well) and are rather easy to digitize.
- We need to think about what data we will provide from our botanical gardens: do we include cultivars, do we provide the original or the current locality, etc.?
- Data quantity has no use without sufficient data quality.

Data rights

- All data shared through Canadensys will be publicly available.
- We need to determine under which license we publish our data (Creative commons?) and advertise how people should cite Canadensys data (a paragraph of instructions is preferred over a long page).

Sensitive data

- The sensitivity of data is often overblown and “denaturing” or “fuzzying” is seldom effective. You can limit what data need protection with the decision tree in Chapman & Grafton ‘Guide to Best Practices for Generalizing Sensitive Species Occurrence Data’ (http://www2.gbif.org/BPsensitivedata.pdf). NatureServe has determined that less than 5% really needs to be protected.
- The above-mentioned document is unclear on the issue of releasing collection data without the consent of the collector (published data might get the collector in trouble, e.g. collected in national park).

Data entry

Before we can share and use our specimen information, it needs to be available in a digital form. The process of digitizing specimen information can be very time-consuming and it is therefore important to make it as efficient and error-proof as possible. Larry Speers discussed this topic at the meeting in his presentation “Data entry (digitization)”.

- See Larry’s presentation for a more in depth overview of the issues related to digitization: software choice, data integrity, fitness for use (= documenting errors in determination, georeferencing, time of identification), optimizing the data quality information chain, workflow, automated processes, staffing, etc.

- The Canadensys community should share data entry experiences via the online discussion group. If necessary, a workshop could be organized.

- Since errors are more expensive to fix the further up the digitization process, it is important to limit the number of errors at the time of data capture. Data capture guidelines are available for herbaria (http://www.nationaalherbarium.nl/virtual/Data-guidelines-NHN.pdf), but not for other collections. These could be developed by Canadensys.

Specify

Specify (http://www.specifysoftware.org) is a free and open source software application for managing biological collection information, developed by the University of Kansas. James Macklin presented Specify 6.0 at the meeting.

- Specify 6.0 is expected to be released at the end of February 2009.

- Several Canadensys participants showed interest or discussed their experience in using the software. Since full technical support is limited to US users only, we should share experience and organize a workshop if necessary.

- Canadensys could contribute to the software by translating the user interface in French (currently in English and some Spanish).

Imaging

Since this topic was too diverse to fully discuss at a first meeting, it was just briefly introduced by Larry Speers in “Introduction to imaging of specimens”.

- Imaging practices depend largely on the kind of collection (flat or 3D specimens) and it is difficult to provide general recommendations. Imaging should probably be covered in different workshops (per kingdom).

- Imaging labels is a recommended practice (depending on the time/cost), since it provides a digital backup or check for text information. Of course, it cannot replace text information, since it cannot be searched or analyzed (see also “What info do we capture”).

- Imaging whole specimens can be very work intensive and should only be considered if it is scientifically valuable.

- A 300dpi image resolution for scanners is sufficient for current applications and uses, but this might change in the future. The Global Plants Initiative uses 600dpi, which can be a challenge for storage.

- Image filenames should not contain human-readable information, but just a unique number (LSID, barcode, UID) so the image can be linked to the database. How multiple images for the same specimen should be named has not been addressed yet.

- Images are generally stored as TIFF files (which can be converted to a lower resolution JPG for web access), but there is a tendency towards JPG2000 recently. JPG2000 allows storing image or specimen metadata in the image file, but it is not supported by a lot of software at this time. We need to ask JPG2000 users for more information before making a decision.

Georeferencing

Georeferencing is the practice of assigning geographic coordinates to specimens, based on their locality information. Larry Speers introduced this topic at the meeting in his “Georeferencing” presentation.

- To maximize the use of our data, Canadensys’ specimens should be georeferenced if possible.

- We should georeference as accurately as possible (no 5 km-grid like in the UK).

- The obtained coordinates should have an uncertainty measure (via the point-radius method) if possible. This “error” is extremely valuable to determine the fitness for use: coordinates with a point-radius error of 10 km might not be useful on a local scale, but could be on a national scale. MaNIS/HerpNet/ORNIS published guidelines and tools to calculate this error.

- Canadensys should reference or publish best practices for georeferencing.

Software

- The BioGeomancer Workbench is an online georeferencing tool developed by the BioGeomancer project (www.biogeomancer.org). The single record georeferencing tool works fine (and also calculates the point-radius error), but batch georeferencing is still insufficient.

- GEOLocate (http://www.museum.tulane.edu/geolocate/) is a desktop georeferencing tool developed by the Tulane University Museum of Natural History. It has single record and batch georeferencing options. Since it was originally developed for fish collections, it uses a box model along a river as an uncertainty measure, instead of the point-radius-method. GEOLocate also allows collaborative georeferencing, using DiGIR to share the data.

- Both BioGeomancer and GEOLocate are good software tools for a first pass, but the coordinates need a lot of human reviewing afterwards. This is mainly caused by insufficient gazetteers for Canada.

Gazetteer sources

Most georeferencing tools are focused on the US and use imprecise gazetteers for Canada. In order to georeference correctly, we need to look for more detailed gazetteers.

- Sources of this kind of information include: the Canadian Geographical Names Service (CGNS) (including historical names), biological stations for sampling places, provincial sources, heritage groups, etc.
- It is often unclear if data from these sources is freely available and how it was calculated (what is the centroid of the town, what is the extent, etc.?).
- Historical data are necessary for older specimens, since locations can change over time (e.g. towns get bigger, Alaska Highway got shorter). It is therefore also important to keep the original locality data in the collection database.

Collaborative georeferencing

- Instead of georeferencing each collection separately, Canadensys should look for ways to georeference collaboratively. This greatly reduces the duplication of efforts and can generate more accurate results, since specimens can be grouped per province/locality in the overall network (instead of per collection), and thus be georeferenced by the people with the most knowledge of these province/localities.
- How we can develop such a collaborative georeferencing network is unclear at this time and should be discussed in more detail. Maybe the Filtered Push and GEOLocate could help here.

The future of Canadensys

Although Canadensys has just started, it is already important to think beyond the five-year project: how can we attract additional funding, participants and users? This topic was discussed during the second day of the meeting.

Future participants

Currently the Canadensys network (mostly) consists of university-based collections, but some other institutes and projects are already interested in collaborating and were present at the meeting.

- Canadensys will not seek for additional collections to join the network at this stage, but will adopt an open door policy, so future participants can be kept updated on Canadensys' decisions and activities.
- Once Canadensys accepts additional collections (phase 2), an invitation will be sent around via the participants' networks. At that time we will need to draft some basic requirements for joining.
- Possible future participants include, but are not restricted to: CBIF-collections, Canadian Museum of Nature, other federal collections, small-scale local collections, Biological Survey of Canada, etc.

Future funding

Canadensys needs to look for future funding opportunities, which can be difficult since databasing and digitization is often seen by Canadian granting agencies as housekeeping rather than providing new raw data for research.

Funding opportunities:

- GBIF: some workshops could be funded by GBIF
- Mellon foundation: digitizing type specimens, as part of the Global Plants Initiative.
- NSERC: Major resources support grant (excellent option to continue funding major research infrastructures beyond initial CFI seed funding)
- California-Canada Strategic Innovation Partnership
- Provincial funds for workshops
- Quebec-Mexico grants
- Funds for linking to international initiatives
- NAFTA (environmental secretariat in Montreal)
- Monarch funding (?)
- EOL funding for taxacentric meetings or as regional node providing Canadian species information.

Future meetings

We suggest to have a workshop in fall 2009 on "Putting recommendations into practice", which could cover IPT, Specify and other. Meetings on taxonomy, imaging and georeferencing were also suggested.

To limit expenses, Canadensys meetings should piggyback on other meetings, like:

- SPNHC meeting Ottawa, summer 2010
- Consortium of Northeastern Herbaria meeting, June 8-9 2009, New Hampshire
- Canadian Botanical Association meeting (May 2009 at Acadia University)
- Canadian Society for Ecology and Evolution meeting (May 2009 in Halifax)
- TDWG is interested in having a meeting in Canada in 2010
- Other society meetings (entomology, botany, etc.)

Future ideas

We need to think about the "product" of Canadensys: what kind of services could make Canadensys indispensable/irreplaceable for the users? Are there ways to make our data even more useful?

- We need to find other entry points to our specimen data beyond names, like checklists, maps, local projects, etc.

- We could combine our information with species lists from biological stations, climate data from weather stations, etc.

- "Checklist idea" of David Shorthouse: Canadensys could host an open-access journal were people can publish checklists, with the requirement to link a part (e.g. 80%) of their taxa with specimens in Canadensys (= vouchered checklist). By assigning DOIs (http://www.crossref.org/) to these checklists, one can automatically trackback in what (further) literature these specimens are referenced. These kinds of statistics are a lot more tangible to administrators than number of web visits.

- We could create (Canadian) species pages, based on our specimen information in combination with existing information from different participants, and serve species page data to EOL.