Putting it all together

Technological integration at the global level

Donald Hobern
GBIF Program Officer for Data Access and Database Interoperability

February 2005
GBIF and TDWG

**GBIF – Global Biodiversity Information Facility**
- Megascience activity involving 42 countries/economies and 28 international organisations
- Secretariat based in Copenhagen, Denmark
- Mission
  - Free and universal access to world’s biodiversity data via Internet
  - Sharing primary biodiversity data for society, science and a sustainable future
- Products
  - Registry of biodiversity data resources
  - Index of biodiversity data
  - Software tools
  - Web portals (http://www.gbif.net) and data services

**TDWG – Taxonomic Databases Working Group**
- Not-for-profit scientific and educational association
- Affiliated to the International Union of Biological Sciences
- Mission
  - To provide an international forum for biological data projects
  - To develop and promote the use of standards
  - To facilitate data exchange
- Products
  - Standards/guidelines for recording/exchanging data about organisms
  - Promotion of use of these standards
  - Forum for discussion (especially annual meeting)
Data resources

Resources vary in interoperability and scope for automated integration

• Images, audio, etc.
  • Require text (metadata) for interpretation
  • Very hard for software to interpret automatically
  • (Normally) link to resource as a unit (e.g. via URL in web page)

• Unstructured text documents (including e.g. PDF, Word, HTML)
  • Can be indexed for full-text search (Google)
  • Hard for software to interpret automatically
  • Normally link to document as a unit (e.g. URL in web page)

• Structured XML documents
  • Can be indexed for full-text search
  • Can be parsed and interpreted by software tools (for known schemas)
  • Can link to document as a unit or to fragments of a document
  • Can use structure to develop highly-specific searches
  • Normally process whole documents (not random-access)

• Databases
  • Can be exposed as structured XML documents
  • Can process any subset of the data (random-access selection of specific records)
Structured XML data

Observation record formatted using the Darwin Core

<?xml version="1.0" encoding="UTF-8"?>
<response>
  <record>
    <darwin:DateLastModified>2003-06-08</darwin:DateLastModified>
    <darwin:InstitutionCode>DGH</darwin:InstitutionCode>
    <darwin:CollectionCode>DGH Lepidoptera</darwin:CollectionCode>
    <darwin:CatalogNumber>DGHEUR_0002976</darwin:CatalogNumber>
    <darwin:ScientificName>Dichomeris marginella (Fabricius, 1781)</darwin:ScientificName>
    <darwin:BasisOfRecord>O</darwin:BasisOfRecord>
    <darwin:Kingdom>Animalia</darwin:Kingdom>
    <darwin:Order>Lepidoptera</darwin:Order>
    <darwin:Family>Gelechiidae</darwin:Family>
    <darwin:Genus>Dichomeris</darwin:Genus>
    <darwin:Species>marginella</darwin:Species>
    <darwin:ScientificNameAuthor>(Fabricius, 1781)</darwin:ScientificNameAuthor>
    <darwin:IdentifiedBy>Donald Hobern</darwin:IdentifiedBy>
    <darwin:Collector>Donald Hobern</darwin:Collector>
    <darwin:YearCollected>2003</darwin:YearCollected>
    <darwin:MonthCollected>06</darwin:MonthCollected>
    <darwin:DayCollected>08</darwin:DayCollected>
    <darwin:ContinentOcean>Europe</darwin:ContinentOcean>
    <darwin:Country>Denmark</darwin:Country>
    <darwin:County>Københavns Amt</darwin:County>
    <darwin:Locality>Merianvej, Hellerup</darwin:Locality>
    <darwin:Longitude>12.538</darwin:Longitude>
    <darwin:Latitude>55.737</darwin:Latitude>
    <darwin:CoordinatePrecision>100</darwin:CoordinatePrecision>
    <darwin:IndividualCount>1</darwin:IndividualCount>
    <darwin:Notes>1 in Skinner trap</darwin:Notes>
  </record>
</response>
Some TDWG Data Standards

Darwin Core
- Simple XML data model to represent taxon occurrence records (only core attributes)
- Extensions to handle e.g. curation details, microbial data, image data

ABCD Schema – Access to Biological Collection Data
- More complex XML data model to represent collection or observation data
- Detailed document structure including features for different communities

DiGIR – Distributed Generic Information Retrieval
- XML protocol for searching remote data resources
- Suitable for use with a wide range of different data models

BioCASe Protocol
- XML protocol for searching remote data resources with more complex schema (e.g. ABCD)
- Derived from DiGIR – new unified DiGIR/BioCASe protocol being developed

Taxon Concept Schema
- XML data model currently under development for exchange of nomenclatural/taxonomic data
- First version to be used for implementation in 2005

SDD Schema – Structured Descriptive Data
- XML data model for descriptive data relating to taxa or specimens (highly generalised)
- Suitable for representation of character tables, diagnostic keys, etc.
Data navigation
Integration strategies

Different approaches to integrating distributed information:

• **Data Warehouse**
  - Bring all data together
  - Data owners can still maintain control over their data
  - Single agreed central data model
  - All data in one place → any complex query can be constructed

• **Federated Network**
  - Data remains with original owners
  - Data owners can manage data as best suits them
  - Agreed data models for exchange
  - Queries can become expensive as number of resources increases

• **Indexed Network**
  - Data remains with original owners
  - Data owners can manage data as best suits them
  - Agreed data models for exchange
  - Most frequently-issued queries can be handled in index

• Selection may be influenced by community culture

• Network may include warehouses/indices for specific areas

• Key driver should be user requirements (use cases)
Central services

GBIF plans to offer the following central services:

- **Service Registry**
  - discover data resources

- **Data Index and Aggregated Data Services**
  - find distributed data

- **Schema Repository**
  - interpret structured data

- **Feedback Mechanisms**
  - add value by connecting users to providers

- **Data Validation Services**
  - report potential inconsistencies within data

- **Usage Reporting Mechanisms**
  - recognise providers for sharing data

- **Globally Unique Identifier Services**
  - provide persistent references to data records
Distributed data access

<table>
<thead>
<tr>
<th>Presentation &amp; Analysis</th>
<th>GBIF Global Portal</th>
<th>Thematic Portal</th>
<th>Regional Portal</th>
<th>Software Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indexing</td>
<td>GBIF Data Index</td>
<td>Entrez</td>
<td>OBIS</td>
<td></td>
</tr>
<tr>
<td>Discovery</td>
<td>GBIF Registry</td>
<td></td>
<td>BCoL?</td>
<td>Network-specific Registries</td>
</tr>
<tr>
<td>Access</td>
<td>XML</td>
<td>XML TCS</td>
<td>XML ABCD DwC</td>
<td>XML SDD</td>
</tr>
<tr>
<td></td>
<td>XML</td>
<td>XML</td>
<td>XML</td>
<td>XML TaXMLit</td>
</tr>
<tr>
<td></td>
<td>URL</td>
<td>URL</td>
<td>URL</td>
<td>URL</td>
</tr>
<tr>
<td>Resources</td>
<td>Sequences / Barcodes</td>
<td>Taxonomy / Nomenclature</td>
<td>Specimens</td>
<td>Descriptive Data</td>
</tr>
</tbody>
</table>

Structured Databases

Unstructured Documents

Structured XML

HTML webpage

PDF

Image
Problems to solve – data served from multiple locations

- Different networks of providers will serve up the same records through different routes, and software tools will not be able to detect all duplicate records.
- Aggregator nodes may in many cases themselves be the sole provider for other important records.
- The same situation will also occur e.g. if a taxonomist serves data for all specimens from many institutions for a single genus, including some that may also have been digitised by the institutions themselves.
Problems to solve – referring to data from outside network

- Data records have no stable URLs on the network (even using constructed DiGIR queries)
- Provider endpoints may change, and specimens may move to different institutions
- There needs to be a reliable way to reference each data record from other web sites and in external media
- The solution needs to handle situations in which data providers move to new locations
- It should ideally also be possible to refer to a particular version of each record in case of modifications
Problems to solve – referring to taxon concepts

- No standardised way to reference taxonomic publications or taxon concepts to allow reliable machine processing of the relationships between them (is "P. Brown, 1949" the same as "Brown, 1949"?)
- Need mechanisms to allow data providers to refer to taxon concepts in a reliable and consistent way
Globally unique identifiers – requirements

- Universal reusable mechanism which can be applied to any type of biodiversity data record (specimen, collection, institution, taxon concept, taxonomic publication, image, etc.)
- Identifiers must refer uniquely to a single data element
- Users must be able to resolve each identifier to locate the data to which it relates
- Identifiers must be resolvable even if data changes ownership or the server is moved to a new endpoint
- Identifiers must be suitable for use by the wider life science community and others
- It should be as simple as possible for researchers
  - to create new identifiers
  - to discover existing identifiers associated with data items
  - to resolve identifiers to retrieve the associated data
Potential model – Life Science Identifiers (LSIDs)

Format

\[\text{urn:lsid:<domainName>::<namespace>::<objectId>[:<revisionId>]}\]

Example: referencing a PubMed article

Example: referencing first version of the 1AFT protein in the Protein Data Bank
\[\text{urn:lsid:pdb.org:pdb:1AFT:1}\]

Potential example: referencing specimen record in GBIF Network (identifiers assigned centrally)
\[\text{urn:lsid:gbif.net:Specimen:2706712}\]

Potential example: name record from IPNI
\[\text{urn:lsid:ipni.org:TaxonName:82090-3:1.1}\]

- Specification for Uniform Resource Name model from EMBL, IBM, I3C and OMG
- "A straightforward approach to naming and identifying data resources stored in multiple, distributed data stores in a manner that overcomes the limitations of naming schemes in use today"
- All LSIDs are formed from 5 or 6 colon-delimited components: \text{urn}, \text{lsid}, the \text{domain name} of the authority for the identifier, a \text{namespace} for the identifier, the \text{identifier} for the object within the namespace, and an optional \text{revision} number
Taxon concept identifiers could be managed in conjunction with a central registry for digital versions of works of biodiversity literature.

- Identifiers can be assigned for publications and concepts.
- The registry will have responsibility for identifying publications which have already been registered and returning the same identifiers each time.
- The registry can provide linkage to any available digital versions of each publication (PDF, online journal, XML summary).

**Taxonomy Server A**

Concept: *Aus bus* sensu Smith, 2003
- *Aus bus* sensu Jones, 1965
- *Aus cus bus* sensu Black, 1926
- *Aus bus* sensu Brown, 1949
- *Aus dus* sensu Brown, 1949

**Taxonomy Server B**

Concept: *Xus bus* sensu J. Owen, 2002
- *Aus bus* sensu P. Brown, 1949

Concept: *Yus dus* sensu J. Owen, 2002
- *Aus dus* sensu P. Brown, 1949

**Digital Library Server C**

Jones, A., 1965: "The genus Aus"

[Link to paper](http://www.diglib.org/jones.pdf)

**User**

1. Register LSIDs for publications and concepts (all new)
2. Register LSIDs for publications and concepts (some matched from existing records)
3. Register digital (non-structured) version of a taxonomic work (matches existing record)
4. Request data for *Aus bus* and find reference to *Aus bus* sensu Jones, 1965 with LSID for publication
5. Request available digital versions for LSID um:lsid:gbif.net:Pub:5 and get PDF from diglib.org
Web-based taxonomy?

Presentation & Analysis
- GBIF Global Portal
- Thematic Portal

Indexing
- GBIF Data Index

Discovery
- GBIF Registry

Resources
- Sequences / Barcodes
- Taxonomy / Nomenclature
- Specimens
- Descriptive Data
- Structured XML
- HTML webpage
- PDF
- Image

Databases
- Structured

Documents
- Unstructured

Taxonomic Collaboration Environment
- Literature
- Sequences / Barcodes
- Taxonomy / Nomenclature
- Specimens
- Characters

Group data
- Network data

Group resources
- Network resources

Taxonomist input