PROTOTYPE FUNCTIONALITY:
INOTAXA - Mesoamerican Portal

Draft, 23 Dec 2005

Read with "INOTAXAPilotScreensRelease1"

CONTENTS

INTRODUCTION

This document describes the functionality of the 'INOTAXA' prototype, currently under construction. INOTAXA ('INntegrated Open TAXonomic Access') will be a web workspace in which taxonomic descriptions, identification keys, catalogues, names, specimen data, images and other resources can be accessed simultaneously according to user-defined needs. It will allow access to data held in multiple servers, and will use a distributed data model. If, in the future, the various nomenclatural Codes permit web publication of new taxonomic names and acts, INOTAXA will be able to integrate single descriptions placed on servers worldwide, so long as they are indexed through a registry such as operated by GBIF.

INOTAXA is built on a set of interoperable XML schemas, and is working with TDWG to ensure that standard schemas are used. These will allow external interoperability with GBIF and access to GBIF-mediated data. INOTAXA is also working with ZooBank[1], and has the potential to serve data in the format required to submit data directly. INOTAXA will provide seamless access from the content to other systems, including GBIF, TROPICOS and Flora Mesoamericana.

The INOTAXA project, although newly-named, was conceived and identified as a priority in a Mellon-funded meeting in 2002 at which a number of major museums and herbaria determined to demonstrate the potential of combining information, literature and research data held within their collections[2]. As a testbed for their ideas they determined to focus on Mesamerican biodiversity, building on a major literature resource, the *Biologia Centrali-Americana*.

In the first project phase the BCA was digitised and made public, and now provides the largest coherent body of taxonomic information available through the internet[3]. At the same time, the project team developed an XML schema for taxonomic literature, 'taXMLit', which is now being developed as a TDWG standard. The INOTAXA prototype will use part of the eBCA, with the addition of other taxonomic literature covering the same taxa, and other data as detailed below.

[1] http://www.iczn.org/new%20ZooBank.htm
[2] AMNH, NHM, RBGK, Missouri Botanical Gardens, NMNH, STRI, Smithsonian Institution Libraries. See http://www.sil.si.edu/digitalcollections/bca/documentation/proposal.pdf
[3] http://www.sil.si.edu/digitalcollections/bca/

Prototype function

a. Test architecture, methodologies and interoperability, both in the initial set-up and as external dynamic links are added (e.g. to GBIF, Tropicos, IABIN). Although the prototype will only deliver data, in the future we plan to extent the functionality to allow upload of information, such as single species descriptions. In this we will link (at least for zoological names) to the GUID system under discussion by the ICZN and GBIF. In any case we anticipate that additional information on the taxa covered will be added, as well as functionality to manipulate specimen-level data.

b. Provide proof of concept to underpin further grant applications. In order to complete the XML markup checking of the other BCA volumes, purchase the server hardware and software for the full version, and implement the full system architecture[4] additional funding will be required. The prototype will allow potential funders to see the potential for the system.

c. Identify user needs through feedback and advisory groups; we will be seeking funds for a workshop to identify how non-technical and technical users approach taxonomic data made available on the web, and how best to present options. In designing any web resource ensuring the users can access the data they require in the most appropriate manner is important.

1. PROTOTYPE CONTENTS

1.1. Project name

The name of the project has been changed from BCAC to 'INOTAXA' (INtegrated Open TAXonomic Access), reflecting a more all-encompassing vision for the project.

1.2. Prototype contents

The prototype will include:

a. The text and images of 2 BCA volumes: Coleoptera Vol IV Pt 3 part and Botany Vol 2;
b. The text and images from one or more modern taxon treatments of taxa covered by the BCA volumes included [to be provided in taXMLit by AW & CL];
c. The gazetteer of localities for the insect collections (AMNH);
d. Images of specimens of species in the BCA & additional treatments included in prototype. Metadata to be specified;
e. Up-to-date synonymic catalogue of taxa covered (to be mediated by GBIF in the future, but for prototype will have relevant subset on the local server);
f. Specimen data (to be mediated by GBIF in the future, but for prototype will have relevant subset on the local server);
g. Samples of prototype data to be available for inclusion on the server by February 2006.

1.2.1. Logos

The INOTAXA logo will be used on all pages

[4] See http://www.sil.si.edu/digitalcollections/bca/documentation/draft-EBCA-BCAC-HLA-final.pdf

Where text or other material from particular source forms all or part of page content the logo of that source (if available) will appear under the INOTAXA logo.

1.2.2. Linkages

Data will all be stored on the server for the prototype. However, some classes of data are being served from other sources (e.g. GBIF), and in the next phase of the INOTAXA project we will link to these sources. Consequently, we will seek to develop mechanisms in the prototype that will enable subsequent linkage to GBIF wherever possible. This will entail using standard TDWG schemas and tools developed by the wider GBIF/TDWG community, wherever possible.

The prototype will benefit from links to some external databases and search engines. These include:

a. Google;
b. Google Images;
c. Harvard Botanist Index Search (http://brimsa.huh.harvard.edu/cms-wb/botanist_index.html);
d. Flora Mesoamericana (see http://www.mobot.org/mobot/fm/intro.html);
e. GBIF (for all collections databases not on the INOTAXA server, including NMNH EMu, NHM EMu (when available), TROPICOS (http://mobot.mobot.org/W3T/Search/vast.html), INBio, CONABIO, etc.

In each case a search will be for the appropriate taxon/name that is referred to on the INOTAXA page (discussed in more detail below), and this entry of the search term should be made automatically so, for example, clicking on a link button on INOTAXA when the screen is showing species *Aus bus* will cause a search of Google for *Aus bus*. *[The choice of right-clicking, left single-clicking, left double-clicking for various features will be tested with users to see what will work most easily for the largest number of users. The options mentioned below may change when this has happened.]*

2. TEXT RENDERING

a. Cascading Style Sheet HTTP rendering with various links (see Section 5);

b. Html formatted or URL taxon locator – *unclear; is this covered under section 4, along with location of other entities? – to be postponed;*

c. Html formatted or URL taxon with highlighted inline referencing (similar to Google cached pages) - *see comments under search results in Section 4 – to be postponed;*

d. PDF rendering for reproducing (next phase);

e. Java thick client for portable device (on CD/DVD ROM, or memory sticks, next phase);

f. Machine-assisted automatic Spanish to English translation for description (next phase);

g. Generating customized output, such as checklists from internet user input (*next phase*?);

h. Dynamically or batch generating distribution maps. [Data will come from: text; digitised specimens, gazetteer, other interpretation. If more than one georeference for a single specimen, these will be used to create a bounding polygon.]

3. INDEX GENERATION

Index files for the content of the literature sources are needed for several purposes. First, they are a means of simplifying and speeding searches for the user. Second, as a means of generating 'cleaned' lists that can be used against authority files, linked to entries in the 'dirty' lists so that searching is more accurate, also to make it possible to list names alphabetically by last name however they are stored in the original data. Those indices which are generated for things without browse functionality are indexed for ease of advanced search capability.

i. Generating author name indices (it will need to link (in time, if not immediately) to the Harvard list of botanists (http://brimsa.huh.harvard.edu/cms-wb/botanist_index.html).

   a. ***Treatment author*** (indexed alone, full browse functionality provided);

   b. ***Taxon author*** (accepted name and all synonyms; indexed with Nomenclatural type and Related taxon authors, but full browse functionality only provided for Taxon author);

   c. Nomenclatural type taxon author;

   d. Related taxon author;

   e. Publication author (indexed with Subheading author, browse functions not provided);

   f. Subheading author;

   g. Introduction author (not indexed);

   h. Postscript author (not indexed);

   i. Author (within Bibliography citation; not indexed);

   j. Taxon circumscription author (not indexed);

   k. Type citation author (indexed with Other citation, Synonym other citation, Related taxon citation, Related specimen citation, and Vernacular name authors, but no browse functionality provided);

   l. Other citation author;

   m. Synonym other citation author;

   n. Related taxon citation author;

   o. Related specimen citation author; and

   p. Vernacular name author.

ii. Generating collection repository name index – there are relatively few of these listed in the BCA, but still it would be useful, and increasingly so for future work, including the other works that will be used in the prototype.

iii. Generating collector name index – these will be indexed by all surnames and by combinations of surnames (working together). We need to explore the issues of having names captured with initials that come first but still sorting on surnames. It will need to link (in time, if not immediately) to the Harvard list of botanists (http://brimsa.huh.harvard.edu/cms-wb/botanist_index.html).

iv. Generating geographic index – this needs to be a hierarchical index which shows valid (or even those proven to be invalid based on later knowledge) choices within a country, state, county or the like. In the prototype this will be a simple index, but in subsequent phases of the project it will develop into a gazetteer, allowing editorial decisions through an interpretation layer.

v. Generating publication name index – this is not just for journals, but also books, book series, etc. It will need to link (in time, if not immediately) to the Harvard list of botanical publications (http://brimsa.huh.harvard.edu/cms-wb/publication_index.html).

vi. Generating taxon name indices (these all need to be separate indices, which can be search alone or all in combination):
    a. Any name anywhere in the volumes (for the most basic search and browse functionality; names which are accepted, listed as synonyms or mentioned anywhere else, including in discussions should be displayed differently (e.g., italics, bold, other fonts));
    b. Names accepted (valid) in treatment;
    c. Names either accepted or listed in synonymy;
    d. New taxa described in the treatment;
    e. Names for which a key is provided (both 'keys to' and 'keys within');
    f. Names for which there is an image; and
    g. Vernacular names.

vii. Type specimens and Type Names cited.

4. BROWSE AND SEARCH FUNCTIONS

Taxon name spell checking recommended for next phase.

4.1. Top-level page

There will be one or more entry-level pages describing INOTAXA and giving access to additional information. These are not discussed here and are yet to be designed. One avenue from the entry-level pages is to the top-level search page.

The proposed top-level search page is presented in Fig. 1. It offers a choice of actions:

a) From this page a user may initiate a simple search on one or more words using the 'enter terms' field and a method of narrowing this search using the 4 boxes 'select taxon', 'select region', 'select work' and 'select context'. See section 4.2.
b) 'Browse Taxon tree', and 'Browse Geographic Tree', reached through the appropriate buttons. Clicking on either of these buttons will change the appearance of the search panel. See section 4.3.
c) Direction to a more complex search using Boolean logic reached through the 'Advanced search' button. Clicking on this button will change the appearance of the search panel. See section 4.4.

d) Clicking on the 'Image Search' button will change the appearance of the search panel. See section 4.2.7.

The Panel also displays the results of the previous simple search by the user (if any). It also has a Home button to take a user to the Top Level INOTAXA page.

4.2. Simple search

4.2.1. Search functionality

The search will be performed from the top-level page described in 4.1. (Fig 1). An explanatory text, not yet prepared, will describe the functionality of the search and in what form the responses will be.

The search panel has four drop-down boxes allowing the user to select a set of refining terms to constrain the search, and the context in which the search can be performed (Fig. 1). The default value for these is 'all', and the user need not change this in order to perform a search. Multiple selection of terms will be possible using the control key. The boxes are:

- Select taxon/taxa:

  This is a high-level list, limiting the search to kingdom, phylum or class and includes some high-level common names. The list is not exhaustive, but uses common names for taxa most likely to be popular choices (Fig. 2a (still incomplete)). The list will be static once completed. For the prototype, the entire list will show, but those selections that are not included will be 'greyed out'.

- Select Region(s):

  This is a list of Countries and sub-national regions represented by works within the prototype. As taxonomic work from elsewhere is included in INOTAXA so the list will grow. For the prototype the list will be as shown in Fig 2b (although this list is incomplete list due to space constraints on this page) with the addition of 'Mesoamerica' (for the countries shown), 'North America' (possibly with subunits 'Canada' and 'USA') and 'South America' (with all countries shown). The list will be static once completed.

- Select work(s):

  This includes the taxonomic works that are accessible through the portal (Fig. 2c). In the prototype they will be loaded onto the same server, but in the following project phase they will be accessible as a distributed resource, indexed perhaps through GBIF (see 1.2.2. above). The list will be generated dynamically from the database.

- Select context:

  This provides a selection of contexts (elements or groups of elements in the schema) within which the search term should be sought (Fig. 2d). The list will be static once completed.

  Although specimen data will be accessible in the system, the simple search will not query these data unless they are cited within an included treatment.

  The default search will be to look for an exact match for a term when this is a single word. In addition, the search term may use wild cards: *xxx (at end of word); *xxx* (within word); xxx* (at start of word). This will require help instructions. The search might operate on a single word (e.g. *Attelabus*), a word string (e.g. "*Attelabus viridans*" – the double quotation marks identify this as a string) or a set of words (e.g. '*Attelabus'*

[and] 'green'). The last of these should produce references to text sections within which both words were used, or where one was in the heading and the other in the subsequent text section. Other Boolean terms are not accessible through this screen, but on the Advanced Search. *[Note that the usual default Boolean operator in places like this is "or". This change is deliberate, and we will need to make it obvious and provide instructions.]*

Fig 3. shows a simple search with the word 'femora' inserted. On clicking on the 'search' button the contents of INOTAXA will be searched for occurrences of this word.

4.2.2. Search results

The search should find words in the text that have diphthongs (e.g. 'æ') or diacritical marks, even though the search is done without those marks. Thus, searching on either 'Salle' or 'Sallé' will enable all instances of 'Sallé' to be found. Curculionidae or Curculionidæ will find all instances of both.

A word might occur in a taxon treatment, a key or introductory (or many other places in) text. However, it is most likely to be under a taxon heading of some description. The results of a search should be displayed under a representation of their original context (not under the heading of the URL). Thus a search on the word 'femora' might produce a screen such as Fig. 4 [in this case, only the first three records have been shown, for the sake of example].

The default sort order used in the results screen for a simple search (or an advanced search where no other order is specified) will be:

a. alphabetically by taxon name, then
b. alphabetically by taxon author, then
c. alphabetically by publication author, then
d. numerically by publication date, then
e. numerically by page number.

The heading for each search result is the heading of the section in which the word occurs, together with the brief citation (author, date, starting page). The URL, date, size and % need not be displayed.

Clicking on the heading will change the display to show the text in context (Figs 5,6,7). The second of the search results in Fig. 4 ("*Attelabus* L. – key – Sharp, 1899, p. 2") leads to an identification key (Fig. 5).

In addition to displaying a frame holding the search results, the screen has two other frames:

a. A toolbar to the left, comprising the options from the original search panel, plus a set of context-specific options (see Section 6 below). The toolbar allows a new simple search to be commenced, while displaying the previous search and its logic (to which the results are displayed) in a text box.
b. The new logo (still in draft form).

These frames are present on all screens other than the search screens and results screens (although the toolbar changes according to the search / browse type).

4.2.3. Display of Keys (Fig. 5)

As noted in section 4.2.2., selecting the second result displayed in Fig. 4 will display the identification key. In the original text (see BCA vol 4 part 3, p. 2) the key is spread over

many pages with taxa between the different lugs. However, in the screen now displayed (Fig. 5) all of the key lugs should be shown in order without the interpolated text. In taxonomic literature some keys have a heading, others do not. In the example shown in Fig. 5 there is no heading in the original text, but one has been generated by the system by prefixing the taxon name in the heading (the treatment heading on p. 1) with "Key to". The citation for the key (Author, date, volume, part, page(s)) is provided.

Had the key been presented in the original text as a simple key with no interpolated taxa, it would have been displayed in the same manner.

In addition to the key the screen has one further frame: a taxon tree, starting three hierarchical levels above the subject taxon but not showing daughter taxa or sibling taxa. This does not function as a full tree, but is a 'reminder' to users where they are in the taxonomic hierarchy. It is generated from the classification appropriate to the text displayed and labelled 'Context' and the source of the context (e.g. BCA for use of that classification, 'Alonso-Zarazaga & Lyal, 1999' for the use of that context). The metadata of the publication that is the subject of the screen needs to carry the data of which classification context is appropriate[5].

4.2.4. Using the key to navigate elsewhere

Navigation from a key will lead to one of two options, depending on the type of key.

a. The key has lugs and couplets separated by other text (shown by the paragraph numbers applying to key lugs in the text not forming a single continuous sequence but interrupted by paragraph numbers that do not refer to parts of a key). The intervening text contains treatments of one or more taxa that are referred to by the key lug by means of a container name.

   i. The taxa within the formal or informal group identified by the lug (in the case of Fig. 5, species names) are listed after the appropriate key lugs, in alphabetic order.
   ii. Each taxon name, if double-clicked, will open the treatment of that taxon (in the publication from which the key is derived). Thus selecting '*Attelabus vinosus*' from Fig. 5 will produce the page in Fig. 6. [Once the taxon is displayed, the taxon tree in the top frame changes to reflect the position of this treatment.]
   iii. In the key shown in Fig. 5 the terminal taxon in each lug does not lead to a separate taxon treatment. Double-clicking on one, therefore, leads nowhere. If there is a treatment available for such a name, however, double-clicking on the name would have caused a new screen showing that treatment
   iv. The names (in the current treatment) are informal names for groups of species. However, they are treated as formal names elsewhere in the literature. Highlighting one and then using the side toolbar for 'Other treatments', 'Images', 'Browse taxon tree' will execute a search within INOTAXA (see section 5 below).

b. The key is continuous (i.e. the paragraph numbers attached to each lug are sequential with no gaps)

   i. The name (or indicator) at the end of the lug links to the taxon treatment heading so marked in the text.
   ii. A screen is produced with the relevant taxon treatment (cf. Fig 6).
   iii. Once the taxon is displayed, the taxon tree in the top frame changes to reflect the position of this treatment.

[5] N.B. this will mean that considerable thought will be needed later about links to ECAT, since this is likely to be dynamic and the texts in INOTAXA fixed. Versioning of ECAT might assist in resolving the problem.

If the key is derived from a publication in which there are no treatments, selecting a taxon will produce a screen with a tree that lists (and provides links to) the taxon treatments covering that taxon that are accessible within INOTAXA.

4.2.5. Display of taxon treatments (Figs 6,7)

Clicking on '***Attelabus vinosus* Sharp, 1899, p.2**.' in the search results shown in Fig. 4 (or double-clicking the name in Fig. 5) will open the full text of the taxon treatment, including the heading (Fig. 6). Note that if a search result is a variety, the full text display will be of the full species treatment, not just that of the variety. The characteristics of the display are that:

- The citation for the treatment is presented. (derived from the stored data)
- The heading is as it is in the publication (text, not format).
- The order of components of the description is in the order that it is in the text of the publication.
- A cropped image of the taxon will appear on the screen if there is one in the BCA.

Clicking on '***Apterocolus* Hamilton, 1998, p.208**' in the search results shown in Fig. 4 will open the full text of the taxon treatment, including the heading (Fig. 7), just as in Fig. 6. However, the tree in the top frame is different. It now shows the modern classification. The appropriate logo appears on the screen.

From this screen alternative treatments can be accessed by clicking on the "Other treatments" button on the toolbar. This will produce a screen with a tree that lists (and provides links to) the taxon treatments covering that taxon that are accessible within INOTAXA.

4.2.6. Display of other results

If the word was a geographical locality, e.g. 'Panama', the search should act as above, and lead to the complete taxon entry as in Fig. 6, not just a "Hab." block in isolation.

If the word or string is an Author, a 'simple' search will only search Treatment Author, Taxon Author and Publication Author. Other authors (see 'Generating Author Name Index') may only be searched using the 'advanced search' feature.

4.3. Image search

If the 'Image search' button on the panel shown in Fig. 1 (or any other panel where it appears), a panel such as that on Fig. 8 appears. A taxon name may be entered in the appropriate box (in the example shown it is '*Pilolabus viridans*'). If desired, the search may be constrained geographically through the use of a 'select region(s)' combo box, with the same selection list as in Fig. 2b.

On clicking the 'search' button a screen with thumbnail images appears (Fig. 8a). Clicking on an image will deliver a full-sized image on a new screen.

Image naming and metadata standards TBD.

Images available for the prototype include:

- Images from the BCA
- Images from other included treatments
- Images of specimens
- Images of species (not associated with any particular specimen)

Other images, such as of collecting localities, people, etc. will be added in later stages of the project.

4.4. Navigation by Browsing on a Hierarchical Tree

The intent is to navigate the content of INOTAXA through a hierarchy, so that information is presented in a form most useful for the user.

Navigation will use a hierarchical structure similar to that employed by Windows Explorer, with clicking on the plus (+) box next to a name will open the hierarchy at the next level down (similarly, clicking on a minus (-) box will collapse the hierarchy at the next level down). A simple (static) form of the tree has already been shown (Fig. 6).

Clicking on a term (rather than the plus or minus box next to it) a single time will highlight the term and allow functionality of the buttons on lower part of the left-hand toolbar, and on the right hand toolbar (if present) to operate on that term (if appropriate). Clicking on the term twice will open another screen with the appropriate details regarding that term (e.g. a taxon treatment for a taxon, or specimen data for a specimen record), should these details be available. If there are no details then the screen will not change (perhaps a temporary pop-up noting that 'There are no further data available' might appear?).

The default order for branches of a tree will be alphabetically within each hierarchical level.

Information relevant to a branch is contained in subfolders; if there is no information in INOTAXA for a given subfolder, that subfolder will not appear. For example, the taxon *Aus bus* in the taxonomy tree might have subfolders for Distribution, Synonymy, Specimens and Treatments. If there is no specimen-level data available the 'Specimens' folder will not appear.

The default method of displaying specimen-level data not generated from taxon treatments in INOTAXA trees is in a separate folder at the appropriate hierarchical level labelled "From Collections" (e.g. see Fig. 30). Within this folder the data from different collections are displayed separately.

When a tree is opened using any search term, the tree will open displaying the search term, the term at the hierarchical level above, and the search term itself opened to show the terms at the level below. Terms at the same level as the search term are shown if there is information within them; otherwise they are not displayed. If there is no information for a term entered as a search term, a screen will open stating 'No information on this term'.

Once a tree navigation has been requested by pressing a 'Browse....Tree' button, a screen as Fig 9 or Fig. 19 appears. The toolbar on the left for browse screens is similar to that in Fig. 4, other than the top of the bar shows on the alternatives 'Simple search', 'Advanced search' and 'Browse [..] tree', where the tree shown is the one not being browsed currently. This toolbar is persistent throughout a tree browse.

Once the tree is opened (see below) the Browse panel disappears, as does the top banner frame, and the INOTAXA logo appears in the top right of the screen.

4.4.1. Navigate by locality/geography

4.4.1.1. The same locality in different records

A single locality may be in the system from several sources, and with several different degrees of accuracy. However, it will not be possible at this stage to unite different 'versions' of the same locality other than where this has been done explicitly in one of the sources. In the BCA the localities given in the text, and the localities on the corresponding specimens, have been clarified and given georeferences in Selander & Vaurie (1962), a gazetteer which

will be accessible as part of INOTAXA. Localities from other sources, even if apparently identical, should only be considered the same if there is an identical georeference (or possibly not even then, depending on the resolution level of the georeference). In INOTAXA, arising from the BCA, other treatments, specimen data, and the gazetteer, the locality 'Bugaba' appears as the following:

a. 'Panama, Bugaba'
b. 'Panama, Bugaba, 1000 feet'
c. 'Panama, Chiriqui, Bugaba'
d. 'BUGABA, CHIRIQUI, PANAMA. Settlement on the Pacific slope about 22 km. north-west of David; 1000 feet; 8° 28', 82° 38'. The locality is sometimes cited erroneously as in Nicaragua'
e. 'BUGABA, NICARAGUA. See Bugaba, Panama.'

All of these are (almost certainly) the same place. However, they will appear at different places in the Geographic Tree. Rather than put in empty placeholders for missing levels, the DetailedLocation will open at a level below the level where there are data. Thus 'a' and 'b' above will open on the level below 'Panama' (Level1) and 'c' will open at the level below 'Chiriqui' (Level2). 'd' and 'e' will also open at the level below Chiriqui; both are from a gazetteer. The reference for Nicaragua will not appear on the Geographic tree.

Fig. 18 shows a way of dealing with this. Localities from the Gazetteer are grouped under 'Gazetteer data' at the 'detailed locality' level. DetailedLocality level data are identified by a prefix 'Locality:'. Level2 data are identified by a prefix 'Province' (or 'District' or 'State'). There are no Level3 data in the example, but as long as they appear together and the DetaliledLocality' data also appear together, a prefix may not be required.

Uniting these so that the records all appear at the same place will require use of the interpretation layer, and will not be implemented in this phase of the project.

The gazetteer table should include an attribution for each record. [In the next phase, a confidence level should be assigned to each record / georeference (cf Beaman & Kahn).]

4.4.1.2. Geographic Levels

The geographic levels accommodated in taXMLit are:

- Level0 - Continent or ocean level data
- Level1 – Political data at Country or equivalent
- Level2 - Political data at first level below Country, e.g. State, Province, District etc.
- Level3 – Political data at second level below Country, e.g. County, Province, District etc.
- DetailedLocation – Any cited location other than the above, including altitude, depth etc.
- Georeference – may include latitude and longitude, or other forms of georeference
- Altitude
- Depth

The top-level geographic term in INOTAXA will be 'World'. This does not exist in taXMLit, which starts with Level0 as continent or ocean. Although the vast majority of records in the system will be from Mesoamerica for this phase, there are some from North and South America in the BCA, and may be some from elsewhere in the other treatments to be included. At Level0, in the prototype we will use regions, initially restricted to 'North America', 'Mesoamerica' 'South America', with countries at Level1. At level 2 we will need

to specify the label given to the locality: 'State' is appropriate to Mexico, for example, but 'District' is appropriate for the equivalent level in Belize, and 'Province' in Panama.

'DetailedLocation' level data may be present in text without Level2 or Level3 data, or without Level3 data. In such cases, as noted above, the locality must appear at the next level without intervening empty placeholders.

4.4.1.3. Use of the geographic tree

Pressing the navigation button from the opening panel (Fig. 1) without any term being selected in the text or combo boxes leads to a screen similar to that in Fig 9, although in Fig 9 a search term, 'Mexico' has been entered in the starting place name.

If a place name is entered into the 'select region' combo box, or in the 'enter terms' box, this is used to identify the base of the tree, which will be one level above the term selected.

If the tree is opened from the toolbar while the screen is displaying a record, the tree should start one level above the most inclusive geographical area for the associated records. Thus if a species record is open with localities in Costa Rica and Nicaragua, the Mesoamerican list will open.

If the term entered is ambiguous (i.e. there are multiple instances of the name in INOTAXA), an interrogation screen will open, providing the alternatives (Fig. 10). In the exemplar screen 'Mexico, Mexico' is noted as '[unspecified]' because it is not clear from the original data what level this locality should be considered.

Clicking on the 'Browse Geographic Tree' box when an unambiguous term has been entered, or when a selection has been made from the panel in shown in Fig. 10, will open a screen similar to that in Fig. 11. The tree figured has been opened for Mexico [country] selected in the screen on Fig. 10. All siblings are shown. The folder for Mexico has been automatically opened at the next level to reveal two folders: "Data for Mexico" and 'Data within Mexico"

At each geographical level several options are available to provide further data. There are two folders for each place name, one to give access to data *for* the place (i.e. data that refer to that geographical level and not within it) and one to give access *within* the place (i.e. data that refer to geographical levels lower in the hierarchy) (Fig. 11). Opening either of these will provide access to data on Collectors and Taxa. In addition, opening the 'data within' folder will also allow access to the next lowest level of the hierarchy (States, in this case), and opening 'data for' will allow access to specimen data (Fig. 12).

If any folder is 'empty' (i.e. there are no data) it will not be visible. In this example Figs 13-17 depict an artificial situation where the system includes only material collected by Klug, Sturm, Salle and Chevrolat; folders (taxa) that do not include these are omitted.

Data for Mexico

The folder "Data for Mexico" includes three folders: "Collectors", "Specimens" and "Taxa" (Fig 12).

- Collectors: This will open a list of collectors who associated with specimens for which the locality data are no more detailed than 'Mexico'.
- Specimens: This will give access to specimen data that do not include more detailed localities than 'Mexico'.
- Taxa: this will give access to taxa for which there are no distributional data, or specimen-associated data, at a more detailed level than 'Mexico'

Opening the 'Collectors' folder reveals a set of initial letters, to sort the Collector names (by family name – see Indexing). These initials might only be necessary if there are many collectors, say above 25.

For each Collector there is a standard set of folders (Fig 13):

- Biographical information (planned as a link to external resources, such as the Harvard site at http://brimsa.huh.harvard.edu/cms-wb/botanist_index.html, where these are available). Clicking on the folder should open a new window with the appropriate material in the source setting [to be developed further].
- Entry in standardised list. This is derived from the index to names discussed above. It should open a new screen displaying the forms and abbreviations for that name [index still to be developed – this functionality to be developed further]
- Itinerary. This is the itinerary of the collector for Mexico [TBD in a later phase: may mean that it is simply a date/set of dates where he was in Mexico, or the full itinerary within the country. I assume it is derived from the specimens collected, in which case at the level under discussion we only know 'Mexico' and not even a date. The itinerary may differ at different levels in the geographical hierarchy (in which case we get information that [the collector] was at [the site] for [a set of dates]. This could be various things, including published itineraries which have been included in INOTAXA as well as the above. To be discussed.]
- Specimens. These data will be for specimens collected by the collector, with locality data no more detailed than 'Mexico'.
- Taxa. This gives information about the taxa collected by the collector identified for the level of the geographical hierarchy specified (i.e. taxa that in the example either have the distribution including 'Mexico' with no more detailed information, or are based on specimens that have no locality data more detailed than 'Mexico'.

The folders have the following properties:

- Specimens. The data will in the prototype be held on the server. In the next phase it will be accessed not only locally but through GBIF. Fig. 15 shows the "Data for Mexico / Specimens" folder opened. In this artificially limited dataset we have recorded only two collectors who recorded specimens at the geographic level of 'Mexico': Sturm and Klug (see below). On opening "Specimens" the tree should automatically open until there is a choice between folders to be made (e.g. if there were specimens both of Coleoptera and Lepidoptera the tree would have opened until the two insect orders were displayed; selection of one of them would have opened the tree further until there was another choice). Specimen data might come from several places: the text of the treatments (in which case data pertaining to a single specimen might be captured from more than one treatment), and specimen data as digitised directly from a museum/herbarium collection (which also might be duplicated in a taxon treatment). In order to minimise duplication of the same data the specimen data are accessible in the context of different treatments and 'from collections'. In Fig. 15 specimen data for Attelabus klugi Gyll. have been recovered from the taxon treatment in Sharp, 1889 and from digitised data in two collections (BMNH, where two specimens have been databased, and NRS, where 5 specimens have been digitised). The same data are available if the "Specimens" folder under "Collectors/K/Klug" is opened (Fig. 14). Repository codens can be taken from standardised sources (e.g. http://hbs.bishopmuseum.org/codens/codens-inst.html for Abbreviations for Insect and Spider Collections of the World; http://sciweb.nybg.org/science2/IndexHerbariorum.asp for plants (we can get a copy of

the latter on request)). [TBD: how the specimen data should be shown; see also below (4.3.2)]

- Taxa. The taxa are accessed through a taxon tree. On clicking the 'Taxa' '+' box, the taxon tree should automatically open until there is a choice between folders to be made (e.g. if there were taxa both of Coleoptera and Lepidoptera the tree would have opened until the two insect orders were displayed; selection of one of them would have opened the tree further until there was another choice). In the case of Fig. 14, this opening has been until a single terminal, *Attelabus*, since there are two specimens available in the database. In Fig. 14 the 'Collectors/K/Klug/Taxa' folder has been opened to show the species *Attelabus klugi* Gyll., which was treated in the BCA on the basis of two specimens, one collected by Klug with the data 'Mexico', the other by Salle with the locality 'Parada' (probably La Parada, Oaxaca, Mexico). The species name should therefore appear at this geographical level for Klug, but it will appear at the level of La Parada, Oaxaca, Mexico in the geographical browse to species. The species was treated by Hamilton (1994) as *Pilolabus klugi* (Gyllenhal). However, Hamilton did not cite the Klug specimen, and his treatment would not appear here. Specimen-level data for the Klug specimen will be generated from the text (as well as from digitised collection data), and will be accessible through the 'Specimens' folder as discussed. Fig. 15 shows the 'Data for Mexico/Taxa' folder opened. In this case it gives access to taxa represented by specimen data, of which there are only two (see above).

Data within Mexico

The folder 'Data within Mexico' includes three folders: "Collectors", "Taxa" and "States".

- Collectors: This will open a list of collectors who associated with specimens for which the locality data are more detailed than 'Mexico'.
- Taxa: this will give access to taxa for which there are distributional data, or specimen-associated data, at a more detailed level than 'Mexico'.
- States: this will give access to the next lowest level within the geographical hierarchy.

There are no 'specimens' folders at this level. This folder will only appear at the detailed locality level, since the utility of these data at a more encompassing geographical level is questionable, and the number of specimens that might be included is very great, and would risk incapacitating the system in download.

Opening the 'Collectors' folder reveals a set of initial letters, to sort the Collector names (by family name – see Indexing). These initials might only be necessary if there are many collectors, say above 25.

For each Collector there is a standard set of folders:

- Biographical information (as discussed above).
- Entry in standardised list. (as discussed above).
- Itinerary. (as discussed above).
- Taxa. This gives information about the taxa collected by the collector which have data more detailed than 'Mexico'. The taxa are accessed through a taxon tree as described above. In Fig. 16 the 'Taxa' folders of the collectors Chevrolat and Salle have been opened to show the species *Attelabus splendens* Gyll., which was treated in the BCA on the basis of at least two specimens, both (all) collected at Vera Cruz, in Mexico. The species name will therefore appear 'Within' Mexico and 'For' Vera Cruz. The species was treated by Hamilton (1994) as *Pilolabus splendens* (Gyllenhal). However, Hamilton did not cite the Salle specimen, and his treatment would not appear under Salle's name.

He did cite the Chevrolat specimens (two of them, with the same data) and thus the name appears under Chevrolat.

Opening the 'States' folder will provide the same choices of subordinate folders as at the country level. In Fig. 17 the State 'Sonora' has been opened; choices for 'within Sonora' and 'for Sonora' are displayed.

4.4.1.4. Desirables at this level:

The lists of taxa and specimens associated with a locality at any level should be downloadable.

4.4.2. Navigation on taxonomy structure tree

Names at all levels in the taxonomic hierarchy should be accessible, including synonyms at family level and below. With each name will be a folder for 'Synonymy' (family-level and below), 'Treatments' (all levels), 'Specimens' (species-level and below) and 'Distribution' (species-level and below') and 'Taxa' (all levels to infra-species).

- 'Synonymy' is built from the PrimaryCitation and the SynonymCitations in the source of the classification. The order is AuthorName, PublicationDate, Publication (full string). If any element in the published work is missing it is not marked by a gap, it is simply omitted as is any separating punctuation. The order shown is PrimaryCitation, then (OtherCitation+SynonymCitation) in date order where all citations are dated, or alphabetically by author, source, then by text order within a work when not all are dated. *[When dates are available in the Interpretation layer, they are used for ordering – to be implemented at a later phase]*
- 'Treatments' is built from the TreatmentAuthor, TreatmentDate and page number of the Treatment. If there is more than one treatment of the taxon within INOTAXA all are shown, in TreatmentDate order. If the user clicks twice on a treatment, the treatment will open, generated from XML.
- 'Specimens' is built from specimen data from the treatment texts and from specimen lists from collections. As in Figs 14 and 15, each collection or treatment that gives rise to specimens is treated separately. Clicking on a tree terminal at specimen level opens a table with just that specimen or specimens listed, with data conforming to the standard decided on (Darwin Core 2 or ABCD, TBD). Clicking on a specimen record derived from a treatment opens a table with just that specimen or specimens listed. If the user clicks twice on 'Specimens', a table opens with all specimen data listed, whether from specimens or the literature, but the source clearly indicated.

If the user clicks twice on a taxon name in the Browse Taxon Tree view, a taxonomic summary screen will open, showing:

- The valid (currently accepted) name of the taxon, its authorship, the primary citation, secondary citations and synonymic citations, and its type (if available). These data are generated from the most recent classification available in INOTAXA, which may (in the next phase) be the one generated from GBIF.

The opening screen (Fig. 19) allows the user to enter the taxon within which they wish to commence the browse. This obviates the need to commence at the highest hierarchical level in the system and work down. A set of instructions is also presented (TBD and we will provide).

If no entry term is set, the first tree that will appear is the top-level one including both animals and plants (Fig. 20). Clicking on the (+) boxes by 'Animalia', 'Invertebrata' and 'Vertebrata' will deliver the tree in Fig. 21, which is derived from the classification used in the BCA. Further clicking on the (+) box by Insecta (and closing 'Vertebrata') will deliver the tree in Fig. 22.

The prototype will contain more than one classification of some of the groups covered, and each of those groups will have different classifications available. The most 'complete' tree for 'all' animals and plants currently is drawn from the BCA itself. Consequently 'Browse Taxon Tree' will open the tree derived from the BCA (and include the eBCA logo). Where a more modern classification is available, a choice will be offered to view the other tree for the section on display (again with logo). For logistical reasons, this choice will only be offered at the level of order and below. In the prototype the choice will only be for the volumes covered (other choices will be 'greyed out' if not included in the prototype, but choices included in provided lists such as is shown in Fig 22).

Although the BCA classification is the most complete overall, it does not contain all possible groups or all terms within groups contained. For example, although 'Curculionoidea' is the standard modern terminology for the taxon including Curculionidae and other weevil families (and will be known to most scientific users of INOTAXA), it is not in the BCA, where the older (and now unused) term 'Rhynchophora' is employed. Thus if 'Curculionoidea' is entered as a search term the BCA classification might not be accessible, although in the case discussed immediately below, the classifications of O'Brien & Wibmer and Alonso-Zarazaga & Lyal would. There may be a case to ensure entry of some synonyms of the more 'important' names in the system to allow the BCA classification to appear in some places where it might be inaccessible (although as soon as a 'Curculionoidea' tree was opened the BCA classification would have appeared as an option on the Browse Taxon Tree toolbar, as discussed in the previous paragraph).

Figure 23 shows the initial taxon tree browse screen with the name 'Curculionidae' entered. Three alternates will be available in the prototype, the default BCA classification and two alternatives:

a. Biologia Centrali-Americana ('BCA')
b. O'Brien & Wibmer, 1982
c. Alonso-Zarazaga & Lyal, 1999 ('AZL')

From the screen in Fig. 23, the screen in Fig. 24 will be shown, presenting the choice available.

Having selected 'BCA Classification' by clicking on the button, the screen on Fig 25 opens. As usual, the tree has opened at one level above the selected term, and one level below. The page has the standard toolbar that appears in a tree browse on the left, and two logos – the INOTAXA logo and the eBCA logo (because the tree shown is generated from the eBCA). It also has a 'Browse Taxon Tree' toolbar on the right (with the same content as the panel on Fig. 24). This is to retain the possibility of examining the alternative classifications available. If a further classification became available during the navigation of the tree, because INOTAXA contained one for that taxon, a pair of buttons referring to that classification would appear). If the tree started with only one classification available (as in Fig. 20), no Browse taxon Tree toolbar would be shown; however, as soon as a second classification became available as the tree is opened, a toolbar presenting the alternatives would appear. The toolbar has buttons below the named alternative treatments allowing comparison between them. Note that older classifications can be compared to more recent, but not the other way

round (newer classifications will generally include more names). This is discussed in section 4.3.2.2 below. The tree can be opened to display the various folders

In Fig. 26 the Taxon tree has been opened to display Taxa within the Curculionidae, and the folders within the subfamily Pterocolinae. In the latter case there are only two folders, Taxa and treatments, since there is no synonymy available.

In Fig. 27 the Pterocolinae have been opened to show the genus *Pterocolus* and species *Ptorocolus auricollis*. The synonymy of *Pterocolus* is generated from the classification source: the BCA (the numbers before the reference will not be seen in the prototype; they were used here solely to order the citations in the software used). The distribution folder at the species level is populated from the specimen-level data in the treatment(s) and from the specimens. When there is more than one treatment, or a treatment as well as specimens, there should be more than one distribution shown; it will be important to identify the source of the distribution data, and this could be done by placing a suffix against the data with "Source: [treatment]" or "Source: specimen data". This differs from the display of information from different sources elsewhere in the tree, but provides a more user-friendly way of portraying the information.

In Fig. 28 the tree has been opened for *Attelabus klugi*. The specimen data have been opened to show the sources (cf Figs 14,15).

4.4.2.1. Standard Toolbar functionality

This refers to the toolbar on the left-hand side of the screen.

If, when a name is highlighted, a button on the lower part of the left-hand toolbar is pressed (e.g. 'distribution map', 'specimens' etc), it will act as if the full treatment were open. If two trees are open, as in Fig. 24, the button will act on the name on the most recent tree. (see Section 6 for discussion of the functionality of the toolbar).

4.4.2.2. Browse Taxon Tree Toolbar functionality

If the button for another tree is clicked on when a tree is being displayed, the tree is replaced by that of the classification selected.

If no taxon on the open tree is selected (highlighted by a single mouse click), the new tree will have as its base the same taxon as did the tree which is currently open, and will appear closed. Thus, if at the screen on Fig. 28 the 'Alonso-Zarazaga & Lyal (1999) classification' button is pressed, because no taxon is highlighted and the current tree is based on the term 'Curculionidae' the tree on Fig. 29 will appear. As on Fig. 25, the name of the next highest rank will appear and the term 'Curculionidae' is opened to the next lowest level. The folder 'Synonymy' is present because there is a synonymy available from AZL; there was not one in the BCA, which is why the folder was not shown on Fig. 25.

As the tree appears the toolbar changes so that the 'BCA Classification' button no longer appears depressed but the 'Alonso-Zarazaga & Lyal, 1999 Classification' (hereafter referred to as "AZL" button does appear depressed. The appropriate logo has also appeared on the screen.

In Fig. 30 the new tree is opened to cover the same species as is opened on Fig. 27 - *Pterocolus auricollis*. It differs in a number of respects from the tree in Fig. 27.

- Because of changes in classification between the BCA and AZL, Pterocolinae is no longer a subfamily in the taxon Curculionidae, but is now a subfamily of Rhynchitidae (which

was a separate subfamily of Curculionidae in the BCA). The user has consequently had to follow a different past to get to the subfamily and species.

- The synonymies are more extensive than those in the BCA, and include the BCA treatments where appropriate. The greater coverage is due to the later date and higher completeness of the later classification.
- For species described in the BCA (such as *Pterocolus auricollis*), the Synonymy folder includes the reference to the original description as the primary citation.
- The Distribution folder has been populated with summary data from four places: Sharp (1890), Hamilton (1998), AZL, and specimens. AZL do not provide specimen data in the database, but do have a field for 'Distribution' which is used to populate the record on the tree[6].
- The specimen data for Hamilton (1998) are very numerous, and are not listed in the Figure; they would be seen in the prototype, however.

Although the tree framework is derived from AZL, this work did not include the full range of data that are presented in Fig. 30. [In fact AZL only treats taxa at the levels of subgenus and above, but a more complete classification derived from it, and incorporating species, will be provided for the prototype; it is this version which is being discussed here]. There will have to be a mechanism within the prototype to allow association of data from different sources. Thus the data from the BCA can be addressed using the classification derived from that source, or more recent classifications (with conditions – see below). Data from Hamilton (1998) can be addressed using the AZL classification but nothing earlier. That said, the systematic position of *Pterocolus auricollis* differs between the BCA and the AZL classifications. We will need to implement internal unique identifiers for each name and concept, and use TCS, to deal with this issue.

4.4.2.3. Comparison of classifications

Others are working on comparison of classifications and we might wait for the standards to settle before implementing this functionality in this phase of the INOTAXA project. However, at least some functionality is required. This is prefigured in the discussion immediately above. Moreover, wherever we wish to select 'Other Treatments' (standard toolbar) or link specimens that are stored under different synonyms of a single taxon for analysis, we will need to be able to correlate the different names applied to a single taxon by different authors (i.e. in different classifications). The schema most appropriate the carry out this functionality is the Taxon Concept Schema, allied with unique identifiers; *this will be tested during implementation of the prototype*.

When there is more than one classification accessible (in the prototype), visual comparison of the placement and status of a taxon between the classifications is desirable. The next few paragraphs suggest a display that might work with a 'Compare Tree' function (TBD exactly what can be included in the prototype).

[6] In this document we suggest that Distributional data might be abstracted from the text of treatments by summarizing all records in 'DistributionAndOrSpecimen' and then presenting only Level0 and Level1 without duplicating the Level0 entry for multiple different records at Level1. However, the Distribution is often summarised (and to a more inclusive extent than justified by the specimens examined) in a GeographicDiscussion paragraph. In the latter case it would make sense to be able to abstract the data completely, but we have not given a means to do so. In the AZL database there would be no alternative available, although in this case there is already a dedicated field for Distribution.

Fig. 31 shows the tree open for the taxon *Allocorynus mollis* in the BCA classification. The species name has been highlighted, thus making the toolbars 'live' for that taxon.

In Fig. 32 the 'Compare Tree' button under 'Alonso-Zarazaga & Lyal (1999) classification' has been pressed (if the 'Alonso-Zarazaga & Lyal (1999) classification' button itself had been pressed the current tree would have been replaced, as described above). The AZL tree appears alongside the BCA tree, open to the same taxon, with the taxon immediately above also visible and can be manually opened to reveal appropriate taxa. Note that in the BCA tree the classification is:

Rhynchophora: Curculionidae: Allocoryninae: Allocorynus: Allocorynus mollis

In the AZL tree the classification is:

Curculionoidea: Oxycorynidae: Allocoryninae: Rhopalotria: Rhopalotria mollis

In this case there is a 1:1 relationship, although this may not often be the case.

The issue of comparing trees will be discussed in more detail in an additional document (to be prepared). The alternatives for constructing the prototype are to not implement this at all or hard-wire in a set of linkages as an example. Developing a functional system is a 'desirable', and for quite a lot of the functionality of the prototype is necessary.

4.5. Advanced (Boolean) Search

The advanced search allows more complex combinations of selectors than does the simple search, and the application of Boolean logic. The starting screen, reached via the 'Advanced search' button on the opening search page (Fig. 1), is shown in Fig. 33.

Five boxes are visible on the search panel, including a 'previous search' read-only box. On the right-hand side of the search panel are three boxes. Each has the same drop-down list (Fig. 34b). These Fields specify the order in which the search results are to be grouped (if required), the order being determined sequentially from top to bottom of the three boxes.

On the left-hand side of the search panel is a single combo box. The drop-down list is of the fields that may be searched (Fig. 34a). These include both indexed and non-indexed fields (see section 3 above) (the list will be extended to include all non-indexed fields that users may need; there are many fields in taXMLit, although not all will need to be there; we will provide ordering for the list. Ultimately (but probably not until the next version) this will also include the appearance of flags to allow users to select values for some attributes of the XML elements.

When a user selects a field from the list, three more items appear on the search panel.

(i) If the selection made in the first box is of an indexed field (see Section 3 above), a combo box will appear next to it, labelled "select term". If the selection made in the first box is of a non-indexed field, the second the box will be text box, labelled "enter term". The default search will be to look for an exact match for a term. In addition, the search term may use wild cards: *xxx (at end of word); *xxx* (within word); xxx* (at start of word). This will require help instructions.
(ii) a combo box with a choice of Boolean operators;
(iii) a 'search' button (Figs. 35, 36).

4.5.1. Non-indexed field

When a non-indexed field is selected the second box is a text box (Fig. 35).

When the user has selected or entered a value into the second box and the Boolean operator box, a new row starts, allowing the user to build a complex query (cf Fig. 30). This repeats until the user hits the SEARCH button.

4.5.2. Indexed field

In this case the potential to use the index to build up the term in the second box is available (Fig. 36). For example, in the case of a Collector Name, as the user begins to enter the name in the second box, the choices decrease. With each letter added a word is proposed from the index (the first alphabetically with that combination of letters). Hitting the drop-down button (or alt+down arrow) on the combo box will show the list as it stands (i.e. before entering any letters the full index would be accessible). One or more words may be selected from the drop-down list (e.g. having entered 'Linn' the user might select 'Linné', 'Linnavouri' and 'Linnaeus'. There are some standard abbreviations, for example 'L.' ideally should automatically select 'L.', 'Linne', 'Linné', and 'Linnaeus', but not 'Linnavouri' or 'L. f.' However, there are potential errors in this, and we will generate a list of standard abbreviations, either for the prototype or in a later phase. Selection of multiple terms will be done using the mouse button and the control or shift keys, as within a Windows directory. Selecting more than one term would automatically develop an 'or' search. If the facility to select a term is not used, letters may be entered until the word sought is presented in the box.

4.5.3. Boolean operators

The operators ([and] / [or] / [not]) will be used. Parentheses will also be used to indicate nested terms, and may be used in the same box in which another operator is given.

4.5.4. Screen changes after 'Search' is selected

Once the query is built the 'search' button is pressed. The screen changes to one similar to that in Fig. 30. The search panel is now on the left-hand side of the screen, in a manner analogous to the way in which the simple search panel was transformed on a search. The panel has the Advanced search components in the top half, and the other toolbar functions in the lower half.

The search argument now appears in a box on the panel labelled 'previous search', and the three boxes are emptied. The toolbar has two header buttons permitting the current search to be refined with the addition of further operators, fields and terms ('Refine search' button), or a new advanced search to be started ('Start new' button). An additional 'select operator' box has appeared, followed by the 'Select field' box. The 'enter/choose term' box is a text box, but depending on the term selected for the first box could change to a combo box. Use of these boxes to refine a search is as follows:

a. a Boolean operator is selected for the top box (as no operator was selected in the previous search this has to be done before refining the search is possible);
b. a field is selected in the second box;
c. a term is entered/selected in the third box;
d. if a further term is to be added the Boolean operator is entered into the fourth box. At this point the argument will be added to the extant text in the 'Previous search' box and the search boxes will clear for the next argument to be entered (the top 'select operator' box will be greyed out, since an operator has been selected);

e. once the argument is complete, the 'Search' button is clicked, the argument added to the extant text in the 'Previous search' box and the search boxes will clear (including the top 'select operator' box, since no operator has been selected). The results of the search will appear on the main screen.

If no Order terms have been specified in the search the output is in the format of Fig 3, since essentially a simple search has been performed.

However, if an Order term is specified a tree such as in Figs 38 appears as an intermediary between the search entry and a list of locations for final results. Not all of the search Order terms will be fully applicable, however. Use of 'Geography' works simply at species level, but is more difficult to apply (and is less meaningful) at higher groups. If 'Geography' is specified as the primary Sort Order for a species search for taxa at a species level the display will list the species in alphabetic order under a geographic tree, resolved in each case to country level, although the tree may be further opened to reveal increasing detail . The 'taxonomy' used, unless there are confining conditions, will be the most recent available in INOTAXA for the groups covered.

4.5.5. Advanced search: Collectors names

Within taxonomic texts a collector's name is usually only presented as a last name ***(Fendler)***. Sometimes, however, there it is 'first initial, last name' ***(S. Hayes)*** or rarely an abbreviation (this will have to be dealt with, it is much rarer than abbreviations of taxon authors, but should be noted as a possibility). Often there is more than one collector's name e.g. ***(Schiede & Deppe or Schiede et al.).*** To deal with this and similar issues with other indexed fields, if the term is entered without quotation marks (e.g. Schiede), all terms in the index with that combination of letters will be searched for (e.g. D. Schiede, Schiede, Schiede & Deppe). If quotation marks are placed round the work (e.g. "Schiede") only that word will be sought. Names with a preceding initial should not appear in a search under the initial (i.e. 'S. Hayes' should not appear in the same search as 'Schiede', but should appear with other collectors whose names begin with 'H'). However, it is important that the integrity of 'S. Hayes' is retained in the database to enable reconstruction of the text in XML. To solve this, a separate field might be constructed by use of rules for searching in which the initial and the name are reversed, or by a rule to do this during a search. See 'Index Generation'. Moreover, 'S. Hayes', 'S Hayes' and 'S.Hayes' should be seen as the same, not three different names. There may also be examples where full specimen citations have three or more collectors spelled out, while a publication has only, for example "Schiede et al.", this will also need to be dealt with at least in future versions. See also 'Index Generation'.

A search for the collector 'Schiede' is presented on Fig. 37. No Boolean operator is selected, but the search is ordered by taxon, then geography, then collection date.

The results of the search on Schiede are presented in Fig. 38. They take the form of a tree, with groupings first by taxon hierarchy, then by geographical hierarchy, and then by date (collecting date, in this instance, as specified in the grouping boxes). The tree has opened from one level in the plant hierarchy above the level with data, and opened to the level below the taxa in which collections were made. Because, as discussed above, there is an operational need to distinguish between specimens cited taxon treatments and specimens listed in specimen databases, the source of the specimens opens in a folder immediately after the taxon. In Fig. 38 all of the sources are treatments, although in each case there is a possibility of an additional folder labelled 'From Collections' (cf Fig. 30).

In Fig. 39 the tree has been opened fully for three species. After the treatment (container for the specimen data) the next level is 'Geography', as per the order specified. The tree then opens as a geographical hierarchy from Level1 to DetailedLocality, as on Fig. 30. The next level is collector (since the collector has been found as 'Schiede' and 'Schiede & Deppe', with details of the collection (using standard botanical codens for repositories from http://sciweb.nybg.org/science2/IndexHerbariorum.asp) and date, as specified in the original sort order.

Double clicking on the collector at the terminal branch will open the specimen data.

4.5.6. Advanced search: Taxa

Fig. 39 also shows a new advanced search being commenced on the side toolbar.

For taxa, three alternatives will be given in the drop-down list: Family-group (Superfamily, Family, Subfamily, Tribe[7], Subtribe), Family, Genus-group (Genus, Subgenus), Genus, and Species-group (Species, Subspecies, Infrasubspecies – variety, form etc), Species, Infraspecies.

Entry of the taxon name in the 'enter term' box will produce a list in the drop-down box as described above as letters are added.

Selection of one or many names will be performed as discussed above. Having entered the terms the 'Start new' button is clicked.

The screen generated from this search is given on Fig. 40. Note that the top text/combo boxes of the search bar have cleared and the search terms now appear in the 'Previous search' box. Because the search did not make use of a Boolean operator, and because there were no search order terms used, the result appears in the same manner as that of a simple search (cf Fig. 4).

Some names occur many times within a single treatment. Thus a search on '*Ophryastes*' would actually produce a list containing considerable duplication (Fig 41a). To improve the functionality of this search, occurrences of a name that are within the same treatment of the name should not be shown separately. Thus of the records in Text-box 1 only the ones shown in Fig 40 would be displayed. In Fig. 40 each record shown refers to a treatment or key, and the separate occurrences of the name are given in context each separated by an ellipsis. For example, '*Ophryastes'* is mentioned twice in the key on p. 88. The first time the text reads: "....Genera *Ophryastes, Tosastes, Caccophryastes,* and..."; the second time it reads "....imperfectly cavernous. *Ophryastes."* See Fig. 40.

Because a genus might be mentioned in the discussion within the Treatment of another genus on the same page, the page number cannot be used alone as the cue. Each of the records above that refer to 'Discussion' is within a separate treatment. Clicking on a heading will open the relevant treatment, key or other section, as described above for a simple search.

For an Advanced Search at the species-group, although the species-group name will be used for the search, the output will include the genus name and author. In the drop-down list to allow selection infraspecific names occur as well as specific names, and in the appropriate alphabetical position. (Fig. 41b)

[7] Within the BCA 'tribe' is used in a non-standard manner broadly equivalent to superfamily. We will need to consider how to deal with this.

Note that the author's name is not obvious for all instances, since species names are sometimes used in the text without the author name and in a context where the author name is not immediately apparent *(if it's there put it in, if not leave it out – this may be changed once interpretation layer is included; TBD)*. In these cases the author name may not be shown in the drop-down list (although development of GUIDs and the attention of a subject specialist could deal with the problem).

To get a list of species of a given genus and see their distribution, the search would be Field: Genus; Term [genusname]; Operator: [and]; Field: species; Term [blank]; Operator [blank]; Order by: Taxon then geography.

4.5.7. Advanced Search: Geography

The selection of search terms works as described above.

Entry of locality at any level in the geographical hierarchy is permitted. Currently we have not specified the Geographic Level at which a search might be carried out, and in some cases specifying a Level could be counter-productive for a user, since in many cases they will not know to what level a name has been assigned in INOTAXA. We might allow searching by each Locality field as well as all of them. A search of all Locality fields on 'Panama' will return results at Level3 (Panama City), Level2 (Panama Province) and Level1 (Panama, the country).

A name that is only found at a lower level in the geographic hierarchy (e.g. DetailedLocation) will be searched to display all examples of the name, irrespective of whether the associated higher-level elements (Level1 and Level2) are the same or not. For example, searching for 'Alamos' will produce instances of the town whether it be the one in Sonora or Chihuahua.

Fig 42 shows the results of an Advanced Search for 'Alamos' (in this case specified at DetailedLocation level). The sort order has been specified as 'geography' then 'treatment author' then 'publication date'. In this case there are (at least) four different DetailedLocation records for 'Alamos' accessible, two from the BCA and two from Selander & Vaurie.

The results show the appearance of the term throughout INOTAXA, including treatments, the gazetteer and specimen-level data from collections. Because specimen-level data do not sort by 'treatment author' and 'publication date' they will appear separately. As noted, the default method of displaying specimen-level data not generated from taxon treatments in INOTAXA trees is in a separate folder "From Collections" (e.g. see Fig. 30). Again, because no order has been specified that applies to specimen data derived from collections, they will be displayed in a default order, which is 'taxon' (genus + species then genus) then 'geographical locality' then 'source collection'.

Selander & Vaurie is the gazetteer for location interpretation for BCA insect records (which apply both to records in the BCA text and specimens referred to in that text, if separate digitised data are available). Although equating a BCA insect locality with the gazetteer locality would be justified (bearing in mind that if it were to be done it should be flagged as an interpretation), this is not justified for localities outside the insect portion of the BCA nor for localities cited in other works if these are not based on the same specimens. In these cases 'Alamos' might actually have referred to the nearest named place rather than the exact locality, or to a place not listed in the gazetteer. Thus all localities should by default appear separately, with their source (see discussion under 'browse taxon tree' above).

Some place names have been corrected in the gazetteer. However, in the prototype the link between the incorrect version and the correct name cannot be made. Thus a search on similar

to the last but on 'Alvares' will present the screen in Fig. 43. Double-clicking on the 'Champion, 1982 p.425' folder will open the relevant treatment (although note that because 'taxon' was not selected as a sort order, what taxon treatment the user will be directed to is not displayed on the tree). Double-clicking on 'Selander & Vaurie, 1962 p. 20' folder will have a different result, specific to only this work within INOTAXA (and ultimately other gazetteers that are included). In this case a separate window will open with the gazetteer entry, and the entry to which there is a cross-reference in Selander & Vaurie, as in Fig. 44. Presentation in this way allows the user to conduct an emended search with the alternative spelling.

4.5.8. Advanced Search: Author

The selection of search terms works as described above. The fields available for searching include "Search by Taxon Author" and "Search by Treatment Author". An explanation must be provided for the user – Taxon Author will lead only to taxa authored by the person, 'Treatment Author' will lead to taxa covered by that author in INOTAXA, whether that author is also the Taxon Author or not.

For example, searching on 'Sharp' as a treatment author, with the sort order 'taxon' will produce (among others) the results in Fig. 45 (note that not all taxa with Sharp as a Treatment Author are displayed here, for reasons of space on the Figure). Although all of the results shown are from the BCA, were there other treatments of these taxa by Sharp in other publications in INOTAXA, these would also be displayed. Double clicking on any treatment reference will open the treatment.

If a search is carried out with Sharp selected as Taxon Author, the search must be more extensive than that above. The following occurrence types will be found (at least!):

a. treatment headings
b. citation/synonymy paragraphs
c. as 'associated taxa'
d. within discussions, both of the same taxon and of different taxa
e. as taxa cited in keys
f. image legends (and the metadata of the images themselves)
g. specimen data cited in taxon treatments (including earlier identifications cited on label data!)
h. specimen data from collections (including earlier identifications cites on label data!)

The variety of possible responses means that this is a search that would be better constrained, and is likely to trigger a message to that effect (if we add the functionality).

Of the above list, 'a'-'d' will appear (in Fig. 46) within 'taxon treatments' folders (noting that the treatment may be of a different taxon from the one for which Sharp is a taxon author), 'e' will appear in a 'keys' folder, 'f' will appear in an 'images' folder and 'g' and 'h' will appear in a 'specimens' folder, the latter under a separate 'From Collections' folder.

In Fig. 47 a refined search has been entered, and on clicking on the 'Search' button the screen on Fig. 48 appears. A species with Jekel as the taxon author has been added to the previous list.

In Fig. 49 a new search has been started on the toolbar. The previous search results have not yet been moved. Because a Boolean operator term has been employed, the screen shown on Fig. 50 will immediately appear, without a button being clicked on. The search term and sort order has been placed in the 'Previous search' screen, and the 'refine search' button appears

depressed. A message appears on the screen informing the user of the status of the query. Also in Fig. 50 the refinement of the search has been entered. On clicking on the 'Search' button the screen in Fig. 51 appears.

5. TOOLBAR FUNCTIONS

As shown on the preceding figures, there is a toolbar on the left hand side of the screen when a search or browse of INOTAXA is in progress. The top part of the toolbar is defined by the nature of the search or browse, and there is a text-box in the middle of the toolbar stating the preceding search (the results of which are displayed). The bottom half of the toolbar comprises a set of buttons that function with respect to what is displayed on the screen. Some function within INOTAXA, some provide external links.

5.1. Links to resources within INOTAXA

The toolbar will have the following buttons enabling links to resources within INOTAXA:

a. 'Other treatments';
b. 'Key(s) within taxon';
c. 'Key(s) to taxon';
d. 'Distribution map';
e. 'Specimens(s)';
f. 'Gazetteer' *to be added*
g. 'Toggle to PDF';
h. 'Toggle to JPEG';
i. 'Image(s)';
j. 'Home';
k. 'Back'

5.2. Functionality

The prototype will not serve all content for all species, but provide an example of functionality.

5.2.1. 'Other treatments'

This will be live if a taxon treatment is open, or if a name is highlighted on a tree browse. In the first instance clicking on the button will allow access to treatments other than the one open, in the second to all treatments of that taxon. Depending on the detail of our content development for the prototype, we would either:

i. link only to taxon treatments with the same name as the one highlighted / on screen, or (preferably):
ii. link to taxon treatments of all synonyms of the name highlighted / on screen *[this will be explained more clearly]*.

The control will only select treatments, i.e. where the taxon name is in the heading or the synonymy of a taxon treatment. If there is only one other treatment in INOTAXA, when button is clicked a new window opens with the full treatment (the old treatment is still available on the extant window). If there is more than one treatment available, a new window opens with a list of the treatments and their authors and dates of publication. One or more of these can be selected, although each will open in a new window.

5.2.2. 'Key(s) within taxon'

This will be live if a taxon treatment is open, or if a name is highlighted on a tree browse (and there is a key included in a TaxonTreatment that matches the name). In either case clicking on the button will open a new window. If there is only one key within the taxon (or within a synonym of the taxon, if we have that functionality), then the window will open to that key, which will have all of the functionality already described for keys. If there is more than one such key, the window will open with a list of the keys and their authors and dates of publication. One or more of these can be selected, although each will open in a new window.

5.2.3. 'Key(s) to taxon'

This will be live if a taxon treatment is open, or if a name is highlighted on a tree browse (and there is a value in "TaxaReferredTo" that matches the name). In either case clicking on the button will open a new window. If there is only one key to the taxon (or to a synonym of the taxon, if we have that functionality), then the window will open to that key, which will have all of the functionality already described for keys. If there is more than one such key, the window will open with a list of the keys and their authors and dates of publication. One or more of these can be selected, although each will open in a new window.

5.2.4. 'Distribution map'

This button will be live if a taxon treatment at species level or below is open, or if a name is highlighted on a tree browse. In either case clicking on the button will open a new window. The map is generated from georeferenced specimen data available, and built from point source data only. The data will be drawn from publications and specimens, but the likely duplication of data will have no effect on the map. Additional functionality is expected later, including the ability to zoom, go to a specific point and see the complete data specimen data associated, etc. (TBD).

5.2.5. 'Specimens(s)'

This button will be live if a taxon treatment at species level or below is open, or if a name is highlighted on a tree browse. In either case clicking on the button will open a new window. The specimen data will be taken from all treatments of the taxon in INOTAXA, and from collection-held specimen data that are available. The various sources will be listed, and all or any of them opened, in the manner discussed above.

5.2.6. 'Toggle to PDF'

This button will be live if the main frame holds the taXMLit-based version of the text of any INOTAXA component. Clicking on the button will open a new window. This frame will show the PDF version of the text concerned, on however many pages of the original it occupied. It may be downloaded.

5.2.7. 'Gazetteer'

This button will be live if the taXMLit-based version of the text of any INOTAXA component, or if a location is highlighted on a tree browse, or if specimen information is displayed in an INOTAXA window. Initially it will only function for the insect BCA volume, as it will be based on the Selander & Vaurie publication. If a location is highlighted in the main screen (i.e. in the BCA 'Hab block') or in specimen data arising from the BCA

collections, clicking on the button will display the Selander & Vaurie data, as in Fig. 43. In future the gazetteer content and applicability will be extended.

5.2.8. 'Toggle to JPEG'

This button will be live if the main frame holds the taXMLit-based version of the text of any INOTAXA component. Clicking on the button will open a new window. This frame will show the JPEG version of the text concerned, on however many pages of the original it occupied.

5.2.9. 'Image(s)'

This will be live if a taxon treatment is open, or if a name is highlighted on a tree browse. In either case clicking on the button will open a new window.

Without other commands, the window will show a set of thumbnail images of the taxon currently on the screen, with the references of the publication concerned, if appropriate. These images might include: EBCA plate, image in BCAC, photo (digital image) of type, photo (digital image) of figured specimen (if different), photo (digital image) of other specimen(s). Clicking on any of the images will produce a larger image, and obscure the rest *(or have them as a slideshow cf Picassa?).* Images from the eBCA should be cropped to the taxon concerned. *Should we also consider geographic locations? Authors? There seems to be no good reason to restrict this, though some could be added in later versions or as such images are added).*

If a figure reference within the text is highlighted, clicking on the Image(s) button will open a new window with the references image displayed as a full-sized image *[we need to discuss image size and resolution at some stage, to develop a set of different sizes and a mechanism of moving between them. Do we allow download?]*

5.2.10. 'Home'

The button is active at all times. It takes the user back to the opening screen of INOTAXA (not opening search screen, the very top screen which is yet to be developed).

5.2.11. 'Back'

The button is active at all times. It takes the user to the previous screen.

5.3. Links to external web resources

The links have been outlined above 1.2.2.

a. [Flora Mesoamericana *Action: CL to discuss with SK when appropriate; Anna to discuss with GD*]

b. Botanical author authority file (Harvard) [http://brimsa.huh.harvard.edu/cms-wb/botanist_index.html]. There will be a button on the toolbar for 'Botanist database'. This becomes live whenever an author or collector name is highlighted in the main frame (ideally only if a plant name is associated). When pressed, the button will open a separate window for the Harvard University Herbarium Index of Botanists record of the individual in question. *[question; the database requires an initial*

selection of 'author' or 'collector'; can this be avoided or, if not, how should it be dealt with?]

c. Google There will be a button on the toolbar 'Search Google'. This will be live whenever text is highlighted in the main screen. It will also be live when a taxon treatment is on the main screen. The search will be performed on the words highlighted, or on the subject of the taxon treatment (scientific name). When pressed, the button will open a separate window for Google with the answers to the search. Google images. There will be a button on the toolbar 'Search Google Images'. This will be live whenever text is highlighted in the main screen. It will also be live when a taxon treatment is on the main screen. The search will be performed on the words highlighted, or on the subject of the taxon treatment (scientific name). When pressed, the button will open a separate window for Google with the answers to the search.

d. GBIF. There will be a button on the toolbar 'Search GBIF' (Prototype only; the next version will have a dynamic link to GBIF). This will be live whenever a scientific name is highlighted in the main screen. It will also be live when a taxon treatment is on the main screen. The search will be performed on the name highlighted, or on the subject of the taxon treatment (scientific name). When pressed, the button will open a separate window for GBIF with the IPR acceptance screen and then, once the user agreement is approved by the user, the response to the search (*CL/AW to discuss with Donald H*.).