

Data quality aspects in biodiversity informatics

3 July 2007

GCP Passport Data Quality
Improvement and Assessment Meeting
Rome, Italy

Helmut Knüpffer, IPK Gatersleben, Germany



GCP Passport Data Quality, Rome, 3 July 2007



Topics

- Intro
- Definitions
- Principles of data quality
- Biodiversity data categories
- Outlook



Intro: Genebanks in the context of biodiversity

Biodiversity collections data

- Different **Biodiversity collections** data describe very similar data objects.
- **Preserved reference collections**, such as those in museums and herbaria.
- **Living collections**, like botanical and zoological gardens, aquaria, seed banks, microbial strain cultures and tissue collections.
- **Data collections**, such as surveys of objects in the field, e.g. floristic observations.

These collections have most of their **attributes in common**, although the **terminology** used to describe them may **differ substantially**.

[<http://www.bgbm.org/TDWG/CODATA/ABCD-Evolution.htm>]



Genebanks in the context of biodiversity

- **Genebanks are biodiversity collections**
- Genebank collections can be compared with botanical garden collections (living) or herbaria (archive)
- *in situ* & on farm – cf. floristic observations (species occurrence in space & time, without reference specimen)
- → Existing information technology approaches for biodiversity collections can be applied also for genebanks



Genebanks in the context of biodiversity

Data categories

- **“Collection level” data**

Metadata about genebanks (collection-holding institutions) and the germplasm collections they hold

e.g. FAO WIEWS, IPGRI Germplasm Holdings DB

- **“Taxon level” data**

Data related to species or other taxa

e.g. Mansfeld DB, GRIn-Taxonomy, Flora Europaea

- **“Unit level” data**

Accession level data for germplasm collections. Genebank accessions share many properties and attributes with other biodiversity specimens

→ e.g. passport data



Data Standards

- Standards and technology for data exchange developed (almost) independently in PGR and biodiversity communities
- TDWG (**Biodiversity Informatics Standards**, formerly Taxonomic Databases Working Group), since early 1980s
 - to provide an international **forum** for biological data projects
 - to develop and promote the use of **standards**
 - to facilitate **data exchange**

www.tdwg.org

- GBIF (**Global Biodiversity Information Facility**), since early 2000s

is a mega-science project with the aim “to make the world’s primary data on biodiversity freely and universally available via the Internet”

www.gbif.org



- **Darwin Core 2 (DwC2)**

- element definitions designed "to support the sharing and integration of primary biodiversity data"
[<http://darwincore.calacademy.org/>]

→ "flat" structure with ca. 48 elements, cf. MCPD

- **Access to Biological Collections Data (ABCD) 2.06**

- "an evolving comprehensive standard for the access to and exchange of data about specimens and observations (ie. primary biodiversity data)"

[<http://www.bgbm.org/TDWG/CODATA/Schema/>]

→ PGR domain

→ preferred standard for PGR



ABCD Access to Biological Collections Data

- **ABCD** is a **common data specification** for data on biological specimens and observations (**including plant genetic resources collections, *in situ* occurrences**).
- The design goal is to be both **comprehensive** and **universal** (about 1200 elements).
- Support from **TDWG/CODATA**, ENHSIN, BioCASE, and GBIF.

[<http://www.bgbm.org/TDWG/CODATA/Schema/>]



Data quality in biodiversity informatics

Arthur C. Chapman, Principles of data
quality

GBIF online publication

TDWG annual meetings, eg 2005 St
Petersburg

Data quality aspects

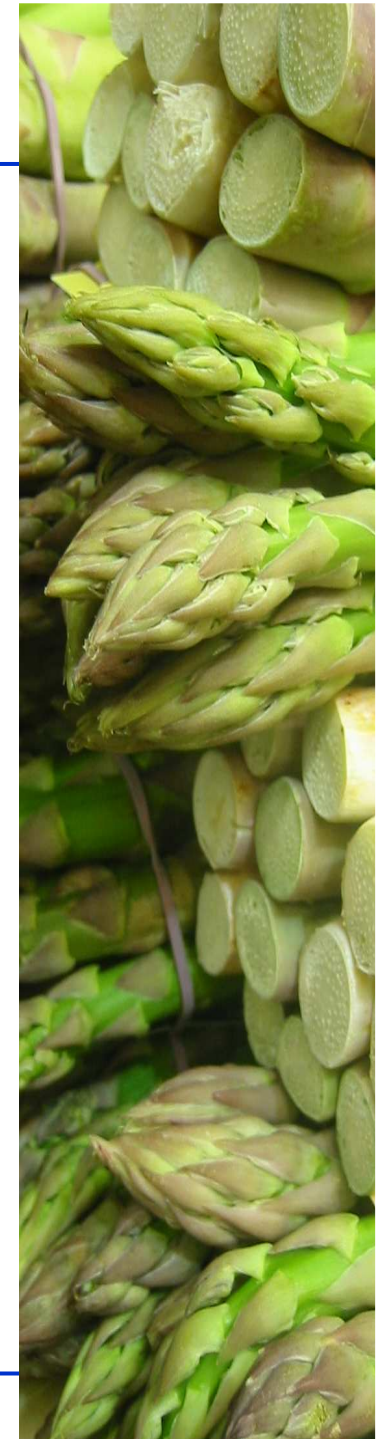
At all stages of data management process:

- Data capture/recording at time of gathering
- Data manipulation prior to digitisation
- Identification of specimen
- Digitisation
- Documentation of metadata
- Data storage & archiving
- Data presentation & dissemination
- Using data (analysis & manipulation)



- Quality – fitness for use
- Many other definitions

- Accuracy and precision
- Uncertainty
- Error – imprecision & inaccuracy
- Validation and cleaning
- Users – feedback



Principles of data quality

Organisations have to develop...

- Vision
 - Think about long-term data & information needs
 - Motivate actions in right direction
 - Sound basis for decision-making, ...
- Policy
 - Think broadly about quality, re-examine day-to-day practices
 - Formalise data management process
 - Provide users with confidence & stability, ...
- Strategy
 - Short-term (6-12 months, easy, immediate use)
 - Intermediate (ca. 18 months, simple in-house methods)
 - Long-term (longer time frame, more sophisticated, collaboration)



Principles of data quality

- Strategy should include
 - Not reinventing info management tools
 - Efficiencies in data collection & quality control procedures
 - Sharing information, data & tools
 - Use existing standards, or develop new robust ones in conjunction with others
 - Networks & partnerships
 - Reducing duplication in data collection & data quality control
 - Looking beyond immediate use, examining user requirements
 - Good documentation
- Prevention is better than cure



Principles of data quality

- The collector has primary responsibility
- The curator has core or long-term responsibility
- User responsibility
 - Feedback
- Building of partnerships
- Prioritisation
- Completeness
- Currency and timeliness
- Update frequency
- Consistency
- Flexibility
- Transparency
- Performance measures & targets
- Data cleaning



Principles of data quality

- Outliers
- Setting targets for improvement
- Auditability – keep track of checks & corrections made
- Edit controls (rules)
- Minimise duplication & reworking of data
- Maintenance of original data
- Categorisation can lead to loss of data & quality
 - e.g. conversion from coordinates to grid cells
- Documentation
- Feedback
- Education & training
- Accountability



Data categories: taxonomic & nomenclatural data

Poor taxonomic data can "contaminate" related areas of studies (Dalcin 2004)

- Taxonomic data:
 - Name (scientific, common, hierarchy, rank)
 - Nomenclatural status (synonym, accepted)
 - Reference (author, publication)
 - Determination (who, when)
 - Quality fields (accuracy, qualifiers)→ e.g. spell-checking using authority files (→Thomas)
- Recording accuracy of identification
 - Degree of certainty, level of expert
 - names derived from other than taxonomic expertise



Data categories: taxonomic & nomenclatural data

- Precision of identification
 - Compare two experts, names of duplicates in different collections
 - Taxonomic level (e.g. family, genus, species, subspecies)
- Bias – systematic error, e.g. by misinterpretation of key structure or use of inappropriate publication (flora for wrong area)
- Consistency – inconsistency by coexistence of two “accepted” names which are mutual synonyms
- Completeness
 - Of files (no records missing)
 - Of records (all fields known for each record)
- vouchers



Spatial data

- Georeferencing
 - BioGeoMancer project (GBIF)
tools expected online 2006
 - Itinerary project (Africa Museum, Belgium – GBIF)

- Quality checking of georeferenced data
 - Checking against other information in the record
(location vs. coordinates)
 - Checking against external reference
 - Outliers (geography, environment)
 - Spatial accuracy
 - False precision and accuracy



Collector and collection data

- Collection author(s) & collector's number(s)
 - Observers experience
 - Collection date/period
 - Collection method (particularly for observations)
 - Associated data
-
- Attribute accuracy
 - Consistency
 - Completeness



Descriptive data



Data entry and acquisition

- Basic data capture
- User interfaces
- Geo-referencing
- Error
 - Errors cannot completely avoided

Documenting data

- Metadata
 - The data must be documented with sufficient detailed metadata to enable its use by third parties without reference to the originator of the data
- Attribute accuracy
- Logical consistency
- Completeness
- Accessibility
- Temporal accuracy
- Documenting validation procedures
- Documentation and database design

Storage of data

- Backup of data
- Archiving
- Data integrity
- Patterns of error

Manipulation of spatial data

- Conversion of data from one format to another
- Datums and projections
- Grids
- Data integration

Representation and presentation

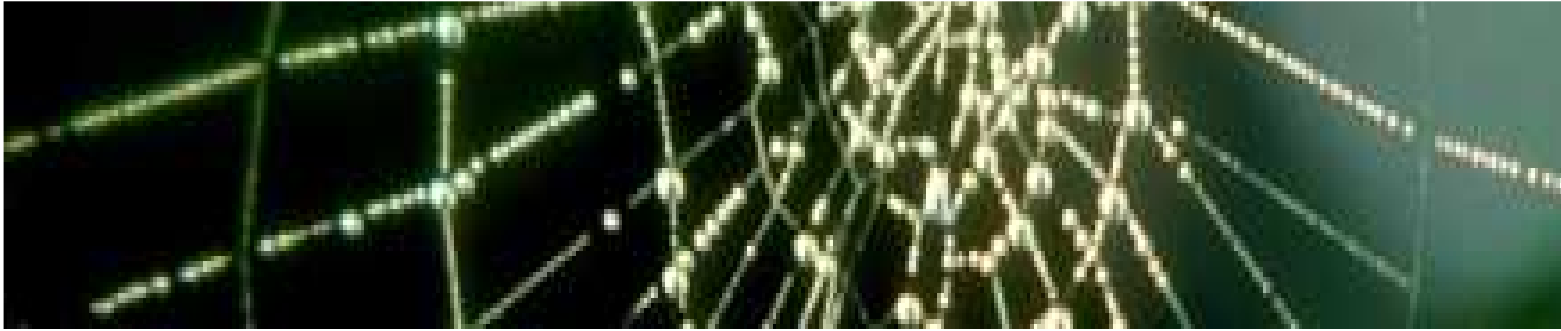
- Determining users' needs
- Relevancy
- Believability
- Living with uncertainty in spatial data
- Visualisation of error and uncertainty
- Risk assessment
- Legal and moral responsibilities
- Certification and accreditation
- Peer review of databases, especially for species databases

Conclusions

- In compilation of guidelines for PGR passport data quality – consider the „best practices“ and other guidelines for data quality in the biodiversity informatics domain
- There is a lot to learn from each other
- Training and capacity building among data creators, data curators, genebank curators, and users
- Feedback from users

acknowledgements

- Arthur D. Chapman
- GBIF & TDWG meeting participants



Thank you for your attention

