From Ink to Electrons
Issues to be Considered
Larry Speers
Agriculture and Agri-Food
Canada
Mammal Networked Information System (MaNIS)
Ornithology Networked Information System (ORNIS)
Herpetology Network (HerpNet)
Australian Virtual Herbarium (AVH) (Hispid)
Nationaal Herbarium Nederland Data Guidelines
What do we mean by digitization?

- Digitization – digitally recording the label data associated with a specimen (Databasing)
- Digitization - the creation of an electronic image of a specimen (Imaging)
  - Note
    - While images are useful, without the supporting information their value is limited
    - Imaging of labels and/or specimens can be an integral part of the databasing process
What databasing collections will not do!

- It **will not** save you money
- It **will not** relieve you of all hand work with the collection
- It **will not** cure the inadequacies of your data
Why database your collection?

- Discovering important specimens
- Enabling the data to be studied in different ways
- Enhancing curatorial activities
- Detecting problems
- Saving curatorial time
- Facilitating efficient information access
- Wider dissemination of Data
  - Contributing to the Global Meta-collection
  - Enhancing the ability of your institution to contribute in areas beyond its traditional remit
The information associated with the specimens in the world’s Natural History Collections is under utilized and as a result collections are under resourced because potential users have no easy way of accessing this information!
USES OF PRIMARY SPECIES-OCCURRENCE DATA

Arthur D. Chapman

Abstract:
This paper examines uses for primary species-occurrence data in research, education and in other areas of human endeavour, and provides examples from the literature of many of these uses. The paper examines not only data from labels, or from observational notes, but the data inherent in museum and herbarium collections themselves, which are long-term storage receptacles of information and data that are still largely untouched. Projects include the study of the species and their distributions through both time and space, their use for education, both formal and public, for conservation and scientific research, use in medicine and forensic studies, in natural resource management and climate change, in art, history and recreation, and for social and political use. Uses are many and varied and may well form the basis of much of what we do as people every day.
Uses of Primary Species Data

- **TAXONOMY**
  - Taxonomic Research
  - Floras and Faunas
  - Field Guides
  - Integrated electronic resources
  - Check lists and inventories
  - Image Databases
  - Phylogenies
  - Parataxonomy
  - Automated Identification Tools

- **BIOGEOGRAPHIC STUDIES**
  - Distribution Atlases
  - Species Distribution Modelling
  - Predicting new species distributions
  - Studying species decline

- **SPECIES DIVERSITY AND POPULATIONS**
  - Species Diversity, Richness and Density
  - Population Modelling — Population Viability Analysis
  - Species Inter-relations
  - Protecting Communities
Uses of Primary Species Data

- **LIFE HISTORIES AND PHENOLOGIES**
  - Life History Studies
  - Phenology

- **ENDANGERED, MIGRATORY AND INVASIVE SPECIES**
  - Endangered Species
  - Invasive species and translocation studies
  - Migratory Species

- **IMPACT OF CLIMATE CHANGE**
  - On Native Species
  - On Primary Production
  - Desertification

- **ECOLOGY, EVOLUTION AND GENETICS**
  - Vegetation Classification
  - Mapping Vegetation
  - Habitat loss
  - Ecosystem function
  - Survey Design - Finding the Gaps

- **Evolution, Extinction and Genetics**
  - Microbial diversity and speciation
  - Archaeological studie

- **ENVIRONMENTAL REGIONALISATION**
  - National Planning studies
  - Regional Planning Studies
  - Marine Regionalisations
  - Aquatic Regionalisations
Uses of Primary Species Data

- **CONSERVATION PLANNING**
  - Rapid Biodiversity Assessment
  - Identifying Biodiversity Priority Areas
  - Reserve Selection
  - Complementarity
  - Ex-situ Conservation
  - Sustainable Use
  - Seed Banks and Germplasm Banks

- **NATURAL RESOURCE MANAGEMENT**
  - Land Resources
  - Water Resources
  - Environment Protection
  - Environmental Monitoring

- **AGRICULTURE, FORESTRY, FISHERIES AND MINING**
  - Agriculture
  - Forestry
  - Fishing
  - Nursery and Pet Industry
  - Mining

- **HEALTH AND PUBLIC SAFETY**
  - Diseases and disease vectors
  - Bioterrorism
  - Biosafety
  - Environmental Contaminants
  - Antivenoms
  - Parasitology
  - Safer Herbal Products
 Uses of Primary Species Data

- **BIOPROSPECTING**
  - Pharmaceuticals
  - Mining and Pollution

- **FORENSICS**

- **BORDER CONTROL AND WILDLIFE TRADE**
  - Border Controls and Customs
  - Quarantine
  - Wildlife Trade

- **EDUCATION AND PUBLIC OUTREACH**
  - School level education
  - University level education
  - Training of Parataxonomists
  - Public awareness
  - Museum displays
  - Image Databases
  - Public Participation Programs
  - Tree of Life

- **ECOTOURISM**
  - Valuing Ecotourism
  - Training Guides and Operators
  - Guide Books
  - Gardens, Zoos, Aquariums and Wildlife Parks
Uses of Primary Species Data

- **ART AND HISTORY**
  - History of Science—Tracking Explorers and Collectors
  - Art and Science
  - Indigenous Art
  - Stamps

- **SOCIETY AND POLITICS**
  - Social Uses of Biodiversity
  - Anthropology and Language
  - Ethnobiology
  - Data Repatriation
  - Biodiversity collecting

- **RECREATIONAL ACTIVITIES**
  - Recreational fishing
  - Hunting
  - Photography and Film-making
  - Gardening
  - Bushwalking, Hiking and Trekking

- **HUMAN INFRASTRUCTURE PLANNING**
  - Risk Assessment
  - Landscaping
  - Wild Animals and Infrastructure
  - Building timbers

- **AQUATIC AND MARINE BIODIVERSITY**
The Natural History Meta-collection:

The whole is greater than the sum of the parts!
Butterflies of Canada - Provincial Collections

- Royal BC Museum
Butterflies of Canada - Provincial Collections

- Alberta Museum
Butterflies of Canada - Provincial Collections

- Sask. Museum
Butterflies of Canada - Universities
Butterflies of Canada - National Collection

Butterflies of Canada - Total Specimen Records

- Specimen Records
  197,000+ records
Primary Data on Species Distributions

It is estimated that approximately 2-3 billion specimens with associated locality and date of collection are held in the world’s natural history collections. Collectively, this massive information resource (the Meta-Collection) provides our most complete record of the biodiversity of our planet.
The formation of Canadensys is a significant step toward realizing the Canadian Digital Meta-Collection.
• 1 billion specimens manually digitized @ 10 min./specimen
• 166,666,667 hours
• 21 million days @ 8 hours /day with no lunch break
• 833,000 people working for 1 full year

James Macklin
Director of Collections and Informatics
Harvard University Herbaria
How do we industrialize the process?

If we are going to succeed we need to consider:

- Quality control issues
- Economies of scale
- Automated processes
- Labor costs
- Community resources
- Workflow
Software selection and/or development?

Recognize that software is transitory!
3 Approaches to Consider

- Write your own
- Hire a consultant to customize especially for you
- Purchase a commercial package
Pros and Cons of Each Approach

- Write your own - Pros
  - Specific to your needs
  - No cash outlay
  - Consistent with other in-house applications

- Write your own - Cons
  - Hidden costs – time and resource
  - Maintainability
  - Fewer features
Pros and Cons of Each Approach

- **3rd Party Consultant – Pros**
  - High degree of customization
  - Adherence to guidelines
  - Maintainability

- **3rd Party Consultant – Cons**
  - High cash outlay
  - Dependence on outside consultant
  - Time investment for requirements
Pros and Cons of Each Approach

- **Commercial Package – Pros**
  - Less costly than 3rd party consultant
  - Maintainability
  - Feature rich

- **Commercial Package – Cons**
  - Dependent on software vendor
  - Not as specific to your needs
  - Cash outlay required
Questions-

• Is the software’s underlying data model robust?
• What is the software’s import capacity?
• What is the software’s export capacity?
• Can the export capacity be easily customized to meet the requirements of different data exchange standards?
Keys to a successful RDBMS

- Integrity of data
- Ease of use
- Scalability
- Interoperability
- Return on investment
  - Development/purchase cost
  - Integration into your workflow
  - Time & money saved
Know your requirements

- What is the intent of your database?
- How does the database fit into your workflow?
- Are there any budget constraints?
- What are your hardware requirements?
- What software exists in your organization?

Who are the subject matter experts?
What types of users?
What applications will access this database?
Understand how the data you choose to track will be used
Know your requirements

- Single vs. multi-user
- User skill level
- Local or remote access
- Security
- Backup strategies
- Conversion of older data
Know your future requirements

- What are the future plans for this database?
  - Size
  - Scope
  - Accessibility
  - Integration with other database systems
Invest the time to implement data integrity. I can’t over-emphasize the importance of this tip. Many of the problems you’ll encounter with inaccurate or erroneous information will be a direct result of poor data integrity. While it sometimes seems like a waste of time to pay so much attention to the many details involved in establishing data integrity, it will actually save you an enormous amount of time in the long run – you won’t have to continually go back to fix things. An interesting fact is that the very people who “…just don’t have the time.” to establish proper data integrity are the ones who usually spend a large amount of time fixing their improperly designed databases. In many cases they will spend up to three times the amount of time it would have taken to design the database properly in the first place! So don’t do it over – do it right!

A multi-part field is a field whose value can be divided into smaller parts, while a multi-valued field contains multiple occurrences of the same type of value.

First normal form: All column values are atomic and non-repeating
Second normal form: All column values depend on the value of the primary key.
Third normal form: No column value depends on the value of any other column except the primary key.
Summary

- Know your requirements
- Select the implementation approach
- Design/choose the database according to your requirements
- Use the database to enforce data integrity
Choosing a Collections Management System

- **Biotica 4.1**
- **Biota**
- **MANTIS**
- **Herbar**
- **BRAHMS**
- **Specify**
- **KE EMu**
- **BioLink**
Quality as applied to data, has various definitions but in the geographic world one definition is now largely accepted - that of “fitness for use” (Chrisman 1983).
Fitness for use

In a database, the data have no actual quality or value; they only have potential value. That value is realized only when someone uses the data to do something useful (English 1999). The quality of data cannot be assessed independently of the users of that data (Strong et al. 1997).
Fitness for use

Data are of high quality if they are fit for their intended use in operations, decision-making, and planning.

(Juran 1964)
What do we mean by “fitness for use”? 

Fitness for use

- Does species ‘x’ occur in Tasmania?
- Does species ‘x’ occur in National Park ‘y’
Documenting Fitness for Use

- In general, error must not be treated as a potentially embarrassing inconvenience, because error or uncertainty provides a critical component in judging fitness for use.

- Uncertainty measures for georeferencing
  - HerpNet Workshop

- Uncertainty measures for taxonomic determinations
“During the revision of *Euscelididia*, a frightening proportion of the borrowed “determined” material was found to be misidentified (62–73%), and a literature search in a *BIOSIS Previews* revealed that the problem is widespread.”

*Meier & Dikow* Conservation Biology, Pages 478–488 Volume 18, No. 2, April 2004
Problem: Misidentification

“For example, of the 1522 rove beetle specimens (Staphylinidae: Coleoptera) in the Struve collection 262 (17%) were misidentified (Rose 2000), and Papp (1978) reports that for a collection of Hungarian Lauxaniidae (Diptera) 28 of the 74 species determined and labeled by Szilády were consistently misidentified.”
“In Euscelidida 13% of all borrowed specimens were classified under an incorrect name, and for a recent inventory of palm collections in botanical gardens, 260 (22%) of the submitted 1208 names were synonyms and 46 (4%) were invalid (Maunder et al. 2001).”
Documenting Taxonomic Determinations

- Several methods exist for documenting taxonomic determinations - none are completely satisfactory
  - Herbarium Information Standards and Protocols for the Interchange of Data (HISPID)
  - Australian National Fish Collection (1993)
  - Several others restricted to one or two institutions

- Proposal – four level:
  - Who determined the specimen and when
  - What was the determination based on: (type specimen, local flora, monograph, etc.)
  - Level of expertise of the determiner
  - What confidence did the determiner have in the determination.
PRINCIPLES OF DATA QUALITY

Arthur D. Chapman

Although most data-gathering disciplines treat error as an embarrassing issue to be
avoided, the very existence of errors in data deserves closer attention and public
understanding. Hence error provides a
critical component in judging fitness for use.
(Chapman 1991)

1. Australian biodiversity information services
P.O. Box 741, Bowen, QLD 4805, Australia
email: aubbi@gbif.org

http://www.gbif.org/prog/digit/data_quality/data_quality
## Principles of Data Quality - Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Definitions</td>
<td>3</td>
</tr>
<tr>
<td>Principles of Data Quality</td>
<td>8</td>
</tr>
<tr>
<td>Taxonomy and Nomenclature—the Classification Domain</td>
<td>19</td>
</tr>
<tr>
<td>Field Data—the Spatial Domain</td>
<td>24</td>
</tr>
<tr>
<td>Field Data—the Collection Domain</td>
<td>26</td>
</tr>
<tr>
<td>Descriptive Data—the Description Domain</td>
<td>27</td>
</tr>
<tr>
<td>Capturing Primary Species Data</td>
<td>28</td>
</tr>
<tr>
<td>Data Entry and Acquisition</td>
<td>31</td>
</tr>
<tr>
<td>Documenting Data—the Metadata</td>
<td>33</td>
</tr>
<tr>
<td>Storage of spatial data</td>
<td>38</td>
</tr>
<tr>
<td>Manipulation of spatial data</td>
<td>37</td>
</tr>
<tr>
<td>Representation and Presentation</td>
<td>38</td>
</tr>
<tr>
<td>Conclusion</td>
<td>49</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>50</td>
</tr>
<tr>
<td>References</td>
<td>51</td>
</tr>
</tbody>
</table>
Issues with QA/QC

- **Legacy Data**
  - Need to deal with what we have
  - Data cleaning tools

- **New data**
  - Do everything in our power to avoid the problems we find with today’s legacy data
Assign responsibility for the quality of data to those who create them. If this is not possible, assign responsibility as close to data creation as possible.

(Redman 2001)
Loss of data quality

- Loss of data quality can occur at many stages:
  - At the time of collection
  - During digitization
  - During documentation
  - During storage and archiving
  - During analysis and manipulation
  - At time of presentation
  - And through the use to which they are put
Data Cleaning

A process used to determine if data are inaccurate, incomplete, or unreasonable and then improving the quality through correction of detected errors and omissions.
PRINCIPLES AND METHODS OF DATA CLEANING

PRIMARY SPECIES AND SPECIES-OCURRENCE DATA

Arthur D. Chapman

Error qua non necessum, apprehendes
An error not resisted is approved.
(Rud. Ditt. & Stud. c. 770).

Keywords:
Data Cleaning, Data Editing.

1Australian biodiversity information services
PO Box 7605, Toowoomba South, QLD, Australia
email: papeco.gbif@arapaima.org

http://www.gbif.org/prog/digit/data_quality/data_cleaning
Staffing.

- Who will do the digitisation?
  - Curatorial staff as part of their regular work.
  - External contract staff/company
  - Volunteer staff?
  - Visiting researchers?
  - Project staff?
Workflow

• assign the different work phases to DIFFERENT people according to the needed skills, costs and possible location of the work
• atomize the work as much as needed/possible
• maintain quality control in all phases
• constantly think about lowering costs without sacrificing the quality/usage
INVOLVED PROCESSES

Bar Coding

Imaging

File Naming

Digitisation from the image
Espeletia brachyaxantho Díaz
subsp. pescana Díaz

Paratypus

Tallo aéreo largo de 4.50 cm. Hojas en roseta erguidas. Inflorescencias con lígulas amarillas, flósculos carmelitos.

COLOMBIA

Depto. Boyacá: Mpio. Pesca; Vereda La peña; Páramo de la Cortaderas. Las Piedras. Alt. 3.750 m.

Трищоргое голубик.474етс
хасан 25.4.8.
коксеба 86
OSUC 0173012
Canadian National Collection
No.: 53-404
Loc.: Juneau, Alaska.
Host: Zonotrichia coronata
Date: 5-V-53 #393
Coll.: R.B. Williams

Canadian National Collection
No.: S-67-63
Loc.: steppes E. Transbaikalia
Host.: U.S.S.R.
Date: nest of sparrow
Coll.:

No.: Ceratophyllum gallinaceae
Det.: G.P. Holland
(C. Schrank)
Automated processes

- File name manipulation
- Image file manipulation software
- Voice recognition software
- Georeferencing tools
Data capture or data interpretation?

- *Data Capture* - Record the data presented on the specimen as written.
- *Data Interpretation* - Alter the data to correct errors, such as incorrectly naming the specimen.
Data Capture Guidelines

Data Guidelines - Collection Data Registration at the Nationaal Herbarium Nederland (March 2007) L.P.M. Willemse / J.B. Mols

http://www.nationaalherbarium.nl/virtual/Data-guidelines-NHN.pdf
3. Collector and number

Field name: COLLECTOR
The name of the person who made the collection (the primary collector), and whose collection number is cited in the field holding the collector’s number.

Figure: In the example the collector name is: Sidiyasa, K.
Figure: In the example the collection was collected on the 13th of May 1998.

In BRAHMS we fill:

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>COLLDD:</td>
<td>13</td>
</tr>
<tr>
<td>COLLMM:</td>
<td>05</td>
</tr>
<tr>
<td>COLLYY:</td>
<td>1998</td>
</tr>
<tr>
<td>DATEINFO:</td>
<td>empty</td>
</tr>
</tbody>
</table>
Dealing with Sensitive Data

GUIDE TO BEST PRACTICES FOR GENERALISING

SENSITIVE SPECIES OCCURRENCE DATA

Arthur D. Chapman and Oliver Grafton
3.1. Is the content and detail of the biodiversity data such that their release would enable someone to carry out a harmful activity upon the taxon or attribute?

<table>
<thead>
<tr>
<th>Yes:</th>
<th>Document using statement 3a with supporting rationale. Go to 3.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>No:</td>
<td>[Data are not sensitive] Document using statement 3b with supporting rationale Go to 4</td>
</tr>
</tbody>
</table>

3.2. Is information already in the public domain, or already known to those individuals or groups likely to undertake the harmful activity?

<table>
<thead>
<tr>
<th>Yes:</th>
<th>Document using statement 3d with supporting rationale. Go to 3.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>No:</td>
<td>Document using statement 3c with supporting rationale. Go to 3.3</td>
</tr>
</tbody>
</table>

3.3. Would disclosure damage a partnership or relationship (especially where the maintenance of which is essential to helping achieve a specific conservation objective)?

<table>
<thead>
<tr>
<th>Yes:</th>
<th>Document using statement 3e with supporting rationale. Go to 3.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>No:</td>
<td>Document using statement 3f with supporting rationale. Go to 3.4</td>
</tr>
</tbody>
</table>

3.4. Would disclosure allow the locations of sensitive features to be derived through combination with other publicly available information sources?

<table>
<thead>
<tr>
<th>Yes:</th>
<th>Document using statement 3g with supporting rationale. Go to 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>No:</td>
<td>Document using statement 3h with supporting rationale. Go to 4</td>
</tr>
</tbody>
</table>
The recommended method for generalisation is:

<table>
<thead>
<tr>
<th>Category</th>
<th>Sensitivity</th>
<th>Georeference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category 1</td>
<td>Extreme</td>
<td>Georeference not released or data may be released by watershed/ bioregion/ county, etc. with no georeference coordinates.</td>
</tr>
<tr>
<td>Category 2</td>
<td>High</td>
<td>Georeference rounded to 0.1 degree</td>
</tr>
<tr>
<td>Category 3</td>
<td>Medium</td>
<td>Georeference rounded to 0.01 degree</td>
</tr>
<tr>
<td>Category 4</td>
<td>Low</td>
<td>Georeference rounded to 0.001 degree</td>
</tr>
<tr>
<td>Not sensitive</td>
<td>Not sensitive</td>
<td>Georeference unrestricted.</td>
</tr>
</tbody>
</table>