

Sketch of BioCorder Meeting Agenda (Orlando, April '05)

Much of what follows may seem very sketchy. That is because there are so many unknowns about how we might achieve what BioCorder sets out to do. The purpose of this meeting is to narrow down some of our options and outline a strategy for the next 6-12 months that will test and implement some of the ideas set out in the BioCorder proposal.

The Basics

This revisits the contents of the proposal but I think it is worthwhile reconsidering the 'big picture' before we consider some of the details. Some of these questions will need to be answered during the meeting; others are intended to raise awareness of certain issues:

What do we want to do?

Most of us have a pretty clear idea of what we want to achieve with this project, but I think it would be worth formalising a set of common goals that we all understand and share. In this sense I think we need to write a mission statement [we need a draft of this before the meeting].

Who are we doing this for?

The project was conceived for the louse community, but with an eye towards all systematic biology researchers. Is this still the case? Most likely we can please some people all of the time, or everybody some of the time. Is a truly generic product possible and what are the barriers to doing this?

What are the core subject themes relevant to the project and can we prioritize them?

Most of these are set out in the proposal and are summarized in the early conceptual schema for the project (<http://www2.flmnh.ufl.edu/biocorder/BioCorderSchema.jpg>). Have these subjects changed, is anything missing from this list and can we identify some priorities?

What other projects are relevant to BioCorder?

Since writing the proposal several web-based projects have emerged that have overlapping goals with BioCorder. What are these projects and can/should we rethink what we are doing in the light of this?

What makes web based projects successful?

Can we identify elements of successful web based projects that we might use? Think HUGE here. To my mind successful web based projects include things like Google, flickr (<http://www.flickr.com/>), Skype (<http://www.skype.com/>), Peer-to-Peer file sharing networks (e.g. Kazaa, Gnutella and the semi defunct Napster), GenBank and PubMed (<http://www.ncbi.nlm.nih.gov/>), SourceForge, BBC News online, Amazon.com, Apple Computers, E-bay and the iTunes music store. Each has attracted a massive following very quickly and there is a reason for this. What have these organisations done right and how might we learn from this?

Philosophy

From the outset BioCorder was intended to be a free, open source, and cross platform software project. Is this still the case? How committed are we to this and should we be prepared to compromise if this will resolve some of our implementation problems?

Implementation Fundamentals

What are the guiding principles on how BioCorder will work and what things should we be thinking about when developing BioCorder? As before, some of these are posed as questions requiring firm answers in the near future, others are broader issues that will be resolved as the project develops.

Modularisation concept

To conceptually manage BioCorder's functions the project was split into discrete modules that might be developed, used and even served (deployed) independently from geographically separate locations. At what level is this division conceptually still relevant to the development and implementation of BioCorder? In other words, how centralized or distributed should/could BioCorder be, and on what levels? Different options pose significant advantages and disadvantages. Things to consider:

a. *At user level:*

Since most modules will use different workflows (see below) it might make sense for users to interact with them independently. Indeed many users will only use a subset of the available modules. However, most modules will have certain dependencies and in many cases module workflows will overlap. All this will need to be transparent to the user.

b. *At the developer level:*

Some modules are more self-contained than others and can be developed in relative isolation (e.g. the bibliographic module). Others will be very tightly integrated with each other and will need very close cooperation between developers (e.g. specimens and molecular data).

c. *At the database level:*

How many databases will there be (one per module/group/user?) and what are the practical limitations on users, groups and authentication across multiple databases.

d. *On a physical level:*

Given the practical constraints (time, money and resources) on this project and the complexities of producing series of databases that are physically located on separate servers, is this vision realistic? Should we compromise on this to centralize the project on a single (mirrored) server and save ourselves the time, effort and development problems of physically mounting BioCorder "instances" (i.e. all modules) or BioCorder independent modules, on physically separate servers? As a compromise, should we develop BioCorder modules so that each communicates via web services but for the moment limit the installation to a single server? Do we need to answer this question right now?

Common functions within and across the BioCorder modules

Regardless of decisions made on how distributed/centralised BioCorder will be, there will be certain commonalities to all modules. Fundamentally, data enters each module (either input from the user, from another module, or from an external database) is stored within the module database, and can be selectively queried and integrated with related data by any user with appropriate permissions or by a connected BioCorder module. Output would be directed to the screen, physical files or other modules as required. We need to devise ways of intuitively capturing the user processes of adding these data to each module (workflows) and ensure that only those people with appropriate permission can access or edit and query these data intuitively. Likewise we need a means for users to perform complex bespoke queries on their data, their collaborators data, and publicly available data, as required. For this to all work data needs to be persistent, we have to be able to track changes, and modules need to communicate with each other. Have I missed any thing from this list? Discuss!

Common data elements within the BioCorder modules

Every piece of data within BioCorder requires a unique and persistent identifier (so called 'Globally unique identifier' or GUID) that can be easily 'resolved' by anyone to locate and access the data (so called metadata) named by each unique identifier. The method that we are likely to use for this will implement Life Science Identifiers developed by IBM (LSID's - <http://lsid.sourceforge.net/#whatislsid>), since these are free, relatively easy to implement, and Rod already has experience with them. We need to work out how LSID's will work in practise for BioCorder. Specifically, how are LSID's issued and by who, what data elements are assigned LSID's, can the data associated with an LSID ever change (what are the implications if this data can/cannot change), how are LSID's resolved and by who? We need some test implementations of this ASAP.

Dependencies with external databases

Examples of BioCorder dependent databases for data input include a Taxonomy Name Server (TNS) (e.g. Rod's Glasgow TNS, ITIS or uBio) and possibly a central LSID issuing authority. Databases that might receive data from BioCorder (e.g. Genbank or TreeBase) will also be relevant but are not dependent upon BioCorder and need not be considered here. Assuming each BioCorder module issues LSID's, are there other dependent databases we need to consider?

Taxon Taxonomy

Instances of BioCorder will be taxon focused, although there might be cases where the taxon focus is exceptionally wide and patchy. Linnaean taxonomy will therefore be integral to BioCorder modules relating to specimens, data and 'products' of these data. There are many options about how BioCorder might deal with taxonomy, each with their own risks and benefits. These relate to where the taxonomy is sourced, whether users can adopt multiple competing taxonomy's, what happens to new taxa / changes in classifications, and how the BioCorder stores taxonomic data.

a. Source Data

Should users be able to specify their own taxonomic classification, or have one forced upon them? If the latter, where should this classification come from?

Options include external taxonomic name servers (e.g. Rod's Glasgow TNS <http://darwin.zoology.gla.ac.uk/~rpage/MyToL/www/>) or via TNS data portals (e.g. <http://darwin.zoology.gla.ac.uk/~rpage/portal>). How might this work in practise?

b. Multiple Classifications

Should we permit users to adopt multiple competing classifications in any one instance of BioCorder? What about across the other BioCorder instances? What are the risks of doing or not doing this?

c. New Taxa and Taxon Changes

Users will need to name new taxa not in any classification and reclassify taxa based on new data, How do we accommodate this and should/could the changes be reciprocally reflected back into the source TNS? Should we treat new taxa differently?

d. Storing Hierarchical Taxonomic Names in BioCorder

How we do this will be dependent upon answers to the previous three questions. Options include not strong names (just nested sets of a taxons LSID's), just storing the terminal name (or it's LSID), and storing a verbose classification string of the actual taxon names. Decisions on this have a major impact on the efficiency of the database/s. We should aim to minimise or eliminate redundancy, but maximise our flexibility.

Data Security – Users, Groups and Permissions

BioCorder as it was originally conceived will be a repository for all sorts of private and unpublished data, along with more refined and assimilated data of publishable or published standard. For a user to have confidence in BioCorder they need to be certain their data is permanent, secure from others that don't have permission to access it, and extractable from the BioCorder network. These are considered below:

a. Data Permanence

This is the least of our worries but is a major reason why we should develop this project within the louse community first. This user base trusts us and has the confidence that we will not disclose or accidentally wipe their data. Other user communities are an unknown quantity and will be less forgiving when we screw up.

b. Security and permissions

We will need to implement a basic security model across all instances of BioCorder. Most likely we should think about a distributed authentication model (such as that offered by Drupal - <http://www.drupal.org/>) to achieve this. This model will most likely need four (possibly three) tiers of security that include an *administrator* (with read and write access to everything in a BioCorder 'instance'; a *user* with read and write access to their own data; *groups* (with read and write access to data marked available to a group of collaborators); and *public* (with read access to publicly available data).

c. Data backups and data extraction

Obviously data in the network will need regular backups. In the long-term we might look to mirroring key servers running BioCorder 'instances' to ensure data permanence and server availability. However, in the meantime to ensure users have absolute confidence our ability to store their data, they need to be

able to extract their data in a format that is familiar to them. Would it be possible for users to trace all the dependent and orphaned data they have contributed to a BioCorder 'instance' and extract this to a local file (e.g. a tab delimited text file) on their personal computer? As a first step could this be a MYSQL dump of their data?

Registering as a Data Provider – BioMoby

Assuming we decide that databases serving BioCorder data will be distributed (i.e. physically located on separate servers), they will need a means of advertising their presence to other data providers and to portals that might want to query their data. BioMoby (<http://biomoby.org/>) is an open source software project that provides a software architecture for doing just this. Data providers (e.g., BioCorder 'instances' or separate BioCorder modules) register their presence and provide instructions for interacting with them on a central server (MobyCentral). When a user or BioCorder module requests data, transactions initially pass through MobyCentral to discover which data providers have the requested data, and receive instructions on how the user's database/client can interact with the data provider (server). All this is invisible to the user. BioMoby is presently being used by many in the genomics community but is relevant to our needs if we decide to physically distribute BioCorder 'instances' or separate BioCorder modules.

GBif and the DiGIR / BioCase Protocols

How do distributed databases serving biological data talk to each other and exchange data? The GBIF organisation that has many goals overlapping with those of the BioCorder project and has done a lot of work on this. Essentially they have selected what they consider to be the common attributes of all biological databases and developed transfer schema so that these data can be recast into a common data exchange format. The first of these transfer schema is called the Darwin Core and is shared with the world using the DiGIR protocol (<http://www.digir.net/>). However, it is rather restrictive in the types of data that can be shared and as a consequence, various extensions to the Darwin Core have been developed. More recently the ABCD transfer schema has become available. This overcomes some of the limitations of the Darwin core and is made available to the world using the BioCase protocol (<http://www.biocase.org/>). We need to evaluate the ABCD schema and BioCase protocol alongside the Darwin2 schema to see whether these make provision for all the data that we are likely to have available through BioCorder. Almost certainly these will be insufficient to do this, but they will hopefully form a starting place for extending these schema to fulfil our needs. BioCorder should become part of the GBif network of data providers, much in the way that MaNIS (<http://elib.cs.berkeley.edu/manis/>) has done for mammal specimen collections. The difference is that BioCorder is focused on data derived from collaborative and taxon focused specimen collections held by lab based systematic biologists, rather than museum collections.

Data Discovery – a data portal

Once part of the GBif network we could conceivably rely on the GBif data portal (<http://www.gbif.net/portal/index.jsp>) to serve up all the data in the BioCorder network. However, there are many problems with this. Firstly, BioCorder will more than likely serve data that is not currently covered in their transfer schema. Also the GBif data portal is a very poor means of discovering data. Users will want to choose whether they query their own data, that of their collaborators, or all publicly available data. They also need a very flexible means of creating bespoke queries for doing this. Examples of what such an interface might look like can be seen from the Seamark demonstration website (<http://www.siderean.com/medidemo.jsp>). Other examples include mSpace (<http://mspace.fm/>) and Flamenco (<http://bailando.sims.berkeley.edu/flamenco.html>). Each is designed to allow users to search vast amounts of information without getting lost, and this concept is readily applicable to BioCorder. All we would need to do is support the data format that search interfaces such as these require.

Application Program Interfaces (API's) and other services

Try as we might to accommodate every users needs to discover and associate data, we won't fully succeed. The solution is to develop an API that allows users to assimilate data within the BioCorder network. API's just specify instructions on how users can access the system. Third parties can then write software or web applications that use these instructions to interact with the site. For example, the photo site flickr (<http://www.flickr.com/>) provides API documentation (<http://www.flickr.com/services/api/>) that allows anyone with limited programming skills to develop all sorts of third part interfaces and tools for getting data into and out of flickr. For example, people have written image uploading programs to upload images and data to flickr (e.g. <http://1001.kung-foo.tv/>), methods of organising photos in the database (<http://www.flickr.com/tools/organizr.gne>), tools to map images to geographic localities (<http://www.mappr.com/>), and ways to visualise the community relationships between flickr users (<http://www.marumushi.com/apps/flickrgraph/>). None of these were conceived by the people that built flickr, all they did was write the API. These third party applications add new levels of functionality to the database helping to get momentum into the project. Other services we need to think about include RSS feeds alerting users to new data as it is uploaded to the network (see for example the RSS feeds of SID) and 'Watch Lists', alerting users to new information on certain taxa or topics in BioCorder. The BioCorder API is not a priority right now, but as we develop BioCorder we need to be aware that these services may need to be bolted on in the future.

Comments and tagging

To establish a means of community involvement in collaborative projects and biodiversity data, users will need a means of commenting on data elements within the BioCorder network. Such comments might relate to the veracity of the information present, or other attributes of a data element that cannot be captured by the formal database schema that we employ. In addition to unstructured text comments on data elements, we might also think about 'tagging'. This is where a user might mark or 'tag' a data element within the network such that they or another user could easily find that data element again. In this sense tagging has been referred to as social bookmarking, and has

become a very popular, although inefficient and chaotic means of data discovery. Because anyone with sufficient permission to see a data element can tag it with any label/s they wish, the method establishes a mechanism for community involvement. This is highly suited to data rich elements of the network such as images, or controversial data elements (such as Linnaean taxonomy). It is being successfully employed in many community based projects sharing data (e.g. flickr for photos [<http://www.flickr.com/>], del.icio.us for webpage's [<http://del.icio.us/>] and technorati for weblogs [<http://www.technorati.com/>]), and although it unsuitable for discovering highly structured data, might easily be implemented throughout the BioCorder network and an informal means of data discovery.

Where are we now and where do we need to be in 6-12 months

The previous section considers what we want to achieve in the long term. We need to assess how we make best use of the web applications we have to date (LouseBase, LSD, SID, PhPBib, the Molecular Lab Notebook) to implement the ideas set out above. We also need to set some priorities for future development. All of us (David and myself especially) have a pressing need to find a repository for the data we have to hand, but I am conscious that rushing ahead with database development without testing some of the fundamentals is likely to set us set us back in the long term. I suspect that much of the work to implement BioCorder will not be in building databases or designing fancy interfaces. More likely it concerns getting the client-server interfaces behind the scenes to work. To my mind the most pressing issues are:

- *Testing and implementing LSID's throughout the databases.* This is probably pretty simple since these are essentially extensions of the primary keys but we need to think carefully about what gets an LSID since this has many implications.
- *Testing and implementing methods of metadata exchange.* Is the ABCD transfer schema and the BioCase protocol sufficient for our needs. If not, can we easily extend it (I think Fred Ronquist has looked into this). We need to test this these schema within our present applications. Perhaps this should be done for LSD as a first step?
- *Taxonomy.* Assuming we've answered the questions set out in the relevant section above, we need to test our chosen implementation within some the applications we have already developed. SID already does some of this.
- *Data availability and the data discovery portal.* Once we have test implementations of the three things above, we can set about finding the best methods of serving up this data. A simple metadata portal along the lines of the examples provided would be useful starting point for this.

Do we all agree on these as our first priorities? Is anything missing from this list?

The next step...

Once these test implementations are complete and we are confident that the concept of BioCorder works, we can set about redesigning the present web based applications and

building new modules as set out in the original proposal. Issues to consider when doing this will be:

- Developing workflows for data input. Close cooperation between the programmers and those entering data (Mark, his RA, Vince and David) will be required here.
- Developing database schemas that are compliant with the ABCD transfer schema and any extensions thereof.
- Development of the web based interface that users will interact with.

At this stage it might be worthwhile discussing the basic elements of what each modules might do. However, in the light what's above such a discussion might be premature. To my mind the most important modules are:

- The Specimen module. This will form the foundation for all of the data modules and some of the resources (e.g., the GIS module)
- The Molecular and Morphological data modules, although the contents of the latter are somewhat nebulous at this stage.
- The GIS module. We need to relate specimen localities to map referenced locations, and perhaps use this party web based mapping software as well.
- The Bibliography. Most of the elements of this are present in PhPBib but there are some issues to be resolved based on a test implementation of this that Rod set up with Bill Piel.
- The images module. In the light of major funding received by the Morphbank project, is it worth continuing with this?

A few Practical Issues

- We need to buy the domain Biocorder.org and if nothing else, put some details about the project on it. This should be done now.
- Simon in Glasgow will need a computer. Can we buy this (say on a P card) from Florida for him in Glasgow? If not, can Rod pay for this up front and David reimburse Rod?
- What resources does Mark and his RA need.
- Server Issues. If we are centralising the project (at least for now) we will need a reliable server that we all have accounts on. For the moment I suspect this needs to be in Florida. Should there also be one in Glasgow?
- We need to establish a framework for regular communication. I would suggest that we aim for weekly video conferences / conference calls (either via iChat or Skype) between David, his programmer, Simon and myself. Rod, Mark and his RA might contribute to these on an ad hoc basis as required. Given the time delay between Glasgow and the US, Friday or Monday mornings (EST) might be best?
- We also need a method for irregular informal communication. I'm not a big fan of Wiki's but I suspect it is the only way of achieving this and keeping a record of what's happening. Does anyone have a better solution?

- Given the overlap between Morphbank, BioCorder and several other IT based biodiversity projects funded in the US and Europe, Fred Ronquist has expressed an interest in applying for a Research Coordination Network grant that will fund regular meetings and establish a framework for collaboration between these projects. Given that it will involve minimal effort on our part to become part of this I think we should support this effort. If nothing else it will save on our travel budget.

Wrap up discussion...

Should we try to set out some landmarks/goals for the next 6 months? If so, what should these be?

Vince Smith, March 2005.

Some useful links and resources:

Test implementations of modules to date:

- Louse Specimen Database
<http://www2.flmnh.ufl.edu/db/>
- LouseBase (another specimen database for the Glasgow louse specimens)
<http://darwin.zoology.gla.ac.uk/~rpage/LouseBase/2/>
- PCR Notebook (a tool for documenting lab based PCR reactions)
<http://www2.flmnh.ufl.edu/pdb/pcr/>
- DNA Primer Database
<http://www2.flmnh.ufl.edu/pdb/>
- Louse Checklist Database
<http://www2.flmnh.ufl.edu/adb/>
- PhPBib (Bibliographic database)
<http://darwin.zoology.gla.ac.uk/~rpage/phpbib/>
- DNA Barcoder
<http://darwin.zoology.gla.ac.uk/~plaid/>
- Specimen Image Database (SID)
<http://darwin.zoology.gla.ac.uk/~sid/>

Taxonomy Name Servers (TNS)

- Glasgow Taxonomy Name Server
<http://darwin.zoology.gla.ac.uk/~rpage/MyToL/www/index.php>
- Glasgow Taxonomic Search Engine (a taxonomy portal)
<http://darwin.zoology.gla.ac.uk/~rpage/portal/>
- uBio
<http://www.ubio.org/>
- ITIS
<http://www.itis.usda.gov/>

Transfer Schema and Communication Protocol Resources

- Darwin Core2
<http://darwincore.calacademy.org/>
- DiGIR
<http://digir.sourceforge.net/>
- ABCD schema
<http://www.bgbm.org/TDWG/CODATA/>
- BioCase protocol
<http://www.biocase.org/default.shtml>
- GBif Project
<http://www.gbif.org/>
- Taxonomic Databases Working Group
<http://www.tdwg.org/>
- BioMoby (for data server registration)
<http://www.biomoby.org/>

Examples of Data Discovery Interfaces

- Seamark Demo Site
<http://www.siderean.com/medidemo.jsp>
- mSpace
<http://mspace.fm/>
- Flamenco Search Interface
<http://bailando.sims.berkeley.edu/flamenco.html>

Resources of Potential Relevance to BioCorder

- Resource Description Framework (RDF)
<http://www.w3.org/RDF/>
- Glasgow Taxonomy Name Server Portal
<http://darwin.zoology.gla.ac.uk/~rpage/LouseBase/2/>
- uBio
<http://www2.flmnh.ufl.edu/pdb/pcr/>
- ITIS
<http://www2.flmnh.ufl.edu/pdb/>
- Nomenclator Zoologicus
<http://uio.mbl.edu/NomenclatorZoologicus/>

Related Projects of Relevance to BioCorder

- Web LSID Resolver
<http://lsid.biopathways.org/resolver/>
- Drupal (web content management platform – distributed authentication?)
<http://www.drupal.org/>
- Creative Commons (for data and image licensing)
<http://creativecommons.org/>
- Science Environment for Ecological Knowledge (SEEK)
<http://seek.ecoinformatics.org/>
- Morphbank
<http://lsid.biopathways.org/resolver/>
- BioImage04 Workshop
<http://www.drupal.org/>
- CIPRes Project
<http://www.phylo.org/index.html>
- Online Map Creation (OMC)
<http://www.aquarius.geomar.de/>