# Integrating DNA Barcoding and taxonomic data

## INOTAXA – how new technology can facilitate Open Access to 300 years of vitally important information

Anna L. Weitzman & Christopher H. C. Lyal



Smithsonian
National Museum of Natural History

NATURAL HISTORY MUSEUM

CONSORTIUM FOR THE BARCODE OF LIFE

Biodiversity Heritage Library

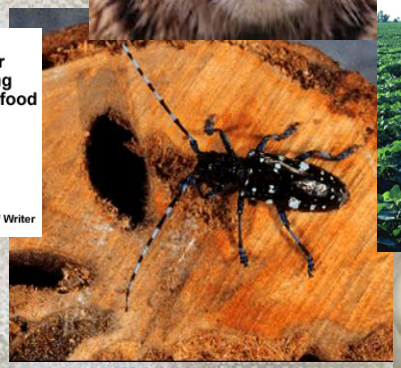
# *Taxonomy (and Systematics)*

[generally interchangeable terms as used in biology]

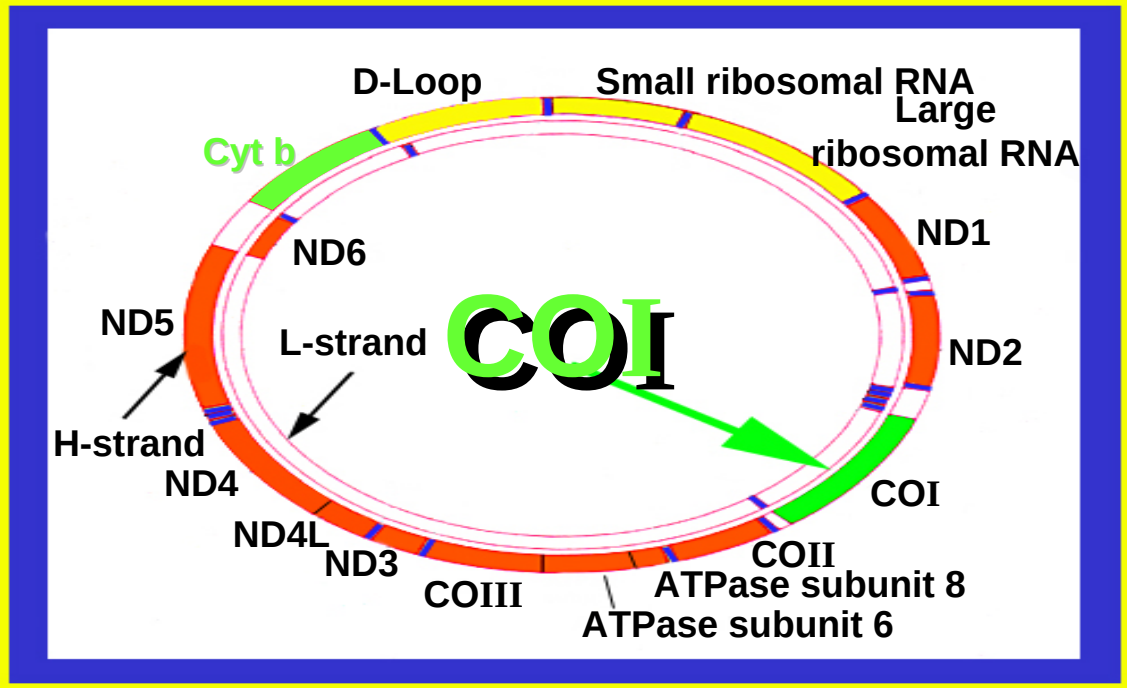
- The study of names and evolutionary relationships of organisms
- Names governed by Codes of Nomenclature
- Evolutionary relationships understood by analyzing shared similarities in morphology and gene sequences
- Some 15 million species (estimates between 5-100 million) believed to exist, only about 1.8 million are currently known to science
- Species knowledge based largely on museum collections estimated at 1.3-3 billion specimens
- Some 300 years of data – much highly structured!
- Understanding all organisms and their evolutionary relationships is vitally important to the future of life (including human) on earth
- Retrospective data and prospective data equally vital!!

***Taxonomy is vital for identifying, understanding, and managing:***

- Endangered/protected species
- Biodiversity conservation
- Agricultural pests
- Invasive species
- Disease vectors/pathogens
- Hazards (e.g., bird strikes on airplanes)
- Environmental quality indicators
- Sustainable development
- Generally understanding the amazing world around us!
- Implementing the CBD

# *DNA Barcodes*




## What are they?

- A DNA barcode is a short gene sequence taken from standard portions of the genome, used to identify species

## How are they used?

1. Research tool for taxonomists:
   - Expand species knowledge to include all life history stages, dimorphic sexes, damaged & partial specimens
   - Test consistency of species definitions
2. Applied tool for identifying regulated species
3. “Triage” tool for flagging potential new species

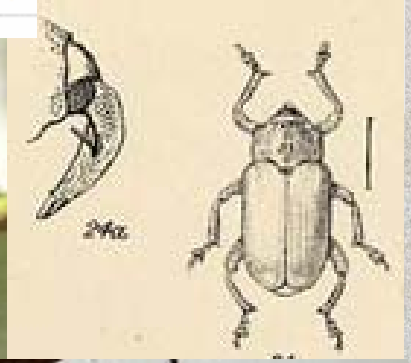
# *DNA Barcodes, Taxonomy and data*

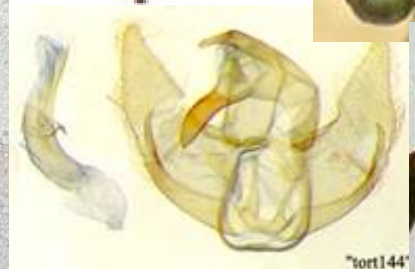
- Barcoding community recognized early that success is dependent on good taxonomy and access to all taxonomic data
- Consortium for the Barcode of Life (CBoL) has active role in standards for biodiversity data
- CBoL catalyzes digitizing taxonomic literature
  - Library-Laboratory meeting in London on electronic access to taxonomic literature
  - Led to formation of Biodiversity Heritage Library (BHL) initiative to digitize all published biodiversity works from libraries worldwide
  - Proactive steps with PubMed to add taxonomic journals to online abstracts
  - Aggressive negotiation with publishers of barcoding papers for Open Access

# *The Taxonomic Impediment: finding the data*

Data are of many types:

- original descriptions
- synonymies
- current treatments
- identification keys
- geographic information
- images of living organisms, type specimens, dissections, organs / parts
- observations
- specimen & associated data



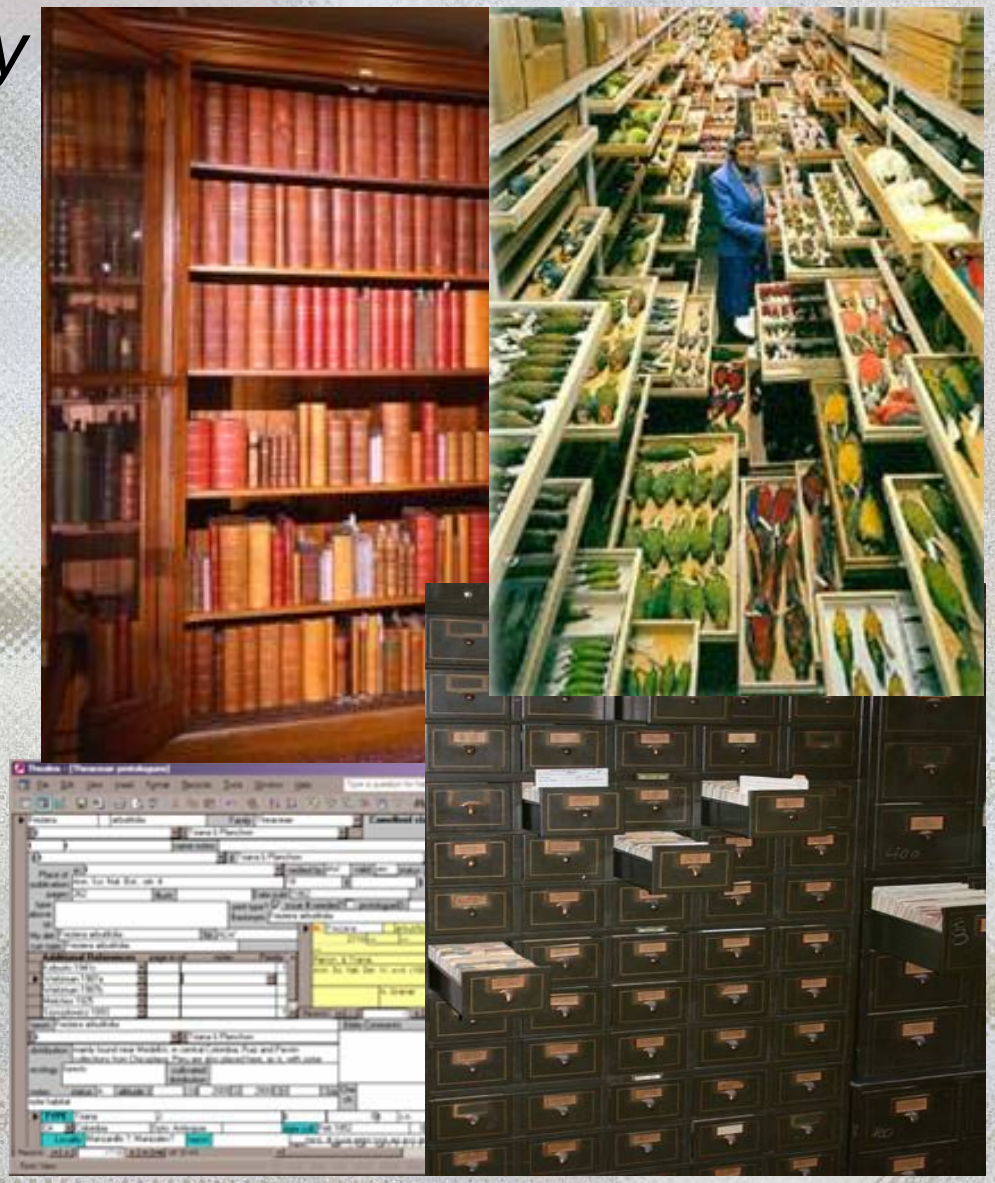



# *The Taxonomic Impediment: finding the data*

*Data can be found in many unconnected places:*

- Specimen collections
- Libraries
- Reprint collections
- Databases
- Publications
- Observations
- ‘grey’ literature
- Index cards
- Field notebooks
- Gene sequence repositories

*And associated with both modern and superseded names*



# *The Taxonomic Impediment: finding the data*

Many taxonomists and other researchers and ‘users’:

- do not know how to find all of the data that they need, and/or
- cannot afford the time or money to access them

<u>Consequently:</u>

Only a limited subset of appropriate literature and potential data are used in most analyses, *limiting the adequacy of scientific results*

***Solution: Leverage existing technology to address the issues…to create a taxonomic web space***

# *Properties of a Taxonomic Web Space*

- All related data (as mentioned above) in interoperable formats
- Accessible from anywhere in the world
- Distributed: accessible through multiple portals
- Flexible so that users may access the data they need in the way that *they* want it
- Analyzable by web and other tools
- Ownership and *IPR* retained by contributors
- Accommodates full taxonomic treatments *and* single species descriptions

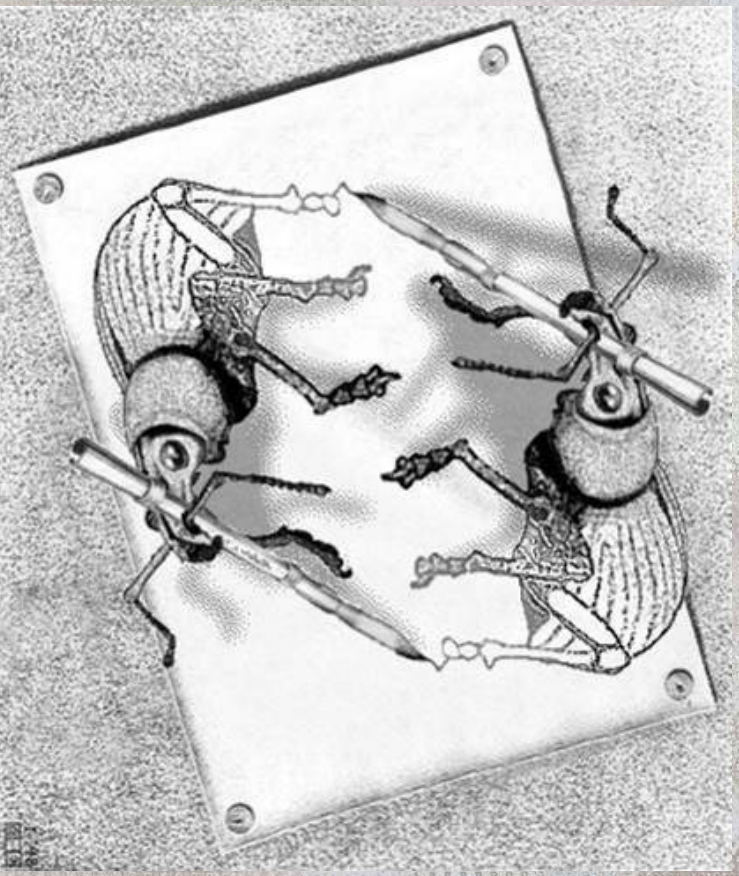
*To enable this, standards must be developed to allow interoperability between different data sets*

**Some functionality is already in place**

# *Uniting data on the web –*

## *Standards permit interoperability*

### The state of play for Taxonomy:



Names & Concepts:

- standards emerging (Linnean Core, Taxon Concept Schema)
- millions of records (CoL, uBio, GBIF, etc)

Specimens:

- standards developing (Darwin Core, ABCD)
- millions of records (BioCASE, GBIF, etc)

Literature:

- library standards for metadata (MODS, etc)
- full works scattered & usually page images or non-standard text formats

Descriptions:

- standards developing (SDD)

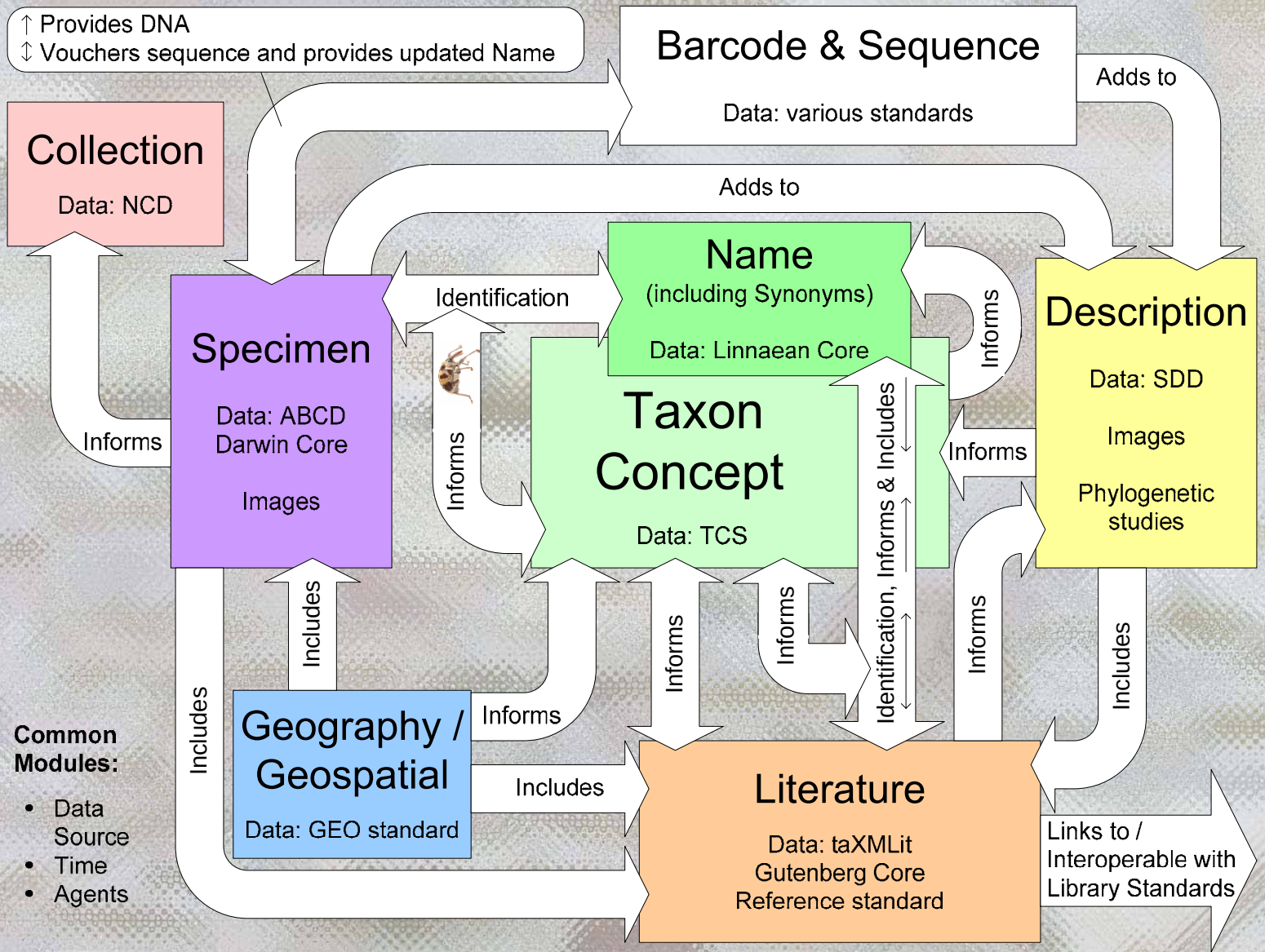
Geography:

- standards elsewhere (FGDC, ADL, etc)
- taxonomy standard (GEO in development)

***GUIDs or LSIDs needed throughout***

# Uniting data on the web

## The complex relationships between data, standards and the way taxonomists use them




**Common Modules:**

- Data Source
- Time
- Agents

# *Uniting data on the web: our focus--literature*

## How can literature be made most useful?

*Current*: images of pages (e.g. jpeg, pdf)

- Improves accessibility
- No easier to find content
- Cannot usually be searched
- Are not interoperable with other data

*Biodiversity Heritage Library aims to create indexed, scanned legacy literature (as copyright allows), which will facilitate the future--*

## ***Future*: text delivered via XML**

- **Can be searched**
- **Can be made interoperable with and link to other datasets as needed by taxonomists and those making biodiversity-related decisions**

# *A test-bed for the vision*

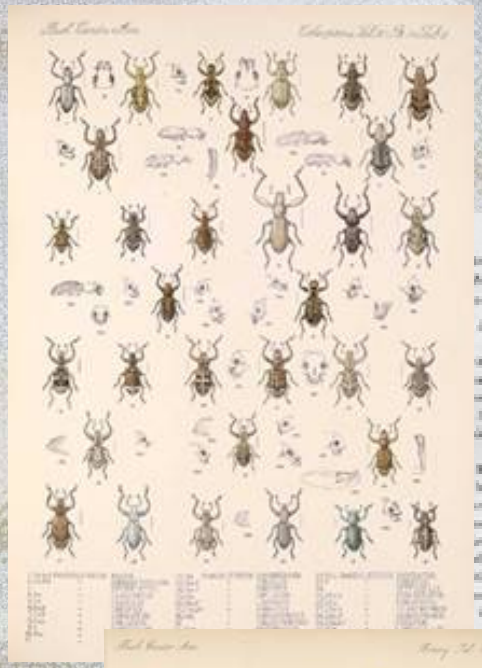
## *The Biologia Centrali-Americana – 1879-1915*

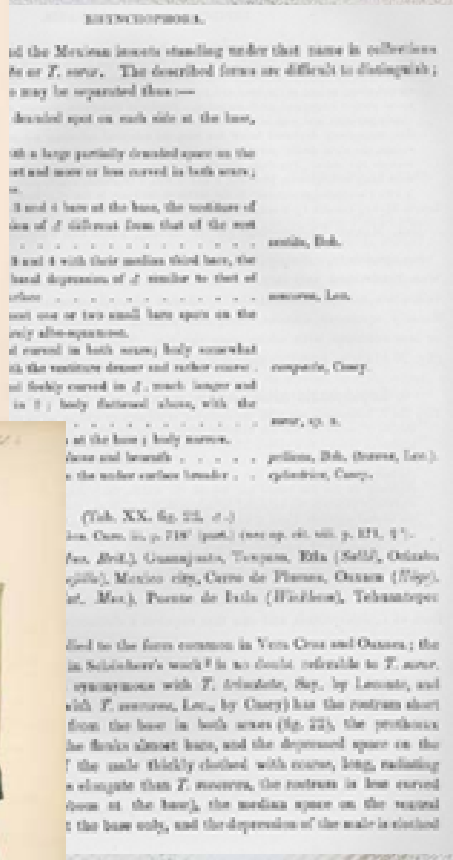
- includes almost everything known at the time about the region's biological diversity
- for many groups still the current state of published knowledge
- privately issued by F. DuCane Godman and Osbert Salvin of The Natural History Museum
- descriptions of 50,263 species of plants, vertebrates, insects, spiders and related invertebrates, and mollusks
- the entire BCA is held by few libraries, mostly Northern; other libraries hold partial sets
- some Central American countries lack a complete set

# The INOTAXA project (formerly *'Biologia Centrali-Americana* Centennial')

Objectives:

- create images in multiple formats of over 25,000 pages of the 58 biological volumes
- create and propose standard structure (schema) for taxonomic literature
- code the full text and other texts in eXtensible Markup Language (XML)
- provide facility to link text elements to specimen, taxonomic and geographic data
- make the entire project freely available on the World Wide Web





http://www.sil.si.edu/DigitalCollections/bca

# *electronic* BIOLOGIA CENTRALI-AMERICANA

*This digital edition of the important and out-of-print Biologia Centrali-Americana makes all 58 biological volumes available. Descriptions of over 50,000 and images of over 18,000 species of animals and plants are now accessible as never before. This is the first step towards an extraordinary new set of electronic resources and knowledge tools for biodiversity studies - the Biologia Centrali-Americana Centennial.*

BCA DIGITAL EDITION | ABOUT THE BCA | PROJECT STATUS | ABOUT THE PROJECT | PARTNERS | KEY CONTRIBUTORS | RESOURCES

Smithsonian Institution Libraries

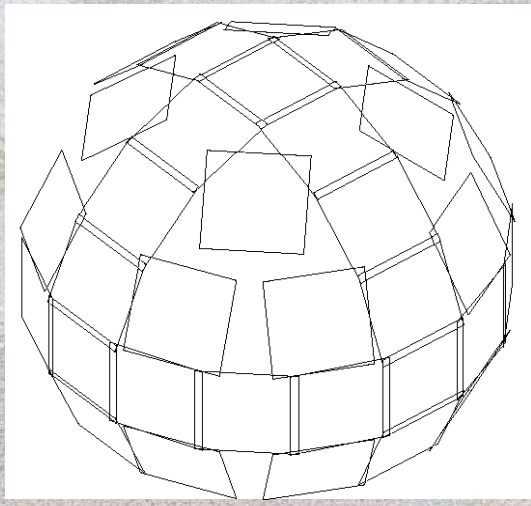
NATURAL HISTORY MUSEUM

Smithsonian
*National Museum of Natural History*

This project was partially funded by
**the Atherton Seidell Endowment Fund of the Smithsonian Institution.**

credits | permissions

# *The INOTAXA project: Taxonomic literature standard (taXMLit)*



*Contents:*

- Name content (names, synonymies, author, literature citations, type information)
- Specimen citations
- Geographic content (distribution & specimens)
- Character content (descriptions & keys)
- Publication metadata

*Will allow:*

- reconstruction of taxonomic text in various formats (species pages, keys, checklists, monographs, etc)
- construction of checklists for geographic areas and taxa
- automatic links to updated to taxonomy because of interoperability with name authority files
- automatic links to updated place names because of interoperability with gazetteers
- linkage to all available related databases

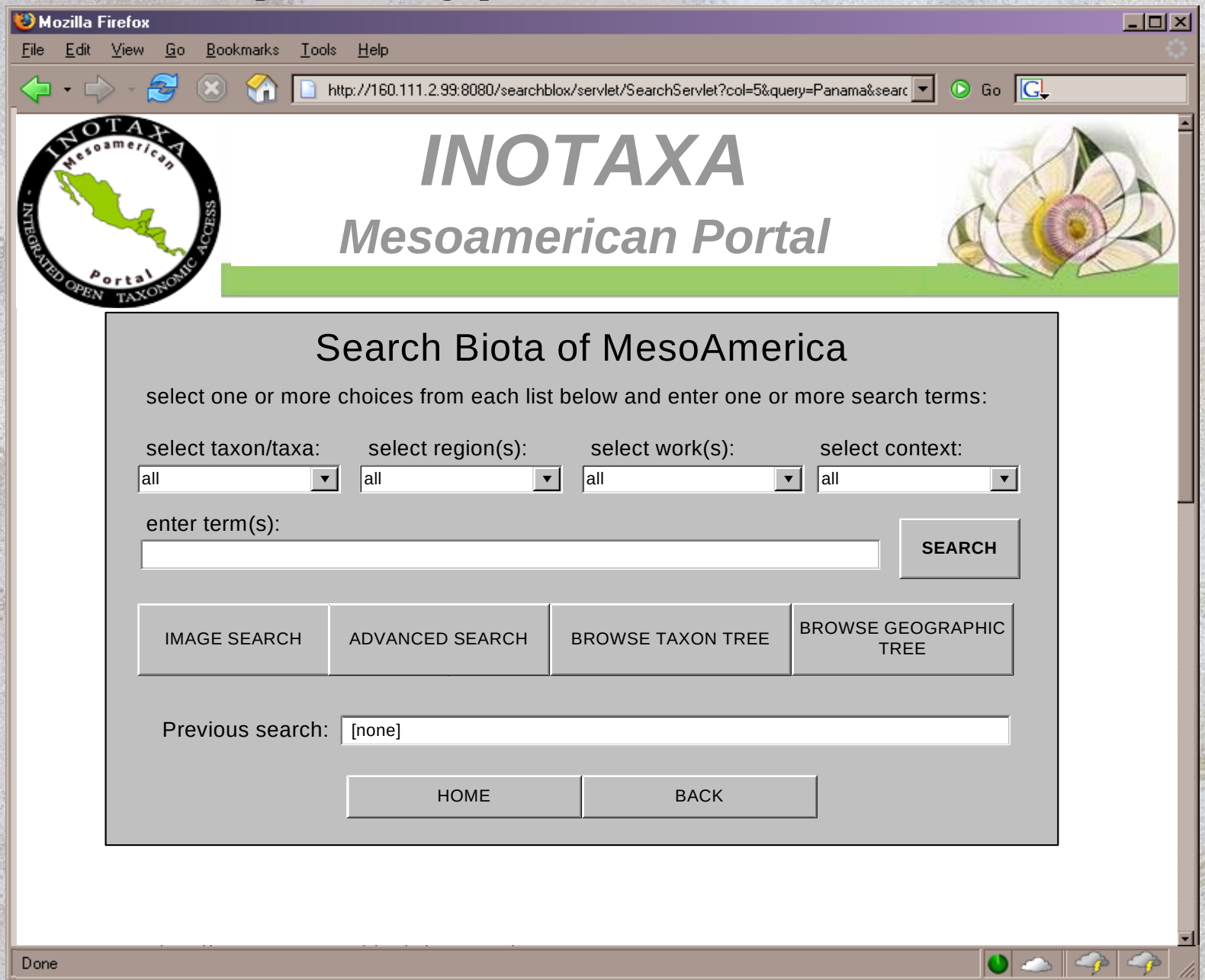
***Comments sought on current draft:***
http://www.sil.si.edu/digitalcollections/bca/status.cfm

***In process of making it a standard through TDWG and GBIF***

# *The INOTAXA project: Next stages*

- BCA biological text and select recent related works (e.g., *Flora Mesoamericana*, recent monographs) available via taXMLit in a prototype web interface (summer 2006)
- Development of interfaces to other data sets (some in prototype in 2006):
  - specimen databases (partner institutions, GBIF, BioCASE, REMIB etc)
  - Taxonomic Name Servers (GBIF, uBio, CoL etc)
  - national and regional checklists
  - images of specimens, species etc
  - web-based analytical tools and other datasets (GIS)
  - locality gazetteer
- Development and addition of interpretation layer schema and integration with INOTAXA, allowing online additions, commentary
- e-publishing facility

# The INOTAXA prototype



# The INOTAXA prototype

Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://160.111.2.99:8080/searchblox/servlet/SearchServlet?col=5&query=Panama&searc

Go

**New Simple Search:**

select one or more choices from each list below and enter one or more search terms:

select taxon/taxa: all

select region(s): all

select work(s): all

select context: all

enter term(s):

SEARCH

IMAGE SEARCH

ADVANCED SEARCH

BROWSE TAXON TREE

BROWSE GEOGRAPHIC TREE

Previous search:

Any field contains "femora" ordered by 'taxon name' then by 'taxon author' then by 'treatment author' then by 'publication date'

OTHER TREATMENT(S)

KEY(S) WITHIN TAXON

KEY(S) TO TAXON

DISTRIBUTION MAP

SPECIMEN(S)

TOGGLE TO PDF

TOGGLE TO JPEG

IMAGE(S)

GAZETTEER

SEARCH HUH BOTANISTS

SEARCH *Flora Mesoamericana*

SEARCH GBIF

SEARCH GOOGLE

SEARCH GOOGLE IMAGES

HOME BACK

Curculionidae
- Attelabinae
  - Attelabus
    - Attelabus
      - Himatolabus
        - Attelabus vinosus

Context in BCA classification

Sharp, 1899, BCA, Coleoptera Volume 4, Part 3 p. 5

INOTAXA Mesoamerican Portal - INTEGRATED OPEN TAXONOMIC ACCESS -

BIOLOGIA CENTRALI AMERICANA electronic

10. ***Attelabus vinosus*, sp. n.**

Rufo-obscurus, pube pallescente sat dense vestitus; prothorace elytrisque fortius sculpturatis; scutello subquadrato, haud transverso.

Long. 5.5 millim.

Hab. MEXICO, Totosinapam, Capulalpam (Sallé), Jalapa (Höge); GUATEMALA, Quiche Mountains 7000 to 8000 feet (Champion).

Closely allied to *A. vestitus*, but distinguished by the dark vinous-red colour and the much more evident sculpture. The sexual distinctions, except in the front legs, are slight. The rostrum is quite short, thick, the head broad, the eyes placed nearly midway between the front of the thorax and the mouth. The thorax is rather coarsely and irregularly sculptured, without any transverse groove. The elytra are even, scarcely at all depressed behind the scutellum, rather coarsely and irregularly sculptured, the striation quite visible. The front **femora** are entirely unarmed.

A specimen of this species in Sallé's collection from Sturm's cabinet is labelled *A. cinnamomeus*, Sturm; but as this name is not a suitable one - being much more applicable to the closely allied *A. vestitus* - I have not used it. The four Mexican examples before me are all in a bad state of preservation; the description therefore is taken from the Guatemalan exponents, six in number.

Done

# The INOTAXA prototype

Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://160.111.2.99:8080/searchblox/servlet/SearchServlet?col=5&query=Panama&searc

Go

SIMPLE SEARCH
ADVANCED SEARCH
BROWSE TAXON TREE
BROWSE GEOGRAPHIC TREE

Previous start:

Pilolabus viridans

OTHER TREATMENT(S)
KEY(S) WITHIN TAXON
KEY(S) TO TAXON
DISTRIBUTION MAP
SPECIMEN(S)
TOGGLE TO PDF
TOGGLE TO JPEG
GAZETTEER
SEARCH HUH BOTANISTS
SEARCH *Flora Mesoamericana*
SEARCH GBIF
SEARCH GOOGLE
SEARCH GOOGLE IMAGES

HOME
BACK

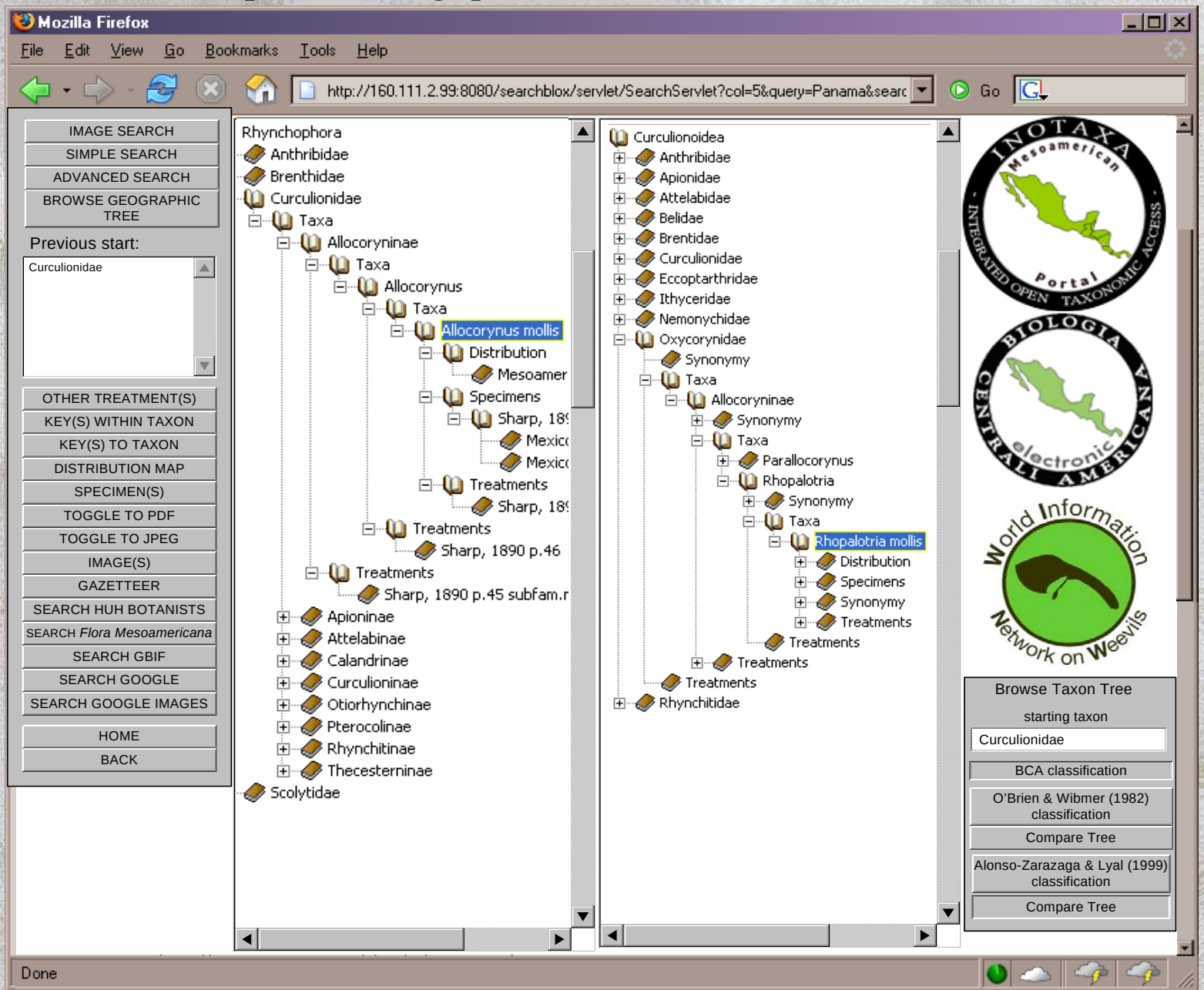
***Pilolabus viridans***

***Pilolabus viridans***

INOTAXA Mesoamerican Portal - INTEGRATED OPEN TAXONOMIC ACCESS -

Done

# The INOTAXA prototype



# *The INOTAXA project*

Includes, but is *not* limited to:

INOTAXA Mesoamerican Portal - INTEGRATED OPEN TAXONOMIC ACCESS -

Smithsonian
*National Museum of Natural History*

Smithsonian Institution Libraries

NATURAL HISTORY MUSEUM

Royal Botanic Gardens, Kew

BIF

Missouri Botanical Garden

INBio
Instituto Nacional de Biodiversidad

Conabio
Comisión nacional para el conocimiento y uso de la biodiversidad

AMERICAN MUSEUM OF NATURAL HISTORY

## *Important URLs:*

- http://inotaxa.si.edu/
  - http://www.sil.si.edu/digitalcollections/bca/bkground.cfm
  - http://www.sil.si.edu/digitalcollections/bca/status.cfm
- http://www.barcoding.si.edu/
- http://bhl.si.edu/

Smithsonian
*National Museum of Natural History*

NATURAL HISTORY MUSEUM