

GoldenGATE

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

(GoldenGATE RC2.5)

Christiana Klingenberg, Guido Sautter, Donat Agosti & Terry Catapano

(klingenberg@smnk.de; sautter@ira.uka.de; agosti@amnh.org; terryc@panix.com)

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

<http://idaho.ipd.uka.de/GoldenGATE/>

January 31, 2007

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 GoldenGATE document editor is intended for the creation of new XML content from plain text data, and for inserting additional markup to document centric XML documents in order to make them more data centric and machine readable. It is based on a data annotation model. An extensive Help-function in the program describes in detail the program and its features, including a glossary of terms.

For the purpose of creating digital documents from printed (legacy) systematics literature, GoldenGATE offers automatic to semiautomatic tools to mark-up its logic (e.g descriptions, nomenclature) content, and to export three versions of TaxonX documents.

The 'purist's' version retains the original sequence of the layout of the printed publications (e.g. figure captions, footnotes, etc. in the middle of treatments), i.e. level 0 or gg0.

The 'data mining ready' version moves all the captions, footnotes to the end of the document, though with an attribute referring to the page within the original publication, i.e. level 1 or gg1.

The enhanced version includes the mark up of more detailed information, such as geographic names, collecting events, people's names, i.e. level 2, or gg2.

A concise version of all the steps needed to create the above mark up is given in chapter 10 (page 25).

1. General introduction to TaxonX

TaxonX is a XML schema for encoding legacy taxonomic literature in order to:

- Create open, stable, persistent, full text digital surrogates of taxonomic treatments,
- Identify taxonomic treatments and their major structural components to enable networked reference and citation,
- Identify lower level textual data such scientific names, localities, morphological characters, and bibliographic citations to facilitate their extraction by, and integration with external applications and resources,
- Study and describe the structure of systematics publications by creating few typical corpora of literature, such as entire journal (eg AMNH Novitates), across taxa (e.g all ant systematics papers post 1995), or faunistic (e.g. all ant systematics paper covering Madagascar ranging from 1758 to 2006).

TaxonX is a lightweight and flexible schema which should be quickly learned and may be applied to the wide variety of formatting present in legacy documents. It permits, and sometimes relies on (see use of MODS for file-level bibliographical metadata), use of external schemata. It has loose content requirements allows for instances to be encoded over time and at many levels of granularity, while maintaining validity through iterations. Additionally, TaxonX contains mechanisms for semantic normalization of the data contained in treatments. The schema can be readily converted to or expressed as an extension of the NLM/NCBI Journal Archiving DTD.

Levels of Markup

Different levels of mark up are possible. For this moment the actual GoldenGATE Version and the TaxonX output is tested up to level 2.

Level 0

Document/treatment metadata; Treatment and nomenclature section (including scientific name) and preserving all layout settings like page breaks, captions, within the treatment, footnotes, etc. As mentioned above, Level 0 documents are not proper for data mining but fulfil the TaxonX XML schema requirements.

Level 1

Document/treatment metadata; treatment and nomenclature section (including scientific name).

Level 2

Structural components of treatments (e.g. Materials Examined, Description, etc...)

Level 3

Phrase level components identified characters, locations, references, etc...

Level 4

Normalization using xid elements.

Element Descriptions and Examples

The following table gives an overview about the TaxonX elements:

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

2. The GoldenGATE Editor

The GoldenGATE document editor is intended for the creation of new XML content from plain text data, and for inserting additional markup to document centric XML documents in order to make them more data centric and machine readable.

Approach

The idea of the GoldenGATE document editor 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 latter includes, for instance, the application of natural language processing (NLP) algorithms.

Key Features

The key features of the GoldenGATE document editor are:

- **XML Editing:** Editing XML data in a flexible environment. In favor of lucidity, 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. This editing mode treats words as the atomic text units rather than individual characters. Since XML tags are highly likely to go between words, this facilitates the selection-based creation of markup by far. In addition, users can configure frequently used actions to become one-click accessible.
- **Text Editing:** Editing the content as plain text. Again in favor of lucidity, 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:** Rule based conversion and transformation of markup. As opposed to other XML editors, GoldenGATE does not apply XSLT or similar standards. It uses some very simple conversions which are easily configured and managed, but still are powerful enough to handle common conversions.
- **Native NLP:** Application of gazetteers and regular expressions. The GoldenGATE editor natively provides components that allow handling and applying gazetteers and regular expressions for automatically creating fine-grained markup.
- **Extensibility:** Straightforward integration of additional functionality. The GoldenGATE editor can include several types of plugin components that serve a variety of purposes. The modules for gazetteers and regular expressions, for instance, are handled this way, although they are provided as part of the native functionalities.
- **Operation Sequencing:** Bundeling of frequently editing step sequences. The GoldenGATE editor allows bundeling frequently applied sequences of editing steps, which makes them accessible as one.
- **Batch Mode:** Autonomously processing a set of files. The GoldenGATE editor is capable of autonomously processing documents, e.g. doing some markup transformations to an entire folder of files in an overnight job.

GoldenGATE is written in Java and thus platform independent. Once GoldenGATE is installed in an appropriate directory, start the program using GoldenGATE.bat, and then follow the steps described below.

3. System Requirements

The minimum system requirements are:

- Operational systems: MS Windows 2000, Windows XP, Mac OS X and Linux (not tested yet)
 - For Windows, the download zip file contains the GoldenGATE.bat to start the editor.
 - For Mac OS and Linux, please open a command prompt, go to the GoldenGATE folder, and use the following command line:
`java -jar -Xms128m -Xmx512m GoldenGATE.jar RUN`
- Random access memory: minimum 512 MB
- Java: version 1.4.2.06 or later, better 1.5 (no guarantee for beta releases)
- Internet connection: ISDN is fine for getting the MODS Header, a broadband connection is better for getting the LSIDs.

4. Document Pre-preparations

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 not further elements are needed. It proved also extremely advantageous to have very clean and spell checked OCR-output, since artefacts have an impact on the mark-up process, and later corrections are very tedious. This has especially an effect on the recognition of names.

5. Workflow to generate a valid TaxonX XML document up to Level 1

In case, the file needs to be saved at any point in-between: ->**file-> save Document to file->select "XML file (all tags) -> save.**

Once the file is saved this way, the setting is saved and in the future "save document" can be used. Beware that all elements are lost unless "all tags" or "selected tags" are chosen to save. Make sure, that the full name is saved, including the extension xml. The document can later be opened as described below.

) **) The steps 11, 15 and 20 require an **internet connection. This might take a moment. To see progress, please switch to the DOS window. If there is no connection available at the moment, these steps can be done later.*

1. Start the GoldenGATE Editor with **GoldenGate.bat**
2. Open the document: open document (File → **load document from file** → **HTM and HTML files (default reader)**)
3. Delete DOCTYPE. **Custom function** → **Delete DOCTYPE**.

During the OCR process the ABBYY Fine Reader generates a document header, which is not needed for the mark up process and do not bear any original information of the publication. (<! DOCTYPE HTML PUBLIC " - // W 3 C // DTD HTML 4.0 Transitional // EN " " http: // www. w 3. org / TR / html 4 / loose. dtd ">). The Pipeline Delete DOCTYPE contains the following functions:

- a) DoctypeAnnotator.annotator*
- b) DoctypeDeleter.markupConverter*

4. Run the Paginator-Pipeline. -> **Custom functions** -> **Paginator**

The Paginator-Pipeline includes the following analyzers:

- a) PageBorderFinder.analyzer*
- b) PageMarker.analyzer*
- c) PageNumberer.analyzer*
- d) PagePartNumberer.analyzer*
- e) PageTitleFinder.analyzer*
- f) CaptionFinder.analyzer*
- g) FootnoteFinder.analyzer*

*Make sure that: all pages are numbered accordingly,
the page titles were found,
the captions were marked,
the footnotes were marked.*

If not, mark it manually (select required text part → right mouse click → Annotate → Annotate (here will follow a detailed description with screenshots how to do that and what to fill in the fields)

5. Correct paragraphs in pages manually with “merge annotations”, (for a better overview use: → edit → slide annotations, set Environment Size = 0 (to exclude tokens of the previous and following pages) -> select “page” -> use next page (next/previous) to move from one page to the next. To merge annotations: **highlight parts of the two -> right mouse button -> merge Annotation**, to split annotations (i.e. insert a new paragraph) **highlight part of the first token of the new paragraph->right mouse button-> split Annotation**. At the end finish the slide viewer.

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 yel] ow pubescence, more abundant on gaster.

</paragraph>

6. Control and correct Page Numbers and Page Titles. Page numbers missing can, but need no be introduced if the page numbers show up as attribute of the paragraph annotations (i.e. <paragraph pageNumber="100">). To check for the page number: -> Edit -> Slide Annotations, set environment Size ="20" -> Slide Annotations -> Next. In case of missing page numbers: → Edit → slide annotations, set Environment Size = 20 to allow to view missing page numbers on the previews page) -> select “page”. In case of missing page numbers: search for the missing page number. If not present, find out where the other page numbers are situated, and then add it in the same position on the adequate page. **Highlight the page number -> right mouse click over highlight -> annotate -> pageTitle; highlight the new “pageTitle tags” -> right mouse click over highlight -> annotate -> pageNumber -> close -> click on <pageNumber> annotation -> enter in Attribute Name “pageNumber”, enter in Attribute Value “the actual page number” -> Add Attribute ->Close.**

If the page numbers reported in the <paragraph> tag do not correspond with the respective page numbers, then a page(s) is missing. In this case, the best way is to either insert the missing pages and restart GoldenGATE, or, in case they are missing, insert dummy pages to fill in the gap and restart the formatting.

7. Clean up the pages. **Custom Functions → Cleanup Pages.**

The “Cleanup Pages” is a pipeline which removes the page tags and normalizes the paragraphs and characters. The page element needs to be removed because they will interfere with elements spanning over page borders, such as <treatment> which might run over a couple of pages The pipeline contains the following functions:

- a) the ParagraphNormalizer.Analyzer (Custom Functions / Analyzer Menu)
- b) the CharacterNormalizer.Analyzer (Analyzer Menu)
- c) the PageTagRemover.markupConverter (Edit → Remove Annotations → page)

8. Mark up all taxonomic names: **Custom function -> Run FAT.**

*The Run FAT function marks all taxa found in the text. FAT is based on an ant species dictionary. 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 **Analayzers → Run → FAT.analyzer.***

9. 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 missing part, annotate it and merge it with the previously annotated part, which was found by the FAT.analyzer.

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

At the end, use MarkupConverters->remove duplicate annotations to make sure, that all taxa are only marked once.

*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: Edit → Select Annotation → taxonomicName.** A list with all marked taxa appears, for removing annotation select the Remove button for the erroneously marked expression.*

10. Complete the Taxa. **Custom function → Attribute Taxon Names**

The **Attribute Taxon Names** pipeline contains the following functions:

- a) the taxonomicNameAttributor.analyzer
- b) the taxonomicNameCompleter.analyzer

*Together with the FAT.analyzer these functions are responsible for the correct taxon treatment. Whereas FAT finds all the taxa, the taxonomicNameAttributor.analyzer assigns the found taxa to there respective status, this is genus, subgenus, species, subspecies, race and variety. If not, mark it manually via the annotator: **click on the tag, a new window opens (edit annotations window), attribute name: subspecies, race, variety, etc.; attribute value: taxon** (e.g. flavus, niger)*

*To assure optimal function of the subsequent steps, make sure that first mention of the names at generic level are annotated with the respective attribute “genus=“GENUSNAME””: **Highlight on the right hand the right taxonomicName button->click on the tag of the first occurrence of the genus name-> add on the right hand of the***

edit annotation menu the taxonomic level (i.e. genus for GENUS) into the “attribute name window” and below the respective GENUSNAME, ->“add attribute” ->close

The taxonomicCompleter.analyzer completes the taxa, e. g. if in the text appears only the species name, the analyzer will search for the last named genus before and complete with the genus name the species. The text will not edit, the information is placed in the tags.

If names are completed with wrong names, go to the beginning where the error occurs, remember where it is, ->Edit->undo->”the last step”, then go back to annotate the erroneous taxa with the wrong completion.

11. Get the LSID: Custom Functions → Get LSIDs for Taxa [*requires internet connection*] *) (**)

With the TnuluLSIDReferencer.analyzer GoldenGATE connects to the Hymenoptera Name Server and get a Life Science ID for each mentioned taxon in the document. Depending on the number of the found taxa in the document, this takes a while. For dubious taxa, a window will appear at the end of the process and the user is called upon to choose between some taxa or to inform the LSID manually. For ant LSID research use <http://atbi.biosci.ohio-state.edu/tnulu.html>.

Alternatively the function can be initiated via **Analyzers → Run → TnuluLSIDReferencer.analyzer.**

12. Get the higher taxonomic names: Custom Functions → Higher Order Taxa

At the end of the taxonomic mark-up process the HigherOrderNameTagger.analyzer will search for taxonomic classifications higher than genus (order, family, subfamily and tribe). Alternatively the function can be initiated via **Analyzers → Run → HigherOrderNameTagger.analyzer.**

13. Split the document in treatments Custom Functions → Mark up Treatments.

“Mark up Treatments” represents a pipeline with the following elements:

- a) *TreatmentStructurerResetter.markupConverter*
- b) *TaxonomicDocumentStructurer.analyzer* (Analyzers → Run → ...)

In case of a second run of the “Mark up Treatments”-Pipeline the TreatmentStructurerResetter.markupConverter deletes all formerly set treatment tags to avoid double marking of treatments and an erroneous XML output.

For a correct document splitting you should carefully read the following rules. Wrong selection or rule ignorance can result in invalid TaxonX outputs. This part require special attentions: it is the only part where the editor can make a subjective selection of the treatment parts.

Alternatively this analyzer can be initiated via **Analyzers → Run → TaxonomicDocumentStructurer.analyzer**

A new window opens and you have to decide what kind of content is in each paragraph. Generally the first element is a nomenclature section which contains the name of the treated taxon. In most cases a description elements follows, after that other elements like biology & ecology, discussion, materials examined, etymology are possible.

The assignation of an annotation to a specific text element is not all the time clear-cut and evident. A possible rule to apply is that in cases of mixed content, e.g. notes on biology and collecting event, the element which is more clearly expressed ought to have priority and due to the nature of the publication nomenclature would prevail over the following ranking material examined, description, diagnosis, distribution biology and ecology, and discussion.

The following options are given: **(for definitions and examples see the GoldenGATE Glossary)**

treatment (start)	biology_ecology; description; diagnosis; discussion; distribution; etymology; materials examined; nomenclature; reference group
treatment (continued)	biology_ecology; description; diagnosis; discussion; distribution; etymology; materials examined; nomenclature; reference group
catalogue entry	without further options
abstract	without further options
acknowledgments	without further options
document_head	document author (including address), document title, other/general
introduction	without further options
key	without further options
materials & methods	without further options
multiple	without further options
reference group	reference, other / general
synopsis	without further options

14. Save the document as XML format (all tags!) e.g. 6200_gg0.xml (HNS-ID_GoldenGATELevel0.xml)

15. Get the MODS Header from the Hymenoptera Name Server: **Custom Functions → get MODS Header.** [*requires internet connection*] *) **)

*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 Analyzers → Run → ModsReferencer.analyzer.

16. Save document as TaxonX XML file: e.g. 6200_tx0.xml (the document will be saved in the TaxonX format and all formats of the original document, such as page breaks are also saved.

This file is a TaxonX Level 0 document.

Close the document.

17. Open the formerly GG document with all tags saved. (6200_gg0.xml)

18. Clean up the pages: **Custom Functions → Dissolve Pages.**

The “Dissolve Pages”-pipeline involves the following functions:

- a) the PageCleaner.analyzer (moves footnotes and captions to the end of the publication)
- b) the PageDissolver.analyzer (*removes page number tags, pages and page borders and discards the original layout, so that the treatments are cleaned up and have no more interruptions by page titles or page breaks*)
- c) the ParagraphNormalizer.analyzer

*You have to check if formerly by page breaks interrupted paragraphs are jointed. **If not, mark the last word of the first paragraphs and the first word of the second paragraph, do a right mouse click: merge annotations.***

19. Check the paragraph structure manually.

You have to check if formerly by page breaks interrupted paragraphs are jointed. If not, mark the last word of the first paragraphs and the first word of the second paragraph, do a right mouse click: merge annotations.

20. Save as XML file (all tags) (eg. 6200_gg1.xml)

21. Get the MODS Header from the Hymenoptera Name Server: **Custom Functions → Get MODS Header.** [*requires internet connection*] *) **)

*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 **Analyzers → Run → ModsReferencer.analyzer.***

22. Save as TaxonX XML file (e. g. 6200_tx1.xml)

Close document

This file is a TaxonX Level 1 document.

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

**) These steps might take a moment. To see progress, please switch to the DOS window.

6. Working with GoldenGate: The Main Menus

The main window of the GoldenGATE Editor provides access to the native functions via the main menu bar. It includes at least 17 menus. **Only the menus needed for TaxonX document generation are explained in detail here.** For information on other menus, please consult the help file or read the general instruction manual for GoldenGATE.

6.1 The File Menu

The File menu provides operations for opening, saving, and closing files, and for closing the editor itself.

6.2 The View Menu

The View menu allows editing the displaying style of the documents currently selected, and refreshing the display.

6.3 The Edit Menu

The Edit menu provides functions for editing XML as well as textual content. This includes several operations for the automated generation of structural markup.

Edit Selection

Character wise edit the plain text currently selected in the Annotation Editor. Since the Annotation Editor treats words as atomic units in order to simplify manually creating XML markup, character wise editing has to be done in a separate dialog. This function exists to avoid switching to the Plain Text Editor for small corrections.

Find / Replace

Do find / replace operations in the document currently selected. This will automatically switch to the Plain Text Editor.

Slide Annotation

Allows checking the marked document in Annotation units (XML tags). Using the slide annotation tool, you can choose which annotations you want to see. The parameters are Annotation Type, Sliding Window Size and Environment Size. With the Annotation Type you can choose which XML tags you want so control. The Environment Size allows seeing the selected numbers of lines after the chosen annotation. The main advantage of this function is, that you won't miss any of the chosen annotation.

Select Annotation

This option opens a Select Annotations window with the chosen annotations. In the window one can choose between removing the annotation, deleting the expression or deleting the expression.

Annotate

Create an Annotation (XML tag) enclosing the text passage currently selected. Since the Annotation Editor treats words as atomic units, the selection is automatically extended to match word borders.

Annotate All

Create an Annotation (XML tag) enclosing every occurrence of the text passage currently selected in the current document. Since the Annotation Editor treats words as atomic units, the selection is automatically extended to match word borders. This function is primarily intended for detail level markup (e.g. location or person names): Select the particular name once, annotate all its occurrences in the entire document with a single click.

Merge Annotations

Merge neighbouring Annotations. Only Annotations of the same type can be merged. This function is intended for joining paragraphs that were erroneously split by OCR software, for instance. At least the end tag of the first Annotation to be merged, and at least the start tag of the last one, have to be selected for this function to work.

Split Annotation

Split an Annotation into two. The inmost Annotation enclosing the currently selected text is determined the target of this operation. It is split at the word border before the selected text passage. Exactly one word or punctuation mark has to be selected for this function to work.

Remove Annotations

Remove Annotations from the document. This operation will only remove the tags, not the textual content of the Annotation. To remove the latter as well, use Delete Annotations instead.

Delete Annotations

Delete Annotations, including their textual content. To remove only the XML tags, use Remove Annotations instead.

Normalize Paragraphs

Apply a built-in paragraph normalizer to the current document. The text within paragraph Annotations is "normalized", i.e. divided words are rejoined, and internal line breaks are converted into spaces.

Normalize Whitespace

6.4 Analyzers

Analyzers are document processors that provide the facilities for integrating external markup tools into the GoldenGATE editor, e.g. NLP components.

The Analyzers Menu

The Analyzers menu provides access to functions for managing and applying Analyzers, and instant access to a set of built-in Analyzers:

- **Annotate Sentences:** Apply a built-in sentence splitter Analyzer to the current document. Sentences are marked with Annotations of the type "sentence".
- **Annotate Paragraphs:** Apply a built-in paragraph splitter Analyzer to the current document. Paragraphs are marked with Document Parts of the type "paragraph".
- **Annotate Sections:** Apply a built-in section splitter Analyzer to the current document. Sections are marked with Document Parts of the type "section", removing the section title from the text and making it an attribute of the section tag.
- **Normalize Paragraphs:** Apply a built-in paragraph normalizer Analyzer to the current document. The text within paragraph Annotations is "normalized", i.e. devided words are rejoined, and internal line breaks are converted into spaces.
- **Normalize Whitespace:** Apply a built-in whitespace normalizer Analyzer to the current document. This operation will remove, e.g., duplicate whitespaces and whitespaces before closing brackets and after opening ones. While this sounds sort of trivial, it is in fact pretty necessary to make the behavior of regular expressions predictable while still keeping them somewhat at least managable.

The Edit Analyzers Dialog

The dialog for Analyzer management is accessible through the Edit function in the Analyzers menu. On the right of the dialog, a list displays all existing Analyzers. In the lower middle, a set of input fields displays the parameters of the Analyzer currently selected in that list:

- **Analyzer Class:** The fully qualified name of the Java class representing the Analyzer.
- **Analyzer Class Path:** The path and name of the jar file containing the Analyzer class. Clicking the label will open a file chooser dialog for selecting the jar file.
- **Analyzer Data Path:** A directory marking the spot in the file system where the Analyzer may deposit its data files. Clicking the label will open a file chooser dialog for selecting the directory.

The Configure Analyzer Processor button triggers the Analyzer component to display a configuration dialog for component specific settings. Since not all Analyzer components have such settings, it might happen that clicking the button induces no action. At the top of the dialog, a line of buttons provides the functionality for managing the Analyzers:

- **Create:** Open a sub-dialog for creating a new Analyzer (integrating a new external component)
- **Clone:** Clone the Analyzer currently selected. This is pretty much the same as creating a new one, beside the fact that the parameter values are copied.
- **Search Types:** Search the jar files in the Analyzer directory for new Analyzer Factories (implementations of the `de.gamta.util.AnalyzerFactory` interface).
- **Search Analyzers:** Search the jar files in the Analyzer directory for new Analyzers (implementations of the `de.gamta.util.Analyzer` interface).

- **Delete:** Delete the Analyzer currently selected. This operation does not delete any program or data files, only the data for binding the external component.

6.5 MarkUp Converters

Markup Converters are document processors capable of converting and filtering the XML markup of documents.

- **Remove DuplicateAnnotations:** Remove duplicate Annotations. Two Annotations are considered equal if they have the same type and mark the same part of the document text. The attributes of duplicates will be unioned.
- **Remove Self-ContainingAnnotations:** Remove Annotations containing another Annotation of the same type.

6.6 Pipelines

The Pipelines are document processors that allow sequencing other document processors (including other Pipelines) so they are accessible as one.

The Pipelines Menu

The Pipelines menu provides access to functions for managing and applying Pipelines, and instant access to a set of built-in Pipelines:

- **Create:** Create a new Pipeline by sequencing some document processors.
- **Edit:** Open the Pipeline management dialog.
- **Run:** Select a Pipeline from a list and apply it to the document currently selected. This function is also available through the Run Pipeline option in the Tools menu.

The Edit Pipelines Dialog

The dialog for Pipeline management is accessible through the Edit function in the Pipelines menu. On the right of the dialog, a list displays all existing Pipelines. In the middle, a table displays the individual document processors in the order they are applied by the Pipeline currently selected in that list. Below and to the left of the table, there are several buttons for managing the document processors that are part of the Pipeline:

- **Remove:** Remove the document processor(s) currently selected in the table.
- **"Add ..." Buttons:** For every type of document processor accessible to GoldenGATE, a button exists which opens a dialog for selecting a processor from a list and adding it to the end of the Pipeline. If the new document processor is a Pipeline itself, an automated cycle detection will ensure that no cycles occur.
- **Interactivity Level:** Below the buttons, there is a selector for the Pipeline's interactivity level, i.e. the frequency documents running through the Pipeline are presented to the user for manual interventions. There are five different interactivity levels:
 - *Feedback / Editing after each step:* Allow the document processors to prompt the user whenever necessary, display the document for manual editing after each individual document processor.

- *Feedback / Editing of result*: Allow the document processors to prompt the user whenever necessary, display the document for manual editing after the Pipeline has finished its work. This is the default.
- *Feedback only*: Allow the document processors to prompt the user whenever necessary; do not display the document for manual editing.
- *Editing of result only*: Force the document processors to work autonomously, but display the document for manual editing after the Pipeline has finished its work.
- *No interactivity*: Force the document processors to work autonomously; do not display the document for manual editing.

- **Up**: Move the document processor currently selected in the table up by one row.
- **Down**: Move the document processor currently selected in the table down by one row.

At the top of the dialog, a line of buttons provides the functionality for managing the Pipelines:

- **Create**: Open a sub-dialog for creating a new Pipeline by sequencing some document processors.
- **Clone**: Clone the Pipeline currently selected. This is pretty much the same as creating a new one, beside the fact that document processor sequence is copied.
- **Delete**: Delete the Pipeline currently selected. This operation does not delete the document processors forming the Pipeline.

6.7 The Window Menu

The Window menu allows editing the configuration of the GoldenGATE editor. This includes access to some built-in resources.

Main

Edit the global settings of the GoldenGATE editor. In particular, these are:

Annotation Editor Default Settings: The upper part of the preferences affects the default settings for the Annotation Editor. In particular there are:

- The font style, size, and color for the textual content
- The font style, size, and color for the XML tags

The behavior of the Annotation Editor when new Annotations or Document Parts are created is shown. Both can be highlighted with a colored background and / or surrounded by newly created tags. While the former is convenient for detail level tags (usually Annotations), the latter would disrupt the optical text structure in most cases. For structural markup (usually Document Part), it is just the other way round. Depending on the use case, however, other configurations may be more convenient.

Global Settings: The initial and maximum memory for the Java Virtual Machine the GoldenGATE editor runs in. Since these settings are arguments to the Java Virtual Machine, changes to them will take a restart of the editor to have any effect.

Document Readers

Document Readers affect the behavior of the GoldenGATE editor when loading a document. In marked-up data, they can filter and rename tags, and decide if a tag will be represented as a Document Part or as a plain Annotation.

Document Writers

Document Writers affect the way the GoldenGATE editor writes a document to disc. They allow filtering which Annotations are written to disc, and which are not. This is useful for documents that contain a lot of (maybe generic or intermediate) markup only a part of which are to be written to the file, e.g. the elements of a particular XML namespace.

Document Upload

The Document Upload Plugin manages Server Logins as Resources and allows uploading documents to servers using these logins.

Custom Functions

Frequently used markup / editing sequences can be bundled to Custom Functions, which makes them one-click accessible via custom buttons in the Annotation Editor. A good example for a useful Functions is marking up the selected text as a paragraph and subsequently applying the structural normalization to the content of the newly created tag.

- **Create:** Create a new Macro by entering an Annotation type and selecting a processor.
- **Edit:** Open the Macro editor dialog.

The Edit Macros Dialog

The dialog for Macro management is accessible through the Edit function in the Macros menu. On the right of the dialog, a list displays all existing Macros. In the middle, it displays the parameters of the Macro currently selected in that list:

- **Macro Label:** The label text for the button representing the Macro in the Annotation Editor.
- **Macro Tooltip:** The tool tip text for the button representing the Macro in the Annotation Editor. This text should provide a one-sentence explanation of what the Macro vaguely does.
- **DocPart Type:** The type of Document Part to create by the Macro.
- **Open for Editing:** If checked, the newly created Document Part will be opened for editing after the processor of the Macro has finished. This is particularly useful for immediately correcting error of the processor, which is not too unlikely if the latter applies NLP.
- **Processor:** Above the buttons, a label displays the type and name of the processor currently selected.
- **"Use ..." Buttons:** For every type of document processor accessible to GoldenGATE, a button exists which opens a dialog for selecting the processor from a list displaying all available processors of the particular type.
- **Summary:** Below the buttons, a label provides a brief summary of what the Macro does in its current configuration. Since this summary is rather generic, it is not too appropriate as a tooltip in most cases, although the Macro editor uses it as a fallback if not explicit tooltip is given.

At the top of the dialog, a line of buttons provides the functionality for managing the Macros:

- **Create:** Open a sub-dialog for creating a new Macro by entering an Annotation type and selecting a processor.
- **Clone:** Clone the Macro currently selected. This is pretty much the same as creating a new one, beside the fact that the parameters are copied.
- **Delete:** Delete the Macro currently selected. The processor of the Macro will not be affected by this operation.

6.8 The Help Menu

The Help menu provides access to help content on all aspects and parts of the GoldenGATE editor, including plug-ins. The main help topics are directly accessible.

7. The Document Editor

Each Document Editor holds an open document. They are managed as tabs in the main window of the GoldenGATE editor. A Document Editor has three sub-editors which serve different purposes:

The Annotation Editor

The Annotation Editor is intended for handling the document mark-up. Since almost all mark-up goes between words, it treats words as the atomic text units instead of individual characters. While this inhibits character wise text editing, it greatly simplifies creating XML tags because they are automatically placed at the word boundaries.

The Editor (center)

The editor is the core part of the Annotation Editor. It displays the document text for mark-up editing, while the other parts (see below) provide options to make the editing process more convenient. However, the editor does not allow editing the text directly. All functions are based on selecting a part of the text and subsequently applying some action to it. These actions are accessible in several ways: Through the Edit menu of the main window, through the button panel of the Annotation Editor (see below), or through the context menu. The context menu is the only one of these ways that changes depending on the current selection. In particular, there are three different context menu states:

- **Part of the text selected:** When a part of the text is selected, the context menu provides options for annotating it (**Annotate**, **Annotate All**, **Document Part**), including instant access to the most recently used annotation types. It also provides functions for splitting the surrounding Annotation (**Split Annotation**), or for merging Annotations at least on tag of which is part of the selection (**Merge Annotations**). The last part of the context menu allows editing the words of the text. This comprises the following functions:

Edit Tokens: Character wise edit the plain text currently selected. This function exists to avoid switching to the Text Editor for small corrections.

Join Tokens: Join the Tokens (i.e. words, numbers, punctuation) currently selected. While this results in subsequent words becoming one word, it will still split the text at clear word boundaries, e.g. before or after punctuation marks.

Copy Tokens: Copy the currently selected words to the clipboard.

Cut Tokens: Cut the currently selected words from the text and store them in the clipboard.

Paste Tokens: Replace the currently selected words with the content of the clipboard.

Delete Tokens: Remove the selected words from the document.

- **Nothing selected, clicked on XML tag:** When opening the context menu on an XML tag, it provides functions for opening the Annotation for direct editing (**Edit**), for removing it (**Remove**), and for transforming it to a Document Part (**Transform**). It also allows handling the content of the Annotation:

Copy Tokens: Copy the content of the Annotation to the clipboard.

Cut Tokens: Cut the content of the Annotation from the text and store them in the clipboard. Since the Annotation itself is empty after this operation, it is removed.

Delete Tokens: Remove the content of the Annotation from the document. Since the Annotation itself is empty after this operation, it is removed.

- **Nothing selected:** When opening the context menu in the text with nothing selected, it allows pasting the content of the clipboard at this location (**Paste Tokens**).

The Button Panel (top)

The button panel is another way to access the most common functions of the Annotation Editor. In particular, it provides the basic functions for handling Annotations:

- **Output Preview:** Allows a quick preview in all possible document formats without saving the document first and then opening it in an editor.
- **Edit Fonts:** Edit the displaying font, font size, and color of the document currently selected. These settings are independent for textual content and XML tags. This is in favor of making content easier to distinguish from
- **Edit Selection:** Character wise edit the plain text currently selected in the Annotation Editor.
- **Find / Replace:** Helps to find and replace a chosen word or formerly selected words in the text.
- **Find Previous:** Finds quickly expressions which were find and replaces in a former action by **Find / Replace**.
- **Replace:** Replaces a selected and marked word.
- **Replace All:** When a expression is selected and marked, this function replaced all such expressions in the text.
- **Find Next:** Looks up for a given word. With “enter” the next location of the given word is found.

The Recent Action Panel (upper-left)

The recent action panel allows instantly repeating the most recently used actions, reusing the original parameters. This avoids, for instance, re-entering the Annotation type / XML element name when creating a new Annotation.

The Custom Functions Panel (lower-left)

The makro panel provides access to Makros. A Makro consists of an XML element name to annotate selected text with, and some resource that is automatically applied to the newly created Annotation. Frequently used markup / editing sequences can be bundled this way, making them one-click accessible. A good example for a useful Makro is marking up the selected text as a paragraph and subsequently applying the structural normalization to the content of the newly created tag.

The Layout Panel (right)

The layout panel allows showing / hiding the individual XML element tags contained in the document, and highlighting the tag content. This is in favor of clarity, e.g. for hiding detail level markup when working on the document structure. The left of the two checkboxes highlights the Annotation content, while the right one shows / hides the respective XML tags. A click on the XML element name / Annotation type will open a dialog for adjusting the color used for this type of Annotation.

8. Glossary

This glossary contains some special words and expressions used in the help of the GoldenGATE editor. In a second section it contains also some special expressions used during the TaxonX XML document generation process.

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 have respective meaning)

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 <i>Proceratium</i> (Hymenoptera: Formicidae). <i>Psyche</i> 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	<i>Proceratium avium</i> 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 <i>berlita</i> from all its congeners: scrobe absent, (...)
Discussion	anything which relates to the description, the nomenclatorial history, the behavior or comments relating to the taxon	<i>Discothyrea</i> 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 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.	A general element in modern scientific publications
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 occiput3
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)	<i>Metapone</i> species: <i>M. leae</i>

9. Further Reading

- Sautter, G., K. Böhm, and D. Agosti. 2006. A combining approach to find all taxon names (FAT) in legacy biosystematics literature. *Biodiversity Informatics* 3, 41-53. (<http://jbi.nhm.ku.edu/index.php/jbi/article/view/34/16>)
- Sautter, G., D. Agosti, K. Böhm. 2007 , Semi-Automated XML Markup of Biosystematics Legacy Literature with the GoldenGATE Editor, in Proceedings of PSB 2007, Wailea, HI, USA, 2007" (<http://psb.stanford.edu/psb-online/proceedings/psb07/sautter.pdf>)
- TaxonX wiki (<http://wiki.cs.umb.edu/twiki/bin/view/Ants/WebHome>)

10. Quick Start: Work flow for a valid TaxonX XML document (Level 0 and 1)

1. Start the GoldenGATE Editor with **GoldenGate.bat**
 2. **Open the document:** open document (File → load document from file → HTM and HTML files (default reader))
 3. **Delete DOCTYPE.** (Custom Functions)
 4. Run **Paginator.** (Custom Functions)
 5. Correct paragraphs in pages manually (Slide Annotations,....)
 6. Control and correct manually Page Numbers, Page Titles, Caption and Footnotes.
 7. Run **Cleanup Pages** (Custom Functions)
 8. Run **FAT** (Custom Functions)
 9. Correct manually the taxa (Select Annotations....).
 10. **Attribute Taxon Names.** (Custom functions)
 11. **Get LSIDs for Taxa** (Custom Functions) [*requires internet connection*] *) **)
 12. Run **Higher Order Taxa** (Custom Functions)
 13. **Markup Treatments** (Custom Functions)
 14. **Save** the document as XML format (all tags!) e.g. 6200_gg0.xml
 15. Get **MODS Header** (Custom Functions) [*requires internet connection*] *) **)
 16. **Save** document as TaxonX XML file: e.g. 6200_tx0.xml
- Close the document.

----- TaxonX Level 0

17. Open the formerly GG document with all tags saved. (6200_gg0.xml)
 18. **Dissolve Pages** (Custom Functions)
 19. Check the paragraph structure manually.
 20. **Save** as XML file (all tags) (eg. 6200_gg1.xml)
 21. Get **MODS Header** (Custom Functions) [*requires internet connection*] *) **)
 22. **Save** as TaxonX XML file (e. g. 6200_tx1.xml)
- Close document

-----TaxonX Level 1

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

***) These steps might take a moment. To see progress, please switch to the DOS window.

11. Content

1.	General introduction to TaxonX.....	2
	Levels of Markup	3
	Element Descriptions and Examples.....	3
2.	The GoldenGATEEditor.....	5
	Approach.....	5
	Key Features.....	5
3.	System Requirements.....	6
4.	Document Pre-preparations.....	6
5.	Workflow to generate a valid TaxonX XML document up to Level 1.....	7
6.	Working with GoldenGate: The Main Menus.....	13
6.1	The File Menu.....	13
6.2	The View Menu	13
6.3	The Edit Menu	13
	Edit Selection	13
	Find / Replace	13
	Slide Annotation.....	13
	Select Annotation	13
	Annotate.....	14
	Annotate All	14
	Merge Annotations.....	14
	Split Annotation	14
	Remove Annotations.....	14
	Delete Annotations.....	14
	Normalize Paragraphs	14
	Normalize Whitespace	14
6.4	Analyzers.....	14
	The Analyzers Menu	15
	The Edit Analyzers Dialog	15
6.5	MarkUp Converters	16
6.6	Pipelines	16
	The Pipelines Menu.....	16
	The Edit Pipelines Dialog.....	16
6.7	The Window Menu	17
	The Edit Macros Dialog.....	18
6.8	The Help Menu	19
7.	The Document Editor.....	19
	The Annotation Editor	19
	The Editor (center).....	19
	The Button Panel (top).....	20
	The Recent Action Panel (upper-left).....	20
	The Custom Functions Panel (lower-left).....	21
	The Layout Panel (right).....	21
8.	Glossary.....	21
9.	Further Reading.....	24
10.	Quick Start: Work flow for a valid TaxonX XML document (Level 0 and 1)	25
11.	Content	26