taXMLit: a vital piece of the puzzle for digitally interoperable taxonomy

Anna Weitzman & Chris Lyal

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

Smithsonian
National Museum of Natural History

THE
NATURAL
HISTORY
MUSEUM

Why do we need a schema?

What’s wrong with what we do now?

- Current system of publishing, storing and accessing taxonomic information is inefficient and hinders progress;
- Finding descriptions & taxonomic acts relies on user, or abstracting services, knowing where they are published and finding the relevant data;
- Literature is frequently not available in countries of origin;
- Probably no taxonomist has access to all relevant literature;
- Literature is held separately from the specimens to which it refers;
- Preparation of complete paper and publication time are rate-limiters in the taxonomic process

Why do we need a schema?

What would it look like if we started from scratch?

- All taxonomic data globally accessible with minimum (no) delay;
- Fully searchable, responsive to user-defined queries;
- Information on non-core taxa of a publication easily found
  - e.g. data on hosts of parasite being revised;
  - data ‘hidden’ by not being referred to in title, abstract or key words.
- Data users can give immediate feedback to institutions holding specimens;
- New data made available to users as they become available;
- Taxonomic changes added simply and clearly to the corpus;
- Changes seen immediately, without detailed search;
- Open to all contributors.

Why do we need a schema?

To do this we need to use XML

Static (HTML / pdf) pages:

- lack flexibility;
- Do not ‘interact’;
- Do not facilitate feedback;
- Essentially digital equivalents of analogue systems.

XML allows:

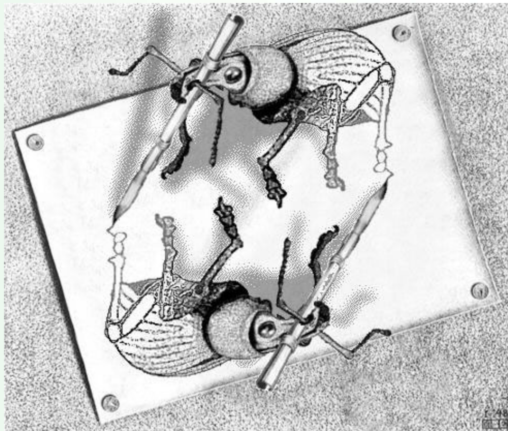
- Interoperability between different datasets;
- Distributed data sources;
- Easy searching through a data registry;
- Transmission of data through system by means of dynamic links;
- Equal accessibility to all marked data irrespective of where they are associated in source.

What are the necessary characteristics?

Interoperability

XML permits interoperability with other relevant data sources:

- Specimen data → ABCD and Darwin Core
- Names → TDWG standard
- References → standard to be agreed (MODS, ‘Gutenberg Core’….)



What are the necessary characteristics?

Accessibility of data

- Addition of new data possible;
- Data stored on servers (distributed system);
- Registry required (GBIF?);
- Useable data within treatments marked so that they can be accessed;
- Data can be downloaded to allow analysis

What are the necessary characteristics?

Addition of data

- Annotation and other comments possible;
- Data can be added at any scale;
- Comments to data provider possible;
- Links to collections and use of GUIDs permits changes at specimen level (e.g. uploading of images, comments on data) to be linked dynamically to treatment.
- (use of GUIDs may also allow collection managers to be updated on status of specimens)

Why do we need a standard?

Clarity of function

- Several schemas in use, serving different functions:
  - Characters
  - Publications
- Character schemas still require refinement, and may be valuable only within taxa;
- Publication-based schemas (may not give full interoperability with other datasets)

Flora of New Zealand – *http://floraseries.landcareresearch.co.nz*

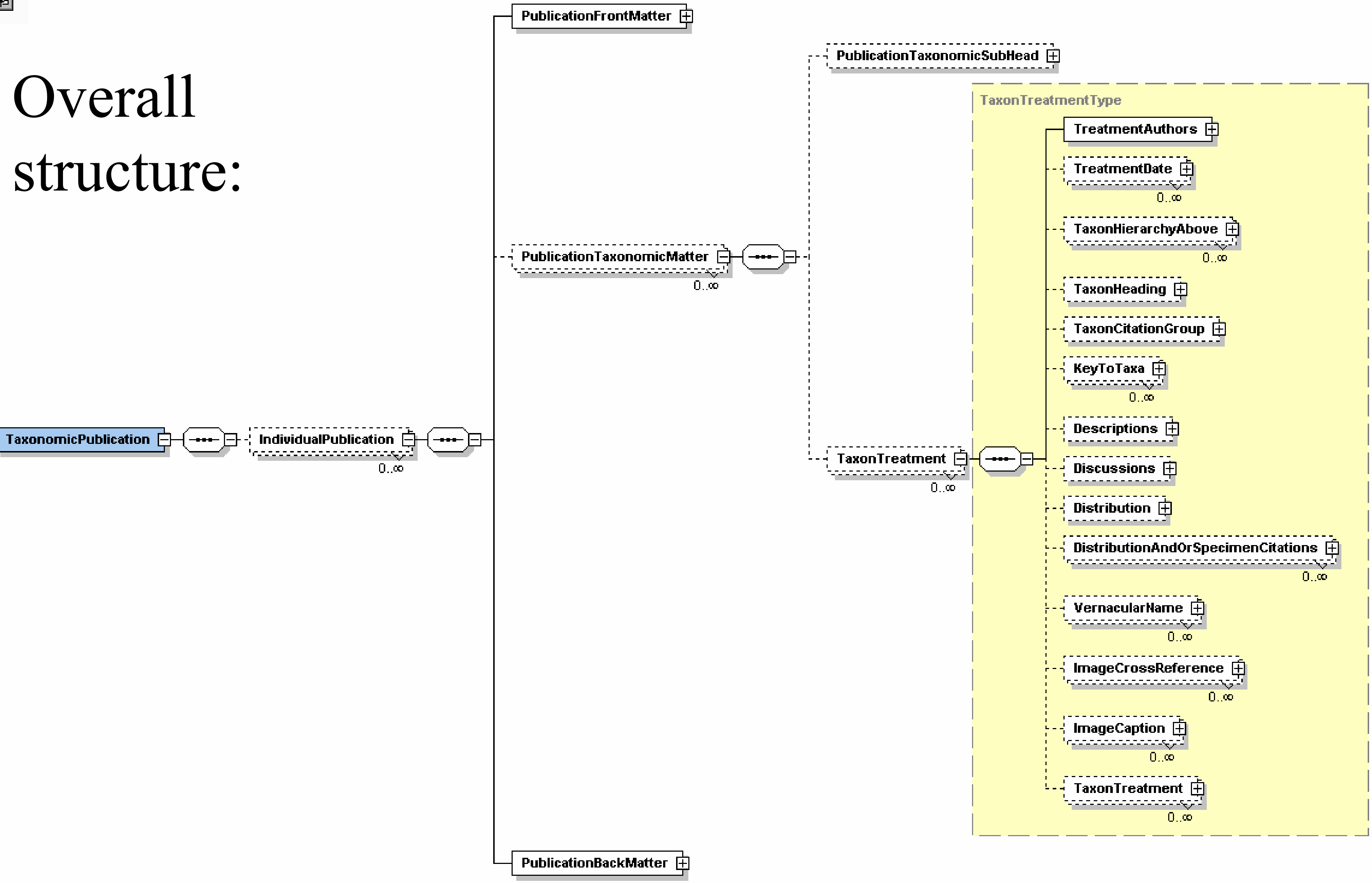
Flora of Australia -

http://www.deh.gov.au/biodiversity/abrs/online-resources/abif/flora/main/about.html

Flora Zambeziaca - *http://www.rbgkew.org.uk/floras/fz/intro.html*

We propose a standard for taxonomic literature:

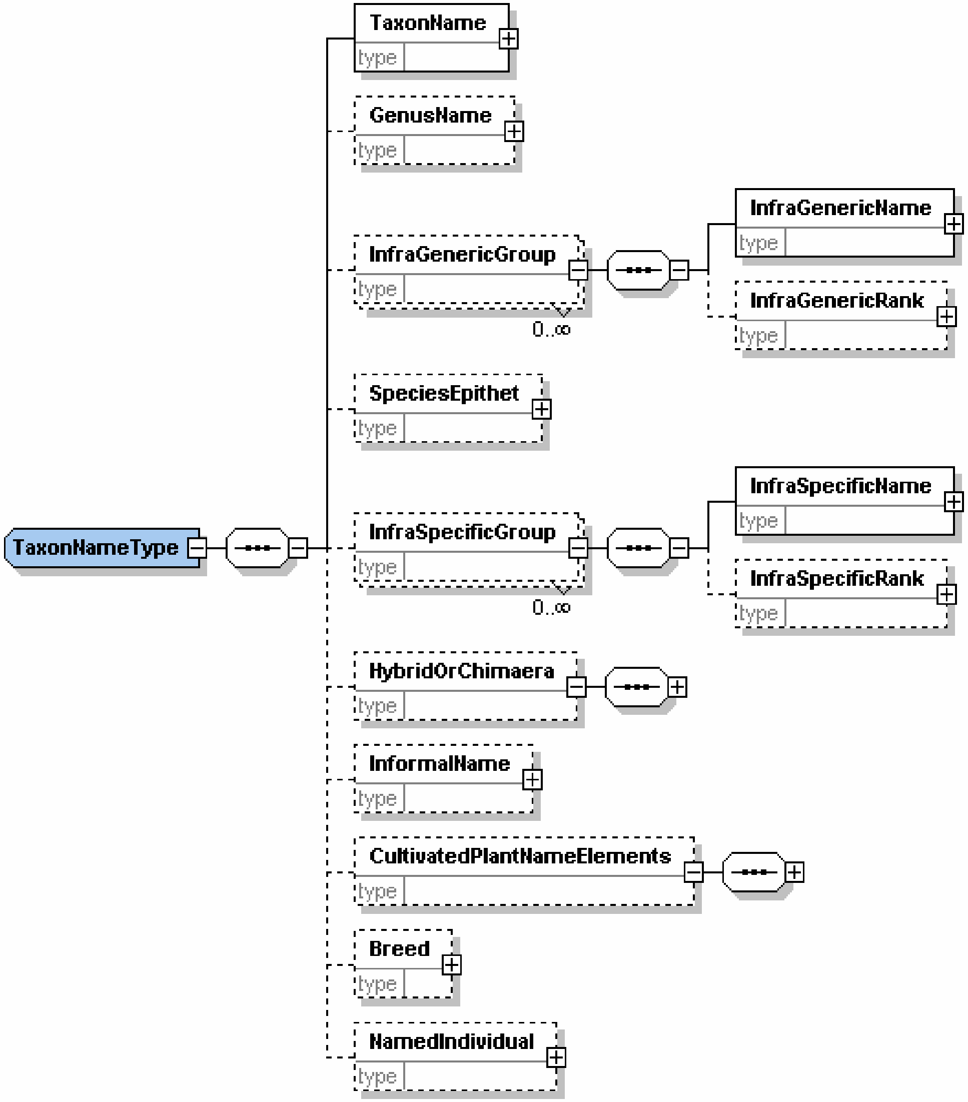
Overall structure:



We propose a standard for taxonomic literature:

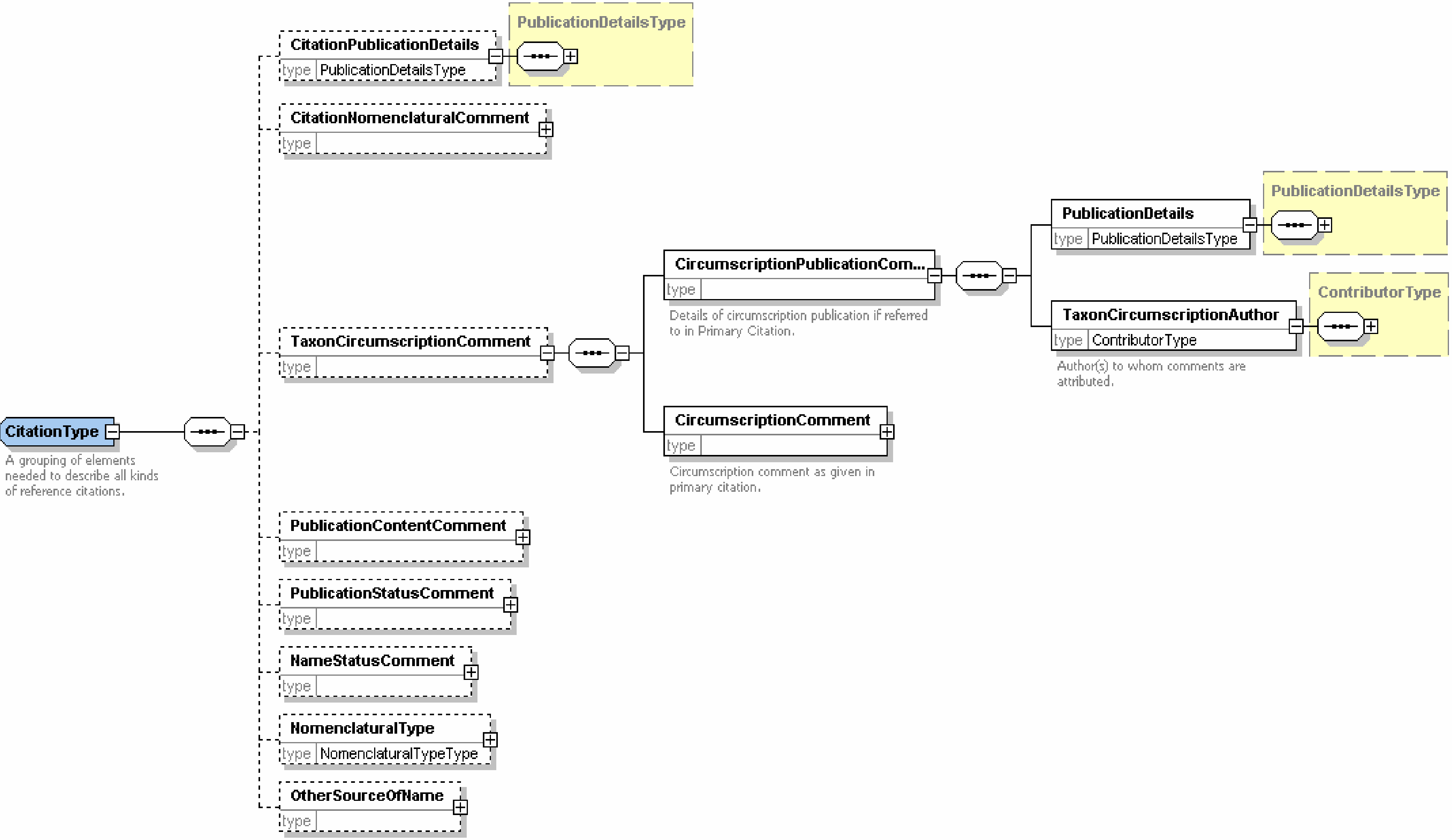
Necessary elements: names

- First element contains full text string. Also holder for suprageneric names.
- Subsequent elements contain atomised information for species and infraspecific names
- Designed for zoological and botanical names so far
- Included elements for rank of infra-generic and infra-specific names (often specified in text)
- Rank also specified outside container



We propose a standard for taxonomic literature:

Necessary elements: citations


CitationType
CitationPublicationDetails
type PublicationDetailsType
PublicationDetailsType
CitationNomenclaturalComment
type
TaxonCircumscriptionComment
type
CircumscriptionPublicationCom...
type
PublicationDetails
type PublicationDetailsType
PublicationDetailsType
TaxonCircumscriptionAuthor
type ContributorType
ContributorType
CircumscriptionComment
type
PublicationContentComment
type
PublicationStatusComment
type
NameStatusComment
type
NomenclaturalType
type NomenclaturalTypeType
OtherSourceOfName
type

A grouping of elements needed to describe all kinds of reference citations.

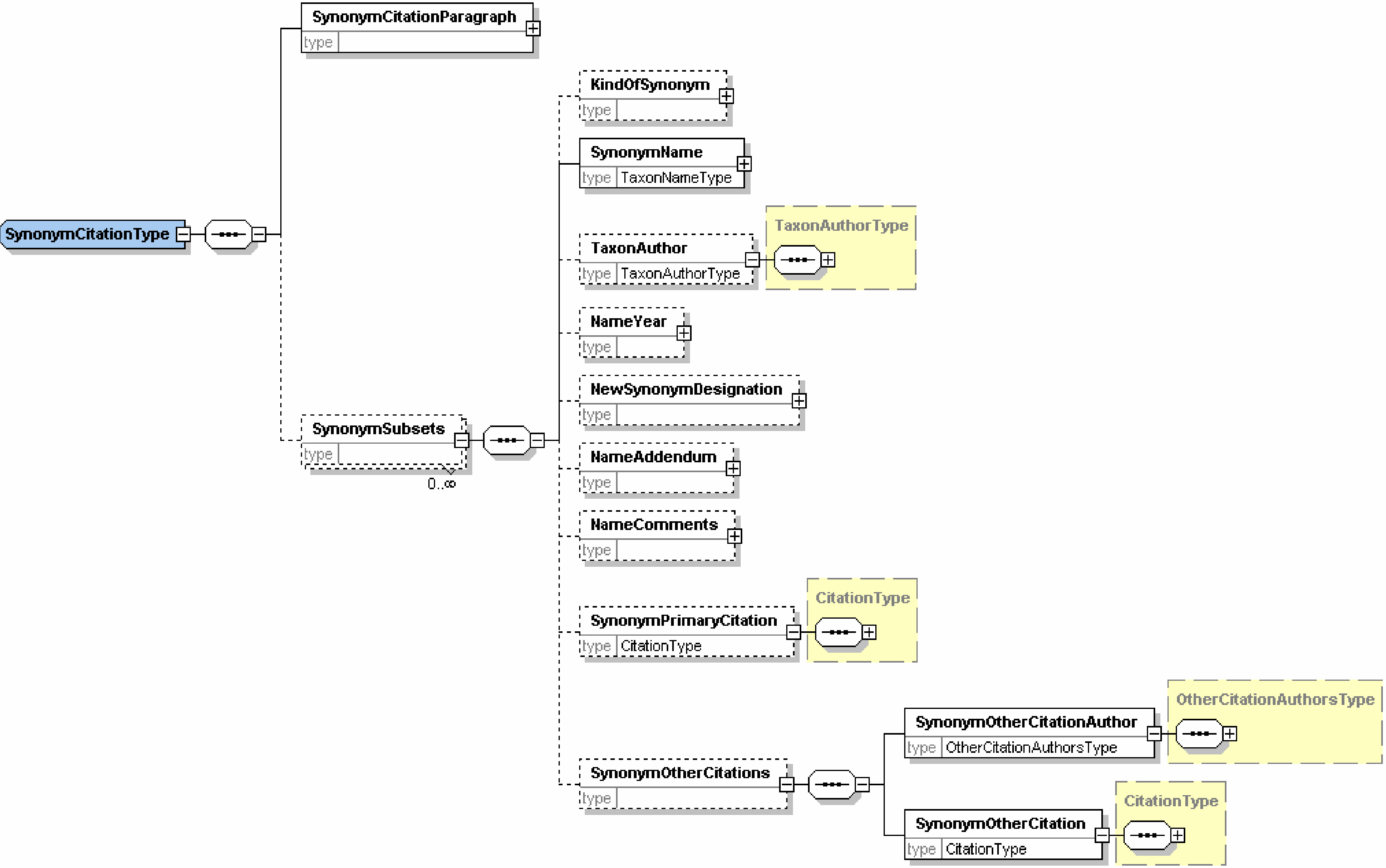
Details of circumscription publication if referred to in Primary Citation.

Author(s) to whom comments are attributed.

Circumscription comment as given in primary citation.

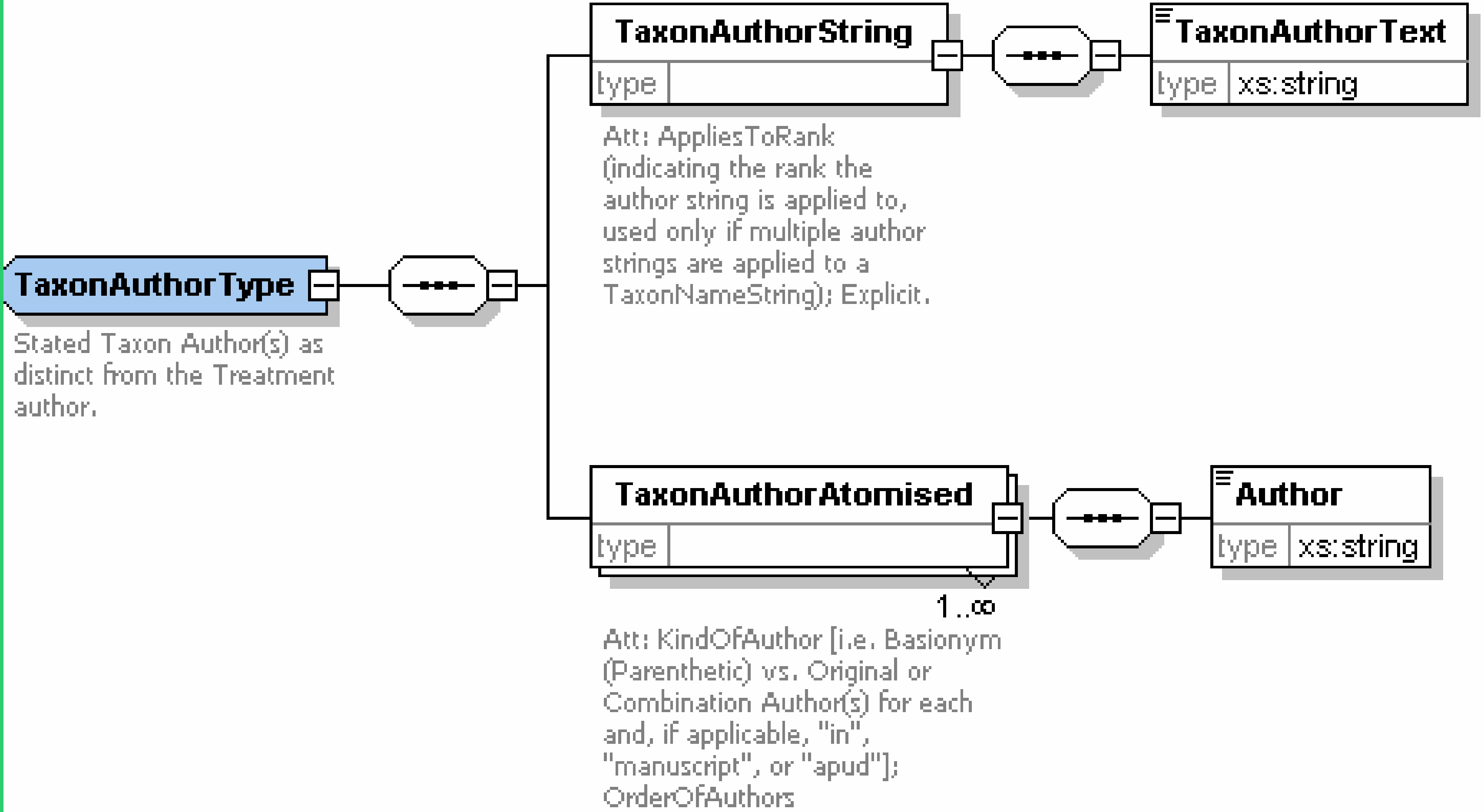
We propose a standard for taxonomic literature:

Necessary elements: synonyms



We propose a standard for taxonomic literature:

Necessary elements: authors



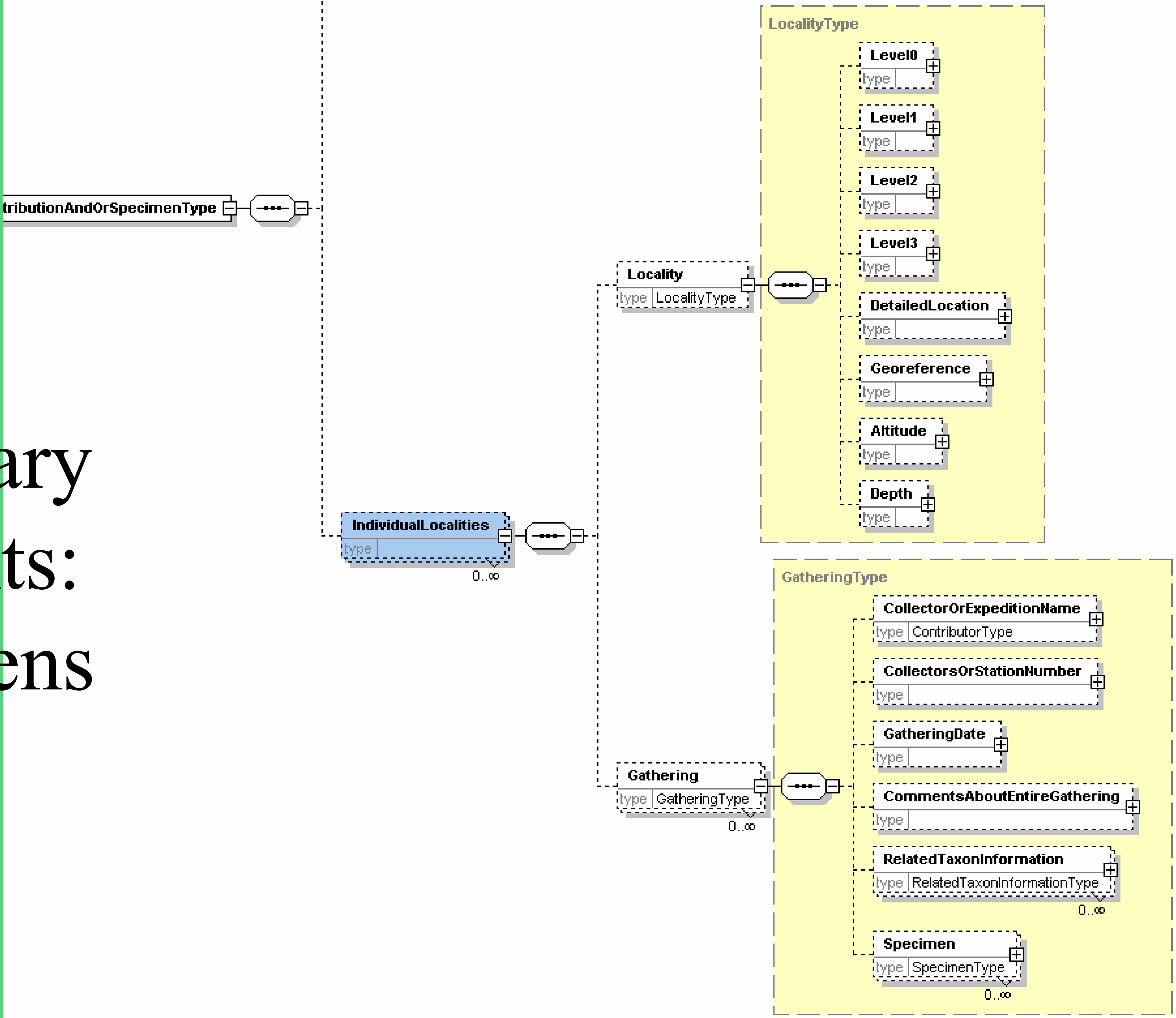
Stated Taxon Author(s) as distinct from the Treatment author.

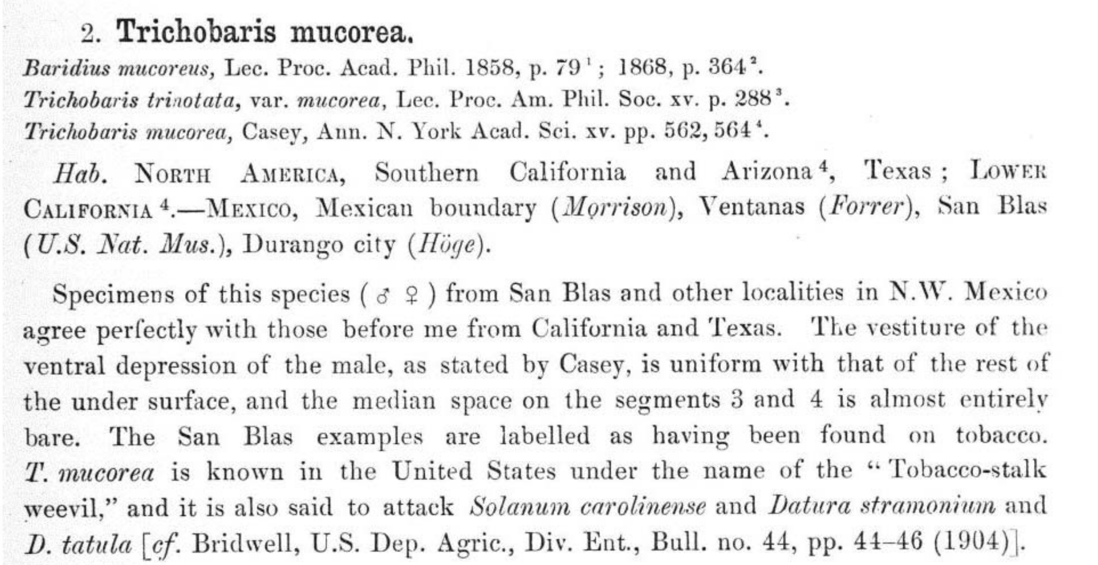
Att: AppliesToRank (indicating the rank the author string is applied to, used only if multiple author strings are applied to a TaxonNameString); Explicit.

Att: KindOfAuthor [i.e. Basionym (Parenthetic) vs. Original or Combination Author(s) for each and, if applicable, "in", "manuscript", or "apud"]; OrderOfAuthors

We propose a standard for taxonomic literature:

Necessary elements: specimens





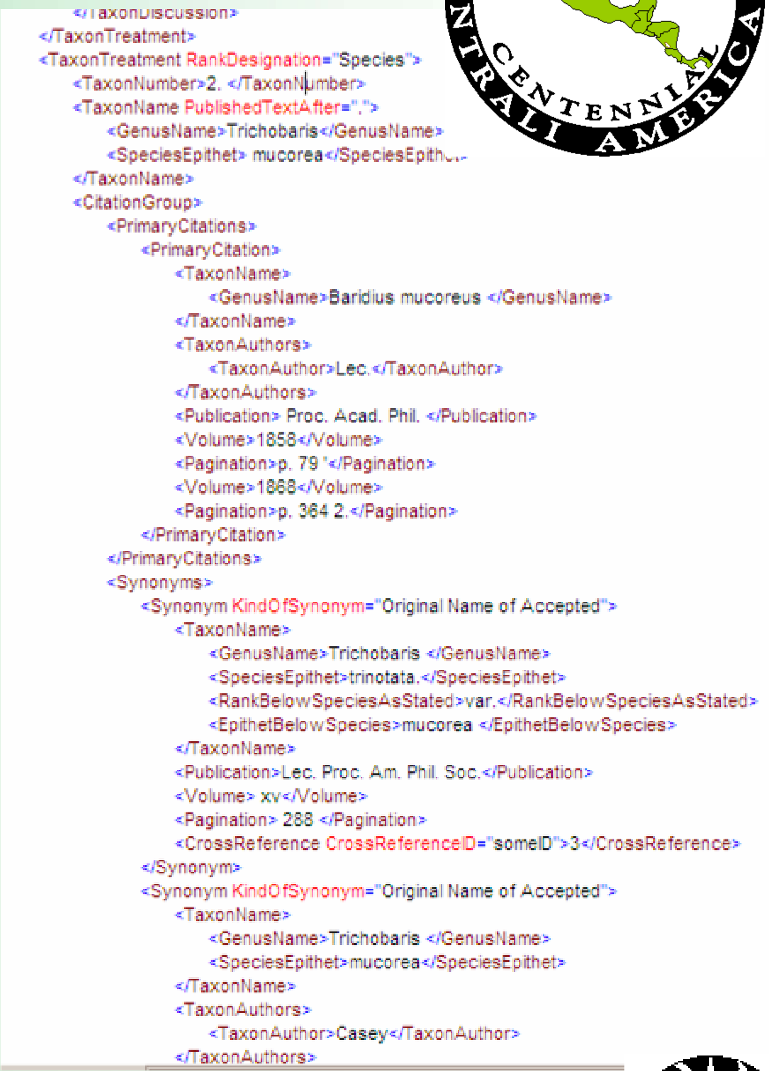
2. **Trichobaris mucorea.**

Baridius mucoreus, Lec. Proc. Acad. Phil. 1858, p. 79 [1]; 1868, p. 364 [2].
Trichobaris trinotata, var. *mucorea*, Lec. Proc. Am. Phil. Soc. xv. p. 288 [3].
Trichobaris mucorea, Casey, Ann. N. York Acad. Sci. xv. pp. 562, 564 [4].

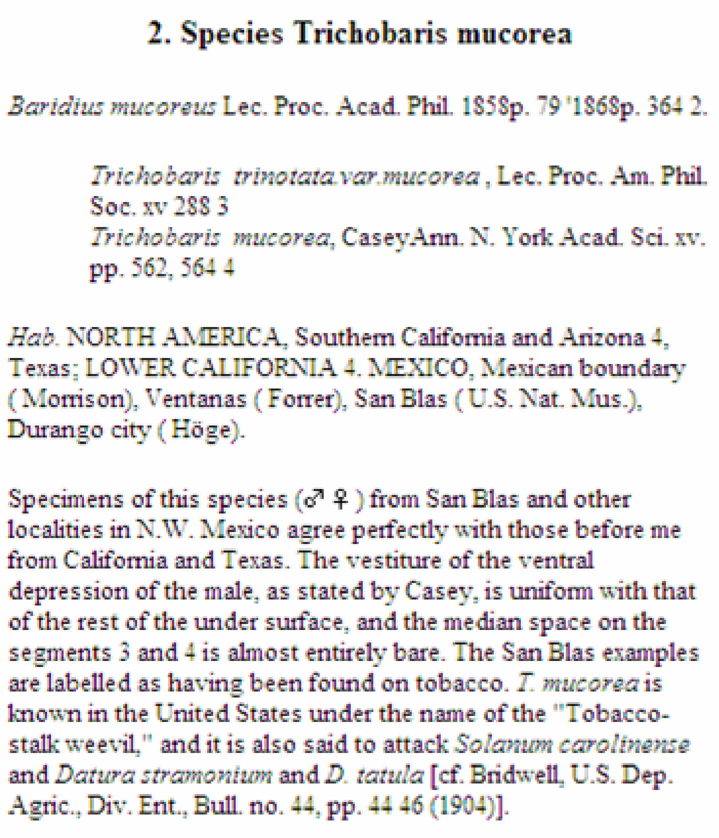
Hab. North America, Southern California and Arizona [4], Texas; Lower California [4].—Mexico, Mexican boundary (*Morrison*), Ventanas (*Forrer*), San Blas (*U.S. Nat. Mus.*), Durango city (*Höge*).

Specimens of this species (♂ ♀) from San Blas and other localities in N.W. Mexico agree perfectly with those before me from California and Texas. The vestiture of the ventral depression of the male, as stated by Casey, is uniform with that of the rest of the under surface, and the median space on the segments 3 and 4 is almost entirely bare. The San Blas examples are labelled as having been found on tobacco. *T. mucorea* is known in the United States under the name of the "Tobacco-stalk weevil," and it is also said to attack *Solanum carolinense* and *Datura stramonium* and *D. tatula* [*cf.* Bridwell, U.S. Dep. Agric., Div. Ent., Bull. no. 44, pp. 44–46 (1904)].

BIOLOGIA CENTRALI AMERICANA CENTENNIAL



```
      </TaxonDiscussion>
  </TaxonTreatment>
  <TaxonTreatment RankDesignation="Species">
      <TaxonNumber>2. </TaxonNumber>
      <TaxonName PublishedTextAfter=".">
          <GenusName>Trichobaris</GenusName>
          <SpeciesEpithet> mucorea</SpeciesEpith...
      </TaxonName>
      <CitationGroup>
          <PrimaryCitations>
              <PrimaryCitation>
                  <TaxonName>
                      <GenusName>Baridius mucoreus </GenusName>
                  </TaxonName>
                  <TaxonAuthors>
                      <TaxonAuthor>Lec.</TaxonAuthor>
                  </TaxonAuthors>
                  <Publication> Proc. Acad. Phil. </Publication>
                  <Volume>1858</Volume>
                  <Pagination>p. 79 '</Pagination>
                  <Volume>1868</Volume>
                  <Pagination>p. 364 2.</Pagination>
              </PrimaryCitation>
          </PrimaryCitations>
          <Synonyms>
              <Synonym KindOfSynonym="Original Name of Accepted">
                  <TaxonName>
                      <GenusName>Trichobaris </GenusName>
                      <SpeciesEpithet>trinotata.</SpeciesEpithet>
                      <RankBelowSpeciesAsStated>var.</RankBelowSpeciesAsStated>
                      <EpithetBelowSpecies>mucorea </EpithetBelowSpecies>
                  </TaxonName>
                  <Publication>Lec. Proc. Am. Phil. Soc.</Publication>
                  <Volume> xv</Volume>
                  <Pagination> 288 </Pagination>
                  <CrossReference CrossReferenceID="someID">3</CrossReference>
              </Synonym>
              <Synonym KindOfSynonym="Original Name of Accepted">
                  <TaxonName>
                      <GenusName>Trichobaris </GenusName>
                      <SpeciesEpithet>mucorea</SpeciesEpithet>
                  </TaxonName>
                  <TaxonAuthors>
                      <TaxonAuthor>Casey</TaxonAuthor>
                  </TaxonAuthors>
```



2. Species Trichobaris mucorea

Baridius mucoreus Lec. Proc. Acad. Phil. 1858p. 79 '1868p. 364 2.

Trichobaris trinotata var.mucorea, Lec. Proc. Am. Phil. Soc. xv 288 3
Trichobaris mucorea, CaseyAnn. N. York Acad. Sci. xv. pp. 562, 564 4

Hab. NORTH AMERICA, Southern California and Arizona 4, Texas; LOWER CALIFORNIA 4. MEXICO, Mexican boundary (Morrison), Ventanas (Forrer), San Blas (U.S. Nat. Mus.), Durango city (Höge).

Specimens of this species (♂ ♀) from San Blas and other localities in N.W. Mexico agree perfectly with those before me from California and Texas. The vestiture of the ventral depression of the male, as stated by Casey, is uniform with that of the rest of the under surface, and the median space on the segments 3 and 4 is almost entirely bare. The San Blas examples are labelled as having been found on tobacco. *T. mucorea* is known in the United States under the name of the "Tobacco-stalk weevil," and it is also said to attack *Solanum carolinense* and *Datura stramonium* and *D. tatula* [cf. Bridwell, U.S. Dep. Agric., Div. Ent., Bull. no. 44, pp. 44 46 (1904)].

THE NATURAL HISTORY MUSEUM

What is happening next?

Current activities and future plans

- Schema and sample encoded text are on http://web4.si.edu/sil/bca/status.cfm
- Schema to be trialled using the *Biologia Centrali-Americana* http://www.sil.si.edu/digitalcollections/bca/
- Comments have been invited from a number of people
- Schema is being put forward as a TDWG standard
- Need to set up working group

THE NATURAL HISTORY MUSEUM

Smithsonian
National Museum of Natural History