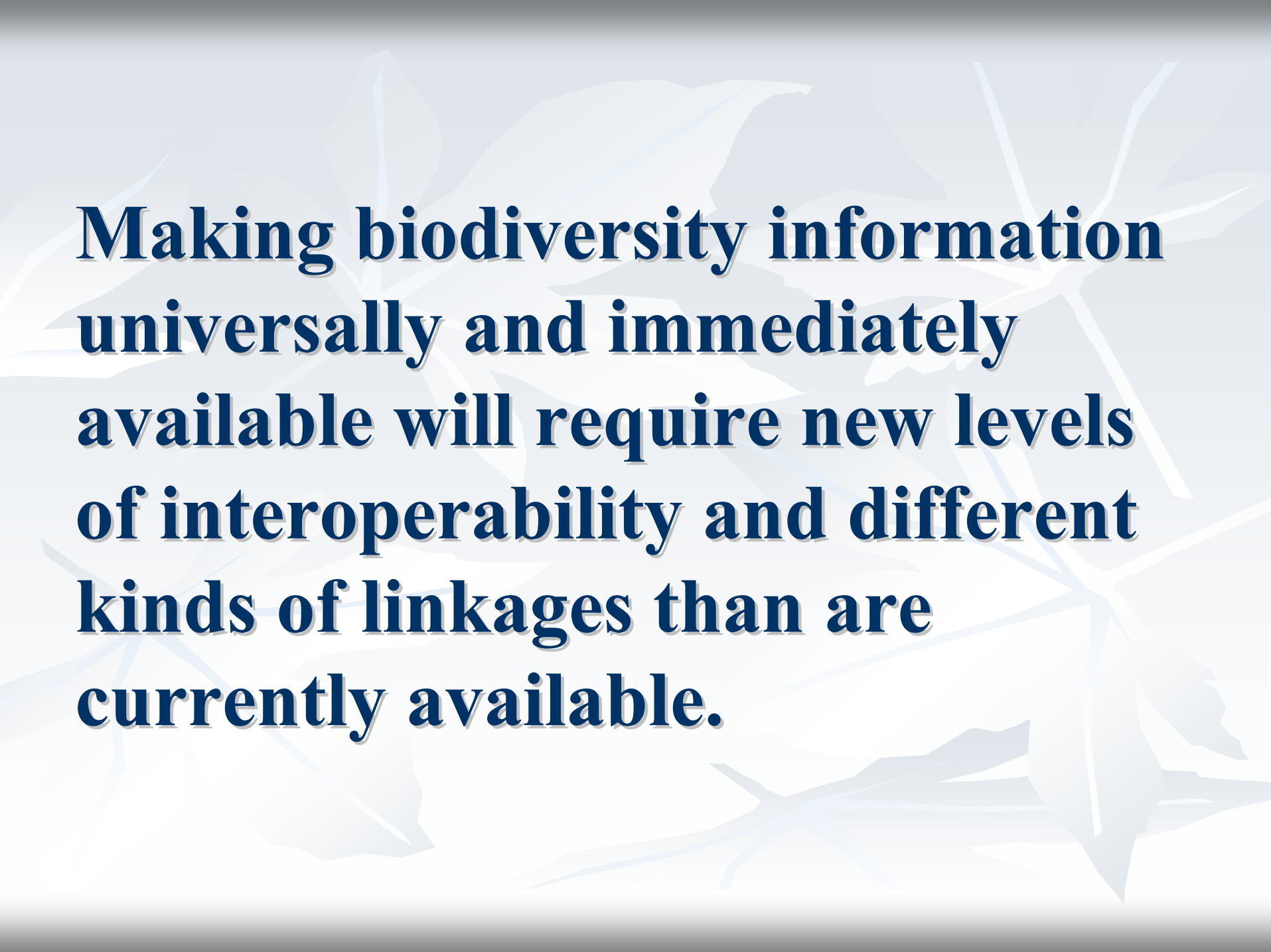# Data standards: objective data, subjective data, and data interchange

## Anna Weitzman & Chris Lyal

presented at the Taxonomic Databases Working Group meeting, Christchurch, New Zealand, 14 October 2004

THE NATURAL HISTORY MUSEUM

Smithsonian
*National Museum of Natural History*

**Making biodiversity information universally and immediately available will require new levels of interoperability and different kinds of linkages than are currently available.**

Context: our work on taXMLit schema and GBIF; discussions on how to parse the kinds of data, linkages, uses, and sources

How to subdivide the data in taxonomic publications so that they are:

- responsive to a variety of kinds of searches
- easily interoperable with other biodiversity data

In that process we considered a variety of links to authority files (e.g., specimen data, gazetteers, bibliographic lists)

We started to include many such in the schema:

- e.g., a link to a standardized or full form of the title of a book, journal, or author's name.

But, each such decision is subjective

As more and more interpreted or subjective data made their way into the schema, we realized that we were confusing the task of capturing original text.

We should capture data in taXMLit as they appear in the publication.

Any interpreted information should be added elsewhere.

This approach allows for a variety of things to be captured and attributed.

# Kinds of interpreted data

- Explicit links to authority files
- Explicit links to other data sources
- Other assumptions/interpretations:
  - opinions
  - hypotheses
  - clarifications
  - observations (about data not organisms)

# Links to authority files

- Links from literature to standardized lists of authors, collectors, publications and journals
- Links among major data sources (e.g. linking taXMLit with Darwin Core/ABCD, Linnaean core, TCS, SDD, Gutenberg core, geography, etc)
- Both automated links and *attributed* (explicitly linked reflecting expert observation) links are important

# Other assumptions/interpretations

## Opinions

- specimen identification
- application of georeference data (or inability to pinpoint) to an older collection where there is uncertainty in the locality
- full publication details for a partial citation

## Hypotheses

- taxon concepts

## Clarifications

- application of georeference data to a specimen where locality is clear
- full publication details for a partial citation

## Observations (about data not organisms)

- place where further information may be found

The more we have thought about this, the more we envision this kind of interpreted data as belonging in separate layers—

- within (explicitly),
- as extensions to schemas and data models, or
- as a linking layer between two or more schemas or data sets.

| data type | example primary elements | example interpreted elements |
|---|---|---|
| **published taxonomic literature** | • authorship of entire publication | • clarification of authorship |
| | • date(s) of entire publication | • clarification or change of date(s) |
| | • cited place of entire publication | • clarification of place of publication or abbreviation and/or link to authority file |
| | • name as cited | • correction to name spelling |
| | | • clarification or addition of name or combination status |
| | • name authorship | • correction to authorship |
| | | • link from abbreviation to full author details and standard form of name |
| | • infraspecific ranking | • change to infraspecific ranking in light of current Code (*Zoo.*) |
| | • cited publications as stated | • clarifications to cited publications as above |

| data type | example primary elements | example interpreted elements |
|---|---|---|
| **published taxonomic literature** | • specimen citations: stated locality | • modern locality (e.g. 'Myanmar' instead of 'Burma') or correction to stated locality (e.g. 'don't know' instead of 'San José')<br>• corrections to specimen citations as above in specimen databases<br>• actual linkage to specimen database records for individual specimen citations |
| | • name(s) of associated species | • correction to name(s) of associated spp. (either in specimen citations or in discussions; either by direct intervention or automated intervention through name server) |
| | • vernacular name, transliteration, language, locality of use, comments | • any of these could be clarified or reinterpreted in light of new work |
| | • various | • various factual corrections (page numbers, image citations, etc) |

| data type | example primary elements | example interpreted elements |
|---|---|---|
| **specimen collections** | <ul><li>Catalogue or accession number</li><li>collector(s)' name(s)</li><li>date collected (if provided in field notes or on label)</li><li>collection locality, habitat, etc, as given by collector(s)</li><li>observations made at the time of collection</li></ul> | <ul><li>clarification of collector(s)'s name(s)</li><li>date collected (as clarification, from itinerary, etc)</li><li>collection locality as interpreted/clarified by someone at a later date (e.g. post boundary changes, town name changes, etc)</li><li>addition of georeference</li><li>addition of altitude</li><li>observations, measurements, made at a later date</li><li>taxon name (with identifier and date)</li><li>type status</li></ul> |

| data type | example primary elements | example interpreted elements |
|---|---|---|
| **taxonomic databases** | <ul><li>name</li><li>basionym (original name) and other homotypic synonyms</li><li>author(s)</li><li>place and/or date of publication</li></ul> | <ul><li>change in spelling of name because of error in original (based on appropriate code)</li><li>taxon concept</li><li>addition or subtraction of heterotypic synonyms</li><li>a later change in authorship based on research</li><li>clarification of place and/or date of publication</li></ul> |

We believe that these kinds of data need to be more explicitly defined and labelled, as well as being a vital part of the discussion around UBIF and the Global Infrastructure and Network for Biodiversity Informatics.