Progress in making literature easily accessible: schemas and marking up

TaxonX / Goldengate & taXMLit / INOTAXA

Terry Catapano, Columbia University
Anna Weitzman, Smithsonian Institution
(presenter for Terry: Donat Agosti)

TDWG Annual Meeting
October 19, 2007

Motivation

- Taxonomy only scientific field that routinely uses literature from 300+ years
- Digitization feasible and finite (BHL taking first step)
- Data used by taxonomists, but also useful for many others:
  - Public interested in world around them
  - Conservation biologists, NGOs
  - Ecologists and other scientists
  - Government agencies
  - Policy makers
- Images only first step
- Need to make information easy to find and easy to use in the ways that the many users need it!

Objectives

- Schema for encoding texts of legacy and new taxonomic treatments
- Maximize interchange and archiving capabilities via XML
- Expose latent data in order to supplement existing data on species
- Assist with breaking down the ‘taxonomic impediment’ by speeding taxonomic work
- For data mining, referencing and retrieval at appropriate degree of granularity

Taxonomic Treatments in Legacy Literature

- TaxonX focuses on Treatment and their major components at two levels:
  - Structural level/Phrase-level
    - Nomenclature/name, status, citations
    - Material examined/collection events, localities
    - Description/character, character states
    - References
- taXMLit focuses on entire works, Treatments, major and smaller components
  - Major components held together for display but also atomized to much finer level for more detailed search and retrieval

TaxonX: Design Issues

- Lightweight:
  - Suited for large scale retro-conversion
  - Potentially lower encoding and processing overhead
  - Uses and transformable into other schema for domain specific mark-up (e.g., TDWG standards; NLM/NCBI)
- Flexible:
  - Allows for progressive mark-up from grosser to finer granularity
  - Can handle mixed content

TaxonX: Transcription and Normalization Layers

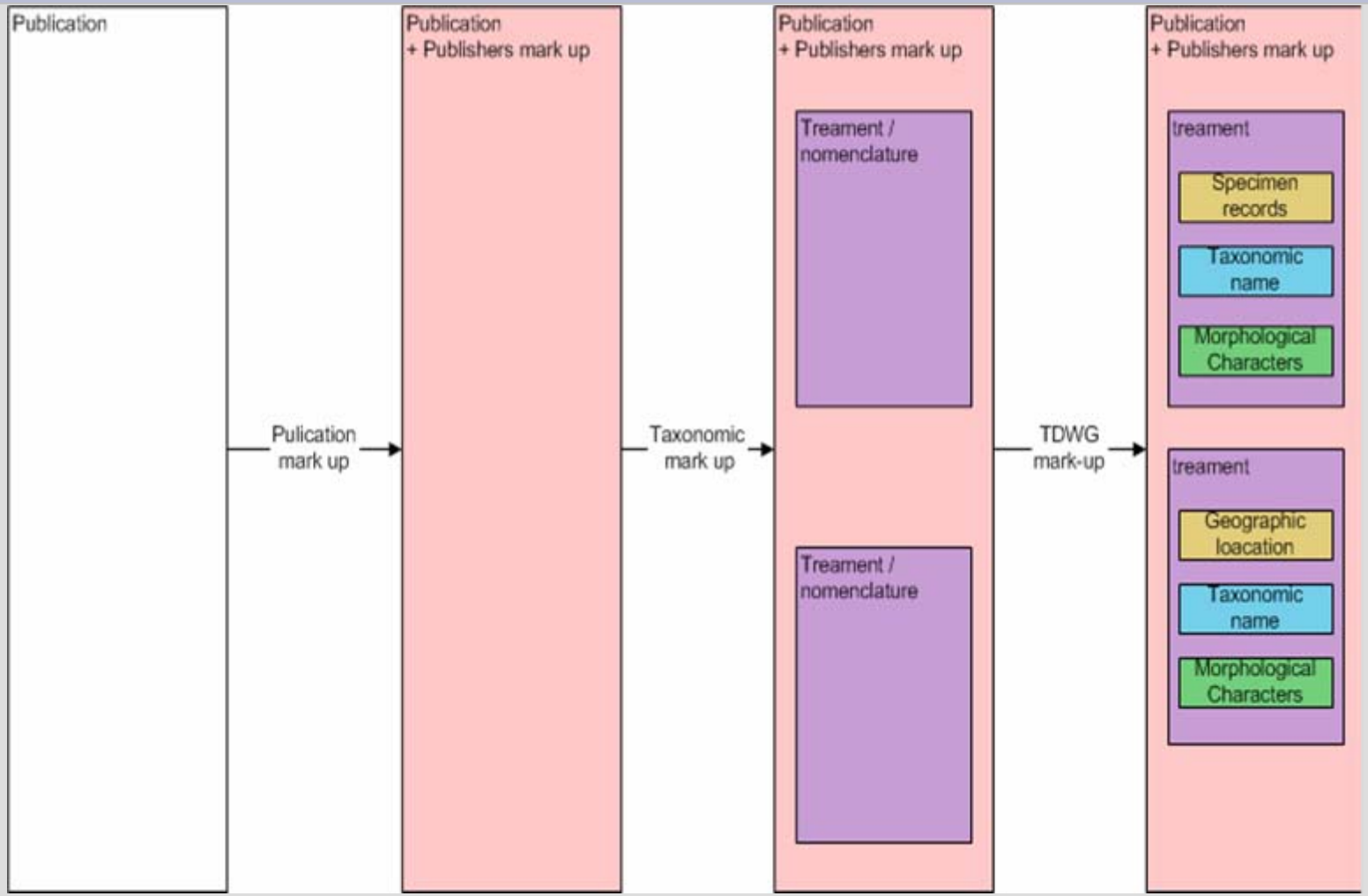
- Original text can be “normalized” via inclusion of data in other schemas, or by pointing to external resources
- Example:

```
…we examined material from the private collection of A. Schulz and
   type material of <tax:name>
<tax:xid source="HNS" identifier="183325"/>
<tax:xmldata>
    <dc:Genus>Temnothorax</dc:Genus>
    <dc:Species>korbi</dc:Species>
</tax:xmldata> T. korbi (Emery, 1922) </tax:name> , …
```

Semi-automated markup: GoldenGate

- Clean OCR as input
- Identifies:
  - Scientific names
  - Localities
  - Bibliographical references
- Performs lookups in external resources
  - E.g., Hymenoptera Name Server
  - Potentially many other similar services
- Outputs
  - Currently TaxonX
  - Other formats could be easily defined

TaxonX Progressive Enhancement



TaxonX Example I

Discothyrea berlita Fisher, sp. nov. Fig 1,4.

TYPE MATERIAL.. HOLOTYPE: Worker. MAURITIUS: Le Pouce Mt., Moka Range, 20°11′55″S, 057°31′44″E, 750 m, closed vegetation, 25 May 2005 (coll. B.L. Fisher et al.) Collection code: BLF12148, specimen code: CASENT0007016 (CASC).

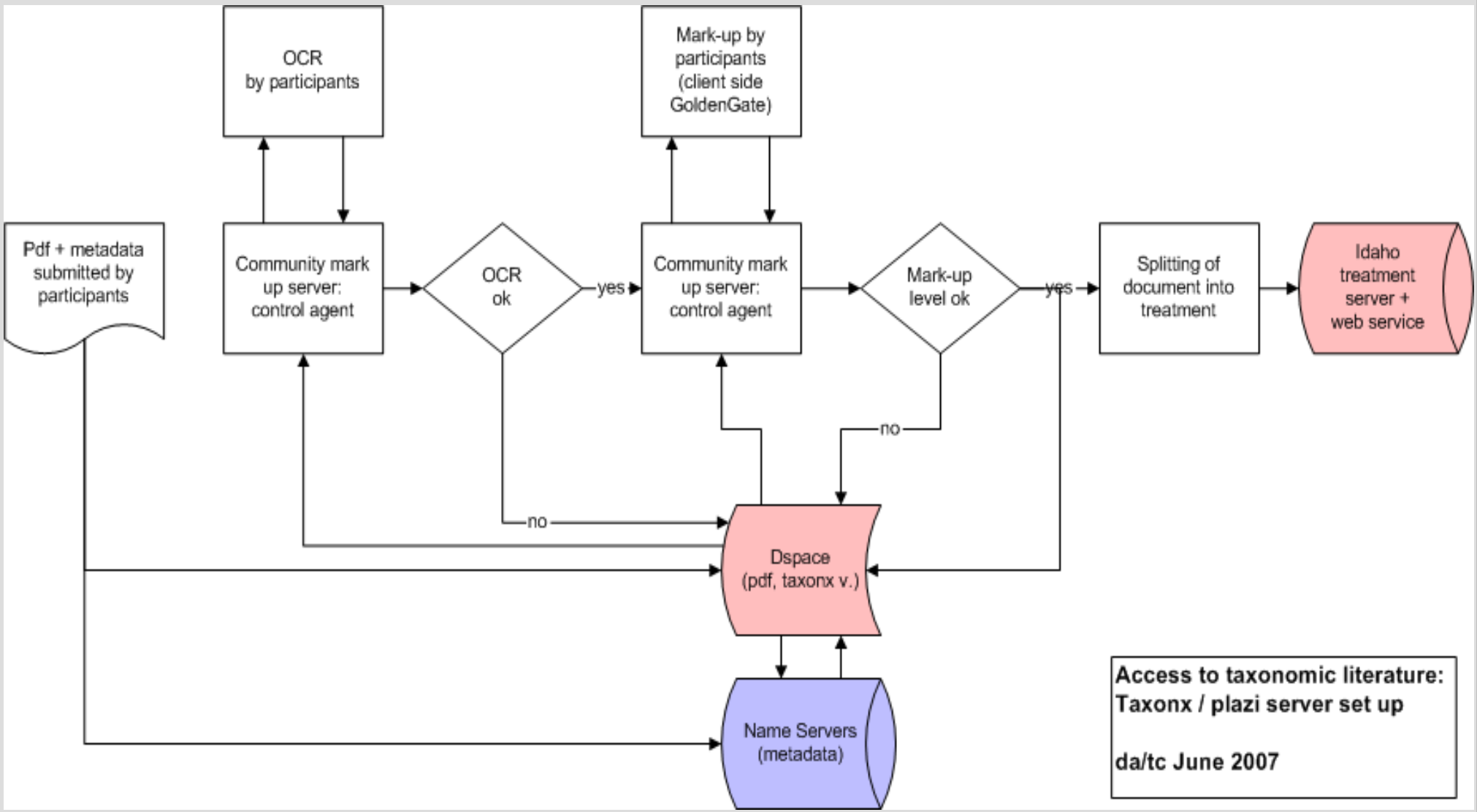
```
<p>Discothyrea berlita Fisher, sp. nov. Fig 1,4.</p>
<p>TYPE MATERIAL.. HOLOTYPE: Worker. MAURITIUS: Le Pouce Mt., Moka Range, 20°11′55″S,
 057°31′44″E, 750 m, closed vegetation, 25 May 2005 (coll. B.L. Fisher et al.) Collection code:
 BLF12148, specimen code: CASENT0007016 (CASC).</p>
```

```
<tax:treatment level="species">
      <tax:nomenclature><tax:name>Discothyrea berlita Fisher</tax:name>, <tax:status>sp.
   nov.</tax:status><tax:figures>Fig 1,4.</tax:figures>
         </tax:nomenclature>
    <tax:div type="materials_examined">
       <tax:p>TYPE MATERIAL.. HOLOTYPE: Worker. MAURITIUS: Le Pouce Mt., Moka Range, 20°11′55″S,
         057°31′44″E, 750 m, closed vegetation, 25 May 2005 (coll. B.L. Fisher et al.)
   Collection code: BLF12148, specimen code: CASENT0007016 (CASC). </tax:p>
    </tax:div>
```

TaxonX Example II

```
<tax:treatment level="species">
      <tax:nomenclature><tax:name>
        <xml:data>
        <dc:Genus>Discothyrea</dc:Genus>
        <dc:Species>berlita </dc:Species>
        </xml:data>
   <tax:xid identifier="HNS153343" source="HNS"/>
   <xid type=”LSID” uri=”[LSID]”/>Discothyrea berlita Fisher</tax:name>, <tax:status>sp.
   nov.</tax:status><tax:figures>Fig 1,4.</tax:figures>
         </tax:nomenclature>
     <tax:div type="materials_examined">
        <tax:p>TYPE MATERIAL.. HOLOTYPE: Worker.<tax:collection_event><tax:xmldata>
             <dc:CollectionCode>BLF12148</dc:CollectionCode>
             <dc:CatalogNumber>CASENT0007016</dc:CatalogNumber>
             <dc:DayCollected>25</dc:DayCollected>
             <dc:MonthCollected>5</dc:MonthCollected>
             <dc:YearCollected>2005</dc:YearCollected>
             <dc:Country>MU</dc:Country>
             <dc:Collector>B.L. Fisher et al</dc:Collector>
             <dc:Latitude>-20.11</dc:Latitude>
             <dc:Longitude>57.31</dc:Longitude>
          </tax:xmldata> MAURITIUS: <tax:locality>Le Pouce Mt., Moka Range, 20°11′55″S,
        057°31′44″E, 750 m,</tax:locality> closed vegetation, 25 May 2005 (coll. B.L. Fisher
   et al.) Collection code: BLF12148, specimen code: CASENT0007016
   (CASC).</tax:collection_event> </tax:p>
     </tax:div>
```

TaxonX Workflow



Lessons Learned

- taxonX
  - Flexiblity
  - Transcription and Normalization “Layers”
  - Modularity
  - Express as module of NLM/NCBI DTD
- GoldenGate
  - Feasibility of semi-automatic markup
  - Could be applied to output of large scale digitization efforts
  - Modular design permits customization and extension

taXMLit: Design Considerations

- Comprehensive
  - Designed to include all contents of taxonomic literature
  - Designed to be interoperable with existing TDWG standards
- Flexible
  - Allows for progressive markup from grosser to finer granularity within a single schema

taXMLit: structure

- Front & back matter included
- Hierarchical structure maintained with IDs
- Main focus on structured data within treatments:
  - Taxon Heading
  - Accepted name & synonyms and their citations
  - Keys
  - Descriptions (not atomised beyond paragraph)
  - Distributions &/or specimen citations
  - Discussions

INOTAXA Workflow

- Start with text marked up in generic TEI-Lite form (with specific extensions for taxonomic literature-specific formats)
- Scripts to parse specific paragraph types and formats to taXMLit Schema
  - Taxon Heading
  - Accepted names & synonyms and their citations
  - Keys
  - Specimen Citations
- As more works are marked up, a parsing script library is built
- Appropriate parsing scripts for a new work will then chosen by human in such a way that the computer will learn (Artificial Intelligence) to do it

INOTAXA: Experience with Mark up

- TEI-Lite mark up done while text of BCA was re-keyed
- Scripts used to parse specific paragraph types and formats
  - Ca. 95% successful for one volume of BCA
- Need to test with more works and build parsing script library

- Additional information may be parsed from discussion if desired.

Lessons learned / next steps

- Explore making GoldenGate more scaleable
- Explore using GoldenGate to do initial parsing for taXMLit
- Test INOTAXA parsing model and Artificial Intelligence with more content