GoldenGATE (Pro-iBiosphere) XML Markup Editor

Introduction and Manual for the Generation of TaxonX-based Legacy Literature Documents using the GoldenGATE Editor

DRAFT 20150305

Jeremy Miller, Donat Agosti, Guido Sautter, Terry Catapano & Christiana Klingenberg

(klingenberg[at]plazi.org; sautter[at]ira.uka.de; agosti[at]plazi.org; catapano[at]plazi.org; jeremy[at]naturalis.nl)

The GoldenGATE release, demo files and detailed instructions are available at:

http://plazi.org

PLAZI taking care of freedom

GLOBAL BIODIVERSITY INFORMATION FACILITY

This version is the preparation for the next V6 of the GoldenGATE manual.

Content

1 Introduction
- 1.1 What is GoldenGATE?
- 1.2 What does GoldenGATE do?
- 1.3 Why xml?
- 1.4 The GoldenGate user interface
- 1.5 System requirements
- 1.6 Document Pre-preparations
- 1.7 Test documents
- 1.8 Making GoldenGATE work

2 The mark up process
- 2.1 Starting with the mark up process
- 2.2 Opening an HTML Document from a local file
- 2.3 Opening an HTML file from GG server
- 2.4 Normalize an HTML document
- 2.5 Opening and post-processing a PDF document
- 2.6 Check spelling
- 2.7 Get or clean MODS: Add metadata of the publication
- 2.8 Parse Bibliography: Bibliographic records and citations
- 2.9 Mark citations
- 2.10 Run Fat: Taxonomic section
- 2.11 Parse Taxon Names: Complete the Taxa.
- 2.11B Markup Taxonomic Name Authority
- 2.12 Markup Treatment Borders: Treatment mark up 1
- 2.13 Treatment Structure: Treatment Markup 2
- 2.14 Get LSIDs
- 2.15 Normalize paragraphs
- 2.16 Mark Up Material Citations (1)
  - Check of the Collection Codes
  - Markup of collecting countries
  - Markup of collecting regions
  - Markup of the entire collecting event of one specimen or specimen group
- 2.17 Attribute Materials Citations (2): Detailed markup of additional information
- 2.18 Geo-Reference Locations
- 2.19 Saving the document and Upload
  - 2.19.1 Save locally
  - 2.19.2 Uploading the document to the Search and Retrieval Server (SRS)
- 2.20 Viewing the document on the SRS
- 2.21 Citing a treatment
- 2.22 Citing a bibliographic reference of citation of a bibliographic reference
- 2.23 Citing a figure caption
- 2.24 Linking via a httpUri attribute

3 Workflows
- 3.1 Workflow for markup html documents
- 3. 2 Workflow for markup pdf documents

4 Editing
- Highlight an attribute
- HIghlight the content of an attribute
- Annotate a token or string
- Annotation of Attributes
- Remove Attributes

- Remove annotations of an attribute
- Annotation of geographic coordinates
- Counting specimens
- Linking to external resources, names or treatments

5 Trouble shooting and special cases
- Footnotes
- Citations / reference groups in catalogues and as part of treatments:
- Template html document
- The new up-dates in an uploaded treatment does not show up on SRS
- HttpUri annotation does not work in SRS-html

6 Glossary
- 6.1 GoldenGate Expressions
- 6.2 TaxonX Expressions
- 6.3 Element Descriptions and Examples
- 6.4 GoldenGate Internal Markup Elements

7 GoldenGATE versions
- 7.1 GoldenGATE “classic”
  - 7.1.1 Master Configuration
  - 7.1.2 pro-iBiosphere.editor
  - 7.1.3 Edaphobase.editor
  - 7.1.4 PensoftImporter
  - 7.1.5 TaxonX.editor
  - 7.1.6 TaxonX-Dev.editor
  - 7.1.7 TaxonX-PDF.editor
  - 7.1.8 TaxonX.markupWizard
  - 7.1.9 ZooTaxa.markupWizard
  - 7.1.10 WebEditor
  - 7.1.11 WebServices
- 7.2 GoldenGATE Imagine

8 Advanced editing and programming
- 8.1 Change configuration
- 8.2 Create analysers
- 8.3 Create pipelines
- 8.4 Programming export from GoldenGATE XML
- 8.5 Programming import into GoldenGATE XML
- 8.6 Annotate document attributes
- 8.7 Annotate Document properties
- 8.8 Add and change publication ID

9 Links to external resources
- 9.1 Links related to GoldenGATE
- 9.2 References

10 Acknowledgments

1 Introduction

1.1 What is GoldenGATE?

GoldenGATE is an editor which allows the creation of new XML content from plain text / html / pdf data. The idea is to support a user in creating XML markup as far as possible. This comprises automation support for manual editing of XML as well as fully automated creation of markup.

The key features of the GoldenGATE document editor are:

- Import of txt, html or pdf files to the markup process
- *XML mark up:* text passages or single phrases or words can be annotated and "tagged" with specific xml elements
- *XML Editing:* individual XML elements can be hidden and shown so that only those tags are visible which are important for the particular editing operation in progress.
- *Text Editing:* GoldenGATE provides a text editing mode that displays the content as plain text, which makes the data way easier to edit because there are no tags needing to be paid attention to.
- *Markup Conversion:* GoldenGATE is highly flexible and it uses some very simple conversions which are easily configured and managed, but still are powerful enough to handle common conversions.
- *Native Natural Language Processing (NLP):* The GoldenGATE editor natively provides components that allow handling and applying gazetteers and regular expressions for automatically creating fine-grained markup.
- *Extensibility:* GoldenGATE can include several types of plug-in components that serve a variety of purposes.
- *Operation Sequencing*: joining of frequently editing step sequences to pipelines.
- *Batch Mode:* Capacity of automatic processing of documents, e.g. doing some markup transformations in an overnight job.

GoldenGATE's strength is that it can be highly customized and versions created that a specific to a particular task such as a domain, a specific journal, or input format. The key for succesful use of GoldenGATE is to choose the right version. The various versions are listed in Annex

1.2 What does GoldenGATE do?

GoldenGATE helps the user with the xml mark up of plain and html text or pdf versions of documents. The flexibility of the editor allows the creation of individual mark up scenarios, containing automatic and semi-automatic mark up workflows. In addition it is a useful tool to prepare taxonomic texts and their contents to be linked to other databases via individual ID referencing.
GoldenGATE also saves the processed texts on the Search and Retrieval Server (SRS at http://plazi.org), so that they can be accessed for data mining. Further, the processed document can be stored locally or at the Community Markup Server (CMS at

http://plazi.org), so that other contributors can continue with a finer markup level.http://harn.ufl.edu/gardens
Finally services exist allowing external users to download and harvest treatments or parts of.

1.3 Why xml?

The Extensible Markup Language (XML) is a general-purpose markup language. It is classified as an extensible language because it allows its users to define their own tags. Its primary purpose is to facilitate the sharing of structured data across different information systems, particularly via the Internet. (from Wikipedia)
An XML schema is a description of a type of XML document, typically expressed in terms of constraints on the structure and content of documents of that type, above and beyond the basic syntax constraints imposed by XML itself. An XML schema provides a view of the document type at a relatively high level of abstraction. (from Wikipedia) When using the TaxonX configuration of GoldenGATE, the processed documents are going to be transformed in the TaxonX xml schema.

1.4 The GoldenGate user interface

The GoldenGATE main window is divided in three parts (Fig. 1). On the left side the custom functions are leading through the entire mark up process (4). On the right side check boxes for the xml tags appear (6). The colours of the tags can be changed by double click. Only the used annotation types appear. The mid part is the text window where the editing text appears (5). It is possible to write within the text to do minor corrections.
The menu bar above (3) offers a series of additional functions which may be used during the mark up process and to facilitate some mark up steps.

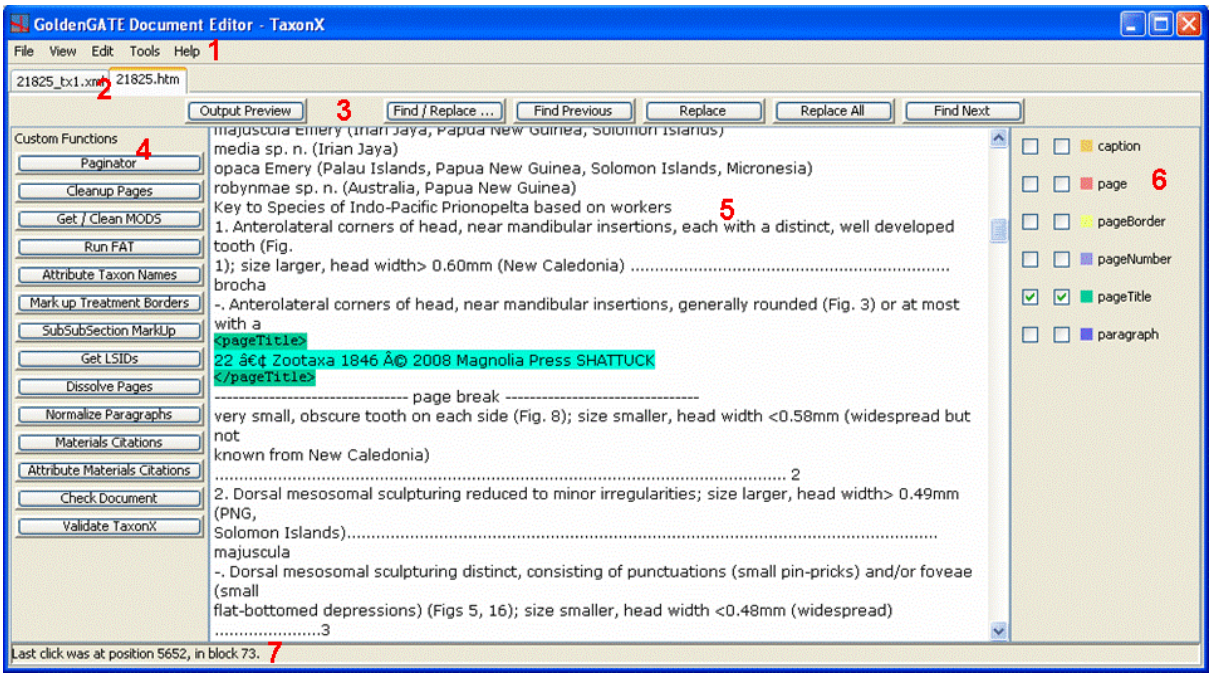


Fig. 1. 1 Basic layout of opening window. 1 General task bar. 2 Indicator of open files. 3 Editing functions. 4. Customized functions following the mark-up process. 5 Text window. 6 Used tags. 7. Position of editing indicator.

The following chapter are structured following the mark up levels.

1.5 System requirements

The minimum system requirements are:

- Operational systems: MS Windows 2000, Windows XP, 7
- Random access memory: minimum 2 GB
- Java: version 1.5 or 1.6 (no guarantee for beta releases)
- Internet connection: ADSL is fine for getting the MODS Header, a broadband connection is better for getting the LSIDs and connection with other databases.

For MS Windows, the download zip file contains the GoldenGATE.exe to start the editor; on all other operating systems, use GoldenGateStarter.jar.

1.6 Document Pre-preparations

Pdf files: For non-digital born pdf, it is recommended that the documents are scanned on a flat bed scanner or equivalent (providing straight text images not showing bent pages) in gray scale at 300dpi. For born-digital pdfs no constraints but size are known but highly dependent on the processing engine.

Html files. It is recommended to prepare the input documents in html format without any formatting but the paragraph elements (<p>) and a solid (<hr>) line to mark page breaks. For the processing of the documents no further elements are needed. It has proven extremely advantageous to have very clean and spell checked OCR-output, since artifacts have an impact on the mark-up process, and later corrections are very tedious. It especially has a negative effect on the recognition of names.
Html files can incorporate either the entire document or just sections of, such as one singel treatment.
For text recognition standard OCR programs can be used (eg. ABBYY Fine Reader). It is important that the output from those programs allows to retain the page breaks and page numbers that are necessary to define the position of each paragraph to be able to refer to a specific page in a document. Alternatively, a simple html document can be prepared by copy pasting the relevant bits into a plain html template (see html template).

1.7 Test documents

This test set of pdfs can be used to start the markup process using the GG pdf import routine.

https://drive.google.com/a/plazi.org/folderview?id=0B_yrQwn4yBySbXRRZWtMaGxqUGM&usp=sharing

1.8 Making GoldenGATE work

Download the latest GoldenGATE release:
http://idaho.ipd.uka.de/GoldenGATE/GoldenGATE.zip or for Pro-iBiosphere
http://plazi.cs.umb.edu/GgServer/Downloads/GoldenGATE-pro-iBiosphere.editor.zip

Unzip the folder on any place you prefer outside the system partition (GoldenGATE only unzips, no installation is required or will occur) and start with **GoldenGateStarter.exe**.

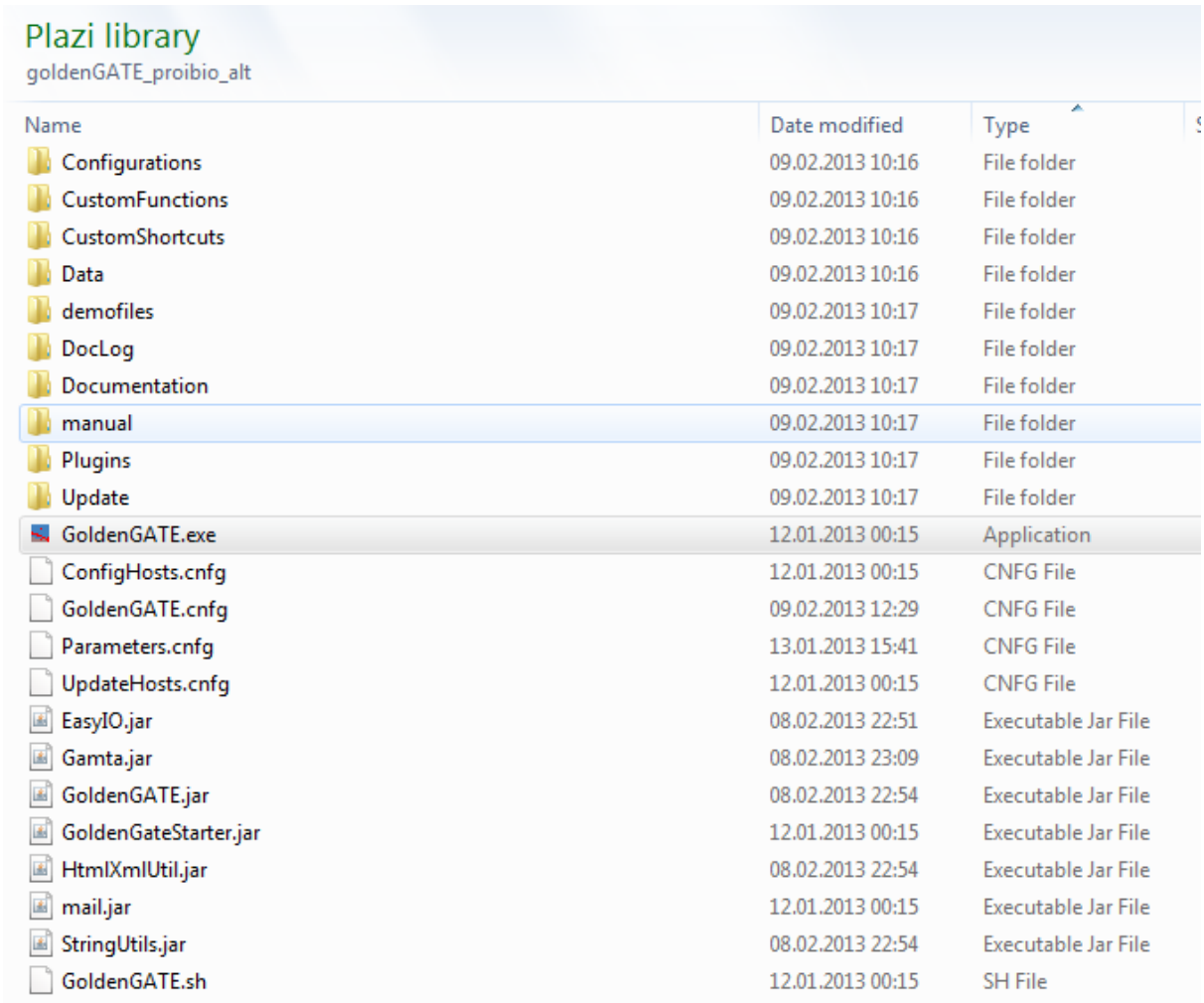


Fig. 2: The unzipped GoldenGATE directory.

Fig. 3: GoldenGATE Editor asks for Internet access.

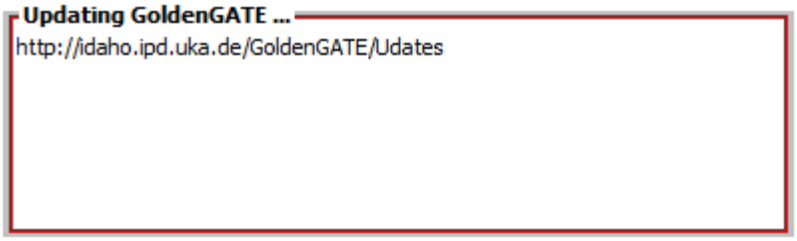


Fig 3a. GoldenGATE is downloading updates

Depending on your internet connection you can choose whether you want to allow internet access. If you have a broadband internet connection, you should obtain for “Ja” (= yes). Even though loading the configurations for TaxonX may take a while. (If you choose “Nein” (=no) see further instructions below.)

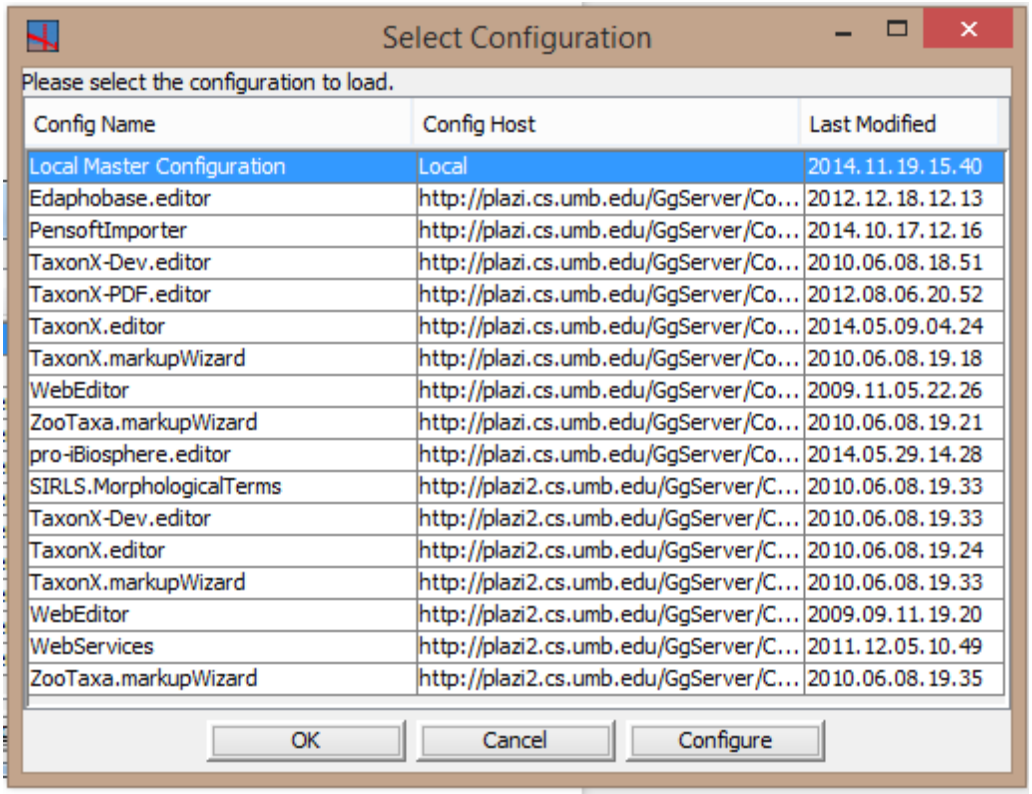


Fig. 4a: The configuration selection menu. The download for all configurations may take some minutes. This is the first time download menu. The vairous conifguations are all customized GG versions.

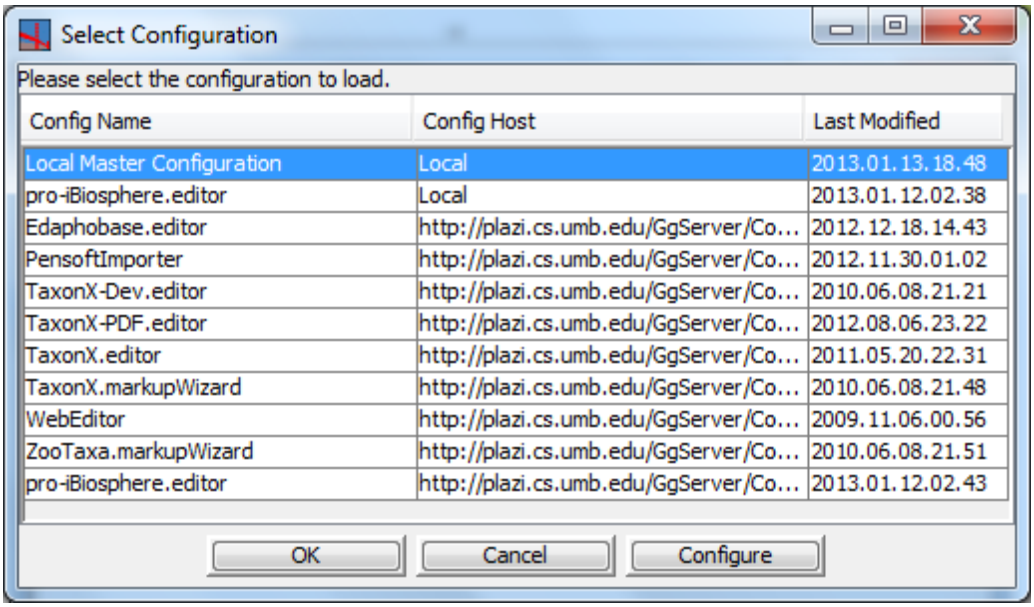


Fig. 4b: The configuration selection menu when opening GG after the initiatl installation process.

Choose “**pro-iBiosphere.editor**” (Fig 4a) and in the subsequent starts chose the the local configuration host (Fig 4b)

.

A window appears and tells you, that the selected configuration is not available locally, but that you can download it, and make it to your local configuration. You should accept. The download may take some time, depending on the Internet connection (ca. 90 MB). During the download the following (Figs 4c-d) notes appear which do not recquire an action.

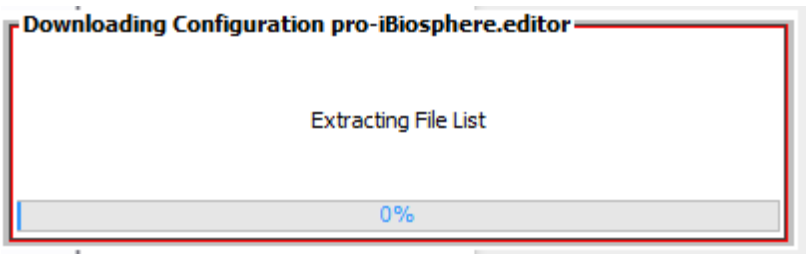


Fig. 4c Progress status window

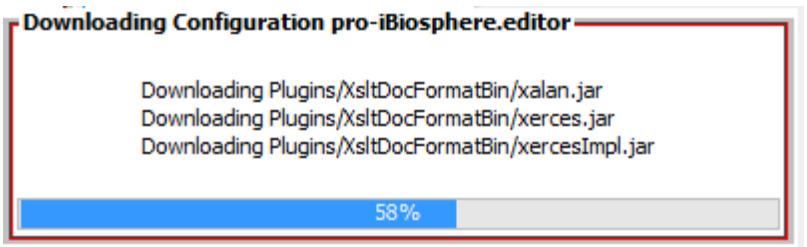


Fig. 4d Progress status window

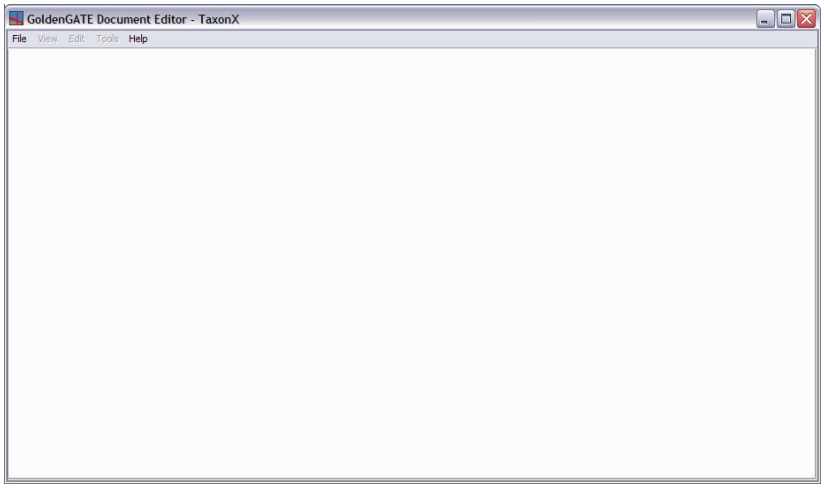


Fig. 5: The initial window after a successful download and installation of GG.

The initial setup is done now and you can start with document processing.

2 The mark up process

The whole workflow of the mark up process can be divided in different logical parts.

a) Cleaning up the document structure and download MODS

All paragraphs should be set up correctly. Page breaks, page title and page numbers should be set and attributed correctly. All further mark up steps depend on the quality of the cleaned up document structure. The MODS data set (Metadata Object Description Schema) includes all necessary metadata of describing the publication.

b) Taxonomic section

The document is going to be scanned for all taxonomic names. Single taxon parts such as species name only are associated with the respective genus.

c) Treatment section

Treatments are taxonomic essays, including formal taxonomic descriptions or any other information about a given taxon. The treatments are often subdivided in several parts such as nomenclature, description, material examined etc.

d) Connection with taxon related databases

All taxa are getting individual IDs (Life Science IDs) by connection with a database, and those taxa unknown to the database are added to the database.

Further mark up

Further mark up beyond includes tagging the material citations (collecting events, collecting locations), so that these are properly identified for use for internet tools like Google Maps. Bibliographic references are able to be tagged as well and associated with corresponding handles/DOIs (if available).

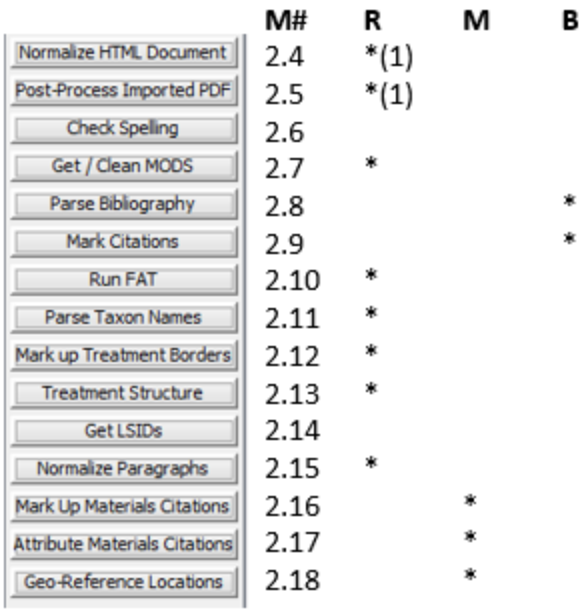


Fig xxx_001_overview. Overview of the Markup process. M reference to the chapter below. R required steps to upload the document to SRS. M markup of the materials citation. B Markup of bibliographic data.

At the end of each mark up process the documents should be saved at the Search and Retrieval Server (SRS).

2.1 Starting with the mark up process

To perform the markup, open a document containing taxonomic information. This can be an html, pdf, xml or txt document. Make sure, that the document contain pagination information such as page numbers. You can load this document from your local machine or from the internet:
File ☐ Load Document (from file) ☐ chose as file type “HTM and HTML files (default reader) or any other file type mentioned above.
In the following chapters the difference between the process of opening HTML or PDF documents is described.

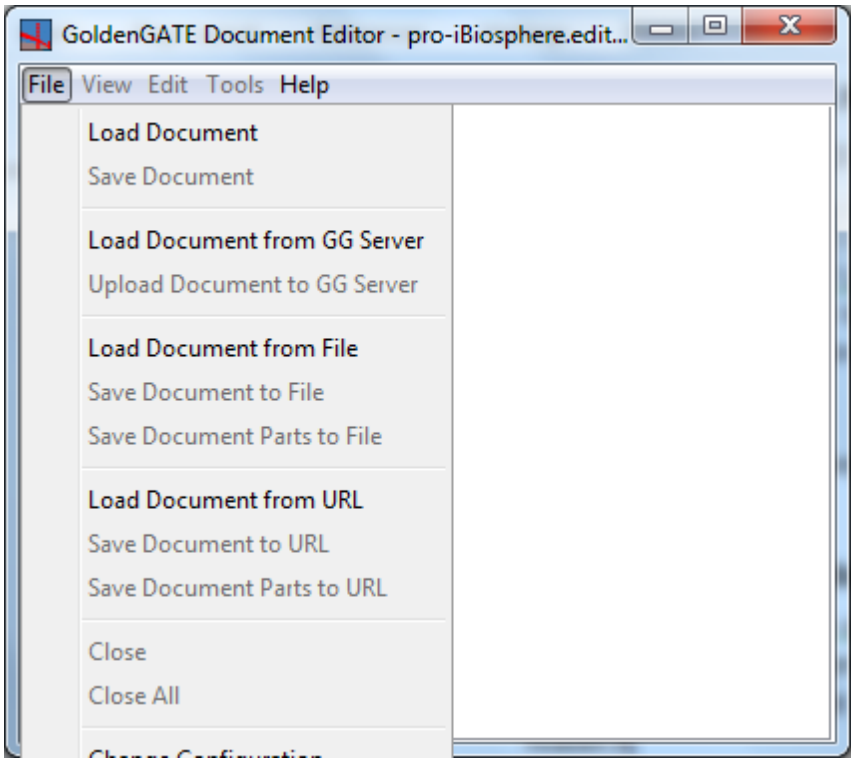


Fig. 6: Opening a file menu: pull down menu

2.2 Opening an HTML Document from a local file

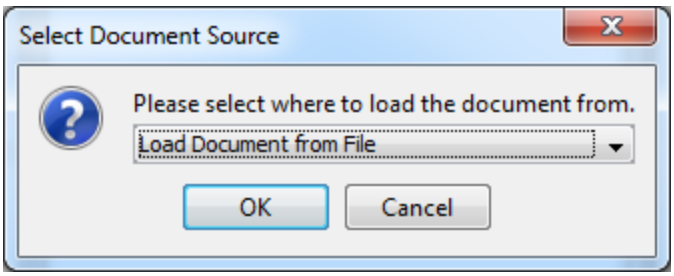
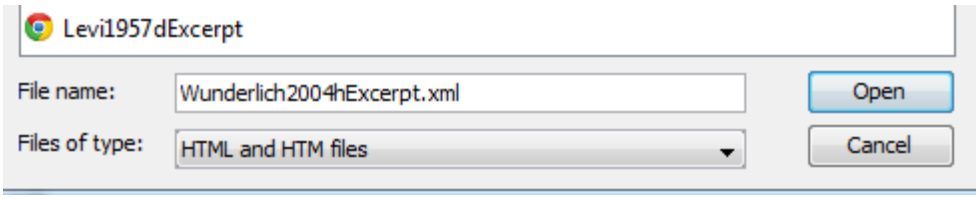


Fig 7_1. File menu. Menu following after selction of the File menu.

For the use of a locally stored file, use the “load document from File” option, and select



HTML and HTM files.

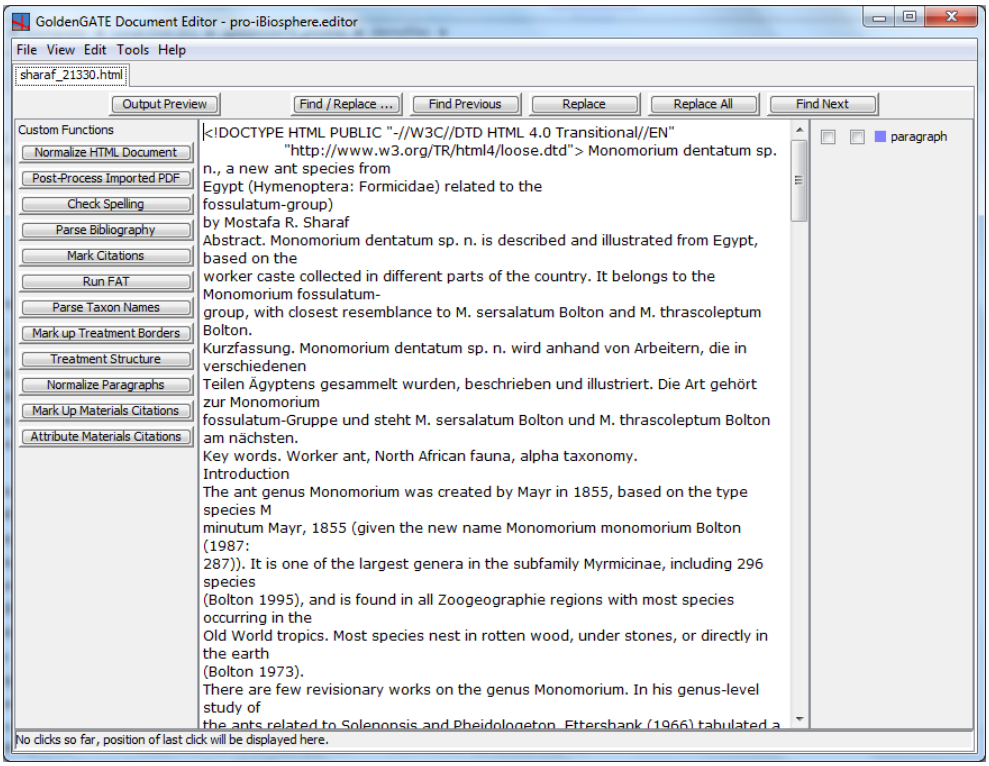


Fig. 8: Opened HTML document.

2.3 Opening an HTML file from GG server

From the file menu, choose “Load document from the GG server” (Fig 6) and enter in the following menu your credentials (Fig 6.2)

Fig 9. A user name and password can be obtained from goldengate@plazi.org

Once the identity is checked, a list of the existing documents will be returned (Fig. 6.3 and 6.4)

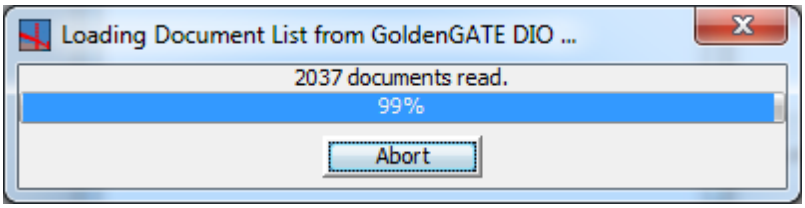


Fig. 10. Status bar during download of available documents.

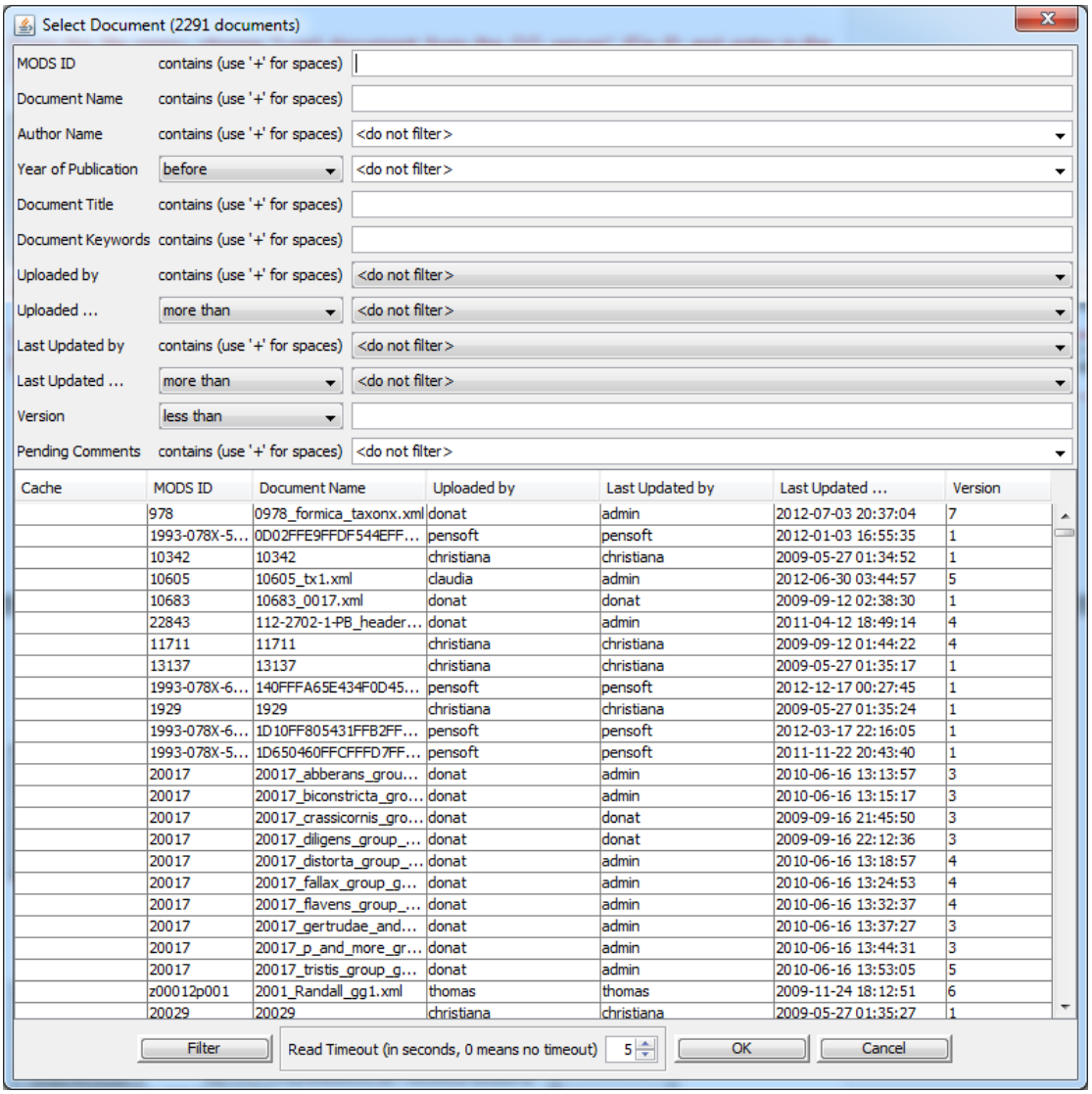


Fig. 11. Menu of available documents and options to deal with them.

To find the requested document, enter data into one of the top fields and then press the “Filter” botton. A reduced number of listed documents will appear of which the desired can be selected by highliting the document and press the “Ok” button .

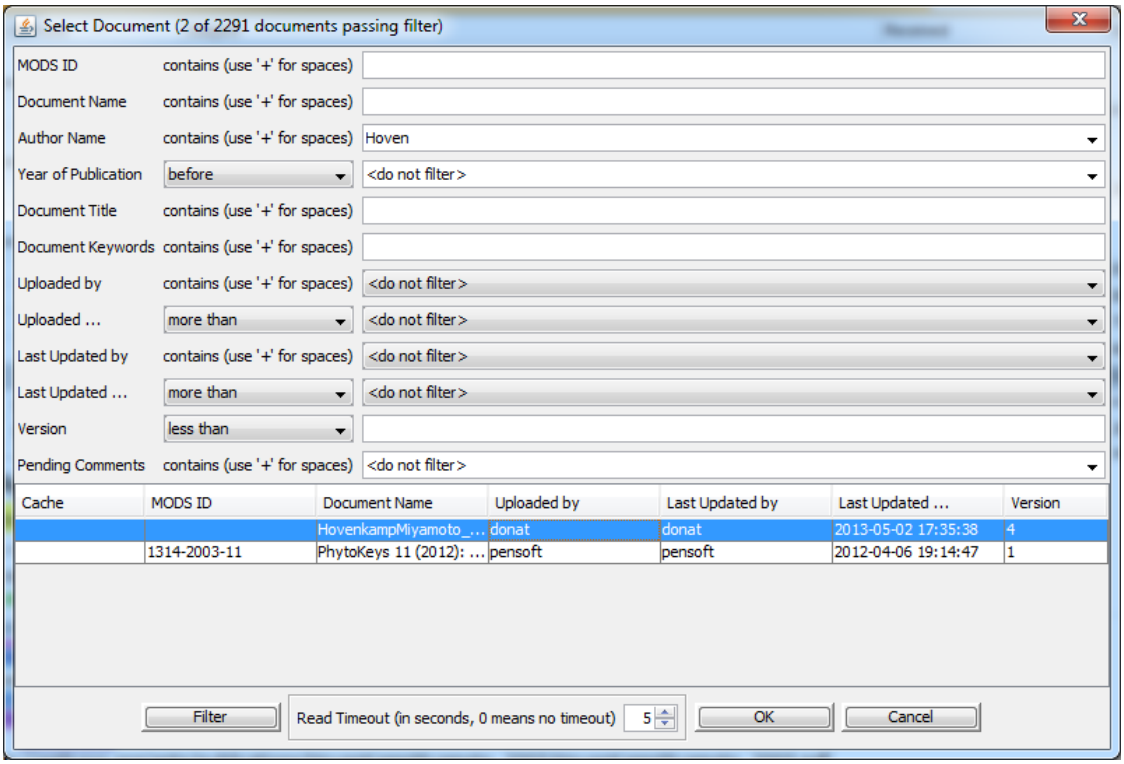


Fig. 12 Menu of available documents and options to deal with them: Filtered documents. The requested document is highlited and ready to download by clicking on the “ok” button. The document will show up.

At this moment, the document is locked on the SRS so that nobody else has access to it. The lock will be reversed at the moment the document has been saved, automatically as a new version, and closed. Do not forget to close the document when you finish your work (see also section 16 (saving).

2.4 Normalize an HTML document

Normalization of an HTML document is only necessary when an HTML-document is the starting point of the markup. For PDF format as starting point use begin at step 2.5

The “normalize HTML document” is a pipeline with several annotators, analyzers and markup converters. Its deletes the “DocType” part, marks pages borders, the pages themselves, the page numbers, the page titles, captions and footnotes. An interface opens that allows checking whether the tags are proberly set. On the the left hand the functions in the pull down menu allow to change the wrong identification of an element.

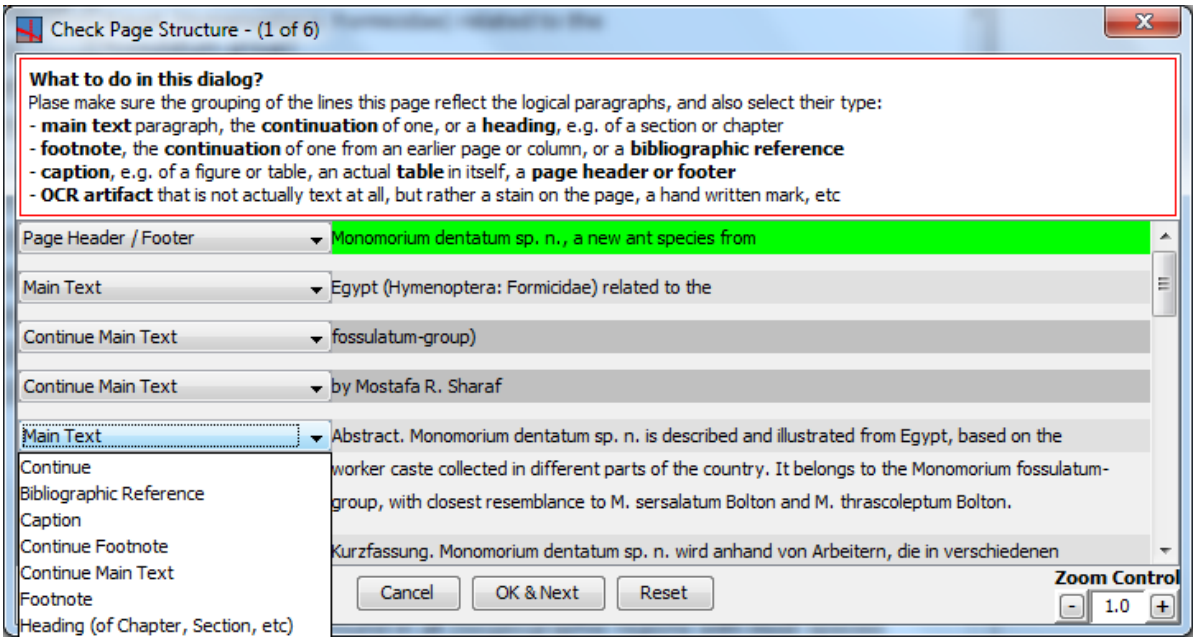


Fig. 18: Dialog visualizing the setting of paragraph borders and its classification

All the main text needs be labelled “main” text and the following lines “continue). In case of paragraphs running over two pages, the paragraph has to be labelled “continue main text”

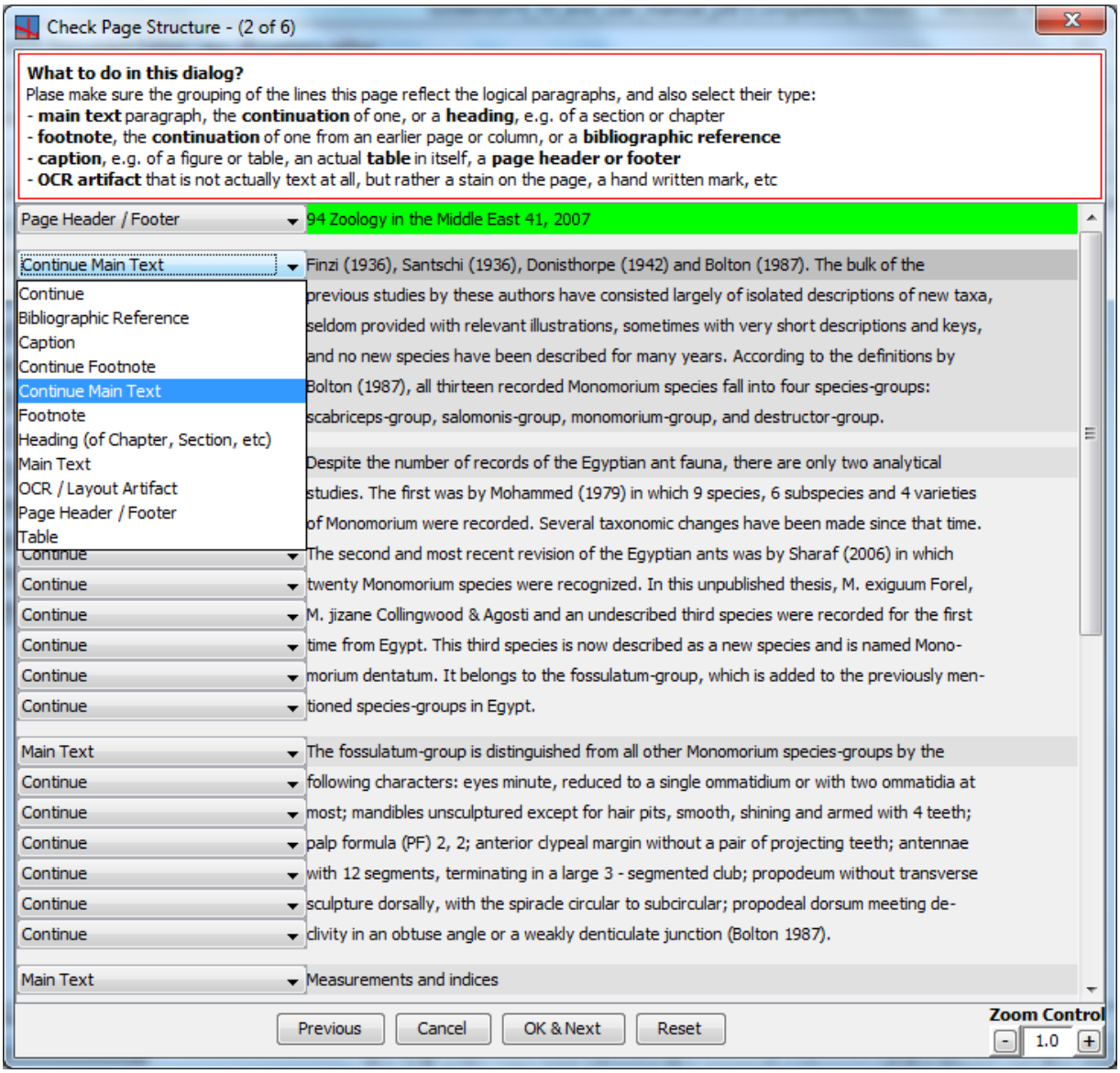


Fig. 19: Dialog visualizing the setting of paragraph borders and its classification

To remove OCR artefacts, select “OCR / Layout artefact”.”

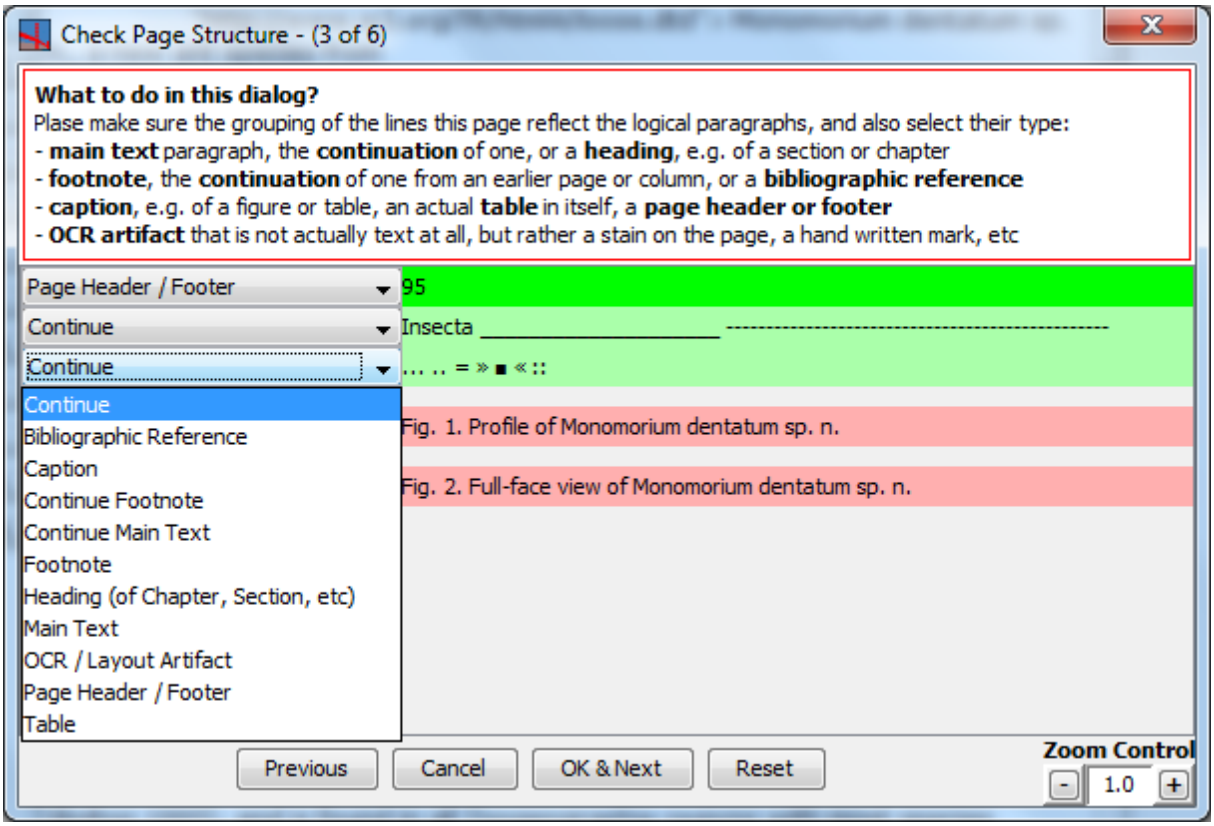


Fig. 20: Dialog visualizing the setting of paragraph borders and its classification: changing settings to remove OCR artifacts.

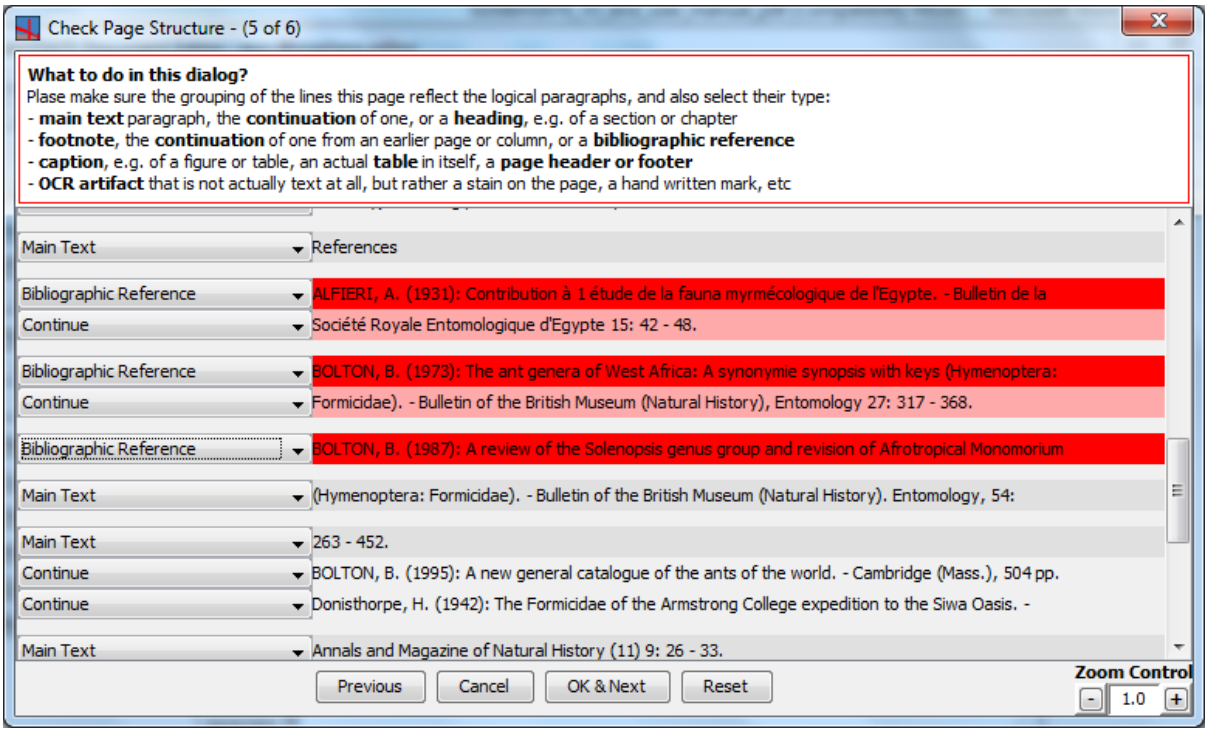


Fig. 21: Dialog visualizing the setting of paragraph borders and its classification: tagging bibliographic references.

Bibliographic references are tagged with “Bibliographic Reference” and “continue” if running over several lines.

If annotations are missing, mark them manually: **select (highlight) required text part ☐ right mouse click ☐ Annotate ☐ Annotate**

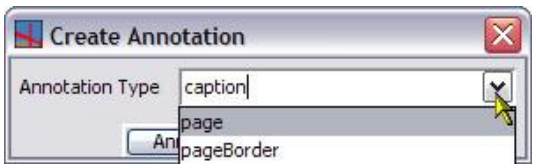


Fig. 12: Creating new annotations. Type the annotation name you want to mark up (here “caption”).

In most cases the annotation type was used before and you have to choose it from the list. If not, type it (example here: “caption”).

Correct paragraphs in pages manually with “merge annotations”. To merge annotations: **highlight parts of the two identical annotations ☐ click right mouse button ☐ choose merge Annotation**, to split annotations (i.e. insert a new paragraph) **highlight part of the first token of the new paragraph ☐ click the right mouse button-> choose split Annotation.**

Make sure that paragraphs before a taxonomic treatment contain only the treated taxon. Higher taxa like subfamily or tribe has to be written in an extra paragraph.

<paragraph>
Formicidae.
</paragraph>
<paragraph>
Subfamily PONERINAE.
</paragraph>
<paragraph>
Ponera grandis, sp. n.
</paragraph>
<paragraph>
[[worker]]. Reddish brown, head darker, mandibles, antennae, and legs lighter. Whole body clothed with sparse yellow pubescence, more abundant on gaster.
</paragraph>

Control and correct Page Page Header.

highlight the page number ☐ click right mouse button ☐ choose annotation ☐ choose annotation title or type it.

The correct page numbers are important, because they will be attached to the metadata of the annotated taxa and treatments, what helps to find them in the whole text during data mining. Note that the paragraph tag, contains the page number as well:

<paragraph pageNumber="417">
A Review of the Ant Genus Mycocepurus Forel, 1893 (Hymenoptera: Formicidae)
</paragraph>

The **paragraph tags at the right side** are the result of the previous box configuration. To check or to edit, activate the boxes. Depending on the text element, the text parts will appear in different colorations.
passages visible.

All mark up steps follow a determined work flow for treatment markup (Could be customized for other mark-up procedures). You can go through this workflow **following the custom functions at the left side**.

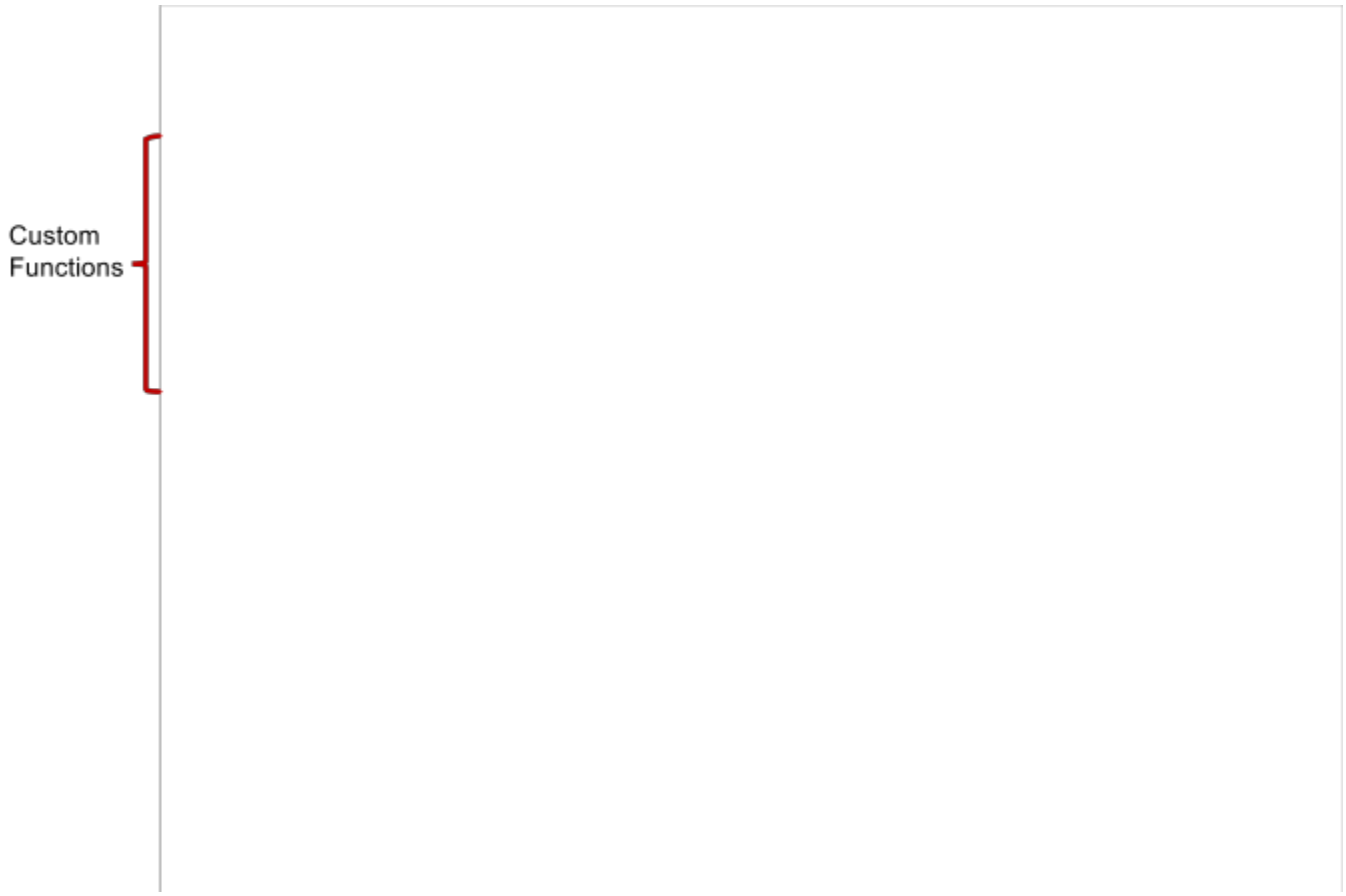


Fig. 22: Custom Functions and paragraph tags.

2.5 Opening and post-processing a PDF document

The opening of a Pdf document will take some time during which the document will be read and depending on the origin (e.g. digital born, scan-based) processed including either extracting the text or OCRing, ready for further conversion.

Fig. 13: Selection menu of documents:
Image Based: scanned PDF document, images will be split up in words during OCR process.
Image Based (text blocks only): scanned PDF document, images will be spilt up in blocks, only some words will be recognized
Image Based (pages only): scanned PDF document, only images at borders will be extracted. It is possible to get a preview which can be analyzed in detail and then be selected, e.g. for only some of the treatments.

From the File menu (Fig 12) select download the respective option that applies. If the pdf is based on a scanned image, then use “Image based”; if it is text based, then use “Text based”.

Fig. 14: An opened pdf document. Each “W” represents an image token.

The button “**Post Process Imported PDF**” is preparing the document for later annotation by cleaning up all the structural and OCR ambiguities and artefacts and assigning each paragraph to a particular class, such as main text.

Fig. 15: The postprocessed PDF file.

Pieces of text are highlighted by boxes that are given a pre-assigned value representing the particular kind of the text within the document. When passing with the mouse pointer on these boxes, new options on “box editing” appear.

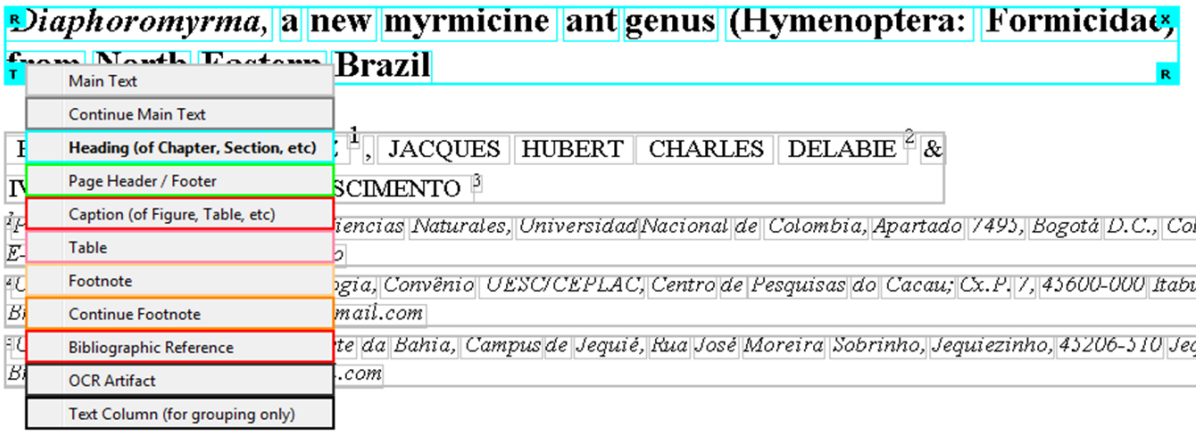


Fig. 16: The box editing menue.

Taking a closer look on the box, in each corner appears a letter:

R: resize the text box
T: transform the text box (instead Heading chose for Footnote, if it is the case)
X: delete the box

In the example above, the publication title is assigned as Heading, what is correct. Never the less, each word of the publication title, the heading, is marked as a separate box. The aim is to have a single heading text box with all words of the publication title. In order to this, first delete all other boxes of the boxes using the X and the use the R to extend the first box over all words in the title.

Make sure, that each recognized box has to have an assigned type, otherwise you will not be able to continue.
To combine various blocks, the blocks to be integrated have to be closed (they then appear in red) and then the integrating block can be reformatted to include the target blocks.

Once all the formatting has been checked, a new popup may appear for special characters and all the character have to confirmed, whether they correspond with the respective ASCII transcripts.

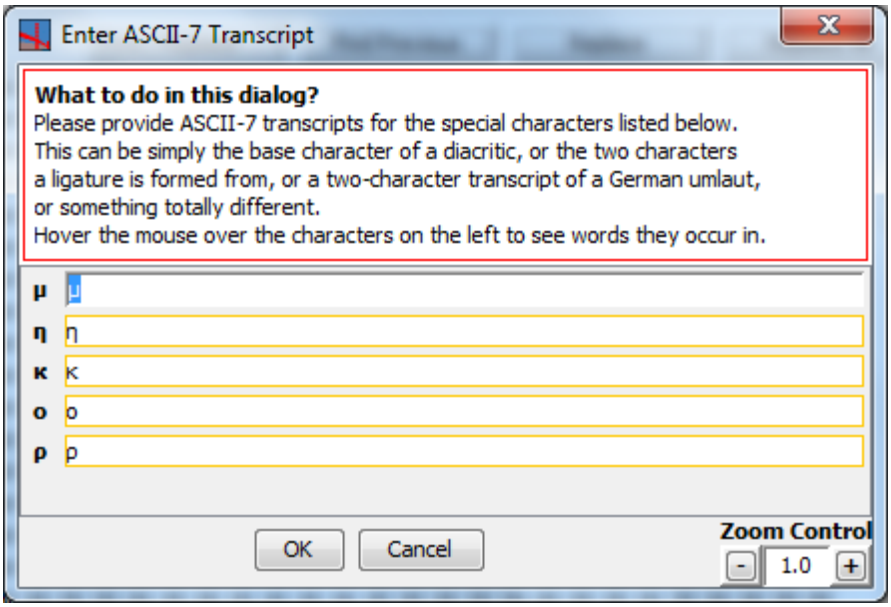


Fig. 17: The special characters popup window

2.6 Check spelling

2.7 Get or clean MODS: Add metadata of the publication

Each document has to be annotated with its bibliographic metadata by adding a MODS-header (Metadata Object Description Schema) Get the **MODS Header**, for example for ants, from the Hymenoptera Name Server: **Custom Functions ☐ Get / Clean MODS** [*requires internet connection*] *) **)

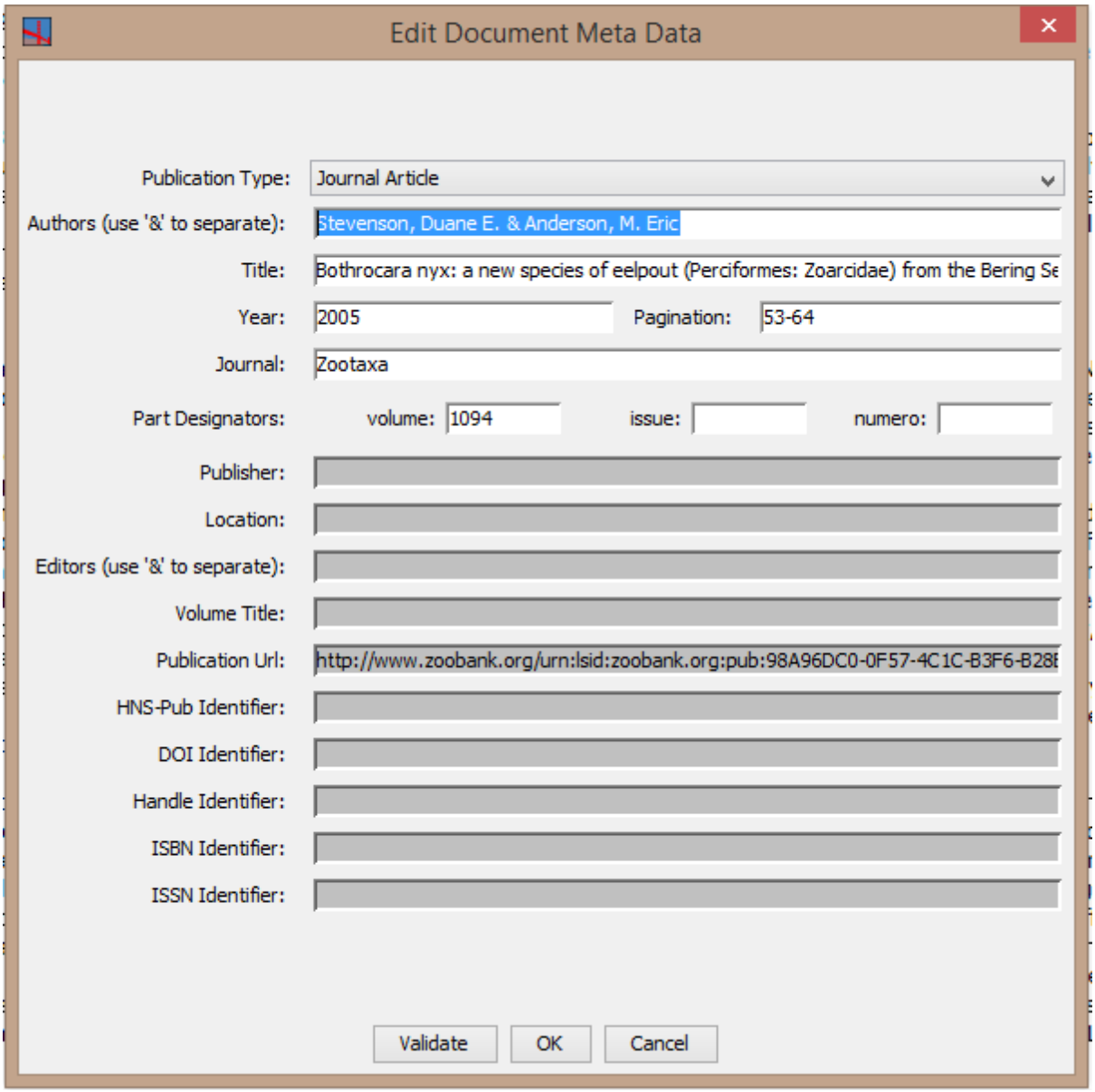


Fig MODS form to enter bibliographic data

Author names should be, preferably, entered as last name, comma, first name (abbreviated or not); individual author names have to be separated by '&'

> *GoldenGATE connects with the Hymenoptera Name Server and completes specific information on the marked-up html-document. The original HNS-ID of the pdf is required. This should be the same number/name of the html-document, but confer;* ***here is a potential error source when saving the edited pdfs in the ABBYY reader!***

Alternatively this analyzer can be initiated via **Tools ☐ Run Analyzer ☐ MRModsReferencer.analyzer.**

The MODS header appears at the top of the document. To see all the included metadata , expand the MODS data at the right window side, below the annotation options. You should not modify the MODS header.

The MODS header requires a link to the document. If GoldenGATE cannot find the link in the internet, search for the handle at http://plazi.org:8080/dspace/community-list or contact the Plazi group for handle creation (http://plazi.org/?q=about)

2.8 Parse Bibliography: Bibliographic records and citations

The bibliographic records can be parsed in its constituent elements (author, publication year, title, etc.) and the citations in the main text can be linked to the respective bibliographic record.

The “Parse Bibliography” menu opens in a pop up menu that proposes the respective assignation of elements. The elements can be changed by highlighting at least one character in the respective string and then using the right click of the mouse to offer the respective menu. To mark an unmarked string, all the tokens (words) of the string must be at least partially highlighted.

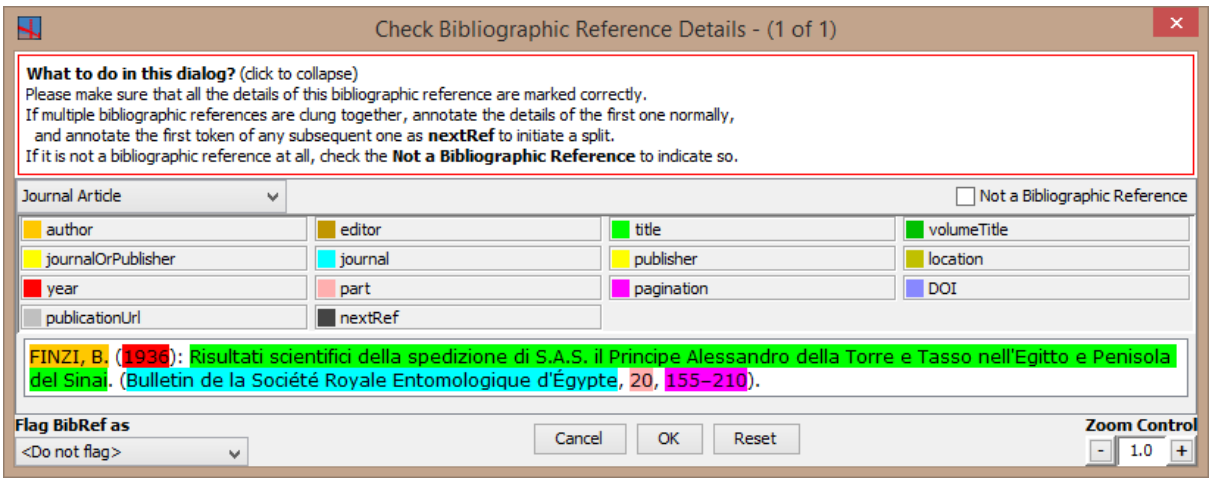


Fig. 26: Parse bibliographic records menu.

The menu on the upper shows the proposed selection of the bibliographic reference, whether it is a journal, books and book chapters,etc.

author = the author(s) of the referenced publication, i.e., article, book, book chapter; mark individually
editor = the editor of an edited volume, usually only present in references to book chapters; mark individually
journalOrPublisher = the name of the journal an article was published in, or the name and/or location of the publisher of a book
journal = the name of the journal an article was published in (detailed version of journalOrPublisher)
publisher = the name of the publisher of a book (detailed version of journalOrPublisher)
location = the location a book was published (detailed version of journalOrPublisher)
pagination = the pagination of a referenced publication that is part of a larger volume, usually present in references to journal articles and book chapters
part = the volume or issue number of a journal article; mark individually if both are present

volumeTitle = the title of the whole book in references to book chapters, conference papers, or articles in special issues of journals that as a whole have a title in addition to the journal name
bookContentInfo = the number of pages, illustrations, etc. in a book; things like “815 pp.” are in this category rather than pagination, as the latter selects a range of pages from a bigger volume
nextRef = the next reference, to be split from the current one; this is to correct cases of two or more references ending up conflated due to missing paragraph bounaries

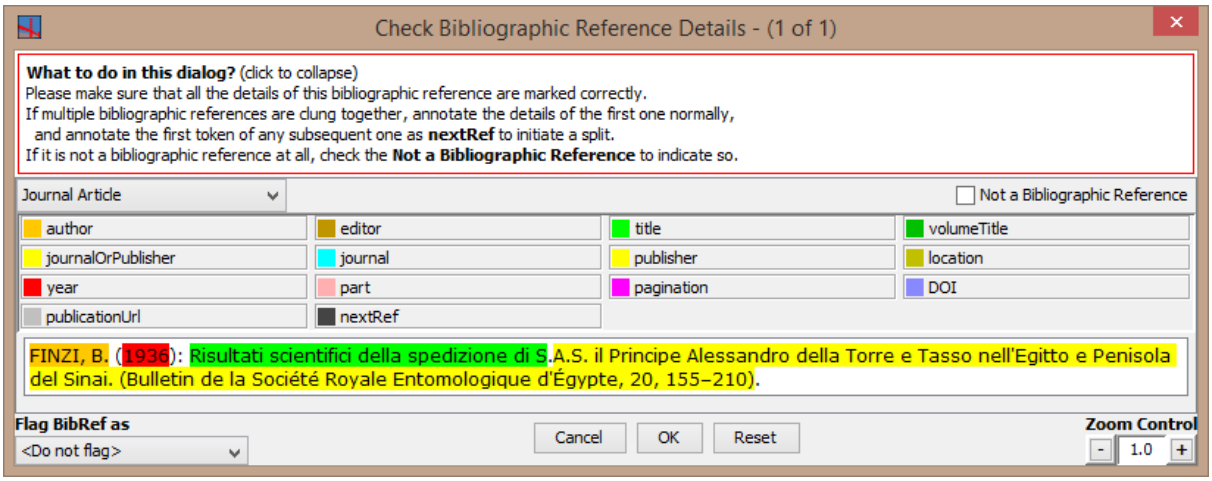


Fig. 26: Parse bibliographic records menu: Edit automatically inferred markup..

To edit proposed mark up, highlight the respective token(s) to remove the annotation for the selected part. Then use the right mouse button to rename the part accordingly.
If possible, avoid including brackets “(“ in the highlights. Alternatively, after the selection, click on the respective element in the menu above the reference and the selection will be tagged accordingly.

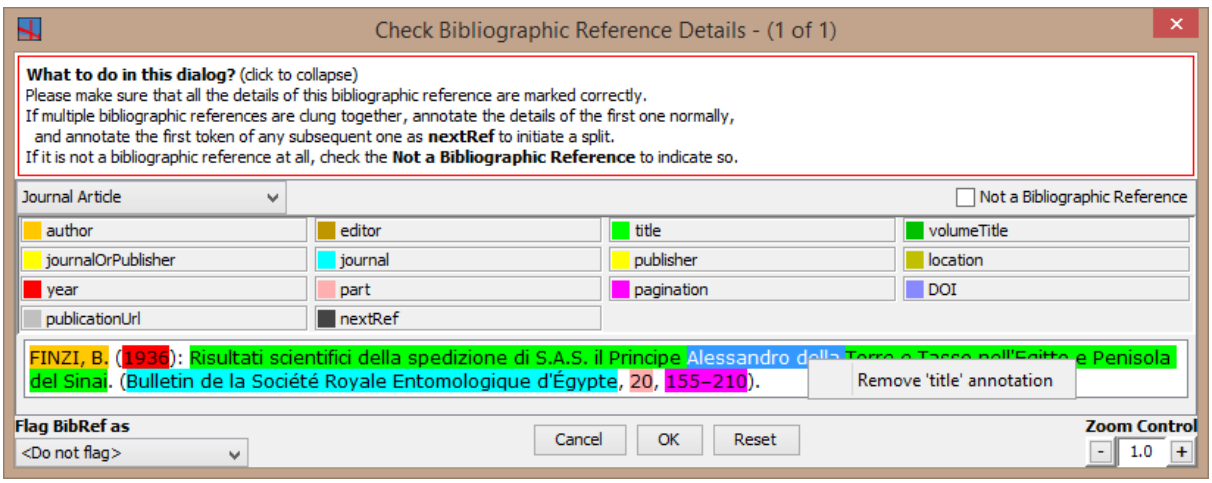


Fig. 26: Parse bibliographic records menu: Edit automatically inferred markup by selecting one to several tokens that include the part to be changed.

To split already marked up parts highlight the token where the split is required, and use the right mouse button. The popup menu will appear with the possible selection. Make sure you hightight only one tolken. “(“ and other symbols are regarded as a token and thus highlighting “(B” will not allow to split.

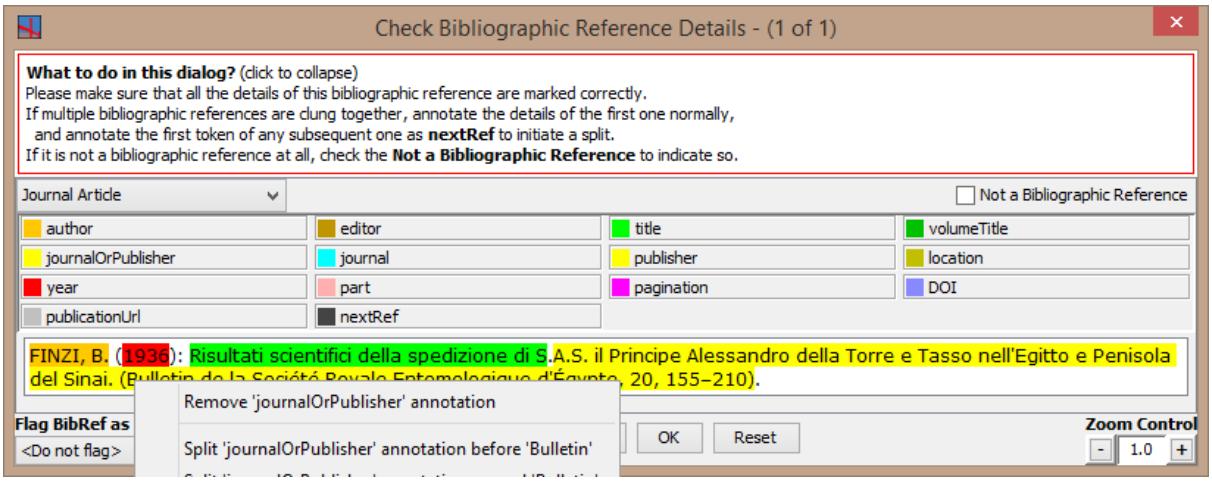


Fig. 26: Parse bibliographic records menu:split annotations by selection one token.

The Bibliographic section can be parsed in one go, or interrupted. In this latter case, the references have to be edited individually.
Click on the requested reference, and a popup window will appear. Select “Parse Bibliography”. If the “Document not Fit for ‘Parse Bibliography’” popup appears, select “Yes” and continue with the markup.

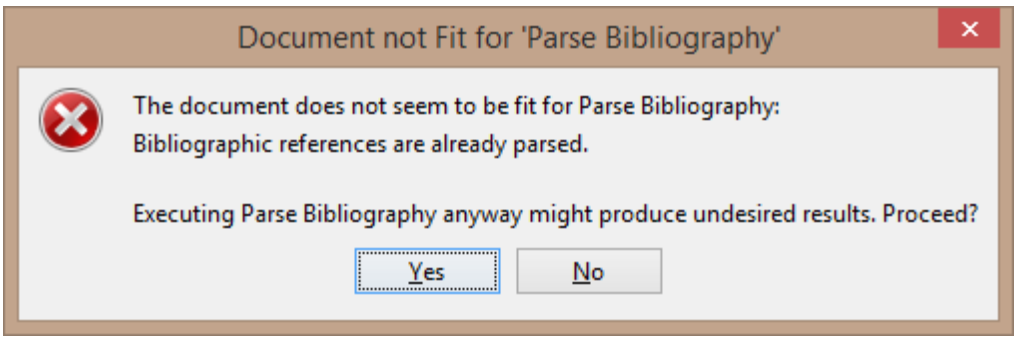


Fig. 26: Parse bibliographic records menu: Error message

2.9 Mark citations

The citations of a bibliographic record in the main text can be linked to its bibliographic record through the “Mark Citation” function. The opening pop up suggests the respective link. Wrong links can be changed using the pull down menu of the offered link (left in the pop up).

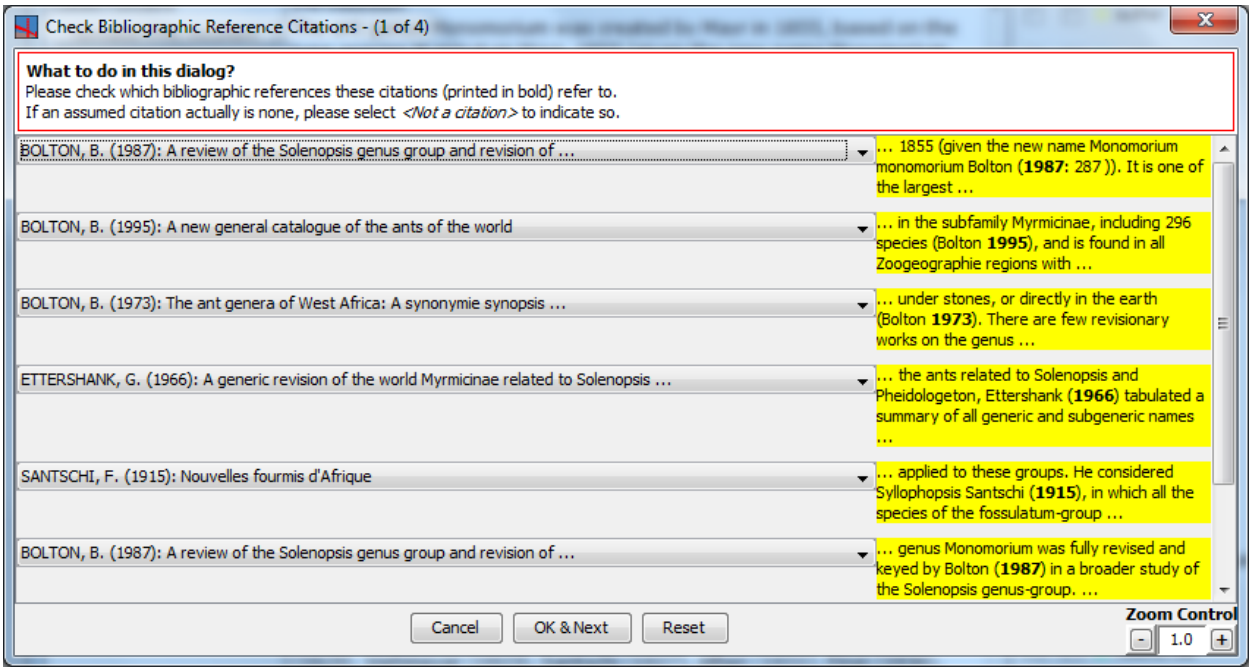


Fig. 27. Linking a citation to its bibliographic record.

If the program has no suggesion but suspects a link, then the string is no colored and only the suspecte year is highlighted but no link is made. The link can be made by choosing from the pull down menu.

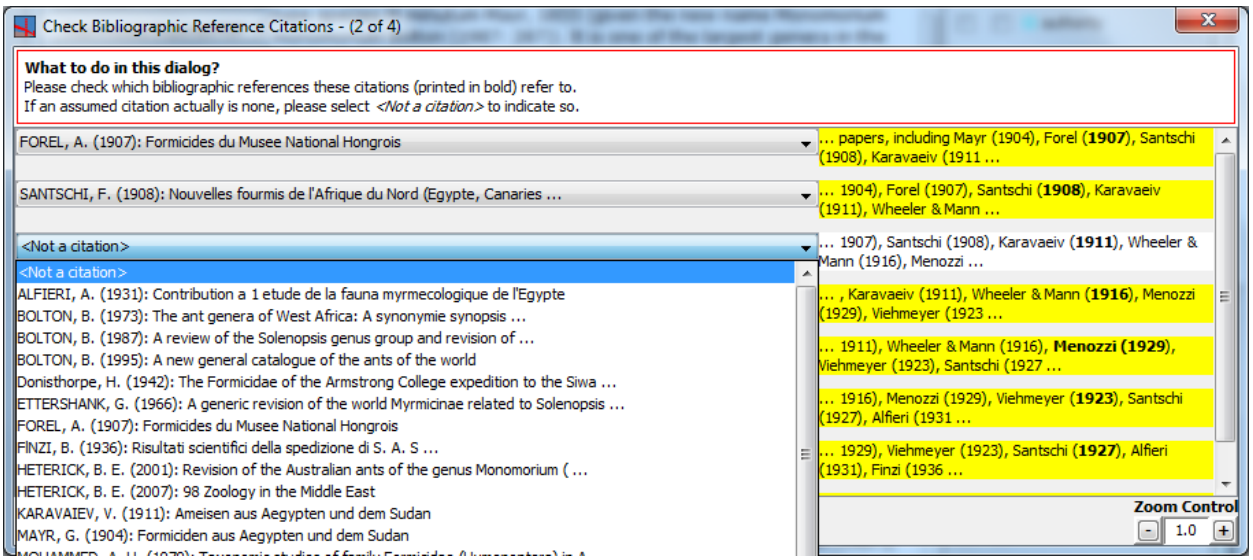


Fig. 28. Linking a citation to its bibliographic record. No selection is offered (no yellow highlight) but a citation is suspected (bold year).

2.10 Run Fat: Taxonomic section

Mark up all taxonomic names: **Custom function ☐ Run FAT.**

The Run FAT function marks all taxa found in the text. FAT is based on specific species dictionaries (right now an ant name dictionary is available). The taxa can be shown when marking the "taxonomicName". The "taxonomicNameLabel" marks the

*status of the taxa (e. g. sp. nov.; gen. nov. etc.) Run FAT can be started alternatively via Tools ☐ Run **Analayzers** ☐ **LFAT.FAT.analyzer**.*

The analyzer will prompt names that are not positivelz identified as taxonomic names. The user has to decide whether a noun is a taxonomic name or not.

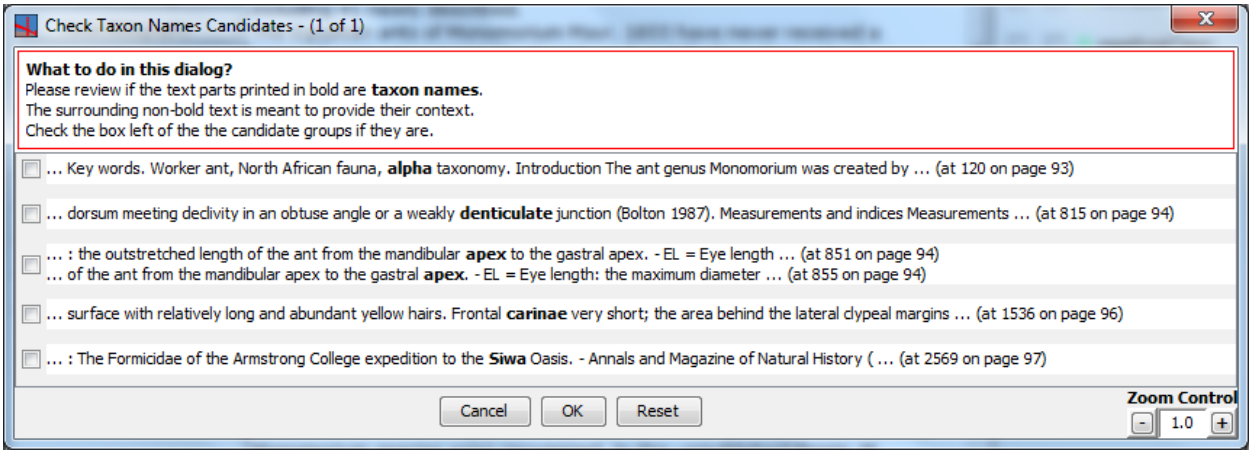


Fig. 23: Check Taxon Names Candidates.

Mark **manually** the taxa which were not recognized.

If taxonomic names are incorrectly annotated (only part of name, e.g. without author, or genus without species, etc.), than select the whole taxon with all it parts.

***Attention!** The status of the taxon should not be annotated (e.g. sp. nov., gen. nov. etc.) The status should be tagged as "taxonomicNameLabel".*

*It is common that GoldenGATE marks some words or letters as taxa which were not. In this case, you have to remove the annotations. This can easily and quickly be done with using **Select Annotations: View** ☐ **Select Annotation** ☐ **taxonomicName**. A list with all marked taxa appears, for removing annotation select the Remove button for the erroneously marked expression.*

Words/terms which where marked erroneously as taxon: ☐ click with the right mouse button on the highlighted term ☐ choose "remove" (only for this single case) or "remove all" (remove the annotations of all equal terms)

2.11 Parse Taxon Names: Complete the Taxa.

Custom function ☐ Parse Taxon Names

Together with the FAT.analyzer these functions are responsible for the correct taxon treatment. Whereas FAT finds all the taxa, the taxonomicNameAttributer.analyzer assigns the found taxa to there respective status, this is genus, subgenus, species, subspecies, race and variety. If not, you can assign the correct status of the given name in the "Please Check Attributes of Taxon Names" window: with the options on the left side, you can choose the correct

category of the taxon, on the right side you can choose alternative names for the taxa, if they are assigned erroneously.

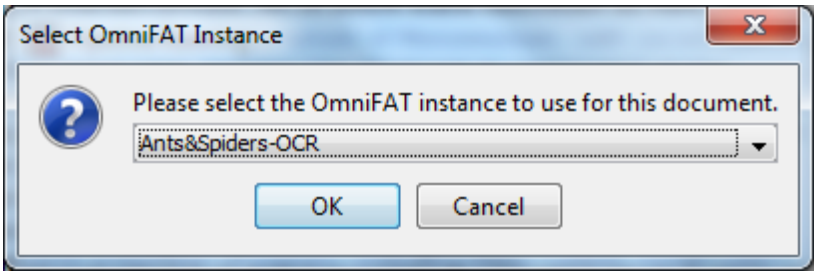


Fig. 24: Pop up menu asking for the domain OmniFAT version.

A pop will show up to ask which specific OmniFAT version ought be used to discover scientific names in the text.

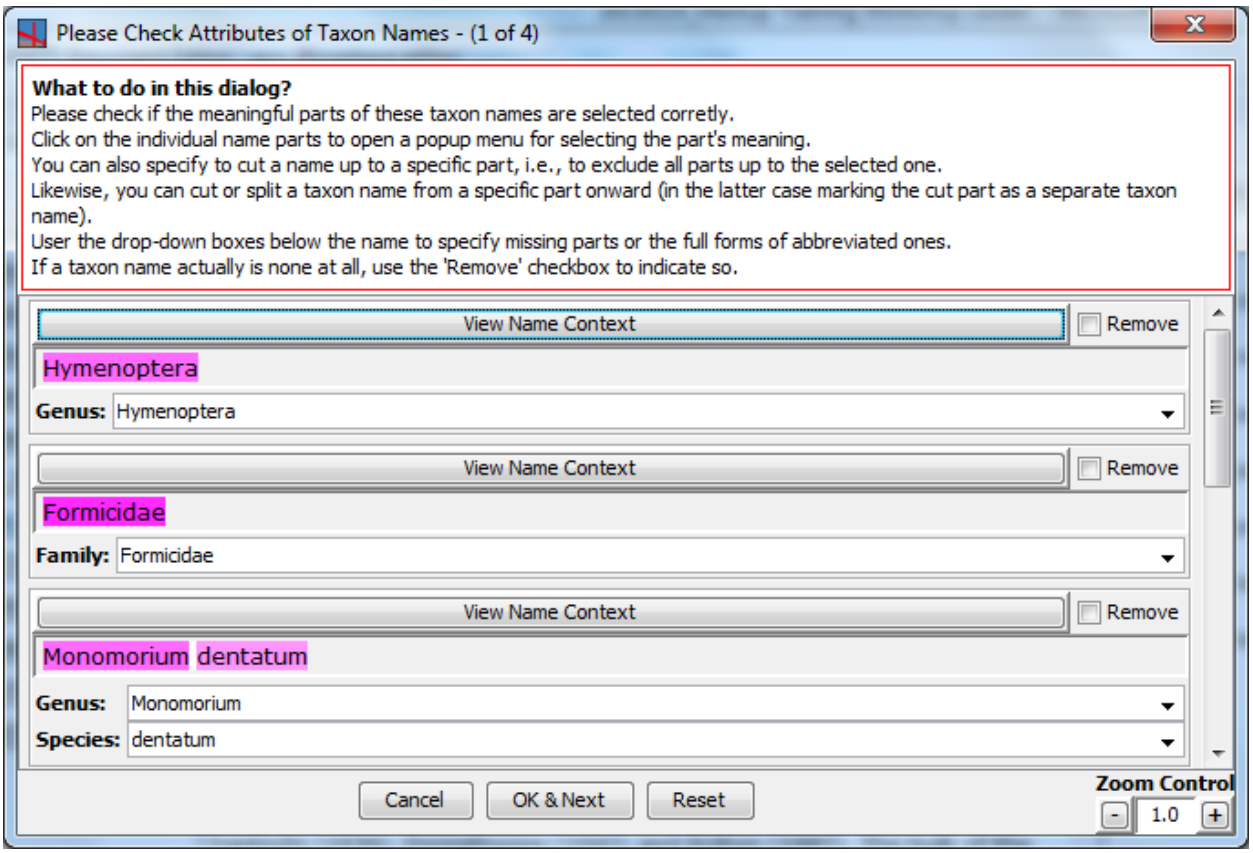


Fig. 25: Attribute taxon names: assign the correct status of the taxon. If click on the cited taxon, the surrounding 10 tokens appear, this might help with the correct attributions o single species, subspecies etc.

Once the candidate names are discovered a menu will display all the names and their rank and combination has to be confirmed or changed accordingly. The context of a names are instances of the same name can be seen by clicking on the “View Name Context” botton.

2.11B Markup Taxonomic Name Authority

See also collector name, author in MODS

Each authority should be marked up individually and not as a string

E.g. Miller, Sautter, Catapano is

<TaxonomicNameAuthority>Miller</TaxonomicNameAuthority>
<TaxonomicNameAuthority>sautter</TaxonomicNameAuthority>
<TaxonomicNameAuthority>Catapano</TaxonomicNameAuthority>

2.12 Markup Treatment Borders: Treatment mark up 1

“Treatments” in this special sense are the individual taxonomic species descriptions together with materials examined, data on natural history, etymology, discussion and reference groups. Definitions of such treatment parts are listed in the glossary at the end of this manual.

The treatment markup is split up in two sections: first you have to mark up the treatment borders, in a second step the individual parts of the treatments are marked up, the here called “subSubSections”.

Mark up the treatment borders: custom **function ☐ Mark up treatments borders**.

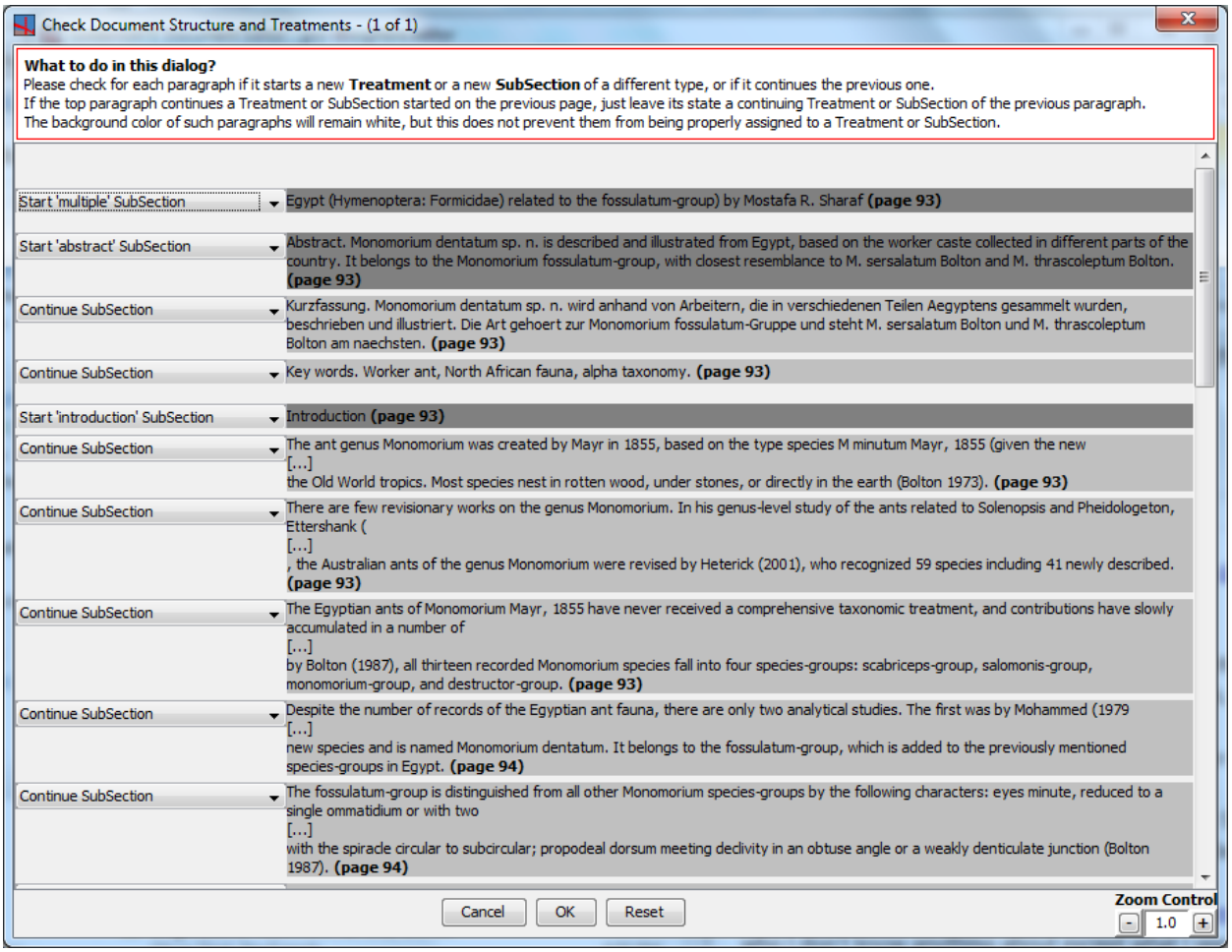


Fig. 29: Treatment border markup.

Please note that the general introduction of the publication is not part of a treatment, neither isolated identification keys, species lists nor bibliographic references at the end of the publication.

After marking the treatment boundaries, the different parts of them should be marked.

2.13 Treatment Structure: Treatment Markup 2

Custom functions ☐ SubSubSection markup

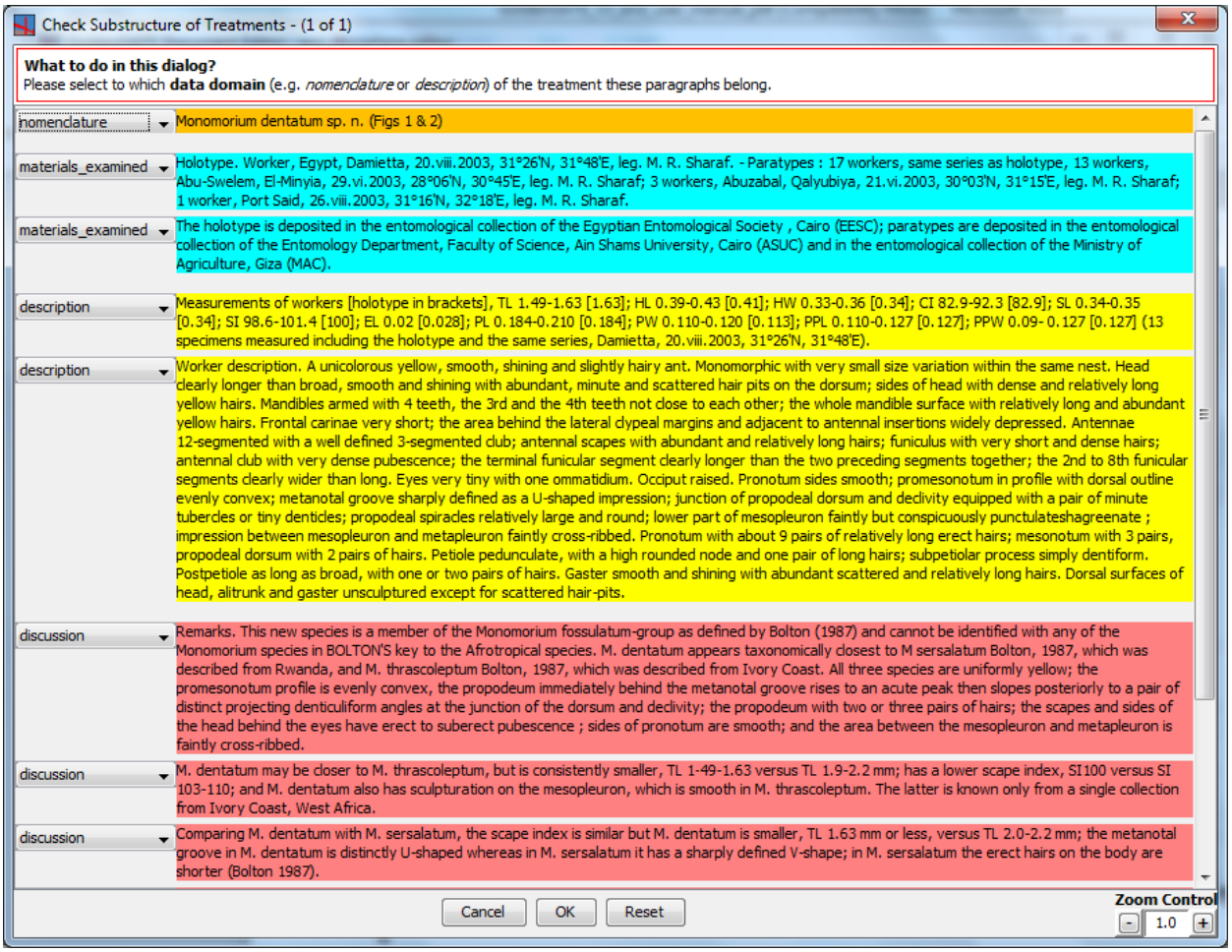


Fig. 30: Treatment substructure markup.

Each treatment is going to processed separately. To go to the next treatment click “ok and continue”. You can stop the mark up by clicking “ok and stop after” to continue later with the workflow.
Here the sections are paragraph based. In this way you have to associate each paragraph to a given subsection which you can choose at the menu at the left side of the window.

Please note that paragraphs with captions and footnotes are still in the treatment and should be marked with the same subSubSection type as the paragraph before or after. The captions and footnotes are going to be set at the end of the entire document during the next markup level (level 1).
The first subSubSection of a given treatment is should always be a nomenclature section and contain a taxonomic name.

2.14 Get LSIDs

This is an advanced feature that allows to get presistent unique identifier for taxon names. Originally build to interoperate with the Hymenoptera Name Server it will be broadened to include UIs from Zoobank and possible other name servers.

This step is not necessary and can be executed later.

2.15 Normalize paragraphs

This step is used to clean up structural layout artifacts that have not been cleaned up in previous steps. This removes most of the layout artifacts that might have an effect later in the display of the treatments.

Run normalize paragraphs by clicking on the button on the left side menu.

2.16 Mark Up Material Citations (1)

In this sense material citations are defined as all information related to a given specimen. This could be the entire collecting event, the type status and the institution where it is hosted. The markup process of the level is only possible after the former markup steps.
The custom function **Mark Up Material Citations** guides through the process. This part of the markup process is treatment based. If there is a button “ok & stop after” the process can be interrupted and continued at any time. Note that only documents with a complete markup level should be uploaded to the Search and Retrieval Server (SRS).

This markup process is divided in five steps:

a) check of the Collection Codes
b) markup of collecting countries
c) markup of collecting regions
d) markup of the entire collecting event of one specimen or specimen group
e) detailed markup of further information such as coordinates, collectors name, specimen code etc.

Check of the Collection Codes
GoldenGATE offers a selection of all found terms which might be collection codes due to their length (3 to 6 characters) and syntax (all upper case). By default GoldenGATE contains a collection code dictionary, but if there are collection codes not automatically marked up and appear in the list (see figure below), please notify Guido Sautter (sautter[at]ira.uka.de) so that he can complete the default GoldenGATE collection code dictionary and it will be available for all users at the next update.

Fig. 31. The user has to check if there are unrecognized and not recorded Collection Codes in the text.

Markup of collecting countries

Each collecting event should contain a collection country. If there is no country to mark, GoldenGATE will offer, by clicking “Ok & Continue” a list of possible countries.

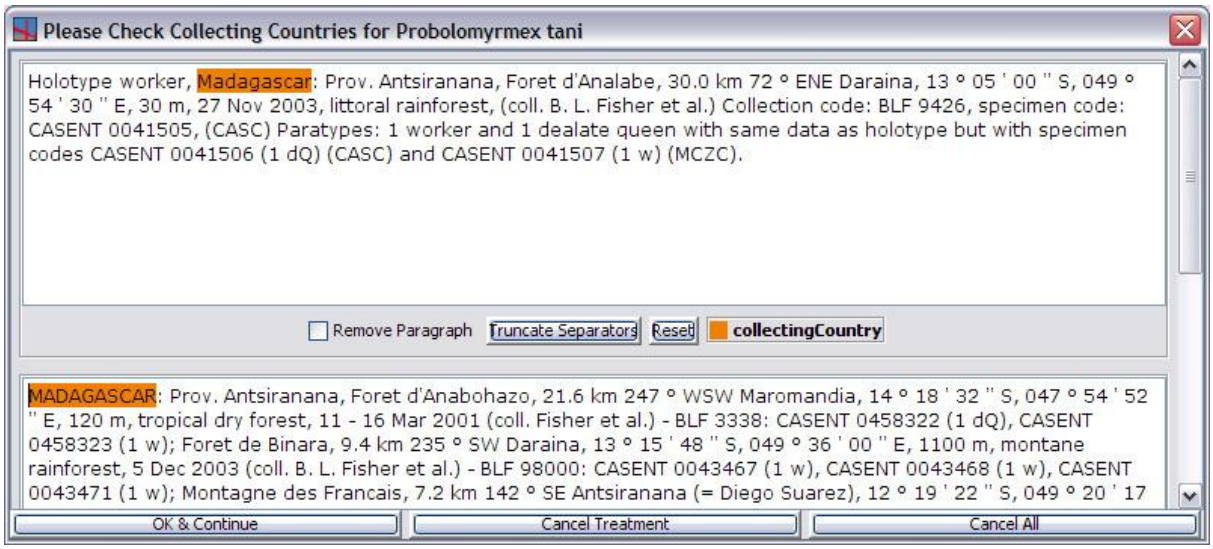


Fig. 32. Marking up collecting countries.

Markup of collecting regions

In most cases the collecting event contains a collecting region. This could be a state or a district.

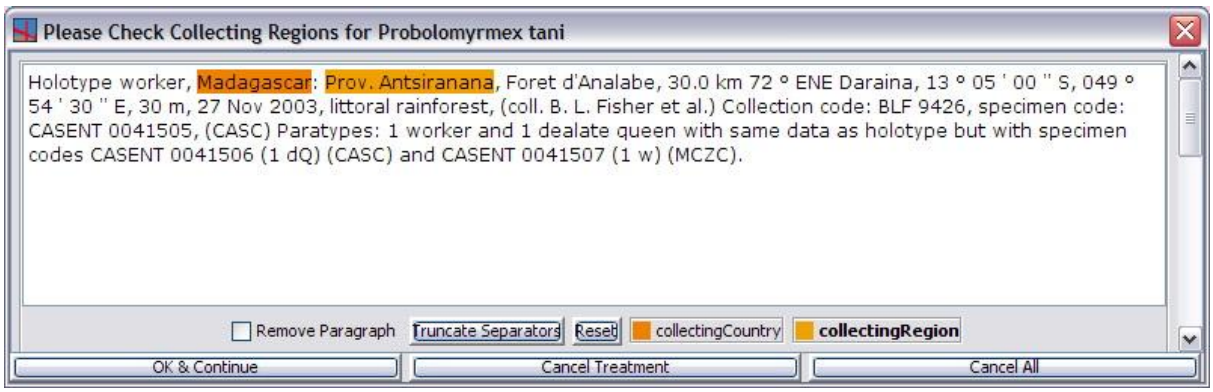


Fig. 33. Marking up collecting regions.

Markup of the entire collecting event of one specimen or specimen group

In many cases the examined material is more then one specimen from more then one collecting event. To differentiate between the single collecting events it is necessary not only to annotate the entire materials citation section in a publication but to annotate each single collection event or information on each single specimen. The type status and the hosting institution make part of the material citation.

Generally GoldenGATE searches for punctuation marks which divided a material citations section in its different parts. Even though it might be possible that a merging and splitting of erroneously marked up parts is necessary.

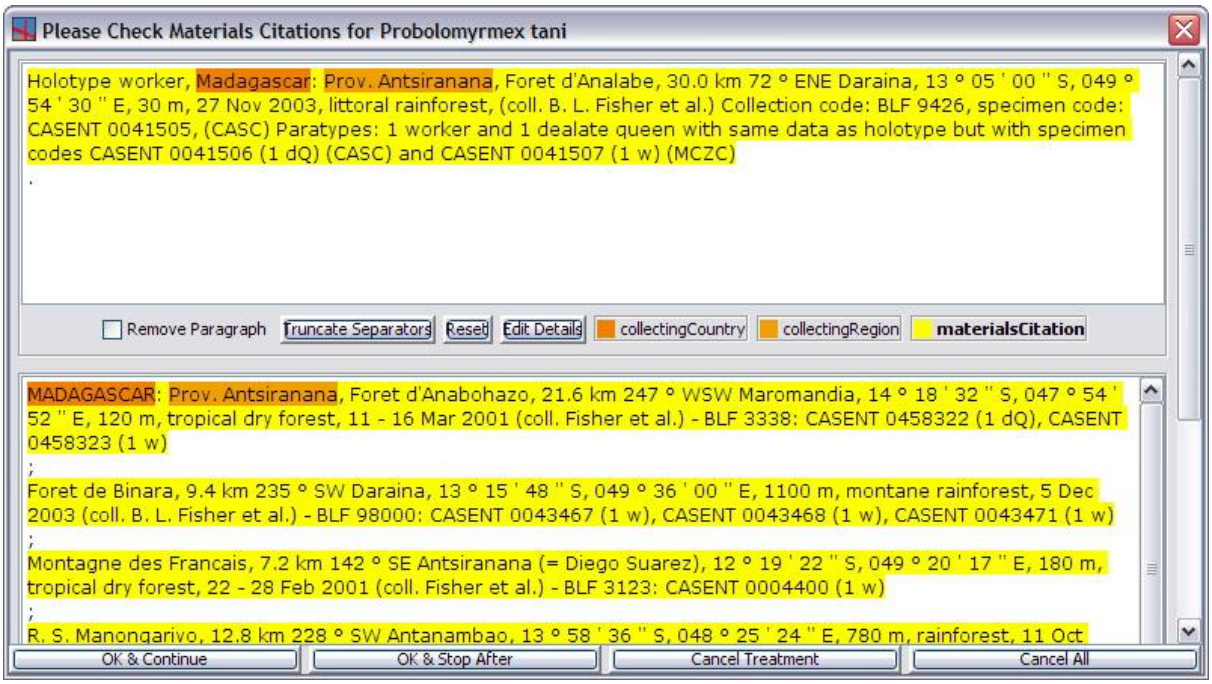


Fig. 34. Checking material citations. Each yellow paragraph displays information of one or more specimen bearing the same information.

2.17 Attribute Materials Citations (2): Detailed markup of additional information

In almost all material citations further information on a specimen are given. This could be geographic coordinates, locality deviations, specimen codes, collecting date etc.

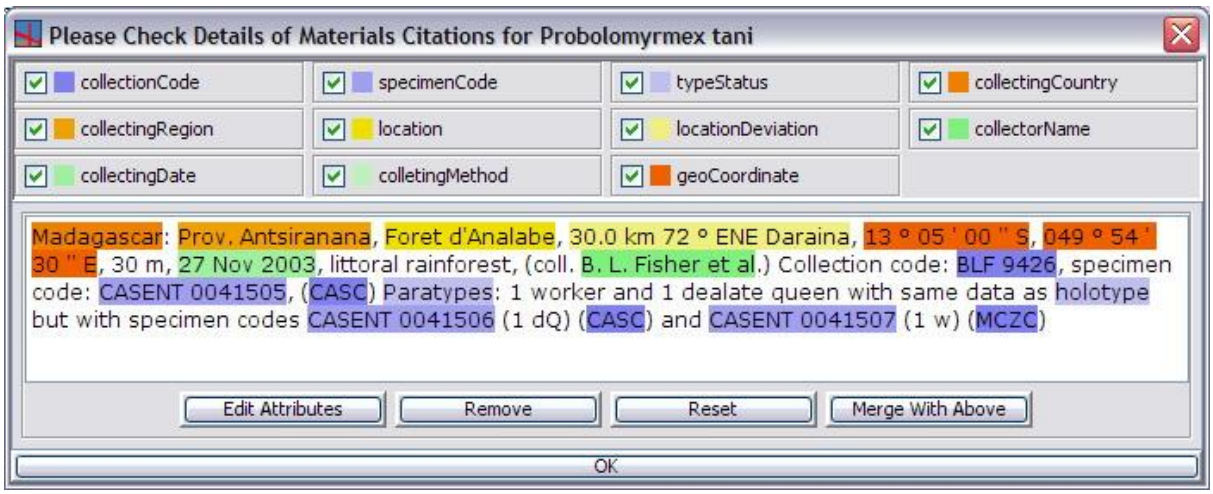


Fig. 35. Fine grained mark up of each material citation.

2.18 Geo-Reference Locations

[explanation to be added]

2.19 Saving the document and Upload

2.19.1 Save locally

The actual workflow requires two saving options. First, saving the processed document in a generic form with all tags (gg0.xml, see above), so that a further mark up can be done later. For this purpose name your document with the document identifier (the document ID you need to get the MODS header) together with the gg0.xml suffix. The gg0 suffix stands for GoldenGATE generic mark-up up to level 0. (e. g. 6200_gg0.xml)
File ☐ save document to file ☐ xyz_gg0.xml (XML file; all tags).

Then make your document TaxonX compatible.
As described above the TaxonX format is highly compatible with other database standards as Darwin Core. For this purpose the tags has to be renamed and restructured. Transform your xyz_gg0.xml document by saving it as xyz_tx0.xml file. Again, name it with the document ID and use the tx0 suffix, which means, that the file is TaxonX compatible and marked-up up to level 0.
File ☐ save document to file ☐ xyz_tx0.xml (XML file; XSLT transformed).

After confirming the saving a new window appears and you have to choose the gg2taxonx.xslt stylesheet.

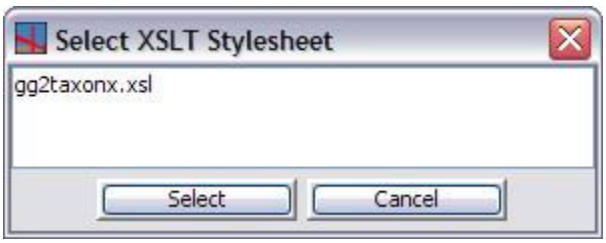


Fig.36: Choosing the XSLT style sheet.

2.19.2 Uploading the document to the Search and Retrieval Server (SRS)

Once a data mining ready version of a document has been reached it should be uploaded to the Search and Retrieval Server. For uploading a user account and password is necessary. Please contact Guido Sautter (sautter[at]ira.uka.de) for account setup.

For uploading the document go to File ☐ upload document to SRS

Fig. 37: Menu for uploading a marked up document to the Search and Retrieval Server.

Fig. 38: For a successful upload a user name and password are required.

Fig. 39: The upload was successful. If errors occur during upload, please contact the SRS administrator (Guido Sautter (sautter[at]ira.uka.de) for further assistance. From now on the document can be searched at the SRS (http://plazi.org:8080/GgSRS/search).

2.20 Viewing the document on the SRS

Once uploaded the document will be almost immediately available on the Plazi Boston server (http://plazi.cs.umb.edu/GgServer/search) and in 5 minutes on Plazi Bern (http://plazi.org:8080/GgSRS/search/). In case the new treatment does not show up, the cache can be renewed by adding "?cacheControl=force" to the URL of the respective treatment

http://plazi.org:8080/GgSRS/html/42E349B8AA3B4430FFC2DBB075CB5271
http://plazi.org:8080/GgSRS/html/42E349B8AA3B4430FFC2DBB075CB5271?cacheControl=force

2.21 Citing a treatment

At the moment of upload, a persistent httpUri identifiers is minted for each treatment.
http://treatment.plazi.org/id/42E349B8-AA3B-4430-FFC2-DBB075CB5271

This identifier can be used to link to this treatment (see linking to treatments). The links produces three different outputs, and html, XML and RDF.

html: http://plazi.org:8080/GgSRS/html/42E349B8AA3B4430FFC2DBB075CB5271

XML: http://plazi.org:8080/GgSRS/xml/42E349B8AA3B4430FFC2DBB075CB5271
RDF: http://plazi.org:8080/GgSRS/rdf/42E349B8AA3B4430FFC2DBB075CB5271

The XML output is GoldenGATE markup XML.

2.22 Citing a bibliographic reference of citation of a bibliographic reference

To link from a bibliographic reference, an httpUri attribute has to be included in the <bibRef> annotation. To link from a bibliographic citation, a httpUri has to be included in the <bibRefCitation> annotation. If multiple HTTP URIs are to be added, use httpUri-0, httpUri-1, etc. as the attribute names.

2.23 Citing a figure caption

A figure caption citation (eg Fig.5) is annotated with a <figureCitation> attribute. To link to an external digital representation, an httpURI attribute has to be added. If multiple HTTP URIs are to be added, use httpUri-0, httpUri-1, etc. as the attribute names.

2.24 Linking via a httpUri attribute

A link to an external resource can be added by using a httpUri attribute. For multiple links from a single annotation, the httpUri attribute has to be numbered as httpUri-0, httpUri-1, etc. It is important that the first numbered httpUri attribute begins with a “0”.

An httpUri attribute can be added to <figureCitation>, <caption>, <treatmentCitation>, <bibRef>, <bibRefCitation>

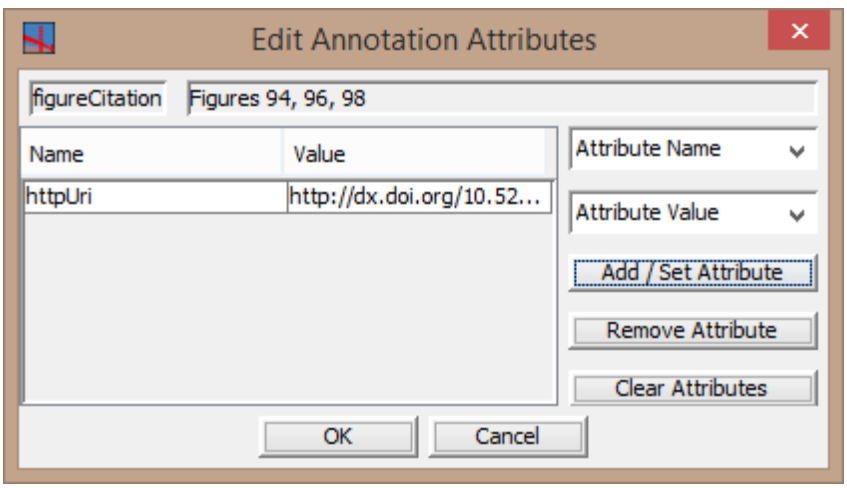


Fig. xx adding a httpUri attribute

Linking to external resources works by including the entire http URL of the identifier, e.g. for a DOI 10.5281/zenodo.15828 the entire http://dx.doi.org/10.5281/zenodo.15828 has to be included.

3 Workflows

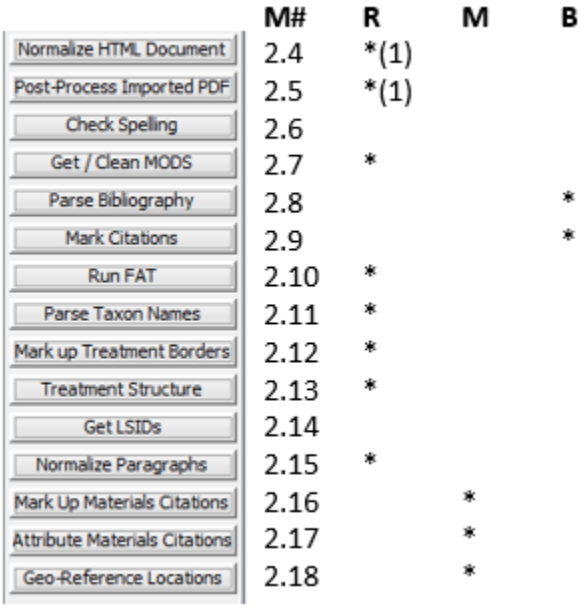


Fig xxx_001_overview. Overview of the Markup process. M reference to the chapter below. R required steps to upload the document to SRS. M markup of the materials citation. B Markup of bibliographic data. See also 2.0

3.1 Workflow for markup html documents

1. Start the GoldenGATE Editor with GoldenGate.exe and decide if you want to allow internet access
2. choose the configuration (for xml mark up of taxonomic literature choose “Proibiosphere.editor (local))
3. Open the document ☐ open document (File ☐ load document from file HTM and HTML files (default reader))
4. Run the Normalize HTML.Document Tab Custom functions ☐ **Normalize HTML Document** and follow the wizard
5. Get the MODS Header. Custom Functions ☐ **Get / Clean MODS** . [requires internet connection] *)
6. **Parse Bibliography**
7. **Mark Citations**
8. Mark up all taxonomic names. Custom function ☐ **Run FAT**.
9. Complete the Taxa. Custom function ☐ **Parse Taxon Names**
10. Split the document in treatments. Custom Functions ☐ **Mark up Treatment Borders**.
11. Mark up the subSubSections of a treatment ☐ **Treatment Structure**
12. **Normalize paragraphs**
13. **Mark Up Material Citations**
14. **Attribute Material Citations**

15. **Save** document / **upload** it to SRS as XML file
16. Close document

3. 2 Workflow for markup pdf documents

1. Start the GoldenGATE Editor with **GoldenGATE.exe** and decide if you want to allow
2. internet access
3. Choose the configuration (for xml mark up of taxonomic literature choose “Proibiosphere.editor (local))
4. Open the document ☐ open document (File ☐ load document from file pdf files (default reader))
5. Run the custom function “**Post-Process Impoted PDF**”
6. Get the MODS Header. Custom Functions ☐ **Get / Clean MODS** . [requires internet connection] *)
7. **Parse Bibliography**
8. **Mark Citations**
9. Mark up all taxonomic names. Custom function ☐ **Run FAT**.
10. Complete the Taxa. Custom function ☐ **Parse Taxon Names**
11. Split the document in treatments. Custom Functions ☐ **Mark up Treatment Borders**.
12. Mark up the subSubSections of a treatment ☐ **Treatment Structure**
13. **Normalize paragraphs**
14. **Mark up Material Citations**
15. **Attribute Material Citations**
16. **Save** document / **upload** it to SRS as XML file.
17. Close document

*) If there is no internet connection available for the moment, these steps can be done later.

.

4 Editing

Highlight an attribute

Issue: How to highlight the attributes listed on the right hand sub-menu of the Edit Annotation menu (see Fig XX).

Solution: In the right hand menu, each of the attributes has to its left two boxes that can be ticked. The right box displays and highlights the annotation, and if edited, all the attributions that have been added.

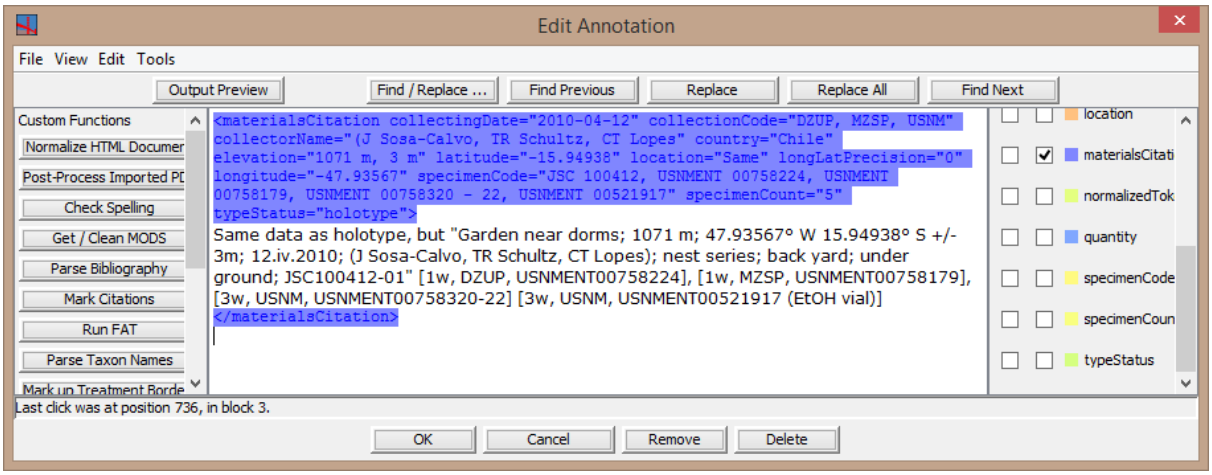


Fig XX highlighted annotation (In this case materialsCitation).

HIghlight the content of an attribute

Issue: how to highlight the content that his enhanced with a specific attribute?

Solution: In the right hand menu, each of the attributes has to its left two boxes that can be ticked. The left box highlights all the content that is included in the document.

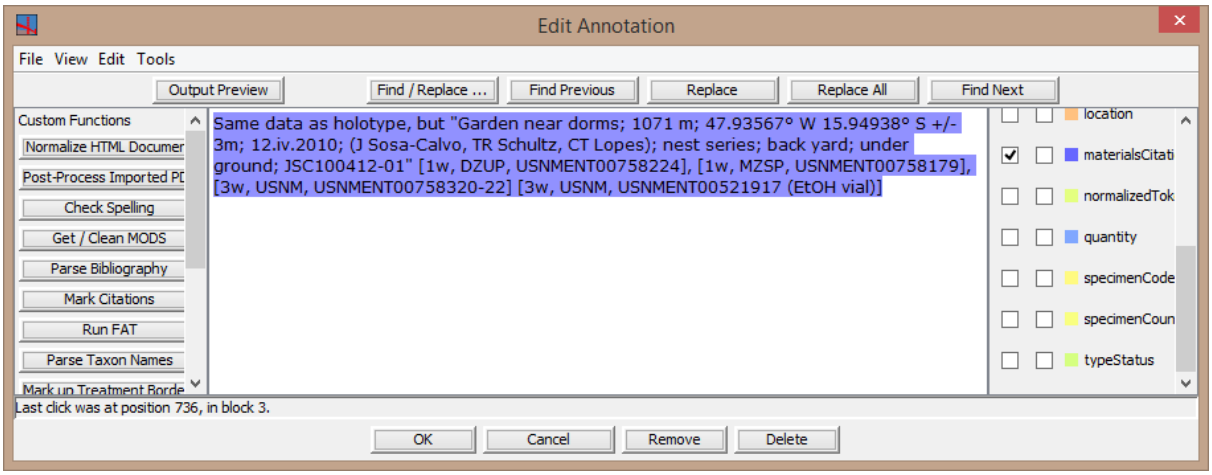


Fig XX. Higlighted string included in the annotation

See also:
Highlight an attribute

Annotate a token or string

Issue: How to annotate a token or string

Solution: Highlight the target token (e.g. Word) or string (sequence of tokens) by holding the left mouse button down and move over the target.

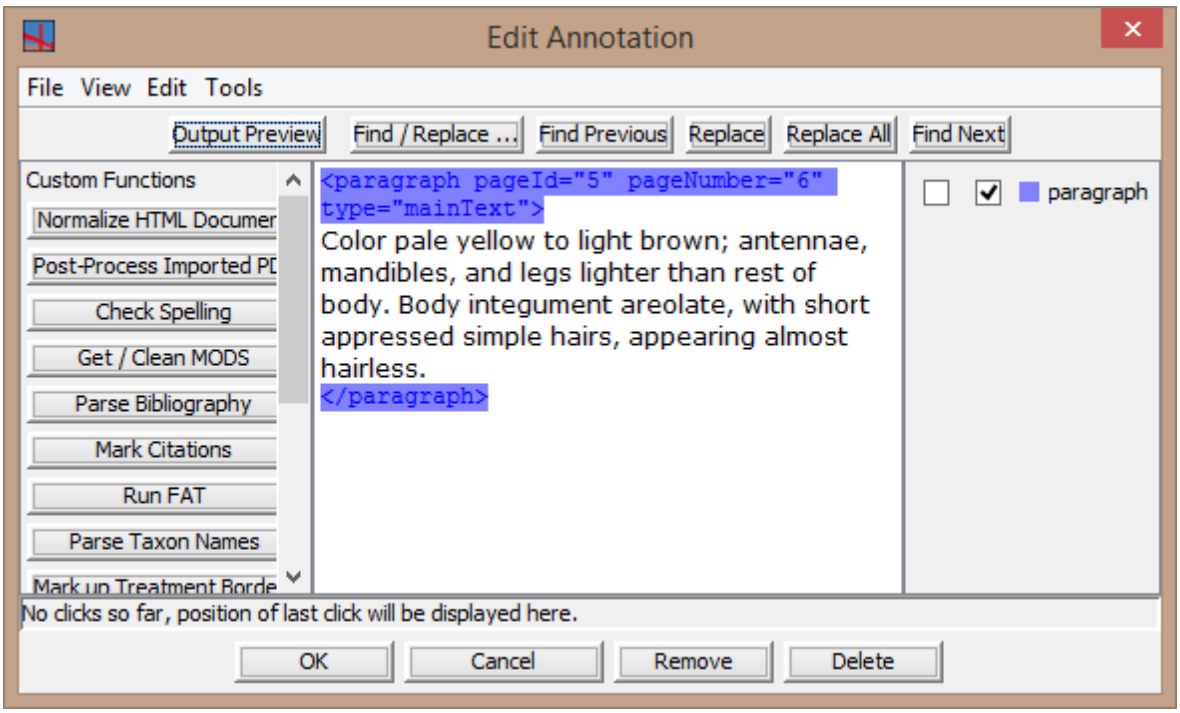


Fig xxx

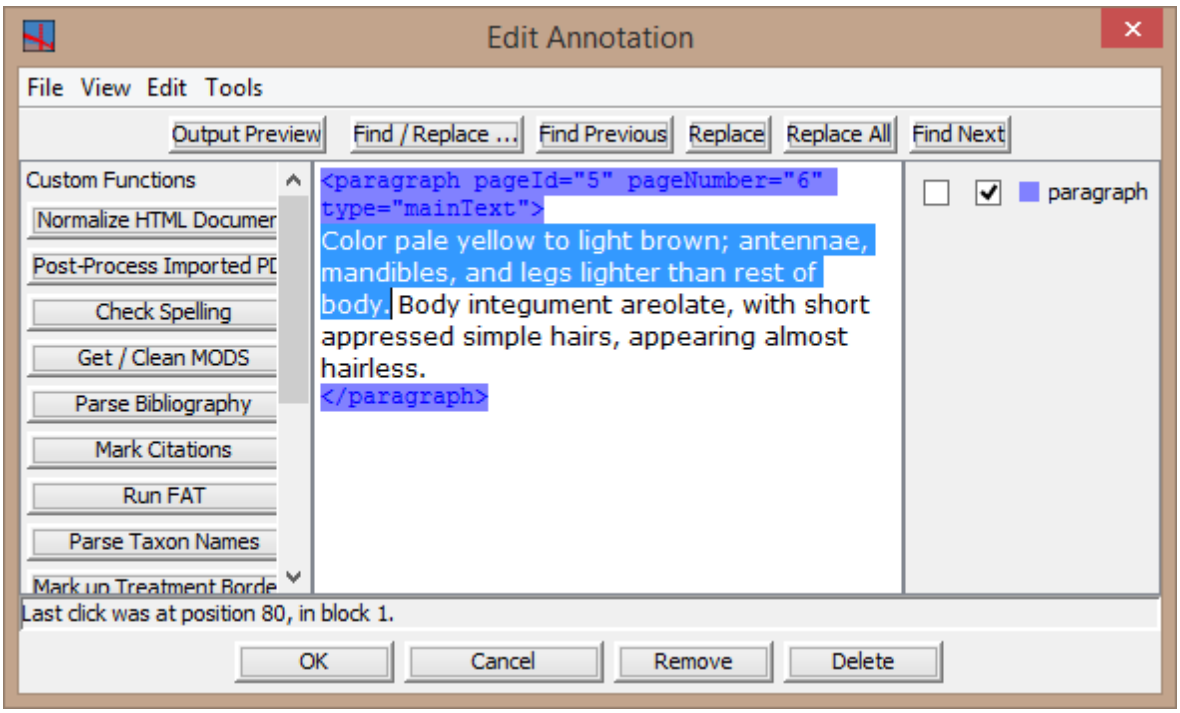


Fig xxx

With the mouse over the highlighted part, use the right mouse button and select “Annotation”.

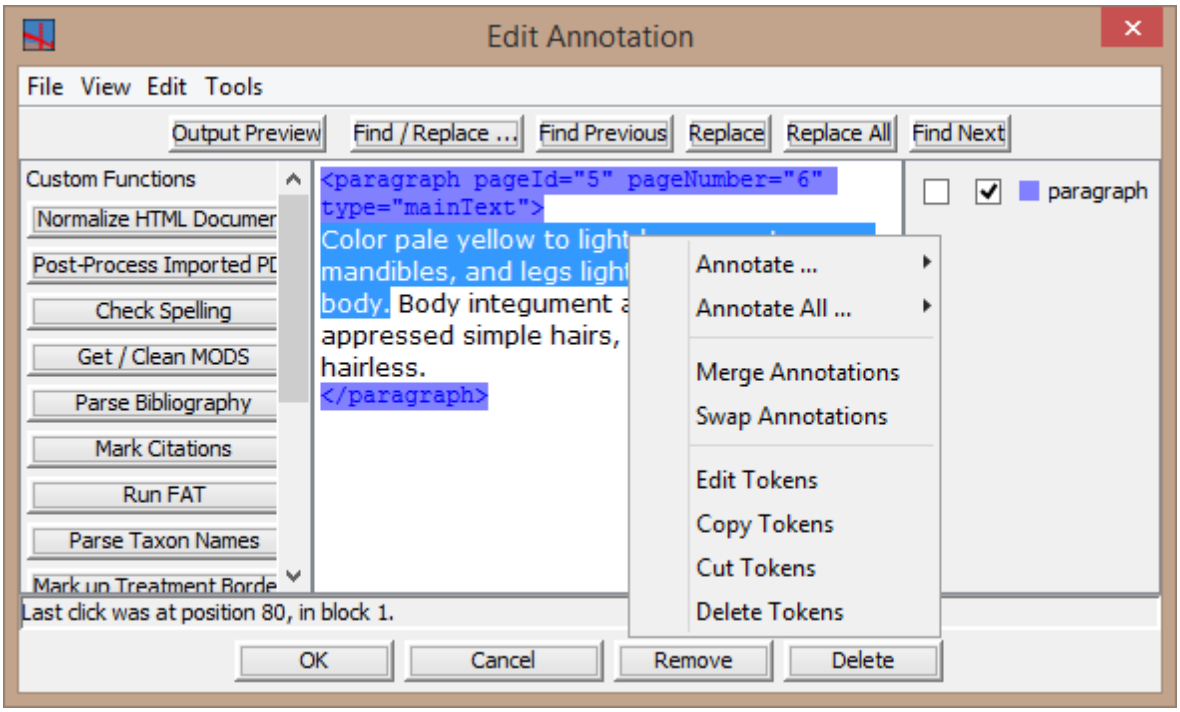


Fig xxx

Select the annotation desired. Only those already used in the text are pre-selected and offered as choice. Alternatively, “Annotate” can be selected and the respective annotation type selected from a pull down menu.

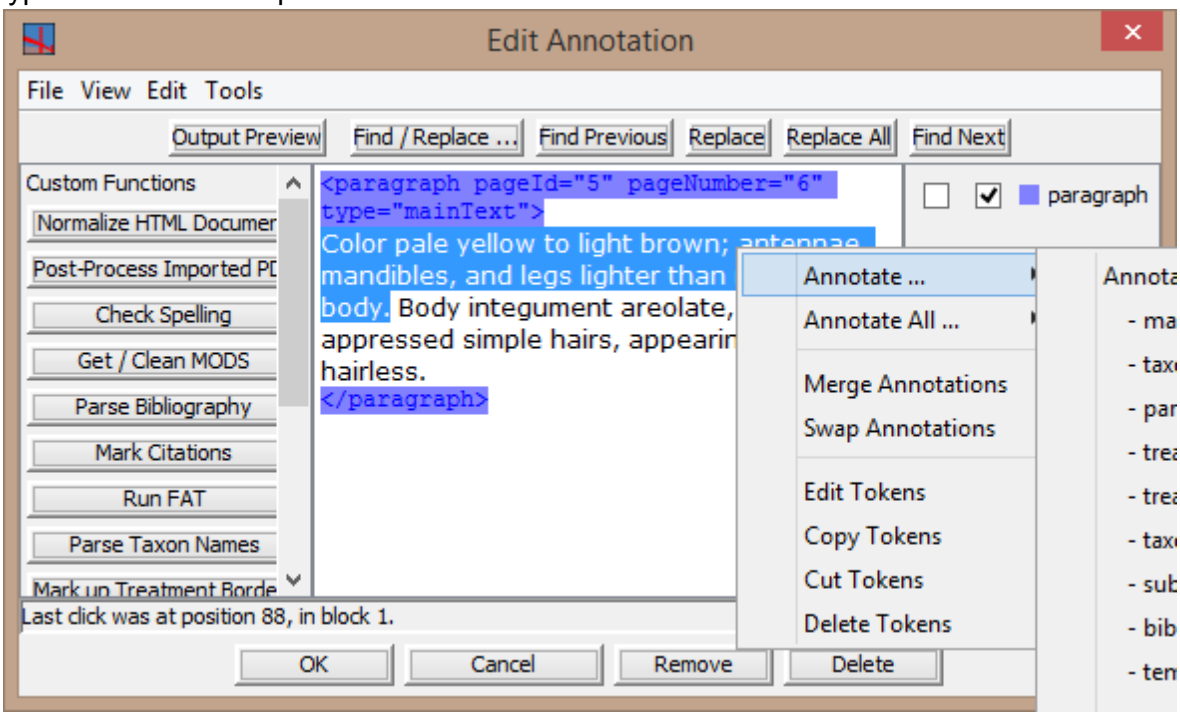


Fig XXX

If the annotation is not offered, select Annotate and type in the requested annotation type. A list of possible annotations is available in Annotation types.

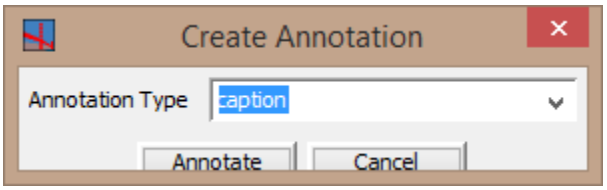


Fig XXX

Advanced: Use annotate All it certain words or strings need to be tagged in the whole text.

Annotations are case sensitive.

If a new annotations is requested, a feature request has to be send to XX to assure that the specific element will be kept and transformed into other XML or output.

Annotation of Attributes

Issue: How to enter an annotation attribute?

Solution: Highlight the annotation.

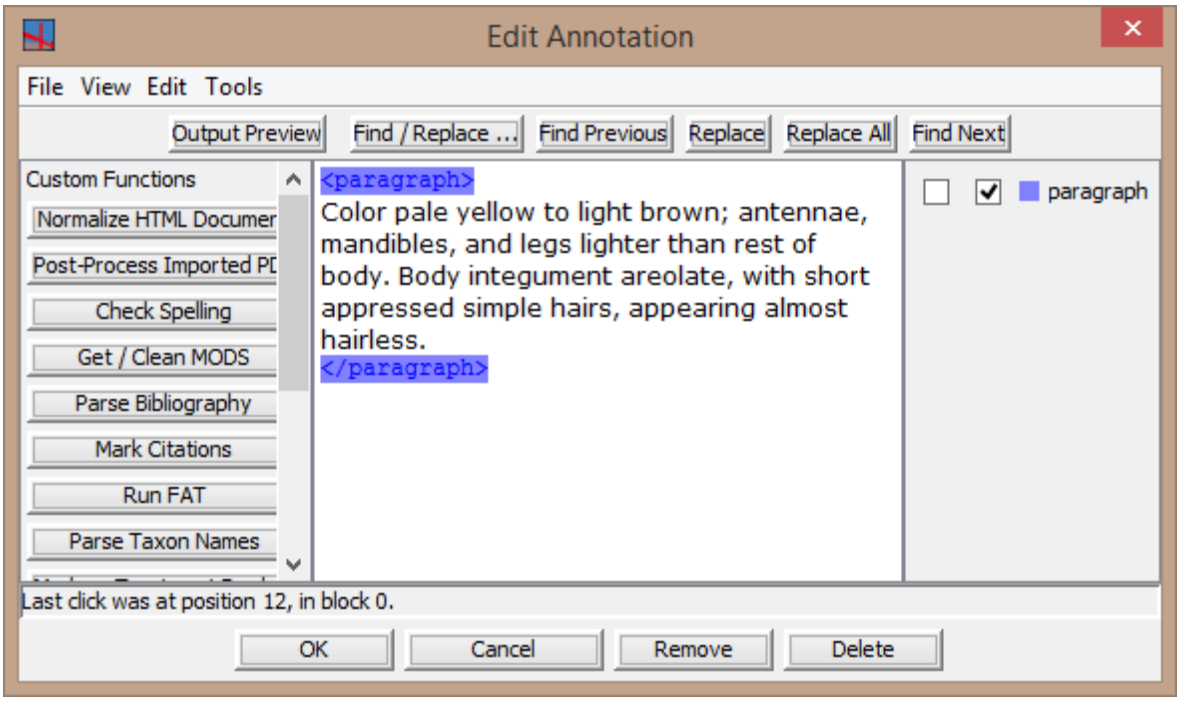


Fig XXX

Right click anywhere in the highlighted annotation

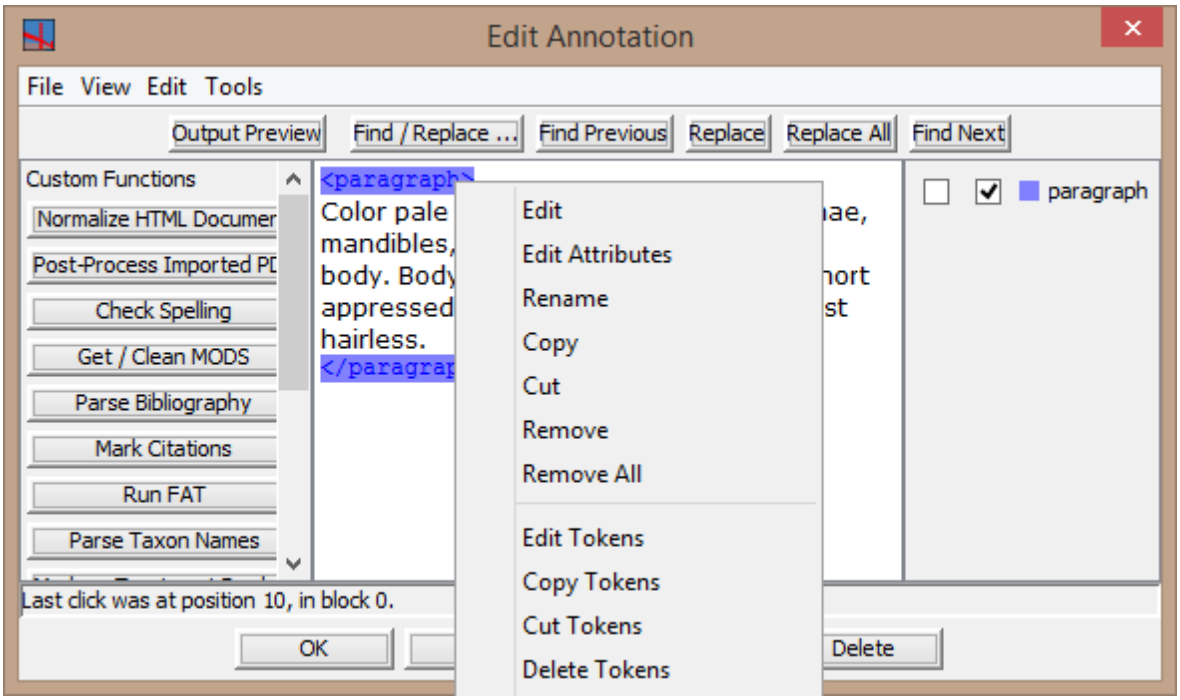


Fig XXX

Select “Edit Attributes”

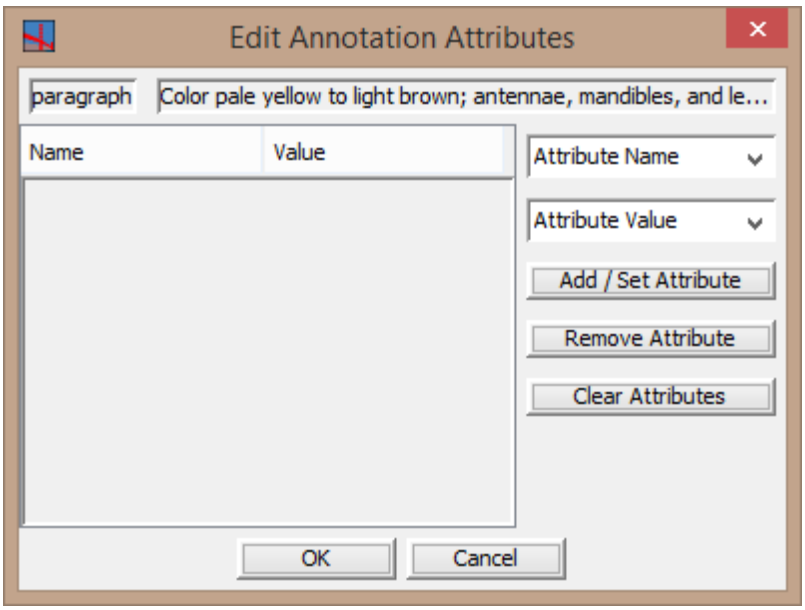


Fig XXX

Enter the Attribute Name; Enter the attribute Value and importantly press “Add / Set Attributes”

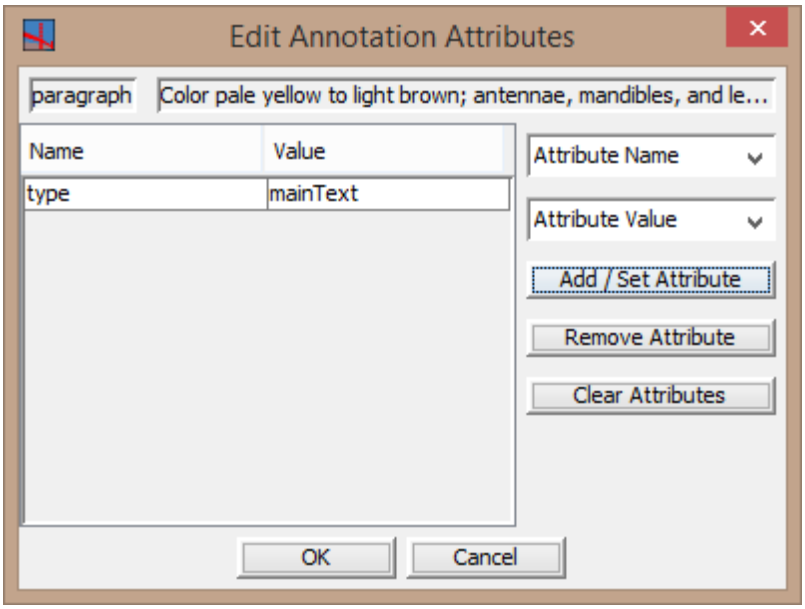
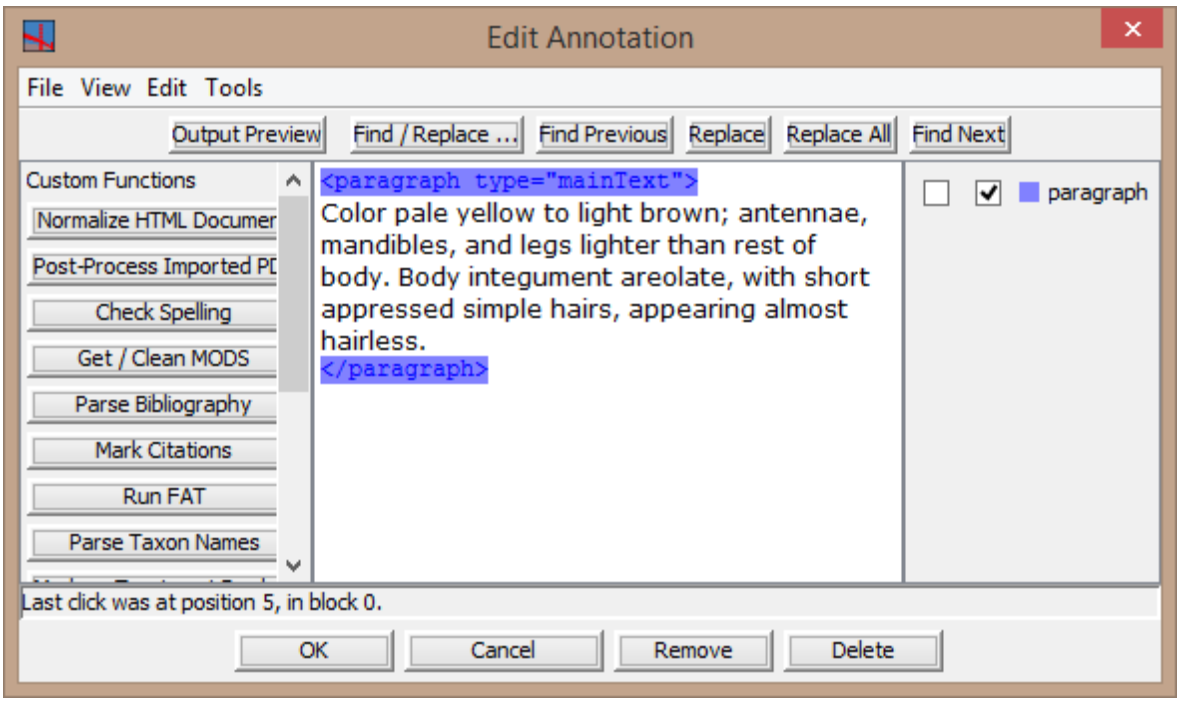


Fig XXX

Press “OK” to close the dialogue and the attribute appears in the annotation



Remove Attributes

Issue: How to remove an attribute?

Solution: Move with mouse over the hightlighted annotation (right box of the respective attributed ticked). Right click with mouse

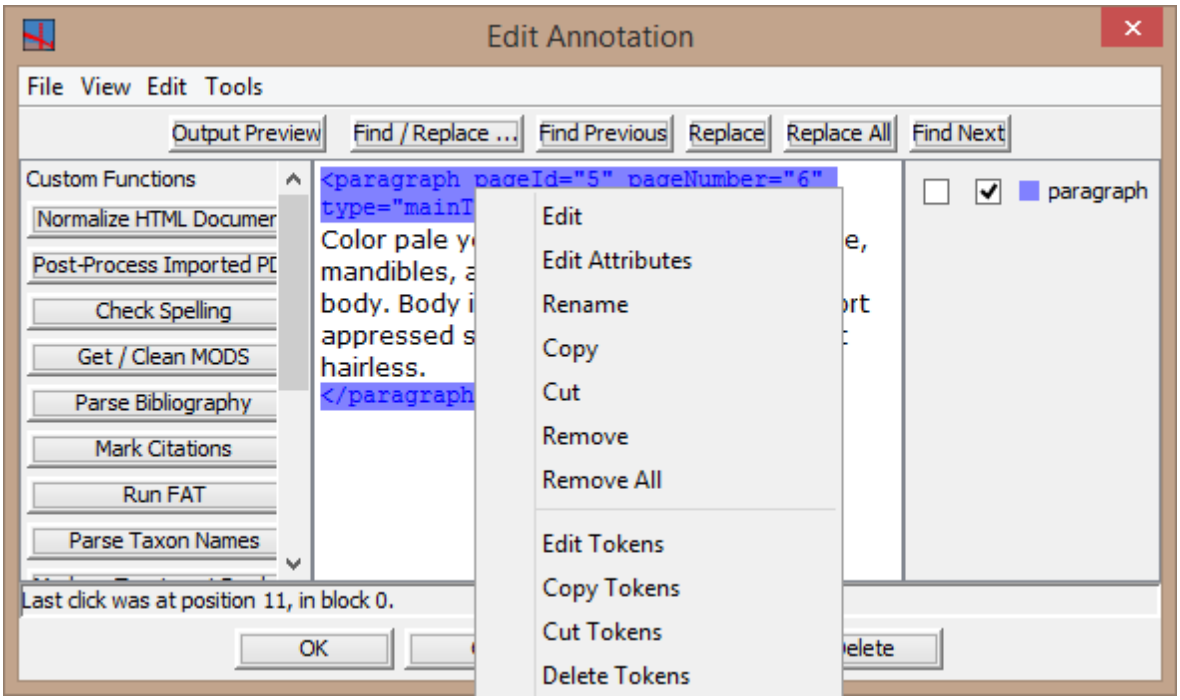


Fig XXX

Select “Remove” to remove the annotation, or “Remove All” to remove all the annotations in the entire document.

Advanced:

Remove annotations of an attribute

Issue: remove the annotation of an attribute

Solution: Move the mouse over the annotation and right click

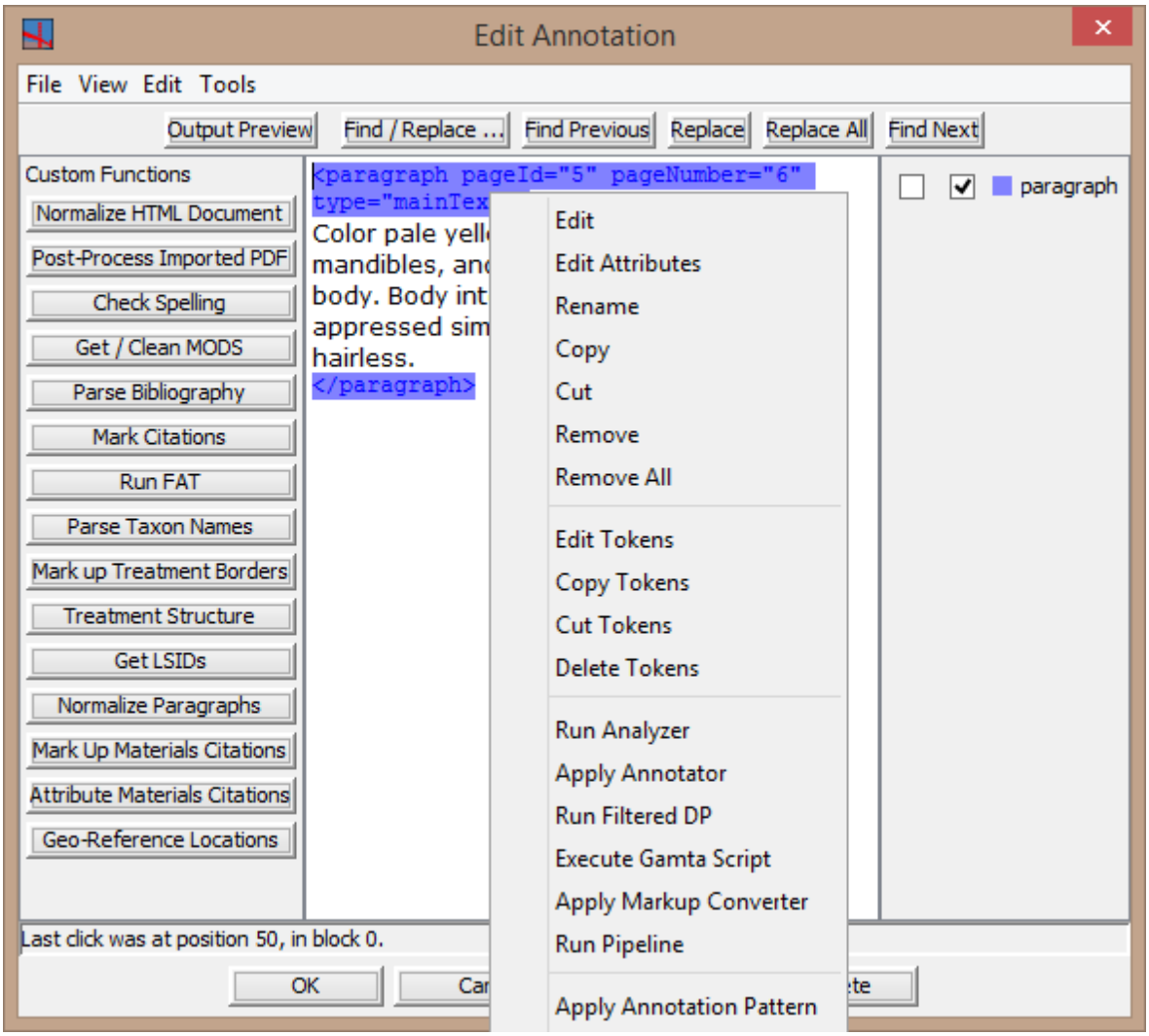


Fig xxx

Select in the menu “Edit Attributes”

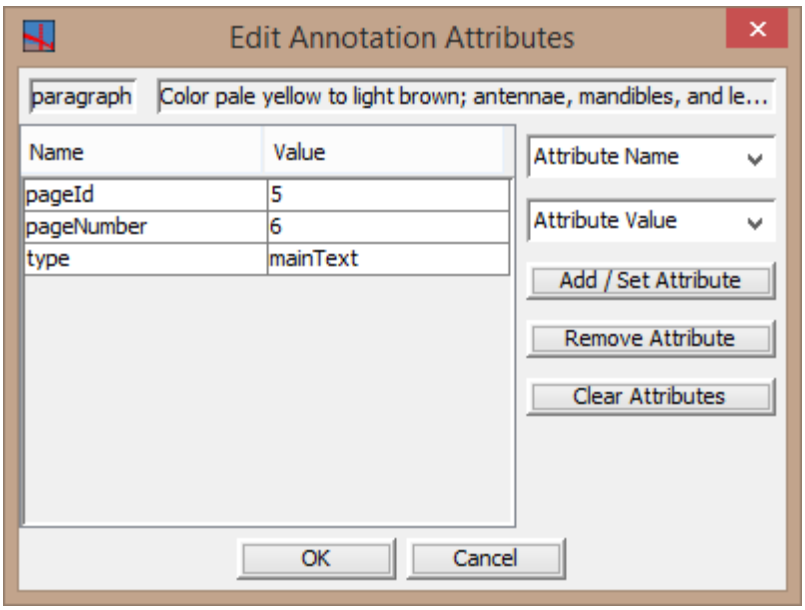


Fig XXX Edit Annotation Attributes menu

Double click with the left mouse button on the name of the attribute, and the element will appear on the right hand of the menu.

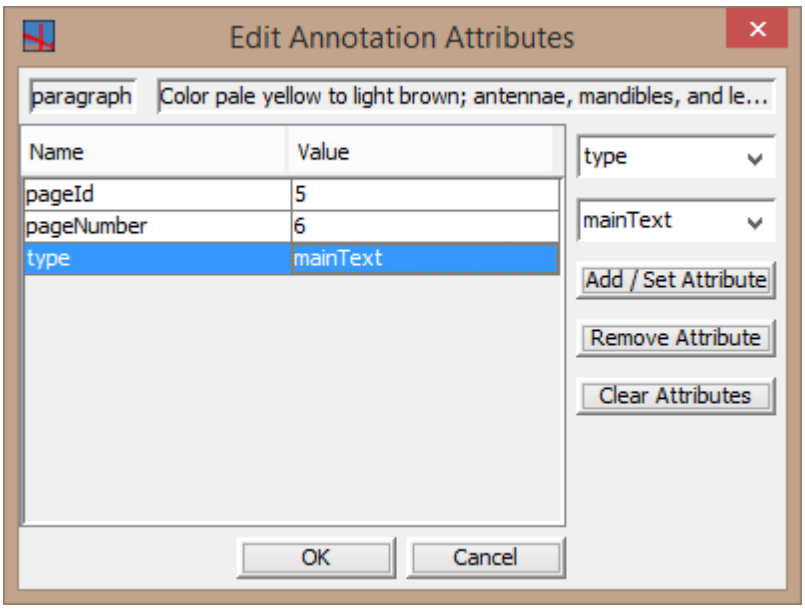


Fig XXX Edit Annotation Attributes menu: attributed selected

Press “Remove Attribute”, and “OK” to leave the menu to return to the editor.

Advanced:

Annotation of geographic coordinates

Issue: To add annotations of geographic coordinates that have been omitted in the markup process ore need be edited. To write the annotations through to the XML document, an additional step is needed.

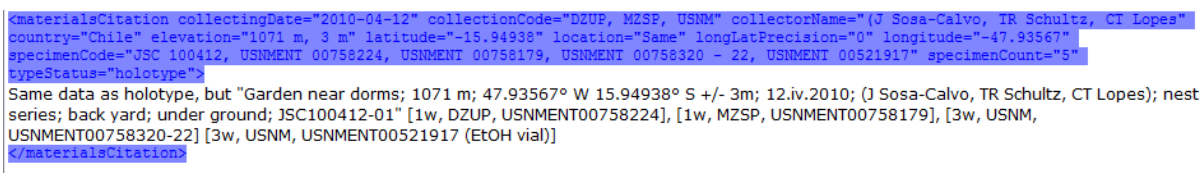


Fig XX MaterialsCitation with a complete set of attributes

Solution:
Open the materialsCitation in its own menu (click on the annotation)

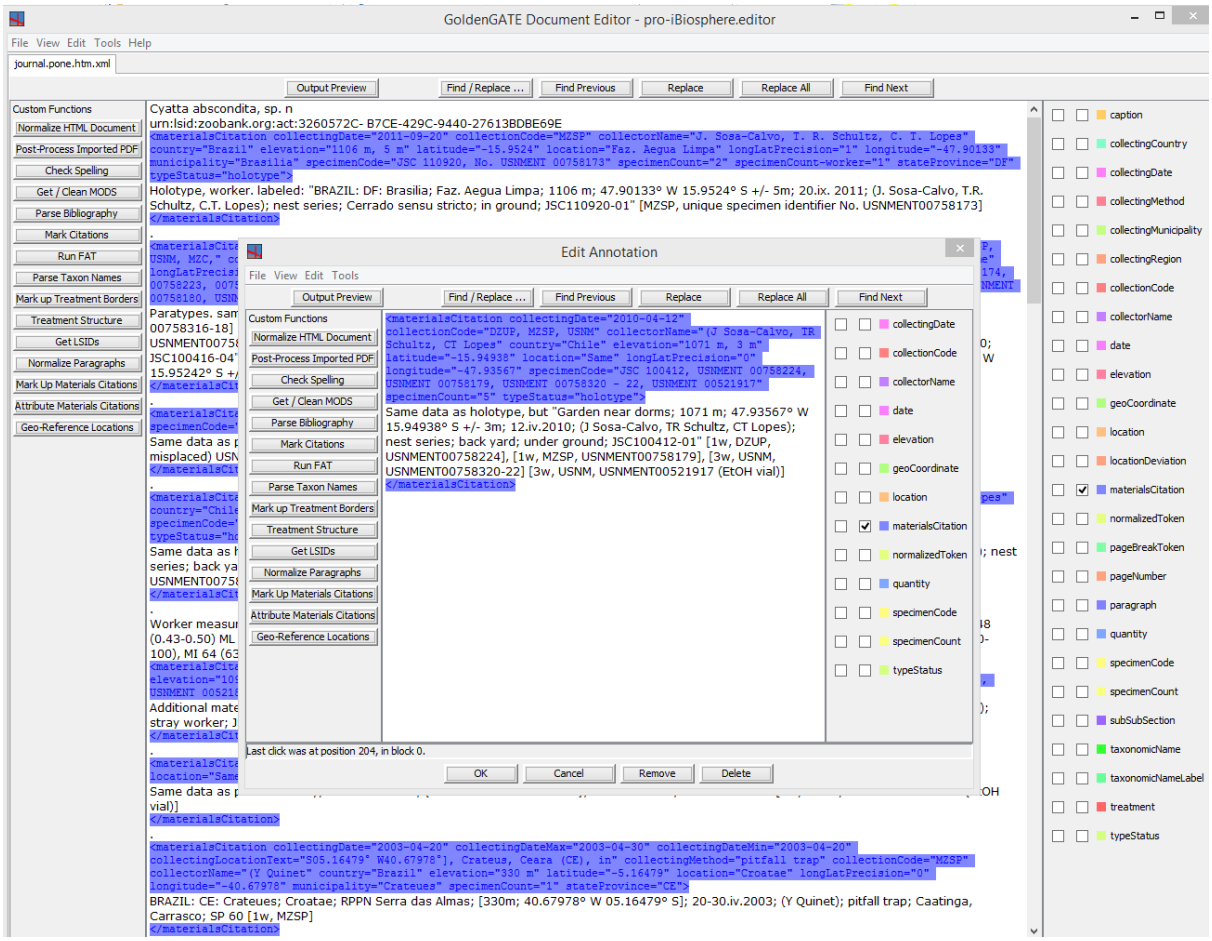


Fig XX MaterialsCitation opened in a popup window

Right click on the annotation

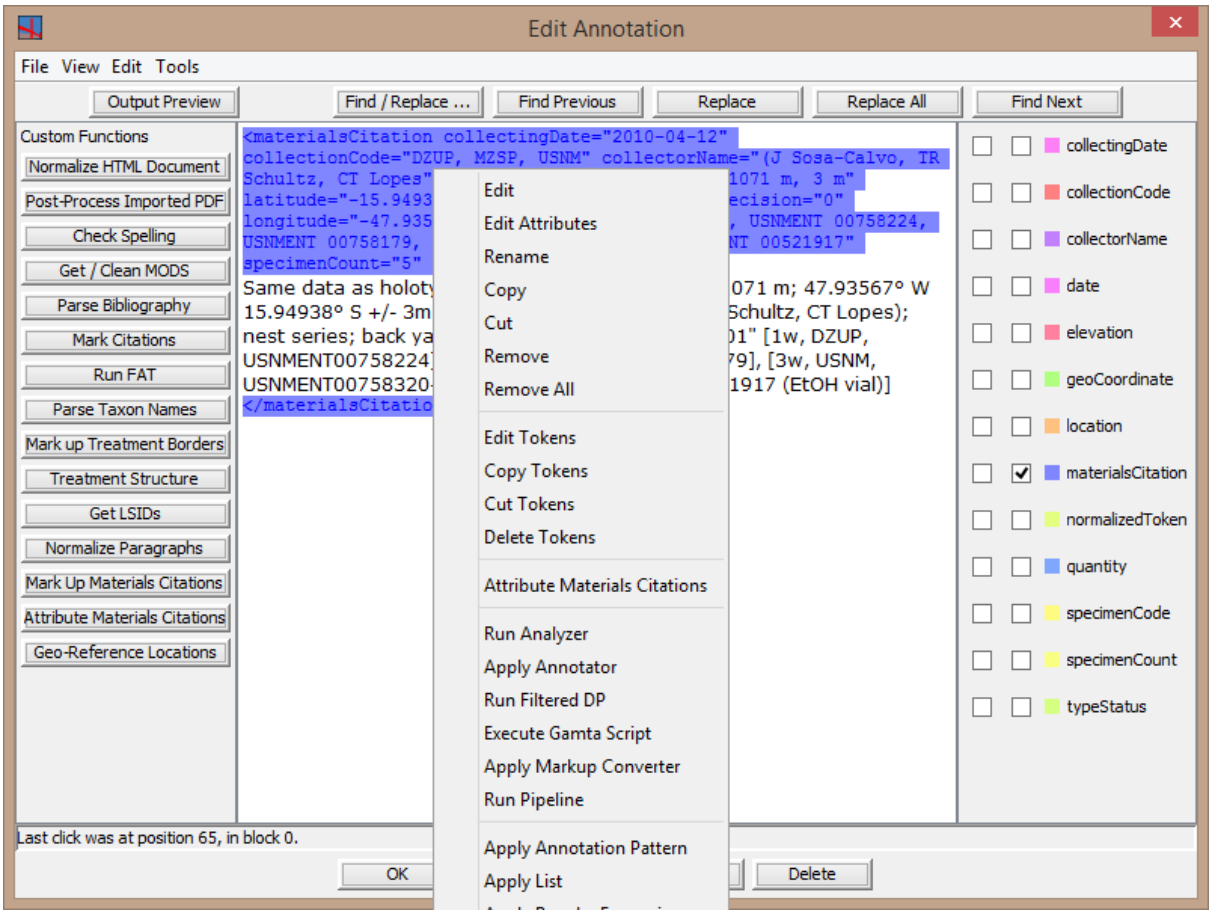


Fig XX MaterialsCitation: menu to select Edit attributes

Select Edit Attributes and insert for geographic coordinates for longitude “longitude” and a decimal degree value, with a “-” for West, select “Add / Set Attribute”, and for latitude“ latitude”and a decimal degree value, with a “-” for South, select “Add / Set Attribute” and close using “OK”. Do not use decimal values with more than 6 digits since such a value is below 1 Meter resolution.

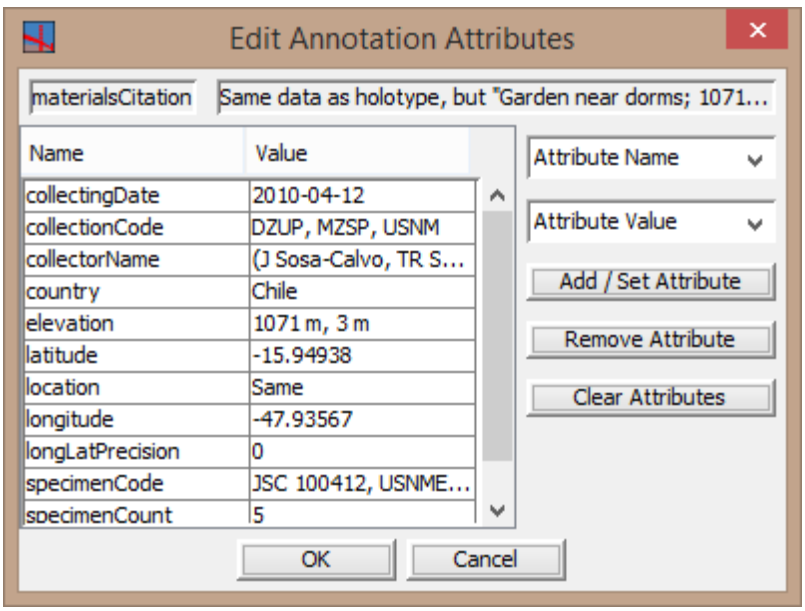


Fig XX MaterialsCitation: Edit Annotation Attributes

Important: To finish the annotation process, open the materialsCiation in a new pop-up window and run “Attribute Materials Citations”.

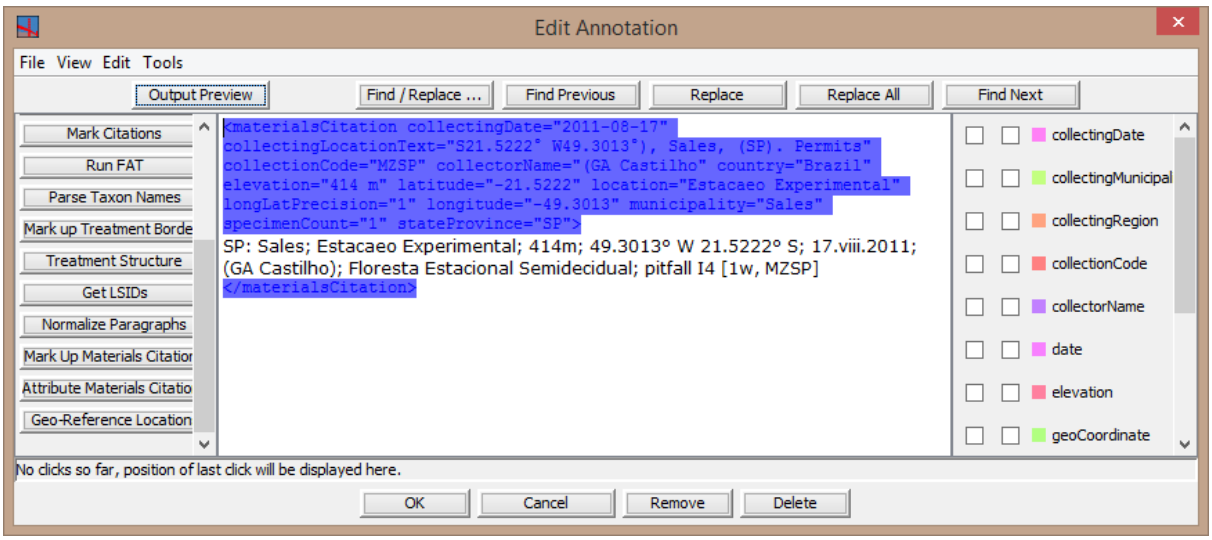


Fig XX

Answer the following pop-up menu with “Yes”

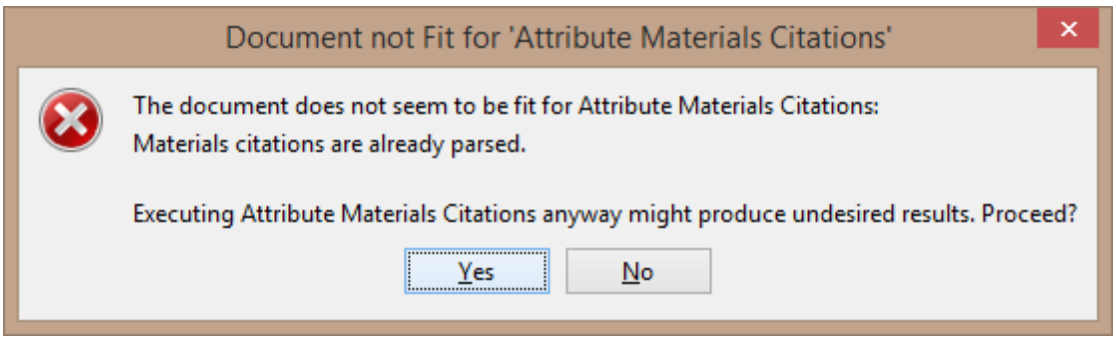


Fig XX

Close the pop-up menu, and the coordinates will appear in SRS googleMaps.

Counting specimens

Issue: Counting specimens in materials citations

Solution:

Linking to external resources, names or treatments

The annotations MaterialsCitation and treatmentCitation can be linked to their cited digital objects by adding the attribute “httpUri” for persistent identifiers such as for treatments (see Citing a treatment below

```
<materialsCitation
      httpUri="http://www.antweb.org/specimen/casent0101073" >
     Madagascar (M. Grandidier)
</materialsCitation>
```

```
<treatmentCitation
      httpUri="http://treatment.plazi.org/id/A31F8BD1-FEE4-FAF5-3F36-B00E6334DEBF">
</treatmentCitation>
```

Names can be linked by adding the following attributes
“LSID-HNS” for Hymenoptea Name Server LSID
“LSID-ZBK” for Zoobank Name LSID

```
<taxonomicName
    LSID-HNS="urn:lsid:biosci.ohio-state.edu:osuc_concepts:235700"
    LSID-ZBK="urn:lsid:zoobank.org:act:B6C072CF-1CA6-40C7-8396-534E91EF7FBB"
```

Anochetus boltoni Fisher
</taxonomicName>

5 Trouble shooting and special cases

Footnotes

Situation:
In some cases footnotes contain taxonomic treatments and extend over more than one page.

GoldenGATE / XML is not capable to accept the following treatment / footnote situation (here the footnote contains a taxonomic treatment): (doc. no. 4018 (http://antbase.org/ants/publications/4018/4018.pdf) , p. 59 – 61). Footnote and treatment are nested!

page 1:
taxonomic treatment (start)
footnote (start)

page 2:
taxonomic treatment (continued)
footnote (continued – end)

page 3:
taxonomic treatment (end)

Solution:
treating such texts as very special case and modifying the work flow as follows: skip the level 0 step and bring forward the “dissolve pages” step (after step 12 “Higher Order Taxa). Now the treatments and the footnotes appear in a correct order and can be marked up normally (it worked!) In this case the gg0 and tx0 documents are lacking.

Citations / reference groups in catalogues and as part of treatments:

Situation:
Taxonomic publications may contain lists with single taxa together with very reduced information such as a bibliographic reference and / or collecting data. In these cases we have to distinguish whether such an entry is part of a treatment (normally appearing as taxonomic history) or if it is part of a list of species or a catalog.

Example (list /catalogue): http://antbase.org/ants/publications/3616/3616.pdf (p. 16/17)

Voici la liste des espèces :
Camponotus (Myrmopelta) vividus (Smith) Santschi, 1921, Ann. Soc. Ent. Belgique, LXI, p. 311, Fig. A. B.
Formica vivida Sm., 1856, Cat. Hym. Brit. Mus, VI, p. 31 [[worker]].
Formica laboriosa Sm., Ibid, p. 32 [[queen]]
Colobopsis vivida Mayr, 1881, Verh. Zool.-Bot. Ces. Wien, XXXV. p. 354.
Camponotus (Colobopsis) vividus, Em., 1889, Ann. Mus. Civ. Genova, XXVII, p. 517.
Camponotus (Orthonotomyrmex) vividus Sants., 1915, Ann. Soc. Ent. France, LXXXIV, p. 264.

Solution:
In this case each entry should be marked as a complete treatment, containing a nomenclature part and at least a reference group:

```
<treatment>
	<subSubSection type="nomenclature">
		Camponotus (Myrmopelta) vividus (Smith)
	</subSubSection>
	<subSubSection type="reference_group">
		Santschi, 1921, Ann. Soc. Ent. Belgique, LXI, p. 311, Fig. A. B.<br/>
	</subSubSection>
</treatment>
<treatment>
	<subSubSection type="nomenclature">
		Formica vivida<br/>
	</subSubSection>
	<subSubSection type="reference_group">
		Sm., 1856, Cat. Hym. Brit. Mus, VI, p. 31 [[ worker ]].<br/>
	</subSubSection>
</treatment>
etc.
```

Example (taxonomic history): http://antbase.org/ants/publications/3616/3616.pdf (p. 18)

Campononotus (Myrmopelta) vividus Sm. v. cato Forel. Fig. C. D.
Camponotus (Orthonotomyrmex) meinerti st. cato Forel 1913, Rev. Zool. Afr., II, p. 336.
Camponotus (Orthonotomyrmex) vividus st. cato Wheeler. 1922, Bul. Am. Mus. Nat. Hist. XLV, p. 248, 976.
Camponotus (Myrmamblys) vividus st. cato Emery, 1923, Cat. Gen. Ins. Formicinae, p. 142.
Cette variété, dont je possède 3 cotypes, diffère du type par sa sculpture un peu plus forte, la face basale de l'épinotum plus courte, mieux bordée et sur un plan plus bas que le promésonotum.

Solution:
Here the listed taxa are making part of a treatment and should be marked as “reference group”. There is no necessity to split them up in nomenclatural parts and citation parts.

```
<treatment>
```

<subSubSection type="nomenclature">
Camponotus (Myrmopelta) vividus Sm. v. cato Forel. Fig. C. D.
</subSubSection>
<subSubSection type="reference_group">
Camponotus (Orthonotomyrmex) meinerti st. cato Forel 1913, Rev. Zool. Afr., II, p. 336.
Camponotus (Orthonotomyrmex) vividus st. cato Wheeler. 1922, Bul. Am. Mus. Nat. Hist. XLV, p. 248, 976.
Camponotus (Myrmamblys) vividus st. cato Emery, 1923, Cat. Gen. Ins. Formicinae, p. 142.
</subSubSection>
<subSubSection type="description">
Cette variete, dont je possede 3 cotypes, differe du type par sa sculpture un peu plus forte, la face basale de l'epinotum plus courte, mieux bordee et sur un plan plus bas que le promesonotum.
</subSubSection>
etc.

Template html document

An html input file can be created using the template below and just copy-pasting the relevant parts, such as the descriptions of a single treatment, into the document.
Each paragraph has to be tagged with the <p> element to retain the structure of the document.
At least two page breaks <hr/> and three page number on the same position in a separate paragraph are needed to allow GG to discover automatically the page numbering.

```
<html>
    <body>
        <p>
            ***first/previous page number
        </p>
        <hr/>
        <p>
            ***second page number
        </p>
        <p>
            *** copy text here; each pargraph in its own <p></p>
        </p>
        <hr/>
        <p>
            ***last/following page number
        </p>

    </body>
</html>
```

The document has to be saved as html document. No additional code should be added. Avoid using MS-Word to create the html file. Best is either a simple text editor or an html editor. Make sure to save as UTF-8 to make sure to keep the special symbols.

The new up-dates in an uploaded treatment does not show up on SRS

Sometimes some of the updates saved the SRS do not show up on SRS immediately. This happens sometimes. It takes some minutes to see the updates on http://plazi.org so for developmental reason, use the Boston server http://plazi.cs.umb.edu/GgServer/ . If this still does not work, add the string “?cacheControl=force” to the end of the treatment URL

http://plazi.org:8080/GgSRS/html/42E349B8AA3B4430FFC2DBB075CB5271
will look like this
http://plazi.org:8080/GgSRS/html/42E349B8AA3B4430FFC2DBB075CB5271?cacheControl=force

HttpUri annotation does not work in SRS-html

when the httpUri link does not work, check for the following:

1. has the httpURI spaces in the URL?

6 Glossary

6.1 GoldenGate Expressions

Word / Expression	Explanation
Token	A word, number or punctuation mark in the document. Treated as the atomic text unit by the Annotation Editor
Annotation	Representation of an XML tag in the GoldenGATE editor's data model
Type of an Annotation	The element name of the XML tag represented by the Annotation
Attribute of an Annotation	Representation of an attribute of the XML tag represented by the Annotation
Value of an Annotation	The textual content enclosed by the XML tag represented by the Annotation
Document Part	Special type of Annotation which can be edited as if it was a document of its own
Resource	Some object fulfilling some purpose related to handling or editing documents, or making editing more convenient
Annotation Source	Special type of resource for analyzing a document and creating Annotations marking some parts of it
Document Processor	Special type of resource for applying some processing to a document, whatever this might be for some particular document processor
Resource Provider	A sub-modul of the GoldenGATE editor managing and providing some type of resource (Annotation Source Provider and Document Processor Provider a have respective meaning)

6.2 TaxonX Expressions

Abstract	a summary of the present publication	A general element in modern scientific publications
Acknowledgements	a listing of acknowledgments to everybody who contributed to the research leading to the present publication, often including a reference to the funding agencies	ACKNOWLEDGMENTS This work was supported by the National Science Foundation under Grant No. DEB-0344731 to B.L. Fisher and P.S. Ward. Fieldwork that provided (…)
Bibref	a bibliographic reference	Brown, 1974 or

		BROWN, W.L. 1974. A remarkable new island isolate in the ant genus *Proceratium* (Hymenoptera: Formicidae). Psyche 81:70–83.
Biology_ecology	content relating to the ecology, biology and behavior of the taxon	(…)this species is a subterranean nester and forager (…)
Citation	a reference to an earlier description or listing of the taxon described	*Proceratium avium* Brown, 1974: 71, figs. 1 and 2 (worker, gyne and male). Mauritius: Le Pouce Mt, 700- 800 m, Native forest, 1 Apr. 1969 (coll. W.L. Brown) [examined] AntWeb MCZTYPE32216 (MCZC)
Description	the morphological (and possibly molecular) description of the taxon	(…)Cranium (excluding median part of clypeus) entirely covered with decumbent / appressed hairs among which standing hairs are scattered [major]; (…)
Diagnosis	the characters which make his taxon unique and separate it from others, and often those allowing to recognize the taxon immediately	The following character combination differentiates *berlita* from all its congeners: scrobe absent, (…)
Discussion	anything which relates to the description, the nomenclatorial history, the behavior or comments relating to the taxon	*Discothyrea* of Madagascar belong to the first group...
Distribution	a summary statement of the distribution of the species. The individual records are listed in "materials examined"	Distribution: North Borneo
Document head	the title, author and there addresses of the publication	
Etymology	the origin of the name of the taxon	The specific name is an arbitrary combination, to be treated as a noun in apposition.
Introduction	the introduction to the present paper; this is an element often marked with a specific title in recent	A general element in modern scientific publications

	publications or then the first general section in older publications proceeding the description of the taxa. Often general issues of the taxa are summarized, which is in modern paper more often to be found in the discussion.	
Key	an identification tool to taxa. In most cases, these are dichotomous, that is a couplet of alternatives referring to the next and finally to a taxon name.	1. Red curved hair on occiput….2 Yellow straight hair on occiput ….3
Materials & methods	a section describing the techniques, measurements and methods used to derive the results in the respective publication	A general element in modern scientific publications
Materials examined	a listing of all the individual collecting events used in the description of this taxon, that is, not necessarily conclusively, a combination of locality, date, collector, sample number, habitat, etc.	Sample No. 4186; type locality: Poring Hot Spring, East ridge, 820 - 860 m a. s. l., Sabah, Malaysia (leg. Annette K. F. Malsch, 16. V. 1998)
Multiple	Generally, anything that can not be assigned to any of the annotation above.	
Nomenclature	any elements pertaining to the naming of taxon according the International Code of Zoological Nomenclature	
Reference group	the section containing the bibliographic references <bibref> in the publication. This is an element hardly known in the old legacy literature	This refers to the “Literature cited section”, A general element in modern scientific publications
Synopsis	A list of taxa (e.g. all the taxa treated in a revision)	*Metapone* species: *M. leae*

6.3 Element Descriptions and Examples

The following table gives an overview about the TaxonX elements. A detailed description is available here http://taxonx.org/documentation/taxonx1.xsd.html#h519182448:

author	The author of the original description of a taxon in a nomenclature or synonymy section of a taxonomic treatment.
bibref	A bibliographical reference

	<tax:citation> <tax:xid identifier="doi:10.1046/j.1523-1739.1998.96177 .x"/> SAFFORD, R.J., AND C.G. JONES. 1998. Strategies for Land-Bird Conservation on Mauritius. Conservation Biology 12:169–176. </tax:citation>
character	A morphological character.
citation	A bibliographic reference.
collection_event	Contains information regarding the collection of a specimen.
div	A block level textual division of a text. Attributes: n, number or name of division; type: type of division. If div occurs inside a treatment, suggested values are: abstract, acknowledgments, biology_ecology, description, diagnosis, discussion, distribution, etymology, introduction, materials_examined, materials_methods, multiple, synopsis.
figure	A figure or graphic.
figures	The statement identifying the figures related to a given treatment. <tax:nomenclature> <tax:name>Proceratium avium</tax:name> <tax:author>Brown</tax:author>, <tax:year>1974</tax:year> <tax:figures>Figs. 5-13.</tax:figures>...
figureCitation	A citation of a figure
head	A heading, such as the title of a section, etc.
locality	A geographical location.
name	A scientific name of a taxon as it appears in the source text.
nomenclature	The heading of a taxonomic treatment containing the scientific name of the taxon described.
note	A note, such as a footnote or endnote, in the source text. Use the place attribute to indicate the placement of the note in the source document (e.g., "foot", "end"). Use the n attribute to contain the number or symbol used to label the note in the source text.
p	A paragraph or other textual block. <tax:p> Venter very glossy. Ostiolar peritreme ligulate, gently curved, quite long, its apex nearly reaching lateral margin of plate. Rostrum reaching onto seventh abdominal sternite. Legs pale yellow, irregularly spotted and blotched with castaneous spots, terminal tarsal segment tending to become rosy. </tax:p>
pb	Page break. Indicates the point in the source text where a new page begins. Use the n attribute to record number of the new page;

	use the url attribute to lnk to an electronic graphical representation of the page.
ref_group	A group of bibliographic references.
seg	Segment. A phrase-level segment of text.
state	A character state.
TaxonX	Contains a single TaxonX document, including a TaxonXHeader and TaxonXBody (see also the documentation for taxonx: http://taxonx.org/documentation/taxonx1.xsd.html#h519182448
TaxonXBody	Contains a single text including at least one taxonomic treatment.
TaxonXHeader	Contains identification and description of the TaxonX document and its source, expressed in the Metadata Object Description Standard (MODS).
treatment	A taxonomic treatment
xid	External identifier. A pointer to an identifier assigned to the parent object in an external system. Contains optional attributes "identifier", the identifier, "source", the system in which the identifier can be found, "uri" a uniform resource identifier.
xmldata	A wrapper element used to include data from an external schema.
year	A year in a citation of a document, a scientific name, or a specimen.

A complete TaxonX documentation can be found at http://taxonx.org/documentation/taxonx1.xsd.html.

6.4 GoldenGate Internal Markup Elements

General Document Structure		
document	root element for the document proper	
section	a document section	
subSection	a document sub section	
subSubSection	a document sub sub section	
paragraph	a paragraph of text, both layout (visual text block) and logical (what would be a paragraph in non-lain-out flowing text)	
page	a page in print layout (removed after abstraction from print layout)	
Special Document Structure		

caption	the caption of a figure, table, etc., usually a single paragraph	
footnote	a footnote, usually a single paragraph, but can be multiple ones	
pageTitle	a page header or footer, usually a single paragraph	
bibRef	a bibliographic reference, usually a single paragraph	
treatment	a taxonomic treatment, same level as subSection	
table, tr, td	a table with its rows and cells	
Generic Word Level		
normalizedToken	a word that had diacritic marks stripped from letters to simplify application of regular expression patterns; original value (with diacritics) preserved in respective attribute	
pageBreakToken	marker for the first word in a page in print layout	
correctedMisspelling	a corrected misspelling or OCR error; original value preserved in respective attribute	
pageNumber	a page number, usually within a pageTitle	
Special Word or Phrase Level		
bibRefCitation	an in-text citation of a bibliographic reference	
figureCitation	an in-text reference to a figure	
tableCitation	an in-text reference to a table	
captionCitation	an in-text reference to a figure or table, or some other element that comes with a caption	
author, title, year, pagination, etc.	the details of a bibliographic reference	
quantity	a quantity, consisting of a number and a unit	
date	a date	
geoCoordinate	a geographic coordinate	
country	a country name	
location	the name of a location	

abbreviation	an abbreviation or code that is short for some detail data	
abbreviationData	a container around an abbreviation and its definition	
abbreviationReference	a reference to an abbreviation, implying the data associated to the abbreviation by its definition	
Taxonomic Word or Phrase Level		
taxonomicName	the scientific name of a taxon	**Formica rufa**
taxonomicNameAuthority	the author of a scientific name of a taxon	Formica rufa **Linnaeus** Formica rufa **Linnaeus, 1758**
taxonomicNameLabel	a qualifier indicating the status of the taxon	**Nov.sp.**, **comb.nov.**
treatmentCitation	citation of an individual treatment, rather than the publication the cited treatment is part of; often a combination of a taxonomic name, its authority, and a page number, with the authority doubling as a bibRef citation	
materialsCitation	a materials citation within a treatment	
collectingCountry	the country a specimen was collected in	
collectingRegion	the region (first level administrative subdivision of a country, e.g. a state or province) a specimen was collected in	
collectingCounty	the county a specimen was collected in	
collectingMunicipality	the municipality a specimen was collected in	
locationDeviation	the distance and bearing from a names location to the location a specimen was collected	
collectingDate	the date a specimen was collected	
collectingMethod	the method or device a specimen was collected with	handnet, pitfall trap, ...

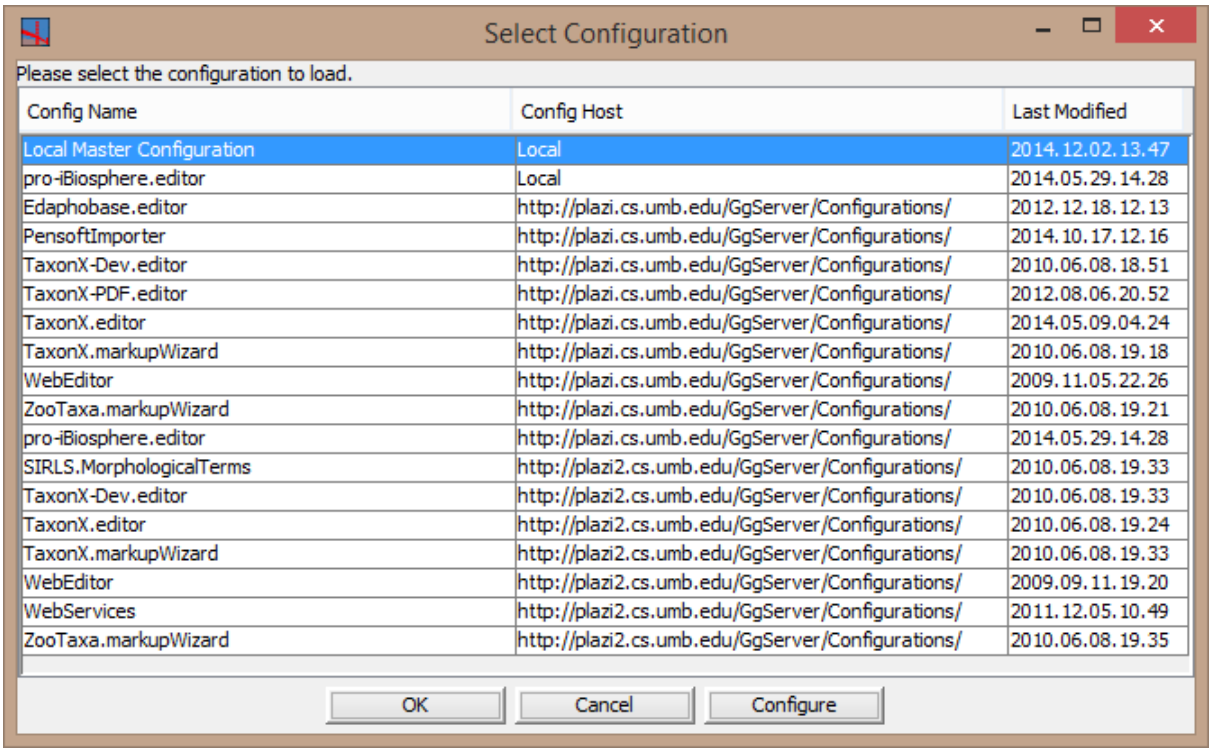
collectorName	the name(s) of the individual(s) who collected a specimen	
specimenCount	the number of individual specimens collected together, often comes as a combination of a number and a specimen type or type status	2 females, 3 paratypes, ...
specimenType	the type / sex / caste of the collected specimens	worker, female, queen, ...
typeStatus	the type status of the specimens	holotype, paratype, ...
collectionCode	the letter code identifying the collection a specimen is deposited in	CASENT
specimenCode	the code identifying an individual specimen	CASENT 0815

7 GoldenGATE versions

7.1 GoldenGATE “classic”

The following versions of GoldenGATE are available uppon starting GoldenGATE. The local versions are those that exist already locally. The other version will be downloaded and installed upon request.

Between the various can be switched using “Configure”.



7.1.1 Master Configuration

This is the master configuration that includes all the analysers and pipelines and other markup automation gizmos; it is not tailored to a specific task or workflow, but rather the union of a markup generation and visualization tools from all workflows.

7.1.2 pro-iBiosphere.editor

This is the current default version for general markup. It supports markup of taxonomic publications starting from HTML, XML, and PDF documents; PDFs can be both born-digital or scanned (imaged based).

7.1.3 Edaphobase.editor

This version supports markup of taxonomic and ecological publications. In comparison to pro-iBiosphere.editor, it includes also helpers for handling elements that not as frequently occur in taxonomic publications, namely table based materials citations and collection data defined and referred to by means of abbreviations.

7.1.4 PensoftImporter

This version is not meant for desktop use; its purpose is to provide the markup generation tools used in the fully automated import of TaxPub documents from Pensoft into the Plazi servers.

7.1.5 TaxonX.editor

This version is the predecessor of pro-iBiosphere.editor. It supports markup of taxonomic publications starting from HTML and XML, but not PDF.

7.1.6 TaxonX-Dev.editor

This version is an extension of TaxonX.editor that includes several experimental markup generation tool. It is sort of obsolete now, as all of the latter have gone into pro-iBiosphere.editor.

7.1.7 TaxonX-PDF.editor

This version is an extension of TaxonX.editor that includes the facilities for starting the markup process from PDF documents, both born-digital and scanned (image based). It is sort of obsolete now, as it has at one point become pro-iBiosphere.editor, and all further development and improvement went into the latter.

7.1.8 TaxonX.markupWizard

This version is not intended for use in GoldenGATE Editor, but rather in GoldenGATE Markup Wizard, an experimental tool that guides users trough a document markup workflow by means of a data driven process definition.

7.1.9 ZooTaxa.markupWizard

This version is a customization of TaxonX.markupWizard, specifically tailored to publications from the Zootaxa journal.

7.1.10 WebEditor

This version is not intended for use in the GoldenGATE Editor desktop application, but was used in an applet based version of GoldenGATE Editor that was embedded in a web page. The latter has been discontinued due to fading support of Java applets, so this version is obsolete.

7.1.11 WebServices

This version is not meant for desktop use; its purpose is to provide the markup generation tools backing the data extraction web services offered by GoldenGATE Web Services via a REST API.

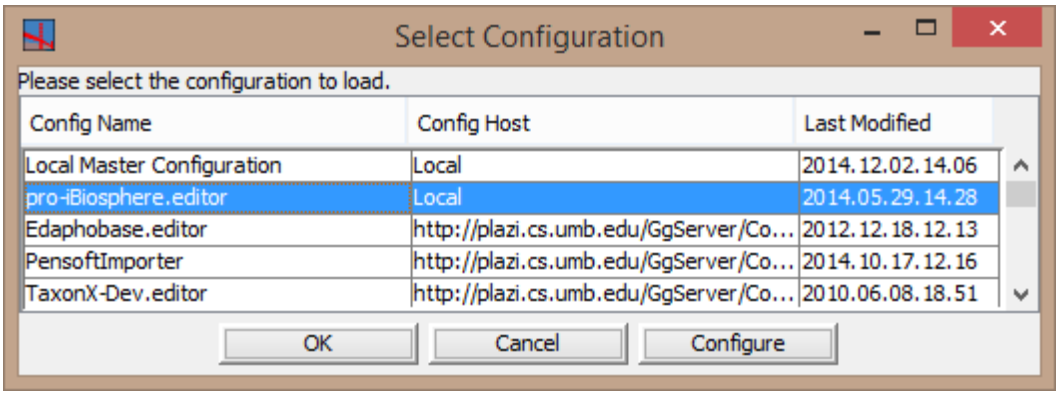
7.2 GoldenGATE Imagine

GoldenGATE Imagine is based on an alternative data model and can not be opened through GoldenGATE Editor. A description and manual is available here.

8 Advanced editing and programming

8.1 Change configuration

The configuration of GoldenGATE can be changed using the “Configure” button



The respective values can be adjusted in the fields in this pop-up menu.

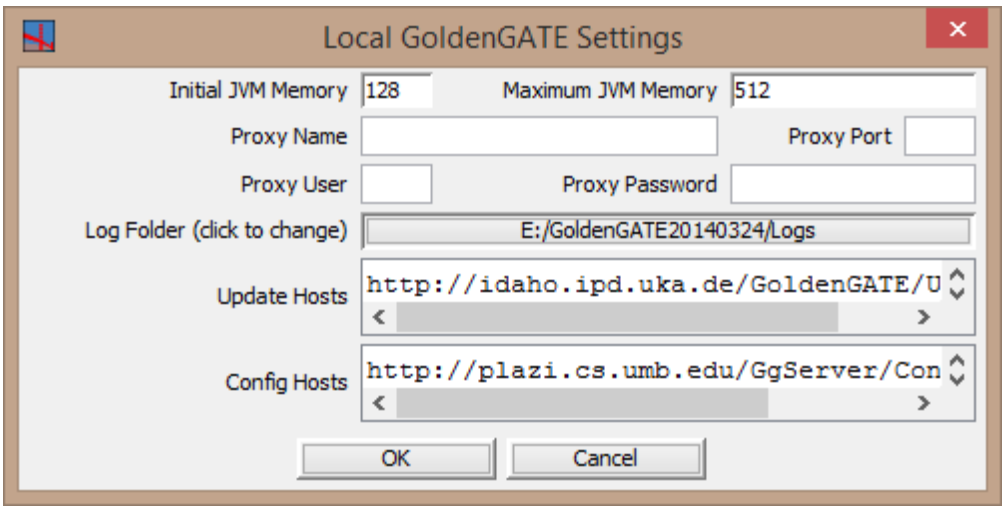


Fig xxx

The allocation of memory to the Java Virtual Memory (JVM) can be adjusted. Default is a maximum of 512. Recommend is a mutliple of it, beginning with 1536MB

Logs for GoldenGATE are assigned be default in a subdirectory of the main GoldenGATE directory.

Update Hosts are the servers that GoldenGATE desktop applications draw updates to their core libraries from.

Config Hosts are the servers that GoldenGATE applications draw the various configurations from, as well as updates for the latter.

8.2 Create analysers

Analyzers are document analysis and markup generation tools implemented in the Java programming language. They handle documents based on the *visitor* design pattern. Being Java classes, Analyzers can use the full API of the data model underlying GoldenGATE to apply whichever changes to a document, both the markup and the text.
Technically, Analyzers are implementations of the *de.uka.ipd.idaho.gamta.util.Analyzer* interface; the JavaDoc in this interface explains the role of the four methods it specifies.
To deploy an Analyzer in GoldenGATE, it has to be packed in a jar file, and the latter be deposited in the *<GoldenGATE>/Plugins/AnalyzerData* folder. Multiple Analyzers can be

packed in the same jar file without problems. There are two abstract convenience implementations of *de.uka.ipd.idaho.gamta.util.Analyzer*: *de.uka.ipd.idaho.gamta.util.AbstractAnalyzer* for rather simple Analyzers that do not use external resources, and *de.uka.ipd.idaho.gamta.util.AbstractConfigurableAnalyzer* for Analyzers that do use external resources, configuration parameters, etc. The JavaDoc of the latter two classes explains this in more detail.
If an Analyzer requires file based resources to work, e.g. gazetteer lists, the latter have to be deposited in the data folder associated with the jar file the Analyzer proper is deployed in. This folder is named the same as the jar file, with the *.jar* suffix replaced by *Data*. This means that the data folder associated with the Analyzers deployed in *MyAnalyzer.jar* is *MyAnalyzerData* in the same folder as the jar file.
If an Analyzer requires third party libraries or jar files, the latter have to be deposited in the binaries folder associated with the jar file the Analyzer proper is deployed in. This folder is named the same as the jar file, with the *.jar* suffix replaced by *Bin*. This means that the binaries folder associated with the Analyzers deployed in *MyAnalyzer.jar* is *MyAnalyzerBin* in the same folder as the jar file.

8.3 Create pipelines

Pipelines are basically concatenations of document analysis and markup generation tools. Their purpose is to make the latter accessible as a whole, and to make sure the latter are always applied in the configured order.

8.4 Programming export from GoldenGATE XML

What do you mean here?
The sort of export you describe in your comment happens in the server, automatically triggered by updates to documents. There are multiple exporters (currently around 10), each one responsible for pushing data to a specific destination in a specific formal. This is nothing you'd implement in GG Editor or GG Imagine.
If you want to pull data (treatments) from the server, you can use the RSS feed (http://plazi.cs.umb.edu/GgServer/xml.rss.xml) that indicates treatment updates to become aware of these updates, and then pull the XML treatments from the server via the URLs embedded in said RSS feed.

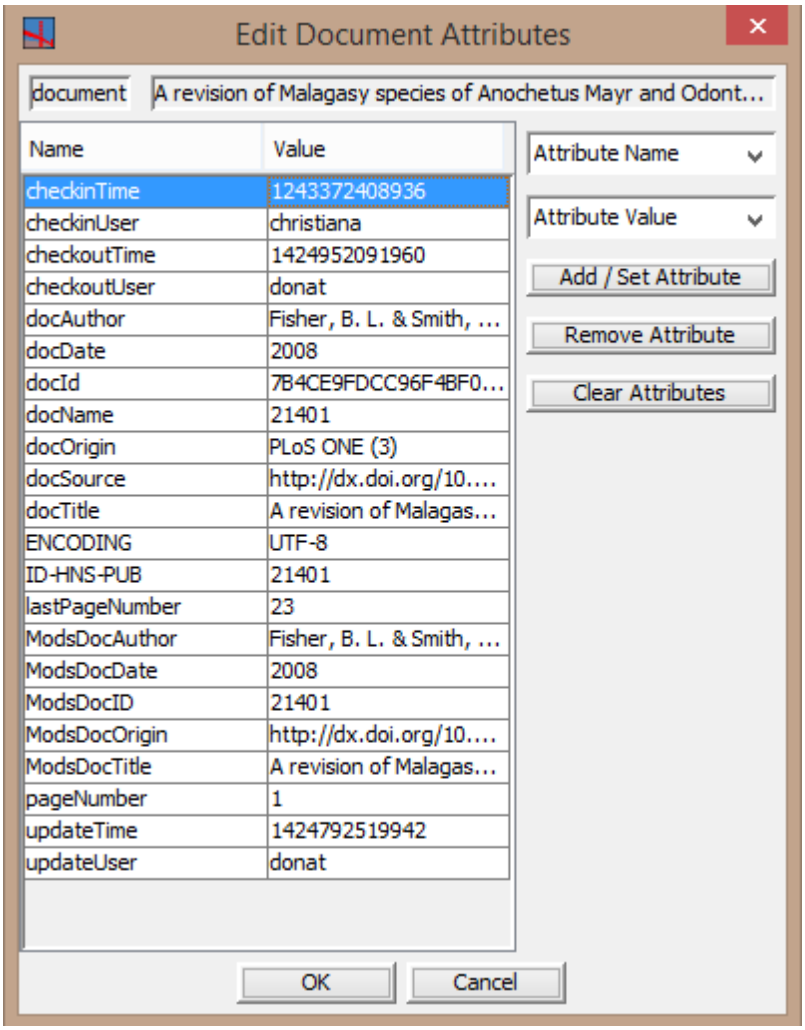
8.5 Programming import into GoldenGATE XML

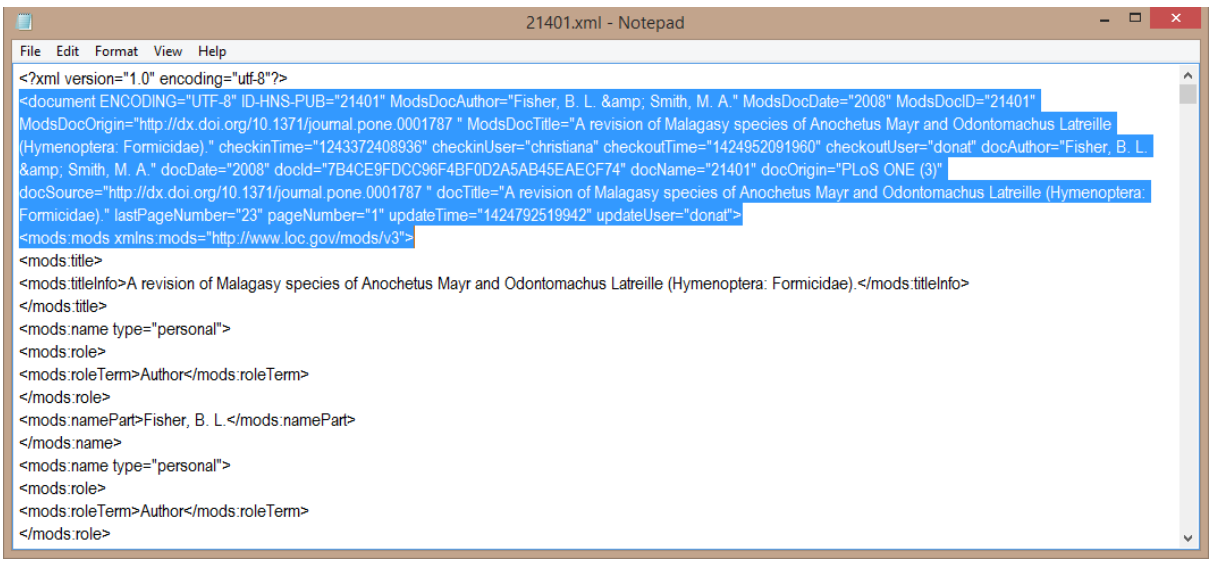
What do you mean here?
The sort of push import you describe is best done via the Java client for the central document archive (class *de.uka.ipd.idaho.goldenGateServer.dio.client.GoldenGateDioClient*). This also requires a user account (login + password) for the server.
The import from Pensoft, on the other hand, work via a pull from inside the server. Such a pull import has to be installed and then executes periodically (every 24 hours in case of Pensoft).

8.6 Annotate document attributes

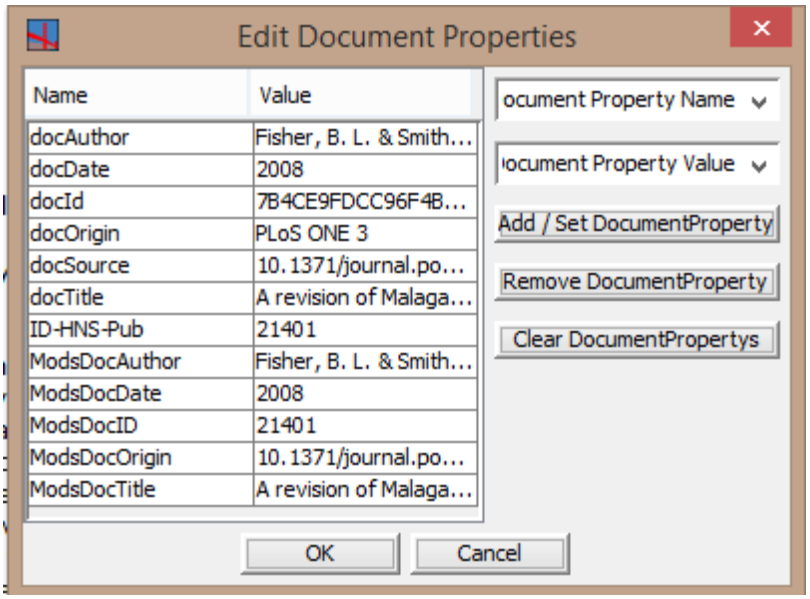
This is a unified way of storing document attributes abstracted from the representation in the document XML.

This makes the sytem independent from the use of a specific metadata schema, such as MODS used by Plazi.





8.7 Annotate Document properties



These are internal vehicles to allow universal access to things like bibliographic meta data.

8.8 Add and change publication ID

1. Open

9 Links to external resources

9.1 Links related to GoldenGATE

GoldenGATE Imagine Manual
SRS Access to treatments on Plazi
Test publications

9.2 References

Plazi publications

10 Acknowledgments

This program is part of a bilateral digital library grant awarded by the Deutsche Forschungsgemeinschaft (DFG BIB47) and US National Science Foundation (IIS-0241229) to Klemens Boehm, Universität Karlsruhe (TH), and Christie Stephenson, American Museum of Natural History, New York The preparation of the manual and its field testing has been supported by GBIF and Pro-iBiosphere (EU-FP7 program).