Models for Integrating TDWG

Taxonomic Literature: Summary and Next Steps

Anna Weitzman
Smithsonian Institution
Christopher Lyal
Natural History Museum, London

TDWG LIT Group

- Literature Group of TDWG is responding to the need for agreed standards for taxonomic literature.
- Earlier work identified 3 key levels: microcitations, metadata and content. These have been modified to an extent over time and during this week.

“Microcitations”

- Further discussion on microcitations has teased these apart into
  - Subcitations – various references to specific portions text (pages, treatments, etc.) [or whatever R Pyle is calling them today]
  - Abbreviated citations – as in synonymy paragraphs where authors & publication names are abbreviated
- For any citation, a means is required to map to both the type of citation employed in bibliographies of modern taxonomic literature and library standards, including considerable metadata, such as MODS.
- Even the orthographic variants used in different library catalogues must be resolvable.

Resolution

- When attempting to resolve all abbreviated citations, the major issue is resolution of its parts.
- Issues include different representations of:
  - author names
  - book and journal titles
  - For both: abbreviation vs transliteration vs translation vs original orthography used in citations.
- Attempts at standardisation exist in published form for various fields
  - Botany standards are ‘existing’ TDWG standards, but not available for machine resolution
  - Zoology standards widespread and partial
    - Zoobank starting to generate
    - UBio has experience in generation
- These all require authority files to be agreed, with alternatives mapped together.

Authority files

- With mapping between all categories of citations/references and appropriate application of LSIDs, it will be possible to:
  - remove a major source of confusion
  - enable unambiguous navigation from the sparsest of citations, using author-employed abbreviations, to the full citation (and then to the full text if required)
  - allow navigation from on-line full text or citations in taxonomic works to library catalogues in order to find locations of printed text
- When marking up literature, many terms and their variants will be encountered – authority files, including all variants, will be required to enable meaning to be applied.

- This is important for all initiatives providing full text digitisation, such as Biodiversity Heritage Library (BHL).

Content/exchange standards

- Taxonomic literature accommodates many different data and information types, including those subject to existing TDWG standards.
  - Specimen / observation data
  - Taxon name data
  - Descriptions
  - Images
  - Ecological / interaction information
  - Geographic / distribution data
  - Bibliographic data
- Users may require navigation between these data.
- Users may require simultaneous access to data from literature and other sources

Currant bun or sticky bun?

Currant Bun – embedded schemas

- Don’t have to reinvent standards (can’t conflict with other standards)
- Expression of data and information in literature often differs from that included in other TDWG standards, and is not easily parsed into those standards.
- As other TDWG standards change, mark-up of literature must change as well to remain interoperable
- Other standards may be complex in ways not necessary in literature—requiring users to learn those complexities in order to apply a literature schema

Currant bun or sticky bun?

Sticky Bun

- Include all data within single schema
- Ensure interoperability between schemas, with a robust ontology to support this.

Are the costs worth it?

- Several goals interact in developing a standard:
  - interoperability across data and information types;
  - maximising cost-effective access to and display of information and data in a manner appropriate for the user;
  - cost-effective mark-up in agreed formats.
- Complexity of mark-up does not impact on search times – this is a product of content storage (native XML vs database).
- The degree of atomisation of the content will impact on:
  - cost of mark-up
  - ease of management as standards evolve
  - breadth of user needs that can be met
- Within a standard, level of atomisation may be optional

Meeting user needs

- Worth exploring potential uses of enhanced atomisation in mark-up:
  - Repurposing of content, enabling re-use of collected data by other users
  - Simpler discovery of latent information (taxonomic, systematic, ecological etc)
  - Literature-based information brought into workflows of biodiversity analysis in other sectors by presenting data in an assimilable fashion
  - Enabling integration of relevant data gathered in non-taxonomic publications through interoperability of schemas

- These uses will only be developed if users want to get engaged:
  - User needs analyses of current practice, use of emerging systems, and interface functionality
  - user-friendly navigation tools will be required

- Development of a literature standard must be in the context of user needs

SHAMELESS REMINDER

INOTAXA demo today at 1:15 and 3:30.