Deliverable 3.3 (D3.3)

Updated release and report on publication data-mining software

M48

Project acronym:	EU BON
Project name:	EU BON: Building the European Biodiversity Observation Network
Call:	ENV.2012.6.2-2
Grant agreement:	308454
Project duration:	01/12/2012 – 31/05/2017 (54 months)
Co-ordinator:	MfN, Museum für Naturkunde - Leibniz Institute for Research on Evolution and Biodiversity, Germany

Delivery date from Annex I:	M48 (November 2016)
Actual delivery date:	M48 (November 2016)
Lead beneficiary:	Plazi
Authors:	Donat Agosti (Plazi), Guido Sautter (Plazi), Terry Catapano (Plazi), Puneet Kishor (Plazi) Lyubomir Penev (Pensoft, Pensoft Publishers Ltd, Bulgaria)

This project is supported by funding from the specific programme 'Cooperation', theme 'Environment (including Climate Change)' under the 7th Research Framework Programme of the European Union

Dissemination Level

PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

This project has received funding from the European Union's Seventh Programme for research, technological development and demonstration under grant agreement No 308454.

1 Executive Summary

1.1 Introduction

The goal of this project is to develop tools to mine published legacy biodiversity literature as well future publications to extract structured information from them. This will speed up the acquisition of records for poorly documented geographic regions and taxa, and make them open for further use in Task 2.5 and WP 4, 5 and 8. Here we describe our software called **GoldenGate Imagine (GGI)**, its use and data production based on it.

1.2 Progress towards objectives

The objective of this task is to create a tool to extract observation records and related data for EU BON based on the published scientific literature. These new data from publications describing an estimated average of 17,000 new species per year as well as a multiple of re-descriptions form a complementary source of data for EU BON. The tool will also allow data mining of hundreds of millions of pages of legacy literature, making them accessible to EU BON partners through the Global Biodiversity Information Facility (GBIF), directly from TreatmentBank, or via the LOD-based Open Biodiversity Management System (OBKMS).

GGI is aimed at taxon-based articles, both recent 'born-digital' as well as the scanned legacy ones. These articles cover recently discovered taxa as well as taxa for which new knowledge is published that may be relevant to EU BON. Target data for GGI are scientific names, treatments and their substructures — for example descriptive, biology, or ecology sections — observation records, and bibliographic references.

The focus of GGI is on PDF-formatted documents. The PDF format is required by the codes of biological nomenclature for archiving all articles including taxonomic works with nomenclatural acts. It is thus the most desirable target format for this tool.

We estimate an average of ~17,000 new species are described per year based on the Index of Organism names (ION) (http://www.organismnames.com/metrics.htm?page=tsj). The journal Zootaxa is the biggest contributor with 12.6% of the new species even though most of the Zootaxa articles are closed-access. The Plazi workflow (Agosti & Egloff, 2009; Egloff 2016) extended with the new version of GGI allows extracting data legally from closed-access publications and making them accessible. For this we focused on the following activities:

Developing the data extraction software GGI. Creation of a stable version of GGI allowing fully automatic extraction of scientific names, treatments and their substructures (such as the descriptive, biology, and ecology sections), observation records, and bibliographic references and figures.

Online editing tool to annotate observation records. Creation of an online editing tool allowing the annotation of observation records and their core elements needed by the users.

Data transfer to GBIF, the EU BON taxonomic backbone, the Biodiversity Literature Repository and RefBank. Creation of an automated pipeline continuously transferring the extracted data from TreatmentBank, the repository of taxonomic treatments extracted with GoldenGate Imagine, to GBIF, the EU BON taxonomic backbone UTIS, and the Biodiversity Literature Repository and RefBank.

Creation of an automated data production workflow. Creation of an automated workflow from scraping the web for new articles, their processing, and transfer to target users.

Testing and production of data. Implementing the automated data production workflow for Zootaxa and extracting data on a daily base, and transferring the data to target users such as GBIF and UTIS. Including a selection of journals to explore the usability of GGI for additional journals.

1.3 Achievements and current status

The development of GGI, the creation of an automated workflow for data extraction and transfer, and its implementation have been achieved. The online editing tool has been developed and released as a Beta version.

GGI source code is available at https://github.com/gsautter/goldengate-imagine under a BSD derivative license.

The software has been written in Java, and its dependencies are:

- servlet-api.jar (by Sun/Oracle)
- idaho-core
- idaho-extensions
- idaho-imagemarkup
- goldengate-editor

GGI implements the Darwin Core Archive (DwC-A) standard for its data transfer protocol. DwC-A makes use of the Darwin Core terms to produce a single, self-contained dataset for species occurrence or checklist data. The format is defined in the Darwin Core Text Guidelines (http://rs.tdwg.org/dwc/terms/guides/text/index.htm).

One template for Zootaxa and two in preparation for fully automated processing of articles for specific journals have been created as of November 28, 2016. For Zootaxa with 17,996 articles published, in a first round 71% of the articles have been processed. With the current adjustments, 14,291 articles resulting in 110,748 taxonomic treatments of which 9,536 are treatments of new species, 38,000 observation records, 120,000 images and more than 200,000 bibliographic references.

A working draft of the documentation for GGI has been created, and is being improved. We have plans to convert this document to a web-based documentation so the latest version can be easily accessed by its users.

A draft manual for GGI (http://plazi.org/resources/treatmentbank/goldengate-editor/) is already online and accessible.

The fully automated workflow is running since June 2016 extracting data from Zootaxa averaging in daily nine articles processed, extracting 93 taxonomic treatments, and 187 observation records.

Additionally, 37 journals are being processed.

Together with our automated workflow for Taxpub XML based publications, GBIF obtained 38,734 observation records and 32,543 taxonomic names, making TreatmentBank one of the main suppliers to the taxonomic backbones of GBIF and UTIS.

More than 120,000 figures are being exported to the Biodiversity Literature Repository (http://biolitrepo.org), a DOI is minted and linked to the treatments.

The following export formats for the data have been developed: Darwin Core Archive (DwC-A) for articles, Resource Description Framework (RDF) and eXtensible Markup Language (XML) for treatments.

In task 3.4 we aim to prepare tools to prepare, extract and mine published biodiversity literature to make taxon based observation data accessible to complement other data sources in EU BON.

1.4 Future developments

Future work will focus on:

- Eliminating any remaining bugs and making the software even more stable;
- Making use of existing analysers in GGI and developing online tools for specific tasks such as parsing extracted strings of observation records, bibliographic data or tables;

- Development of improved web-based documentation;
- Continued development of a user friendly online editing tool and promoting its use;
- Expansion of coverage by adding more journals from which data are extracted.

Table of content

2 GoldenGate Imagine (GGI)

2.1 Introduction

The goal of the deliverable is to develop tools to prepare, extract and mine published legacy and prospective biodiversity literature to speed up the acquisition of records for poorly documented geographic regions and taxa, and to make it open for further use in Task 2.5 and WP 4, 5 and 8.

In task 3.4 we aim to prepare tools to prepare, extract and mine published biodiversity literature to make taxon based observation data accessible to complement other data sources in EU BON. The open source semi-automatic text markup and data extraction tool - GGI - is aimed at taxon based articles, both recent born-digital and scanned legacy ones, with the current focus on the former. These articles cover recently discovered as well as taxa for which new knowledge is published that can be relevant for EU BON. Target data are scientific names, treatments and its substructures (e.g. descriptive, biology, ecology sections), observation records, and bibliographic references. The data can be stored locally, and it can be exported to Plazi's treatment server TreatmentBank, from where GBIF harvest Darwin Core Archives for further EU BON processing in WP4.

This task has the intention to unlock the huge pool of data and information on biodiversity published in the taxonomic and ecological literature, including each year some 20,000 new species and a multiple of that in re-descriptions. Though this literature is increasingly published using semantic markup (e.g., Plazi's TaxPub extension of the US National Library of Medicine/National Center for Biotechnology Information Document Type Definition (DTD) now known as the Journal Article Tagset (JATS) as used by Pensoft), allowing machines to harvest the included data automatically (e.g. import by Plazi into its Treatment Server), by far most of the hundred millions of printed (analog and digital) pages are unstructured and made for human consumption.

The targeted "data" is primarily observation records of a given taxon, secondarily traits; specifically, observation records and traits of a particular taxon published in a specific article, all of which can be cited. Taxonomic and faunistic/floristic literature is the main carrier of this data and thus the target corpus. This literature includes treatments, sections of a text that explicitly cover one particular taxon (Catapano, 2010), often including observation data. By assigning a persistent identifier for each publication and treatment, relations among them can be modeled, and the sources cited.

This data including target groups for WP5 in EU BON will then be open for further use in Task 2.5 and WP 4, 5 and 8, but its origin will always be explicit (**Fig. 1**).

The current focus of GGI is on born-digital PDFs, though it can process scanned legacy literature as well. A large part of the targeted recent literature is published in PDF format. These PDF publications are - from a nomenclatural point of view - considered the canonical version of the articles, not those in other formats such as XML or HTML. The intention is to provide a tool that allows harvesting of articles immediately after publication in order to make available the most recent, relevant data.

Regarding legacy publications, which are generally available as digital images only (e.g., from Biodiversity Heritage Library and an increasing number of research projects), OCR errors still pose a non-negligible challenge, especially regarding efficient detection and user-assisted correction. A prototypical clustering based approach exists for desktop use, but deploying such approaches to the cloud and crowd requires further development.

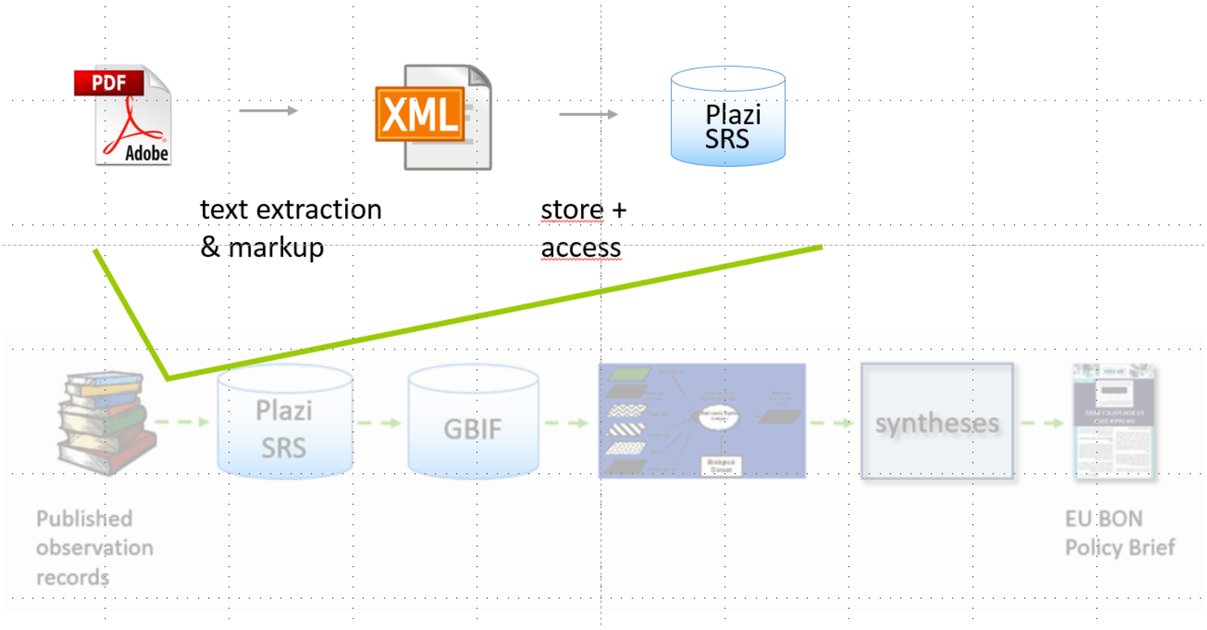


Figure 1. Schematic overview of the GGI data extraction software within the EU BON project

Building a specific tool for reading a born-digital article and converting it into XML became apparent since no open source tool exists that would fulfill the specification we had for the task. This includes decoding of embedded fonts, extraction of text, images, tables, formatted text, sequence of the text, treatments, parsed bibliographic references and parsed observation records. Thus the goal became to develop a tool that allows conversion of unstructured data in a two-level process into semantically enhanced documents: first, a conversion into cleaned up, ready to mine text, and then, through a series of automated and semi-automated processes, into an XML document in which the classes of data described above are identified and delineated.

A further requirement is that the markup should be able to be performed by untrained third parties such as members of the EU BON team. These users would likely benefit greatly from a graphical interface to visualize marked up elements, both for processing as well as quality control and, where required, correction.

After semantic enhancement, the converted PDF documents need to be stored, both for further enhancement or later corrections, and for data processing, linkup, integration and export to third parties. The former requires a document storage format that represents all aspects of a document, so it may be re-opened for editing and enhancement to proceed at a later time. This requires a versatile data format compatible with the tools and services the extracted data is intended for: Darwin Core Archives for uploading occurrence data to GBIF as part of the EU BON workflow, XML for making taxonomic treatments accessible via the Plazi's treatment server TreatmentBank, PNG for further processing of extracted figures and images, and CSV for importing extracted tables into analysis and data mining tools.

Copyright issues of the workflow have been addressed by Agosti & Egloff (2009) for the entire workflow, data linked to names (Patterson et al., 2015) and images (Egloff et al., 2016) with the conclusion that the mining and providing access to the data is legal.

To lower the bar for data extraction from PDF documents, both born-digital and scanned, as far as possible while maintaining high extraction accuracy, we have designed and built an application called *GoldenGATE Imagine* (GGI) (https://github.com/plazi/GoldenGATE-Imagine/releases). GGI opens PDF documents, extracting or rendering page images, performing OCR or decoding embedded fonts where required, and finally segmenting pages into columns, blocks, paragraphs, and lines. Thereafter, it offers a wide range of semi-automated tools and manual markup and editing functionality.

2.2 Software

2.2.1 Client based

Name: GoldenGATE-imagine (GGI)

Programming Language: Java

Application Overview: GGI is made up of several parts all working together (**Fig. 2)**. There is the *GgImagine.jar*, the GGI core, wrapped around a GoldenGATE Editor core, including plug-in and resource management infrastructure, convenience and helper classes for plug-in implementation; GGI user interface; Servlet backing browser based HTML + CSS + JavaScript user interface to GGI, supporting a variation of authentication and document storage mechanisms. The *GoldenGateImagineStarter.jar* holds the startup and live update facilities for GGI application.

There is also the *GoldenGateImagineWeb.zip* made up of *GgImagine.jar* (see above) plus the web application specific configuration and resource files.

The *GoldenGateImagineBatch.jar* is an add-on for GGI application, enabling console based fully automated batch runs plus its configuration and resource files.

Finally, the *VersionPacker.imagine.jar* is a utility for zipping up GGI together with configurations.

Software Architecture: GoldenGATE Imagine is a pure Java application for converting PDF and other image and page-based documents into semantically enhanced XML documents. Holding only PDF decoding and general Image Markup File (IMF, **Fig. 3**) document input-out (IO) functionality in its core, along with document display facilities, GGI draws all editing functionality from plugins to maximize flexibility. The PDF decoding facilities also integrate two non-Java tools (via Java's internal command line): ImageMagick for converting the wide variety of bitmap image formats found in PDF documents into the open and widely supported PNG format, and the Tesseract OCR engine for text extraction from scans.

Plug-ins come in bundles called *Configurations*, which may be designed for a specific type of input documents such as taxonomic journal articles, and for a specific task such as the extraction of bibliographic metadata and references, citations, taxonomic names and treatments, and observation data. The user selects such a configuration on startup, and the application core then loads the plugins contained in it and seamlessly integrates them in the user interface.

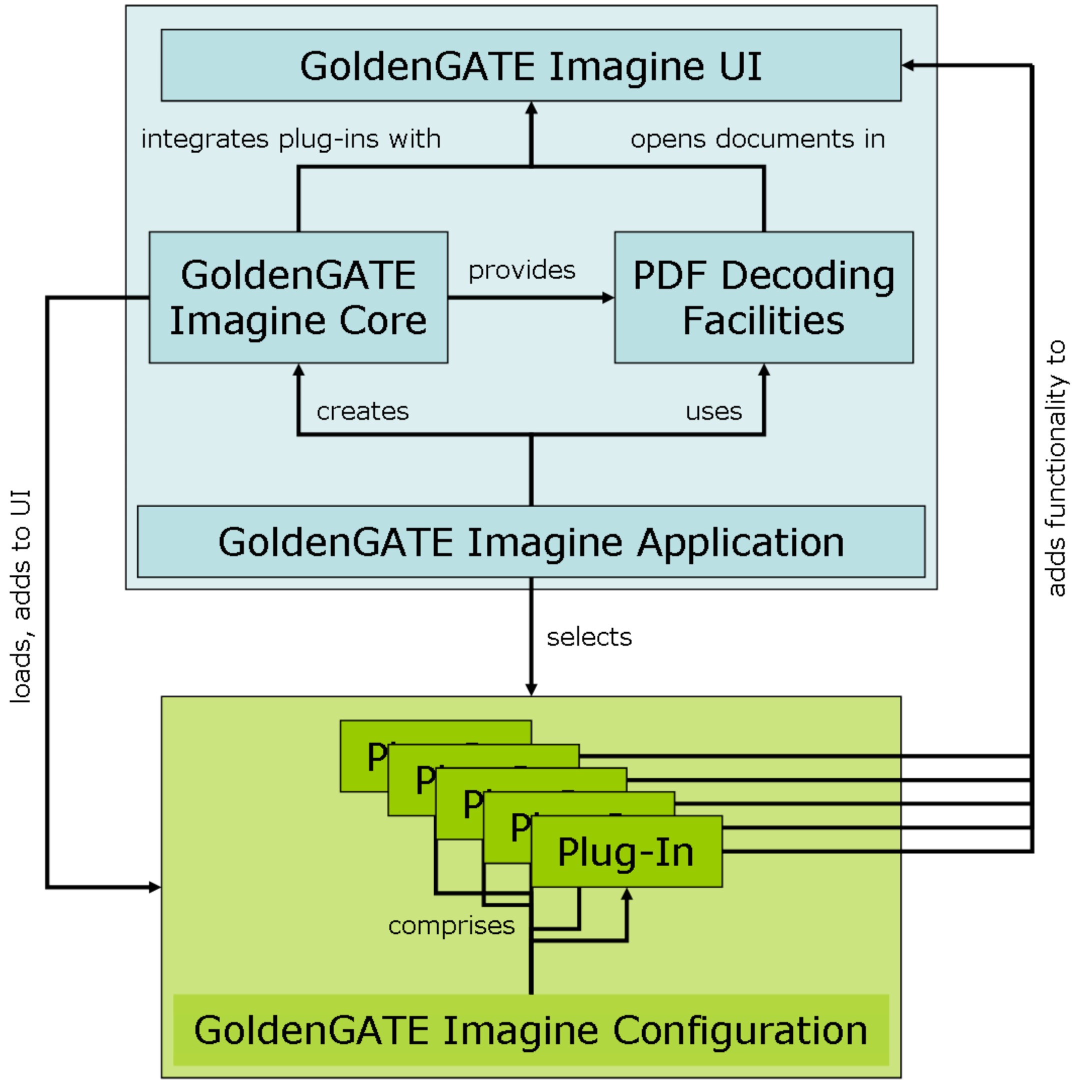


Figure 2: Schematic Architecture of GoldenGATE Imagine. Individual plug-ins can provide a wide range of functionality, for instance:

- Extraction of bibliographic metadata and a respective editing dialog
- Facilities for interactive editing of annotations and document structure
- Integration with a remote document storage facility like TreatmentBank
- Specialized document views like XML or listings of specific markup elements
- Interactive correction of font decoding and OCR errors
- Automated document structure analysis, identifying page headers and footnotes as well as figures, tables, and their associated captions, and in-text citations of figures and tables
- Integration of external Java components that add further task specific functionality such a recognition and parsing of bibliographic references, taxonomic names, occurrence data, or any third-party Java-based natural language processing (NLP) components
- Export of specific document contents such as figures and tables, or of annotated semantic details, to a Darwin Core Archive (see below) for use in data mining applications
- Coupling captions to their in-text citations so they may immediately and automatically reflect a change to the caption in the respective citations as well

Plugins can be built and integrated by anyone, providing many other functionalities, depending on the documents to process and the task at hand.

For managing configurations, and for configuring basic data extraction resources such as regular expression patterns and gazetteer lists, GGI relies upon its older, XML centered sibling GoldenGATE Document Editor.

License: BSD derivative

URL: https://github.com/plazi/GoldenGATE-Imagine

Last commit: Nov 29, 2016

Dependencies:

- servlet-api.jar (by Sun/Oracle)
- idaho-core
- idaho-extensions
- idaho-imagemarkup
- goldengate-editor

Plug-ins:

- goldengate-plugins
- goldengate-imagine-plugins
- idaho-analyzers

Manual:

- A draft User Manual https://github.com/plazi/GoldenGATE-Imagine/blob/master/GoldenGATE_Imagine_V1_end_user_manual.pdf
- Issue tracker https://github.com/plazi/GoldenGATE-Imagine/issues

Source code:

- Build of GGI, with all the afore mentioned functionality https://github.com/plazi/GoldenGATE-Imagine/releases
- The underlying source code https://github.com/plazi/GoldenGATE-Imagine
- Source code for the dependencies listed above is available from their respective repositories.
- The individual taggers and parsers are also available:
    - as web services at http://tb.plazi.org/GgWS/ws
    - API documentation at http://tb.plazi.org/GgWS/API.txt
    - test form at http://tb.plazi.org/GgWS/ws/test

Customization for Specific Document Types: Tasks like the automated extraction of bibliographic metadata from the head section of a scientific article are hard to solve in a generic fashion that works with every given article, and respective solutions tend to be to inaccurate to produce acceptable data quality without manual correction. On the other hand, such tasks are fairly easy to solve for individual layouts, which put all the data elements in specific spots and use specific font styles and sizes for their rendering. To customize metadata extraction, as well other layout specific tasks like detection of headings and captions, which are all governed by journals' layout guidelines, GGI can make use of respective Document Style Templates. These templates are open-ended lists of parameters that describe journal style specific aspects of document layout, for instance:

- the on-page positions and font styles of metadata elements in article headers
- the alignment and font styles of headings at various hierarchy levels
- the (minimum) amount of whitespace between text blocks
- the positioning of captions relative to the figures and tables they describe

- the font size and lead words of captions
- the font size and on-page position of footnotes

The values of document style parameters are bounding boxes (describing an area on a document page), strings, regular expression patterns, numbers, and boolean values.

Batch Processing: Document Style Template (DST) based customization of GGI facilitates increasing the accuracy of fully automated PDF conversion and data extraction to a level that allows fully automated runs over hundreds or thousands of articles from a specific journal and use of the extracted data without any human intervention. GGI Batch is a command line tool that allows to run an instance of GGI without the user-interface on a whole folder of PDF documents from a specific journal in fully automated mode and without any user interaction.

2.2.2 Server based online editing

To foster community involvement in data extraction and quality control, TreatmentBank provides some basic editing functionality right in the website. Just like the Wikipedia, users can perform edits without prior login, registration, or creation of an account, to lower the bar for active participation as far as possible. The website does ask for an optional user *screen name* for attribution purpose, but users do not have to provide one if they prefer to remain anonymous.

In particular, website users may edit the metadata of the article from which a treatment was extracted, mark and edit taxonomic names and occurrence data (**Figs. 17, 18**), and link occurrence data to digitized specimens. The edit history of a treatment is available via a respective menu item, and edits may be undone until they are written back to the main document (whole articles) that contains the edited treatment. Additionally, users may leave comments on a treatment, as well as reply to other comments.

2.2.3 Document Storage

Converted PDF documents are stored as Image Markup Files (IMF). Akin to, and inspired by Darwin Core Archives, an Image Markup File (IMF) is a ZIP archive with specific contents as follows (also see **Fig. 3**):

- a CSV table (“document.csv”) containing document attributes, e.g. the document ID and bibliographic metadata
- a PNG image of every page in the PDF (rendered for born-digital PDF documents, enhanced scans for scanned ones)
- a CSV table (“pageImages.csv”) listing the page images, together with attributes like their resolution, which is particularly important for scanned PDF documents
- a CSV table (“pages.csv”) listing attributes of pages, e.g. page numbers
- a CSV table (“words.csv”) listing individual words, each anchored to a specific position in a specific page; words are chained together in reading order; they bear a Unicode string, as well as attributes like font size and style
- a CSV table (“regions.csv”) listing rectangular document regions like columns, blocks, and lines, as well as tables and table columns, rows, and cells; like words, they are anchored to a specific position in a specific page
- a CSV table (“annotations.csv”) listing annotations that mark the semantics of (sequences of) words, e.g. location or person names, taxon names, materials citations, etc.; they are anchored to their first and last words
- a CSV table (“fonts.csv”) listing embedded font characters and their transcription to Unicode; only exists if the original source PDF was born-digital, as there are no embedded fonts in scanned PDF documents

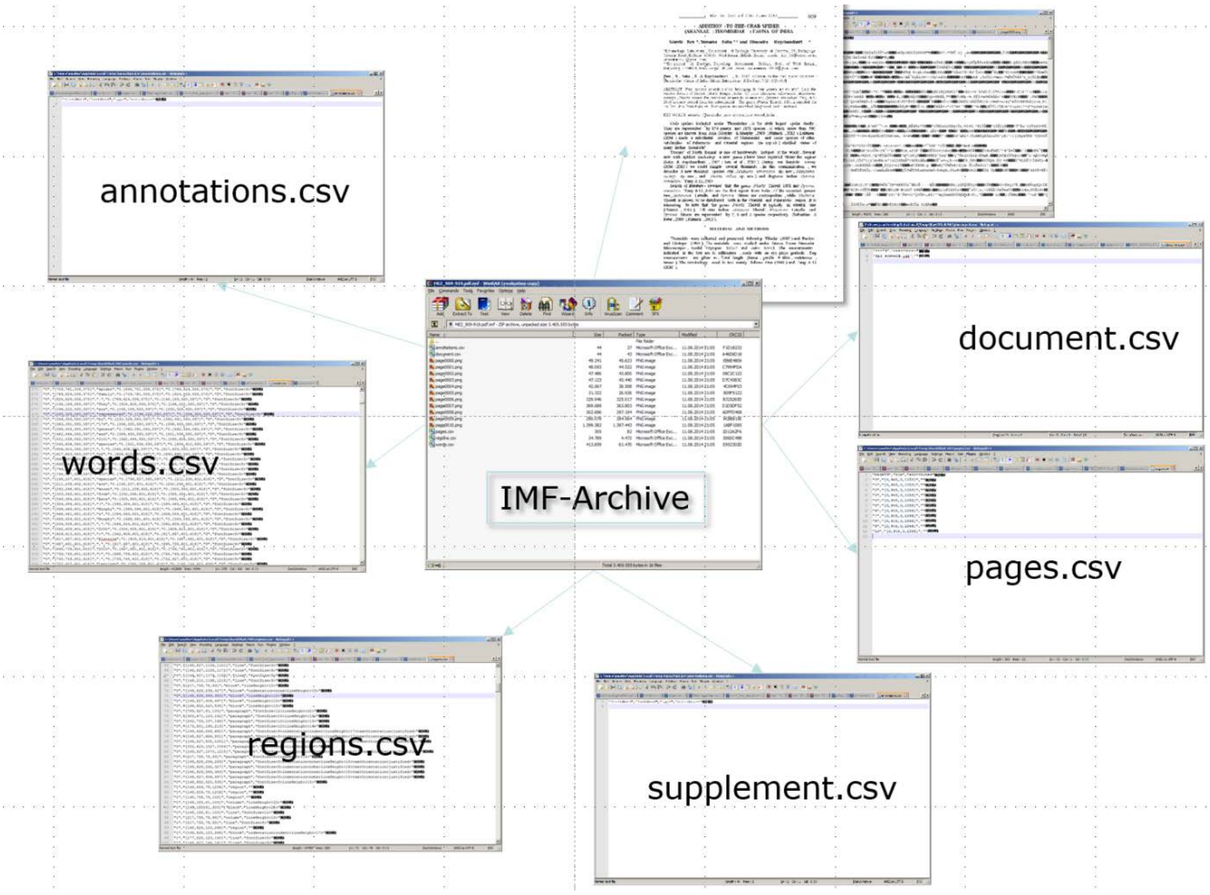


Figure 3: Image Markup File (IMF) structure.

To make IMF extensible, they can further include supplements with unique names and bearing further attributes; the supplement bodies may contain arbitrary binary data. For instance, figures extracted from born-digital PDF documents are stored as supplements, with their resolution and their in-document position as attributes and a PNG of the image as the binary body.

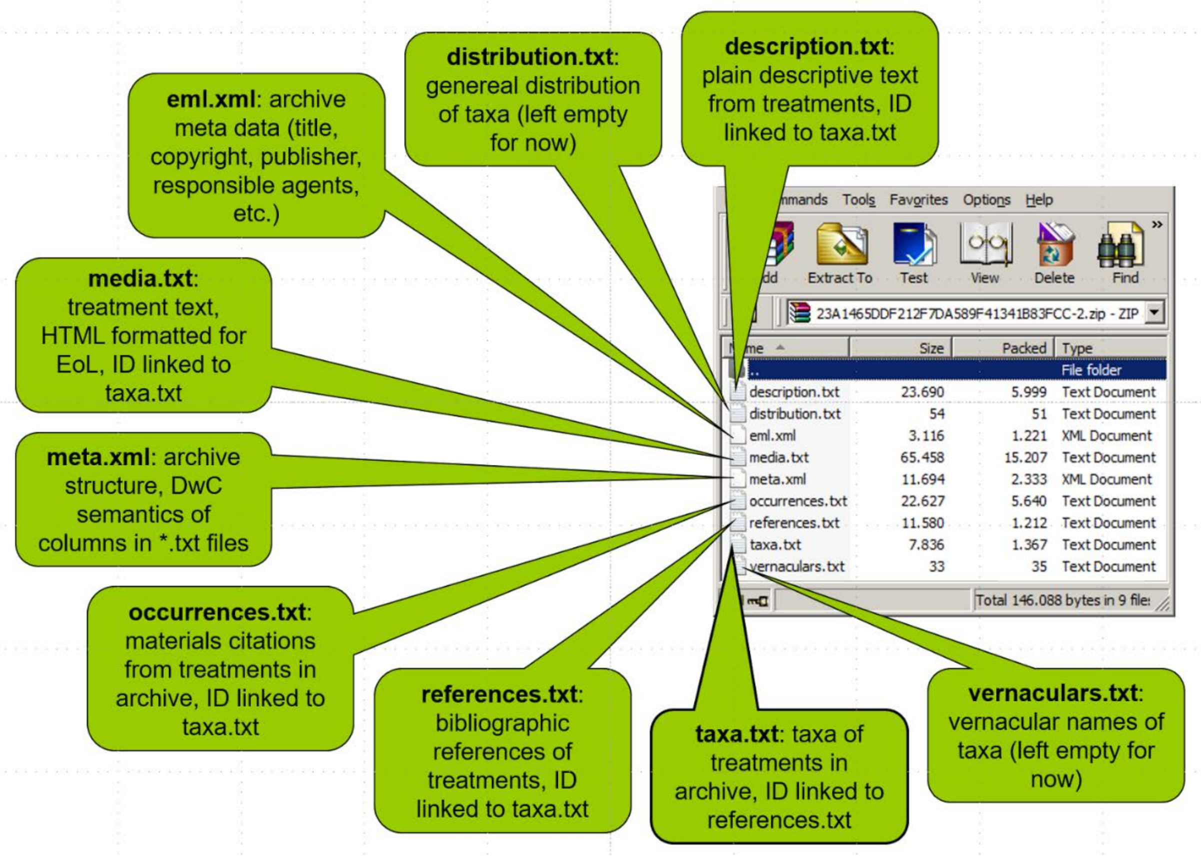
To keep track of the existing supplements and their attributes, they are listed in a dedicated CSV table named “supplements.csv”.

2.2.4 Document & Data Export

For use in other applications, converted PDF documents can also be exported in a variety of other data formats. As opposed to IMF, these formats do not store the document completely, but cover only the aspects required for their specific applications. In particular, GGI offers the following export data formats:

- XML: exports the document text, annotations, and regions into an XML document; there are several different options for how to fold sequences of chained-together words into one another, ranging from raw layout order, akin to the original PDF document, to strict reading order, which concatenates chains of words one after the other; this is to provide the best suited folding for any given application.
- XML to Treatment Server: exports the document text, annotations, and tables as an XML document to Plazi’s Treatment Server; the latter then goes on to export a Darwin Core Archive (Fig 13) with treatment and occurrence data to GBIF and Encyclopedia of Life, and bibliographic metadata and references to RefBank.

- Figures & Tables: exports all figures and tables to a ZIP archive, together with the respective captions; tables can be exported as CSV or as tab separated text
- Darwin Core Archive (**Figs. 4, 5**): exports the taxonomic treatments marked in the document to a Darwin Core Archive, along with the detail data marked inside them; among other things, the latter includes materials citations. The Darwin Core vocabulary is used for materials citation (occurrences).


mmands
Tools
Favorites
Options
Help
dd
Extract To
Test
View
Delete
Find
»
23A1465DDF212F7DA589F41341B83FCC-2.zip - ZIP
Total 146.088 bytes in 9 file:

Name	Size	Packed	Type
..			File folder
description.txt	23.690	5.999	Text Document
distribution.txt	54	51	Text Document
eml.xml	3.116	1.221	XML Document
media.txt	65.458	15.207	Text Document
meta.xml	11.694	2.333	XML Document
occurrences.txt	22.627	5.640	Text Document
references.txt	11.580	1.212	Text Document
taxa.txt	7.836	1.367	Text Document
vernaculars.txt	33	35	Text Document

eml.xml: archive meta data (title, copyright, publisher, responsible agents, etc.)

distribution.txt: genereal distribution of taxa (left empty for now)

description.txt: plain descriptive text from treatments, ID linked to taxa.txt

media.txt: treatment text, HTML formatted for EoL, ID linked to taxa.txt

meta.xml: archive structure, DwC semantics of columns in *.txt files

occurrences.txt: materials citations from treatments in archive, ID linked to taxa.txt

references.txt: bibliographic references of treatments, ID linked to taxa.txt

taxa.txt: taxa of treatments in archive, ID linked to references.txt

vernaculars.txt: vernacular names of taxa (left empty for now)

Figure 4: Darwin Core Archive structure.

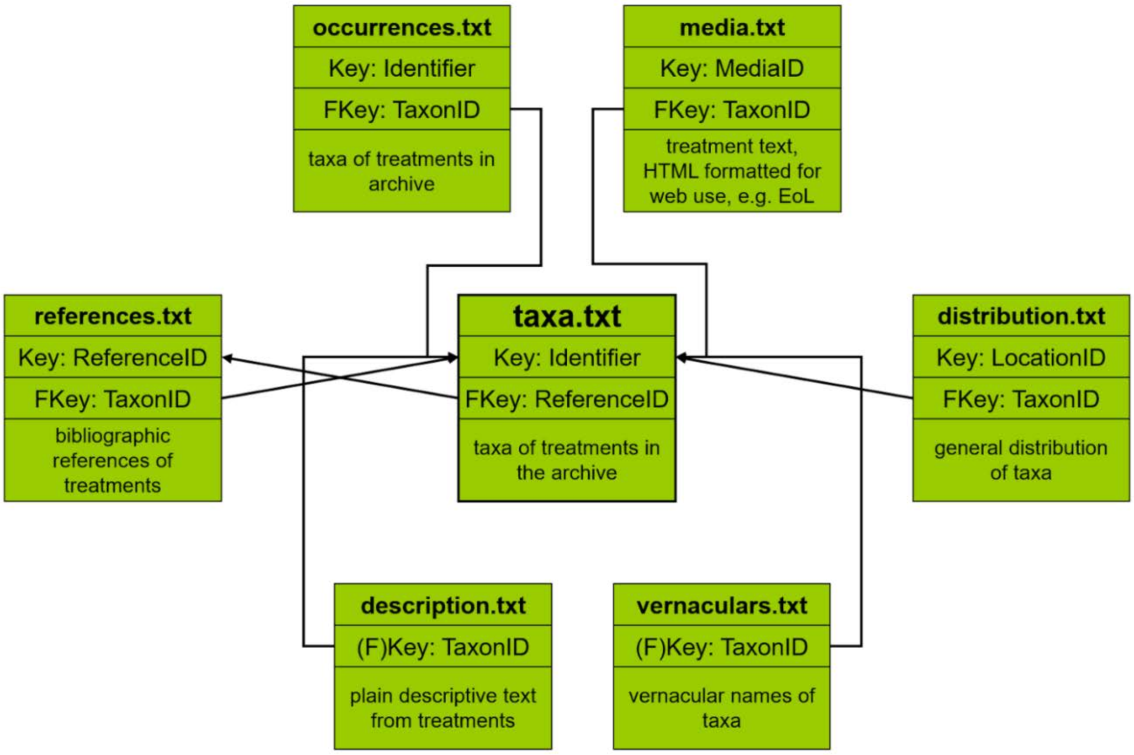


Figure 5: The GBIF checklist format used in Darwin Core Archive

2.2.5 Ongoing development

Such automated tools always evolve and thus are subject to continuous development. This is because

- such tools inevitably use fuzzy logic and statistical inference routines, for which there is always some room for improvement, refinement, and additional training and adaptivity
- the number of journals, and thus the number of different layouts and all the other peculiarities that might exist, is just too large to have been modelled a priori
- advances in PDF obfuscation techniques constantly present new challenges for information extraction software
- new third-party ontologies, catalogs, and services might come into existence, offering new opportunities for further reduction of error and thus the users' correction efforts.

Finally, as use of GGI expands, greater numbers of new users will inevitably come up with ideas and functionality requests beyond what a relatively small number of alpha users could conceive, and such requests will be considered in our future planning and developments. Users are encouraged to submit bug reports and feature requests to the project's issue tracker at https://github.com/plazi/GoldenGATE-Imagine/issues.

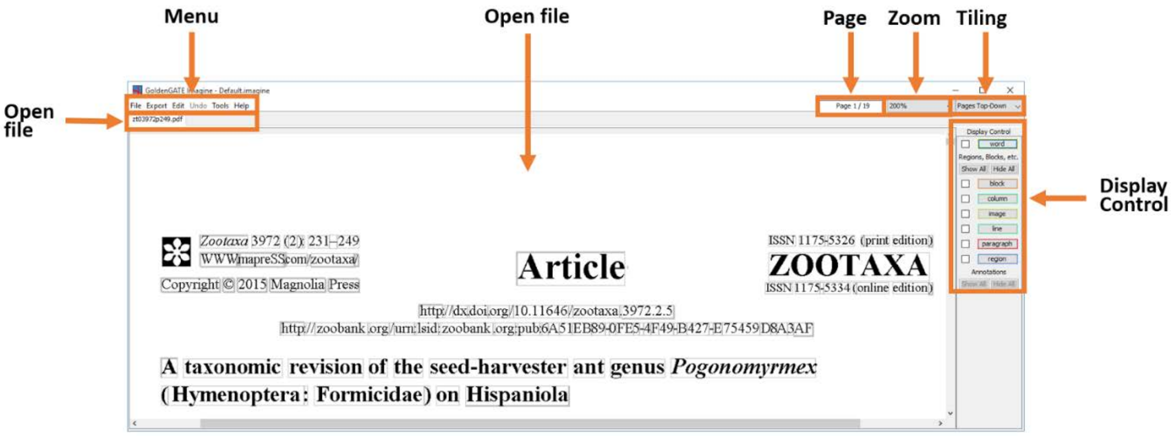
2.3 Markup process

2.3.1 Client based

The GGI enabled markup process includes the following tools:

- automated document structuring, comprising elimination of page headers, extraction of figures and tables, together with their respective captions, detection of footnotes, and detection of headings and their hierarchy (**Fig. 6**),

- semi-automatic document metadata extraction (**Fig. 7**),
- detection and parsing of bibliographic references (**Fig. 8**), using RefParse (Sautter & Böhm, 2014),
- detection, atomization, and reconciliation of taxonomic names, backed by Catalogue of Life, GBIF, and IPNI (**Fig. 9, 10**),
- markup of generic structure (**Fig. 11**), taxonomic treatments (**Fig. 12**) and their inner structure (**Fig. 13**),
- extraction (**Fig. 14**) and parsing of materials citations (**Fig. 15**), including the extraction and normalization of dates and geographic coordinates and normalization of country names using ISO-encoding, as well as country names, and
- tagging of trait terms, backed by ontologies (**Fig. 16**).



Zootaxa 3972 (2): 231–249
WWWmapreSScom/zootaxa/
Copyright © 2015 Magnolia Press

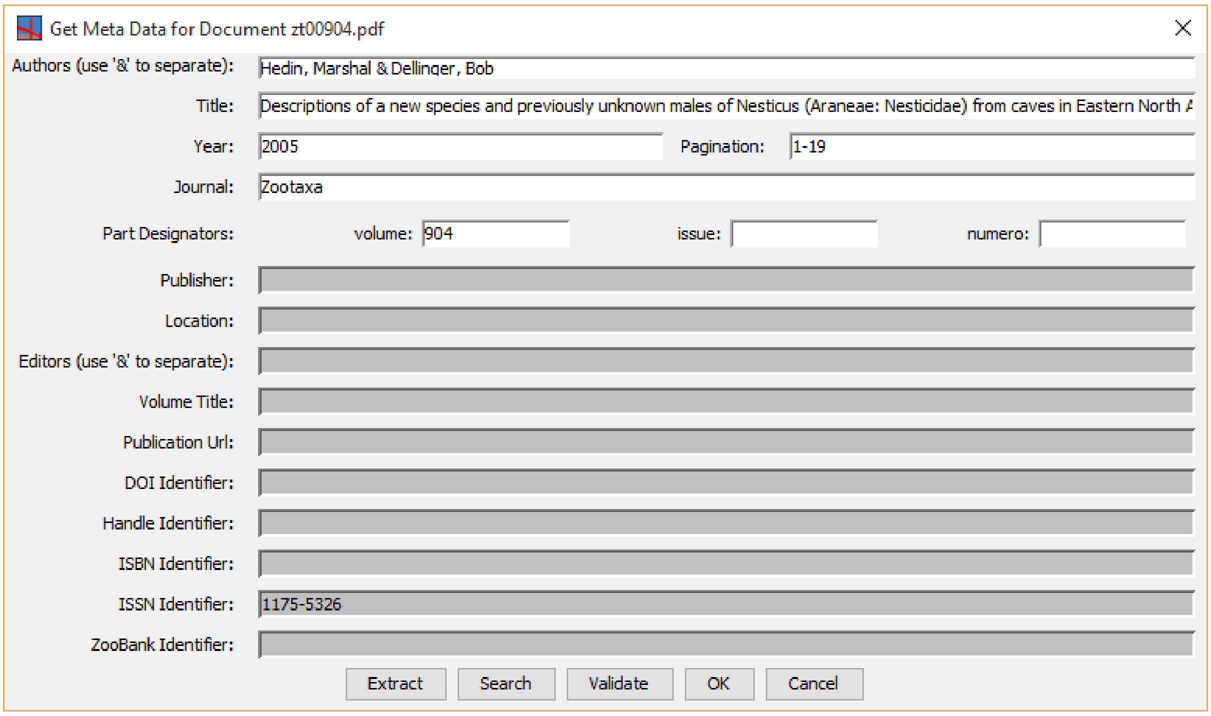
Article

ISSN 1175-5326 (print edition)
ZOOTAXA
ISSN 1175-5334 (online edition)

http://dx.doi.org/10.11646/zootaxa.3972.2.5
http://zoobank.org/urn:lsid:zoobank.org:pub:6A51EB89-0FE5-4F49-B427-E75459D8A3AF

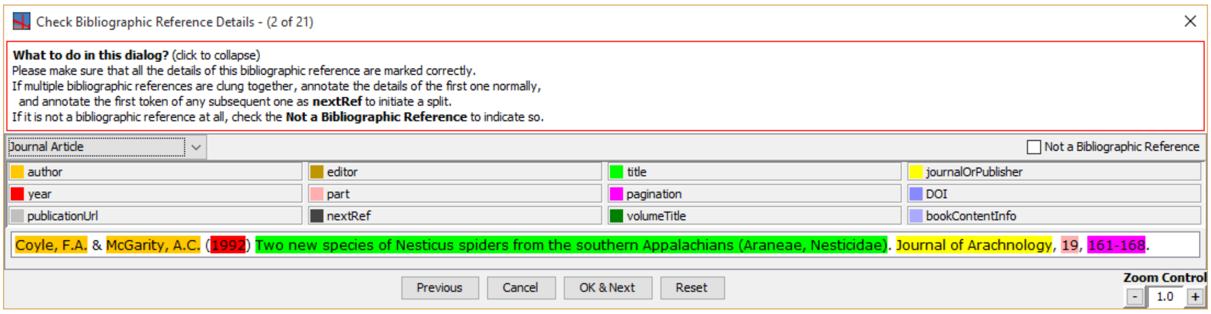
A taxonomic revision of the seed-harvester ant genus *Pogonomyrmex* (Hymenoptera: Formicidae) on Hispaniola

Figure 6: GoldenGATE Imagine user interface


Get Meta Data for Document zt00904.pdf

Authors (use '&' to separate): Hedin, Marshal & Dellinger, Bob
Title: Descriptions of a new species and previously unknown males of Nesticus (Araneae: Nesticidae) from caves in Eastern North A
Year: 2005 Pagination: 1-19
Journal: Zootaxa
Part Designators: volume: 904 issue: numero:
Publisher:
Location:
Editors (use '&' to separate):
Volume Title:
Publication Url:
DOI Identifier:
Handle Identifier:
ISBN Identifier:
ISSN Identifier: 1175-5326
ZooBank Identifier:

Figure 7: Document metadata extraction.



Check Bibliographic Reference Details - (2 of 21)

What to do in this dialog? (click to collapse)
Please make sure that all the details of this bibliographic reference are marked correctly.
If multiple bibliographic references are clung together, annotate the details of the first one normally,
and annotate the first token of any subsequent one as **nextRef** to initiate a split.
If it is not a bibliographic reference at all, check the **Not a Bibliographic Reference** to indicate so.

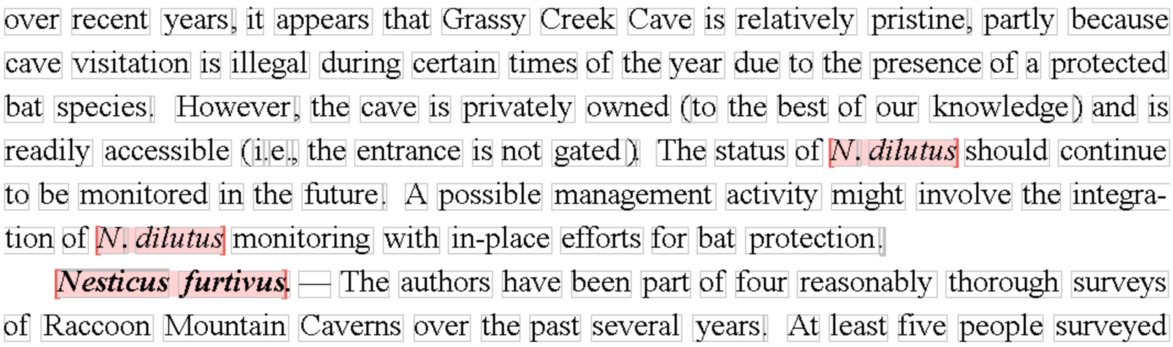
Journal Article

□ Not a Bibliographic Reference

🟧 author	🟫 editor	🟩 title	🟨 journalOrPublisher
🟥 year	▨ part	🟪 pagination	🟦 DOI
▨ publicationUrl	⬛ nextRef	🟩 volumeTitle	🟦 bookContentInfo

Coyle, F.A. & McGarity, A.C. (1992) Two new species of Nesticus spiders from the southern Appalachians (Araneae, Nesticidae). Journal of Arachnology, 19, 161-168.

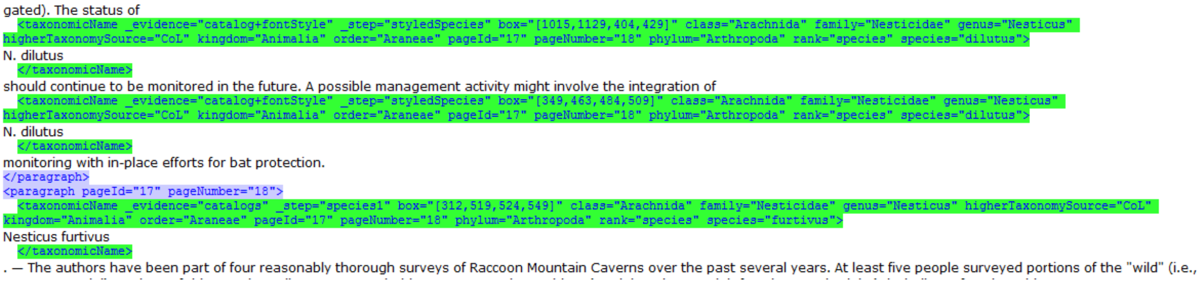
Figure 8: Bibliographic references parser



over recent years, it appears that Grassy Creek Cave is relatively pristine, partly because cave visitation is illegal during certain times of the year due to the presence of a protected bat species. However, the cave is privately owned (to the best of our knowledge) and is readily accessible (i.e., the entrance is not gated). The status of *N. dilutus* should continue to be monitored in the future. A possible management activity might involve the integration of *N. dilutus* monitoring with in-place efforts for bat protection.

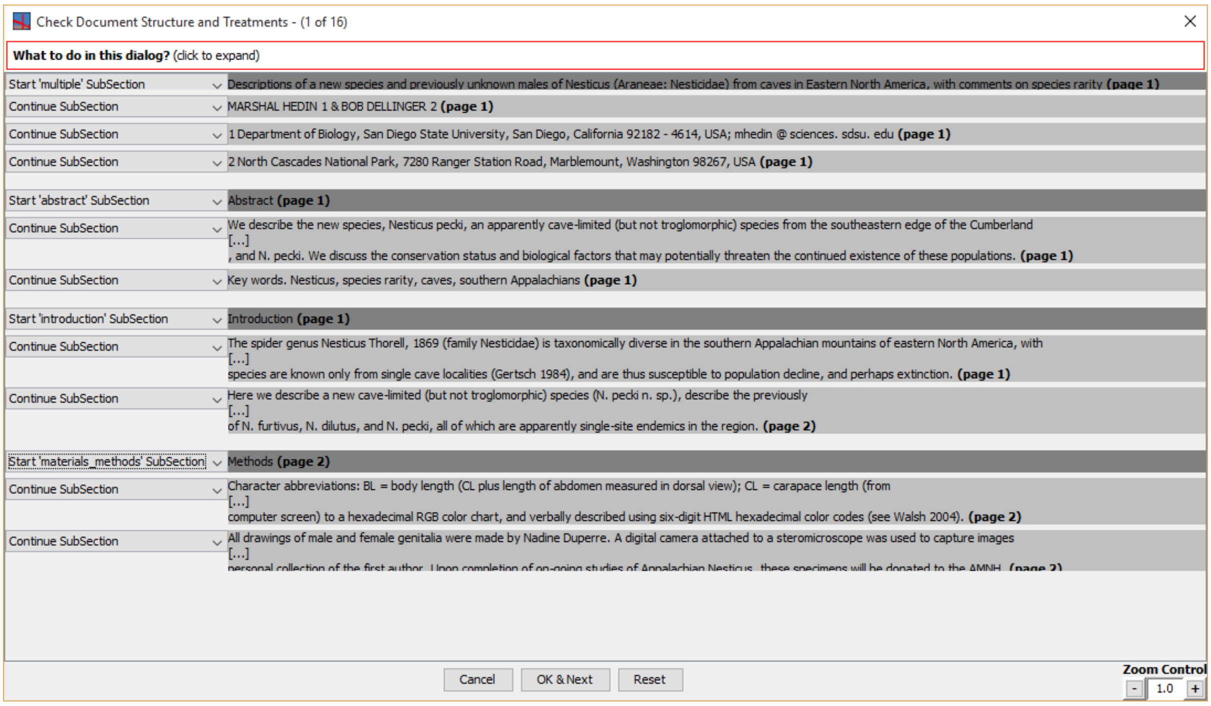
Nesticus furtivus. — The authors have been part of four reasonably thorough surveys of Raccoon Mountain Caverns over the past several years. At least five people surveyed

Figure 9: Markup of taxonomic names, backed by Catalogue of Life (COL) and GBIF as external catalogs



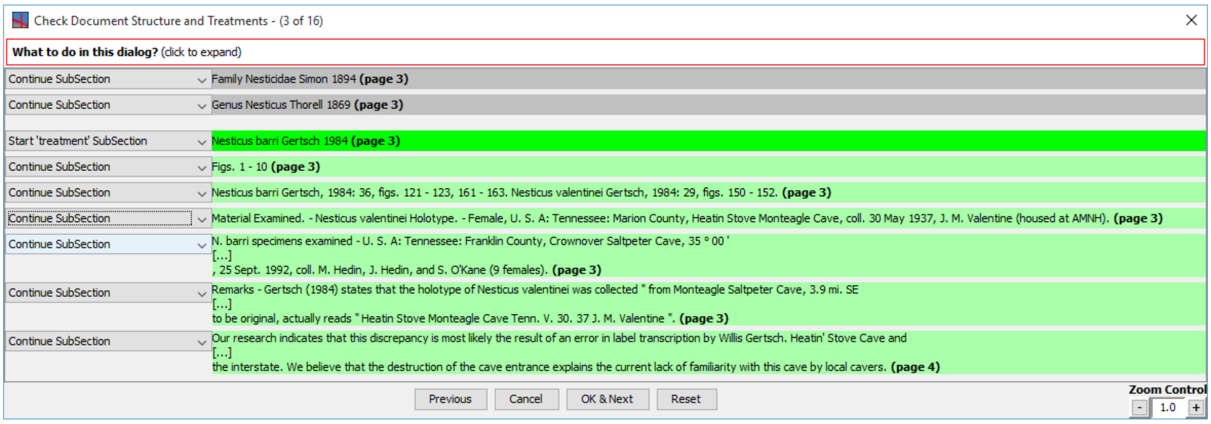
```
gated). The status of
  <taxonomicName _evidence="catalog+fontStyle" _step="styledSpecies" box="[1015,1129,404,429]" class="Arachnida" family="Nesticidae" genus="Nesticus"
higherTaxonomySource="CoL" kingdom="Animalia" order="Araneae" pageId="17" pageNumber="18" phylum="Arthropoda" rank="species" species="dilutus">
N. dilutus
  </taxonomicName>
should continue to be monitored in the future. A possible management activity might involve the integration of
  <taxonomicName _evidence="catalog+fontStyle" _step="styledSpecies" box="[349,463,484,509]" class="Arachnida" family="Nesticidae" genus="Nesticus"
higherTaxonomySource="CoL" kingdom="Animalia" order="Araneae" pageId="17" pageNumber="18" phylum="Arthropoda" rank="species" species="dilutus">
N. dilutus
  </taxonomicName>
monitoring with in-place efforts for bat protection.
</paragraph>
<paragraph pageId="17" pageNumber="18">
  <taxonomicName _evidence="catalogs" _step="species1" box="[312,519,524,549]" class="Arachnida" family="Nesticidae" genus="Nesticus" higherTaxonomySource="CoL"
kingdom="Animalia" order="Araneae" pageId="17" pageNumber="18" phylum="Arthropoda" rank="species" species="furtivus">
Nesticus furtivus
  </taxonomicName>
. — The authors have been part of four reasonably thorough surveys of Raccoon Mountain Caverns over the past several years. At least five people surveyed portions of the "wild" (i.e.,
```

Figure 10: XML code of the reconciled taxon names


Check Document Structure and Treatments - (1 of 16)
What to do in this dialog? (click to expand)
Cancel
OK & Next
Reset
Zoom Control
-
1.0
+

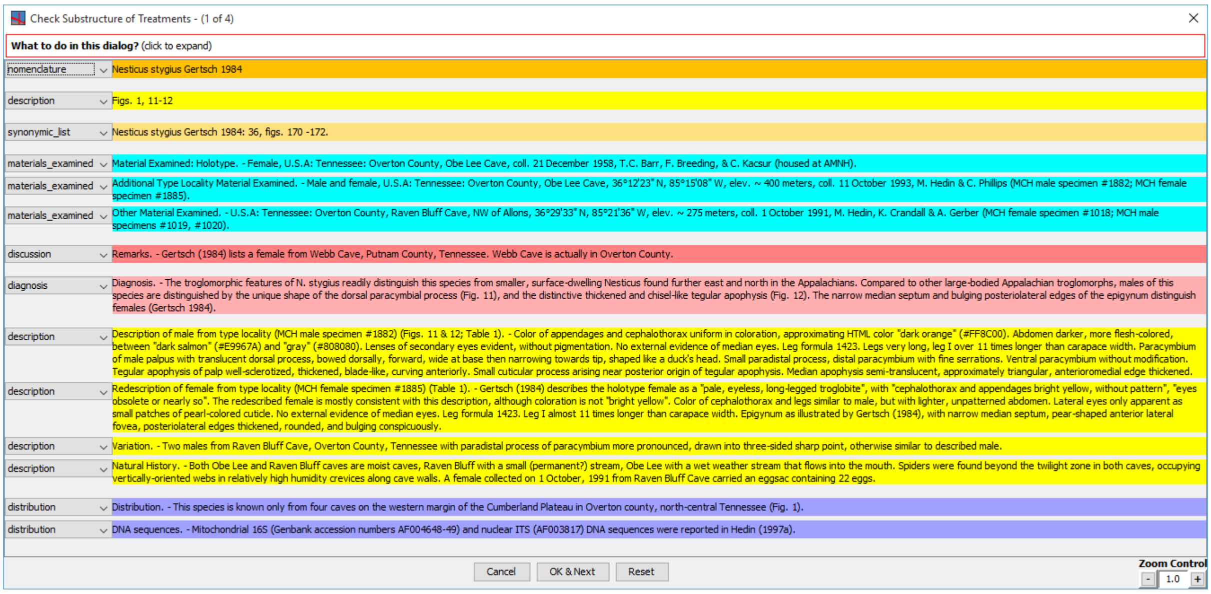
Start 'multiple' SubSection	Descriptions of a new species and previously unknown males of Nesticus (Araneae: Nesticidae) from caves in Eastern North America, with comments on species rarity **(page 1)**
Continue SubSection	MARSHAL HEDIN 1 & BOB DELLINGER 2 **(page 1)**
Continue SubSection	1 Department of Biology, San Diego State University, San Diego, California 92182 - 4614, USA; mhedin @ sciences. sdsu. edu **(page 1)**
Continue SubSection	2 North Cascades National Park, 7280 Ranger Station Road, Marblemount, Washington 98267, USA **(page 1)**
Start 'abstract' SubSection	Abstract **(page 1)**
Continue SubSection	We describe the new species, Nesticus pecki, an apparently cave-limited (but not troglomorphic) species from the southeastern edge of the Cumberland […] , and N. pecki. We discuss the conservation status and biological factors that may potentially threaten the continued existence of these populations. **(page 1)**
Continue SubSection	Key words. Nesticus, species rarity, caves, southern Appalachians **(page 1)**
Start 'introduction' SubSection	Introduction **(page 1)**
Continue SubSection	The spider genus Nesticus Thorell, 1869 (family Nesticidae) is taxonomically diverse in the southern Appalachian mountains of eastern North America, with […] species are known only from single cave localities (Gertsch 1984), and are thus susceptible to population decline, and perhaps extinction. **(page 1)**
Continue SubSection	Here we describe a new cave-limited (but not troglomorphic) species (N. pecki n. sp.), describe the previously […] of N. furtivus, N. dilutus, and N. pecki, all of which are apparently single-site endemics in the region. **(page 2)**
Start 'materials_methods' SubSection	Methods **(page 2)**
Continue SubSection	Character abbreviations: BL = body length (CL plus length of abdomen measured in dorsal view); CL = carapace length (from […] computer screen) to a hexadecimal RGB color chart, and verbally described using six-digit HTML hexadecimal color codes (see Walsh 2004). **(page 2)**
Continue SubSection	All drawings of male and female genitalia were made by Nadine Duperre. A digital camera attached to a steromicroscope was used to capture images […] personal collection of the first author. Upon completion of on-going studies of Appalachian Nesticus, these specimens will be donated to the AMNH. **(page 2)**

Figure 11: Detection of the generic structure of a document


Check Document Structure and Treatments - (3 of 16)
What to do in this dialog? (click to expand)
Previous
Cancel
OK & Next
Reset
Zoom Control
-
1.0
+

Continue SubSection	Family Nesticidae Simon 1894 **(page 3)**
Continue SubSection	Genus Nesticus Thorell 1869 **(page 3)**
Start 'treatment' SubSection	Nesticus barri Gertsch 1984 **(page 3)**
Continue SubSection	Figs. 1 - 10 **(page 3)**
Continue SubSection	Nesticus barri Gertsch, 1984: 36, figs. 121 - 123, 161 - 163. Nesticus valentinei Gertsch, 1984: 29, figs. 150 - 152. **(page 3)**
Continue SubSection	Material Examined. - Nesticus valentinei Holotype. - Female, U. S. A: Tennessee: Marion County, Heatin Stove Monteagle Cave, coll. 30 May 1937, J. M. Valentine (housed at AMNH). **(page 3)**
Continue SubSection	N. barri specimens examined - U. S. A: Tennessee: Franklin County, Crownover Saltpeter Cave, 35 ° 00 ' […] , 25 Sept. 1992, coll. M. Hedin, J. Hedin, and S. O'Kane (9 females). **(page 3)**
Continue SubSection	Remarks - Gertsch (1984) states that the holotype of Nesticus valentinei was collected " from Monteagle Saltpeter Cave, 3.9 mi. SE […] to be original, actually reads " Heatin Stove Monteagle Cave Tenn. V. 30. 37 J. M. Valentine ". **(page 3)**
Continue SubSection	Our research indicates that this discrepancy is most likely the result of an error in label transcription by Willis Gertsch. Heatin' Stove Cave and […] the interstate. We believe that the destruction of the cave entrance explains the current lack of familiarity with this cave by local cavers. **(page 4)**

Figure 12: Taxonomic treatments

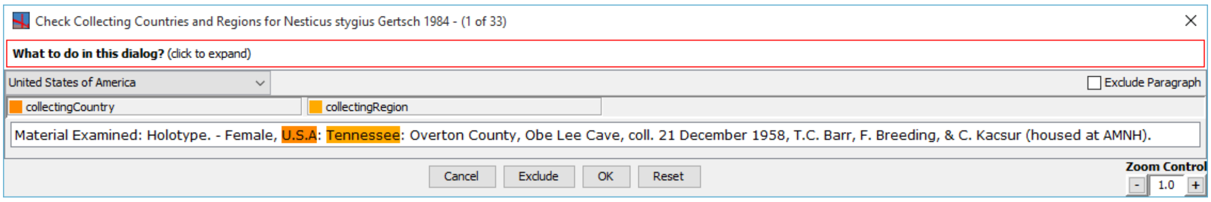


Check Substructure of Treatments - (1 of 4)

What to do in this dialog? (click to expand)

nomenclature	Nesticus stygius Gertsch 1984
description	Figs. 1, 11-12
synonymic_list	Nesticus stygius Gertsch 1984: 36, figs. 170 -172.
materials_examined	Material Examined: Holotype. - Female, U.S.A: Tennessee: Overton County, Obe Lee Cave, coll. 21 December 1958, T.C. Barr, F. Breeding, & C. Kacsur (housed at AMNH).
materials_examined	Additional Type Locality Material Examined. - Male and female, U.S.A: Tennessee: Overton County, Obe Lee Cave, 36°12'23" N, 85°15'08" W, elev. ~ 400 meters, coll. 11 October 1993, M. Hedin & C. Phillips (MCH male specimen #1882; MCH female specimen #1885).
materials_examined	Other Material Examined. - U.S.A: Tennessee: Overton County, Raven Bluff Cave, NW of Allons, 36°29'33" N, 85°21'36" W, elev. ~ 275 meters, coll. 1 October 1991, M. Hedin, K. Crandall & A. Gerber (MCH female specimen #1018; MCH male specimens #1019, #1020).
discussion	Remarks. - Gertsch (1984) lists a female from Webb Cave, Putnam County, Tennessee. Webb Cave is actually in Overton County.
diagnosis	Diagnosis. - The troglomorphic features of N. stygius readily distinguish this species from smaller, surface-dwelling Nesticus found further east and north in the Appalachians. Compared to other large-bodied Appalachian troglomorphs, males of this species are distinguished by the unique shape of the dorsal paracymbial process (Fig. 11), and the distinctive thickened and chisel-like tegular apophysis (Fig. 12). The narrow median septum and bulging posteriolateral edges of the epigynum distinguish females (Gertsch 1984).
description	Description of male from type locality (MCH male specimen #1882) (Figs. 11 & 12; Table 1). - Color of appendages and cephalothorax uniform in coloration, approximating HTML color "dark orange" (#FF8C00). Abdomen darker, more flesh-colored, between "dark salmon" (#E9967A) and "gray" (#808080). Lenses of secondary eyes evident, without pigmentation. No external evidence of median eyes. Leg formula 1423. Legs very long, leg I over 11 times longer than carapace width. Paracymbium of male palpus with translucent dorsal process, bowed dorsally, forward, wide at base then narrowing towards tip, shaped like a duck's head. Small paradistal process, distal paracymbium with fine serrations. Ventral paracymbium without modification. Tegular apophysis of palp well-sclerotized, thickened, blade-like, curving anteriorly. Small cuticular process arising near posterior origin of tegular apophysis. Median apophysis semi-translucent, approximately triangular, anterioromedial edge thickened.
description	Redescription of female from type locality (MCH female specimen #1885) (Table 1). - Gertsch (1984) describes the holotype female as a "pale, eyeless, long-legged troglobite", with "cephalothorax and appendages bright yellow, without pattern", "eyes obsolete or nearly so". The redescribed female is mostly consistent with this description, although coloration is not "bright yellow". Color of cephalothorax and legs similar to male, but with lighter, unpatterned abdomen. Lateral eyes only apparent as small patches of pearl-colored cuticle. No external evidence of median eyes. Leg formula 1423. Leg I almost 11 times longer than carapace width. Epigynum as illustrated by Gertsch (1984), with narrow median septum, pear-shaped anterior lateral fovea, posteriolateral edges thickened, rounded, and bulging conspicuously.
description	Variation. - Two males from Raven Bluff Cave, Overton County, Tennessee with paradistal process of paracymbium more pronounced, drawn into three-sided sharp point, otherwise similar to described male.
description	Natural History. - Both Obe Lee and Raven Bluff caves are moist caves, Raven Bluff with a small (permanent?) stream, Obe Lee with a wet weather stream that flows into the mouth. Spiders were found beyond the twilight zone in both caves, occupying vertically-oriented webs in relatively high humidity crevices along cave walls. A female collected on 1 October, 1991 from Raven Bluff Cave carried an eggsac containing 22 eggs.
distribution	Distribution. - This species is known only from four caves on the western margin of the Cumberland Plateau in Overton county, north-central Tennessee (Fig. 1).
distribution	DNA sequences. - Mitochondrial 16S (Genbank accession numbers AF004648-49) and nuclear ITS (AF003817) DNA sequences were reported in Hedin (1997a).

Figure 13: Internal treatment structure detection interface.



Check Collecting Countries and Regions for Nesticus stygius Gertsch 1984 - (1 of 33)

What to do in this dialog? (click to expand)

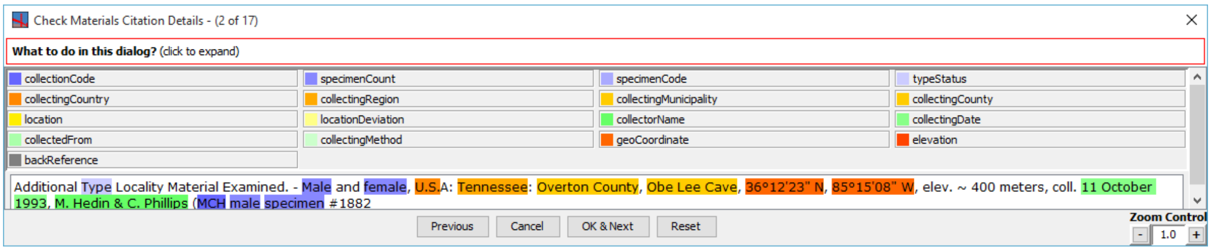
United States of America

□ Exclude Paragraph

🟧 collectingCountry 🟧 collectingRegion

Material Examined: Holotype. - Female, U.S.A: Tennessee: Overton County, Obe Lee Cave, coll. 21 December 1958, T.C. Barr, F. Breeding, & C. Kacsur (housed at AMNH).

Figure 14: Detection of materials citation interface



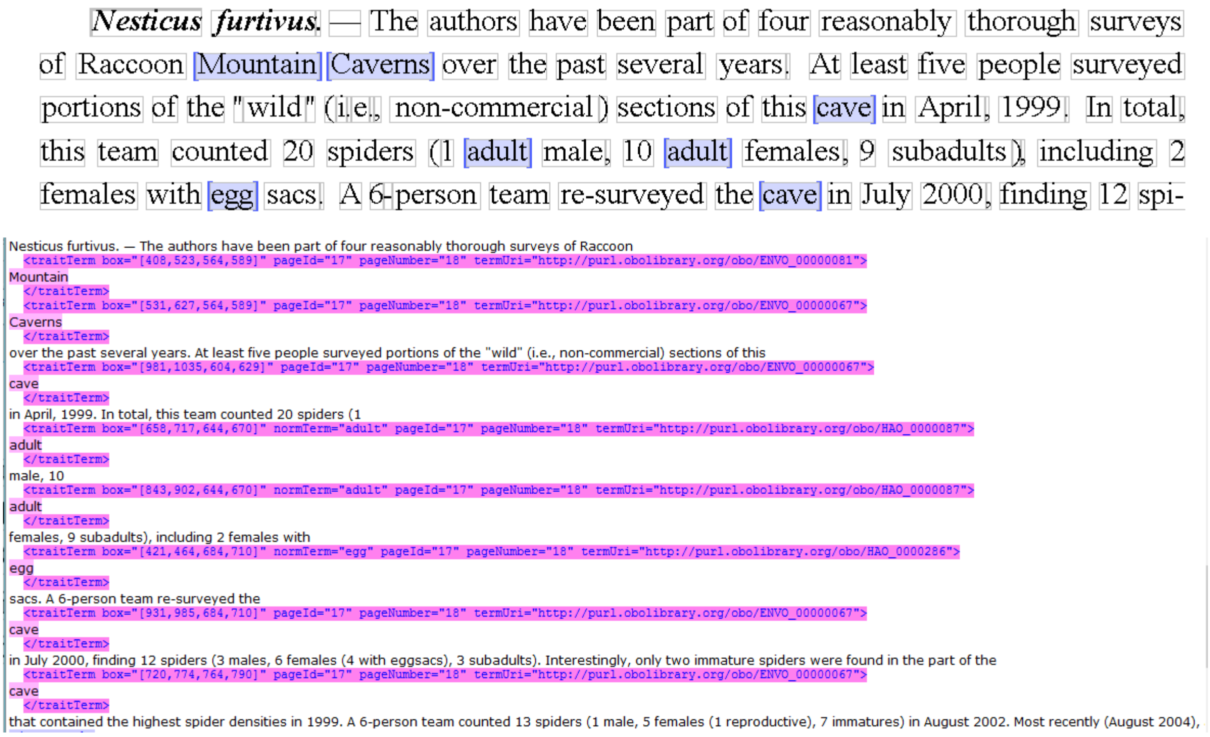
Check Materials Citation Details - (2 of 17)

What to do in this dialog? (click to expand)

🟦 collectionCode	🟦 specimenCount	🟦 specimenCode	🟦 typeStatus
🟧 collectingCountry	🟧 collectingRegion	🟨 collectingMunicipality	🟨 collectingCounty
🟨 location	🟨 locationDeviation	🟩 collectorName	🟩 collectingDate
🟩 collectedFrom	🟩 collectingMethod	🟧 geoCoordinate	🟥 elevation
⬛ backReference			

Additional Type Locality Material Examined. - Male and female, U.S.A: Tennessee: Overton County, Obe Lee Cave, 36°12'23" N, 85°15'08" W, elev. ~ 400 meters, coll. 11 October 1993, M. Hedin & C. Phillips (MCH male specimen #1882

Figure 15: Parsing of a materials citation interface



Nesticus furtivus. — The authors have been part of four reasonably thorough surveys of Raccoon Mountain Caverns over the past several years. At least five people surveyed portions of the "wild" (i.e., non-commercial) sections of this cave in April, 1999. In total, this team counted 20 spiders (1 adult male, 10 adult females, 9 subadults), including 2 females with egg sacs. A 6-person team re-surveyed the cave in July 2000, finding 12 spi-

```
Nesticus furtivus. — The authors have been part of four reasonably thorough surveys of Raccoon
   <traitTerm box="[408,523,564,589]" pageId="17" pageNumber="18" termUri="http://purl.obolibrary.org/obo/ENVO_00000081">
Mountain
   </traitTerm>
   <traitTerm box="[531,627,564,589]" pageId="17" pageNumber="18" termUri="http://purl.obolibrary.org/obo/ENVO_00000067">
Caverns
   </traitTerm>
over the past several years. At least five people surveyed portions of the "wild" (i.e., non-commercial) sections of this
   <traitTerm box="[981,1035,604,629]" pageId="17" pageNumber="18" termUri="http://purl.obolibrary.org/obo/ENVO_00000067">
cave
   </traitTerm>
in April, 1999. In total, this team counted 20 spiders (1
   <traitTerm box="[658,717,644,670]" normTerm="adult" pageId="17" pageNumber="18" termUri="http://purl.obolibrary.org/obo/HAO_0000087">
adult
   </traitTerm>
male, 10
   <traitTerm box="[843,902,644,670]" normTerm="adult" pageId="17" pageNumber="18" termUri="http://purl.obolibrary.org/obo/HAO_0000087">
adult
   </traitTerm>
females, 9 subadults), including 2 females with
   <traitTerm box="[421,464,684,710]" normTerm="egg" pageId="17" pageNumber="18" termUri="http://purl.obolibrary.org/obo/HAO_0000286">
egg
   </traitTerm>
sacs. A 6-person team re-surveyed the
   <traitTerm box="[931,985,684,710]" pageId="17" pageNumber="18" termUri="http://purl.obolibrary.org/obo/ENVO_00000067">
cave
   </traitTerm>
in July 2000, finding 12 spiders (3 males, 6 females (4 with eggsacs), 3 subadults). Interestingly, only two immature spiders were found in the part of the
   <traitTerm box="[720,774,764,790]" pageId="17" pageNumber="18" termUri="http://purl.obolibrary.org/obo/ENVO_00000067">
cave
   </traitTerm>
that contained the highest spider densities in 1999. A 6-person team counted 13 spiders (1 male, 5 females (1 reproductive), 7 immatures) in August 2002. Most recently (August 2004),
```

Figure 16: Tagging of trait terms, backed by ontologies (i.e. ENVO: Environmental Ontology, HAO: Hymenoptera Anatomy Ontology). Top: ENVO terms highlighted in text; bottom: marked up text

2.3.2 Editing

Manual editing and markup functionality helps correct errors where the automated tools produced errors. Among others, this includes:

- assisted revision of blocking, paragraph breaks, and text flow
- association of tables and figures to respective captions
- connection of tables spread out over multiple pages
- corrections in table structure, e.g. merging or splitting rows or columns
- correction of OCR errors or mis-decodings of embedded font characters

2.3.3 Saving and export

At any stage of document processing, users can choose to export documents as XML. As soon as document metadata is extracted and taxon names and treatments are marked, they can also export a Darwin Core Archive or upload the document to the Plazi treatment repository, which allows them to share their markup work with the public. Further, users can export tables and figures, together with their captions. The two following sections explain this in more detail.

2.3.4 Document Storage

Converted PDF documents are stored as Image Markup Files (IMF).

To keep track of the existing supplements and their attributes, they are listed in a dedicated CSV table named “supplements.csv”.

2.3.5 Document & Data Export

For use in other applications, converted PDF documents can also be exported in a variety of other data formats.

- XML: exports the document text, annotations, and regions into an XML document; there are several different options for how to fold sequences of chained-together words into one another, ranging from raw layout order, akin to the original PDF document, to strict reading order, which concatenates chains of words one after the other; this is to provide the best suited folding for any given application.
- Figures & Tables: exports all figures and tables to a ZIP archive, together with the respective captions; tables can be exported as CSV or as tab separated text
- TreatmentBank offers individual treatments for download in RDF-XML format, as well as a ZIP archive containing RDF-XML representations of the whole treatment collection. The RDF comprises taxonomic and bibliographic information, links to cited figures that have (HTTP) URIs, cited treatments, and cited occurrence data. The latter also includes links to any digitized specimens if available.
- Darwin Core Archive (Fig 13, 14): exports the taxonomic treatments marked in the document to a Darwin Core Archive
- EU BON Taxonomic Backbone (UTIS) imports taxonomic names and associated bibliographic information from TreatmentBank by means of purpose-generated RDF-XML (D1.4: http://tinyurl.com/hp9fade).

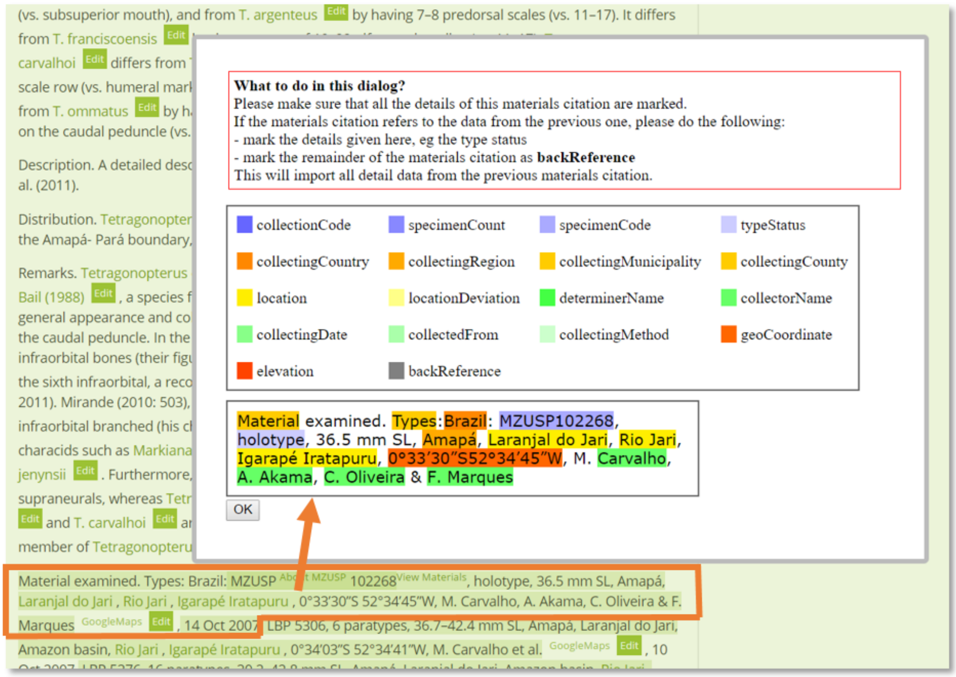
2.3.6 Manual

A manual for using GGI is available online http://plazi.org/resources/treatmentbank/goldengate-editor/ It includes an introduction to the setup, the markup process, trouble shooting and a glossary.

2.3.7 Server side, online editing

Online markup is limited to taxonomic names and observation records. If they are not already marked up, they can be highlighted and annotated. Observation records are automatically parsed (**Fig. 17**) and the annotations of specific elements suggested. They can be accepted, rejected or changed.

Once these elements are annotated, they can be edited (**Fig. 18**).



(vs. subsuperior mouth), and from T. argenteus Edit by having 7–8 predorsal scales (vs. 11–17). It differs

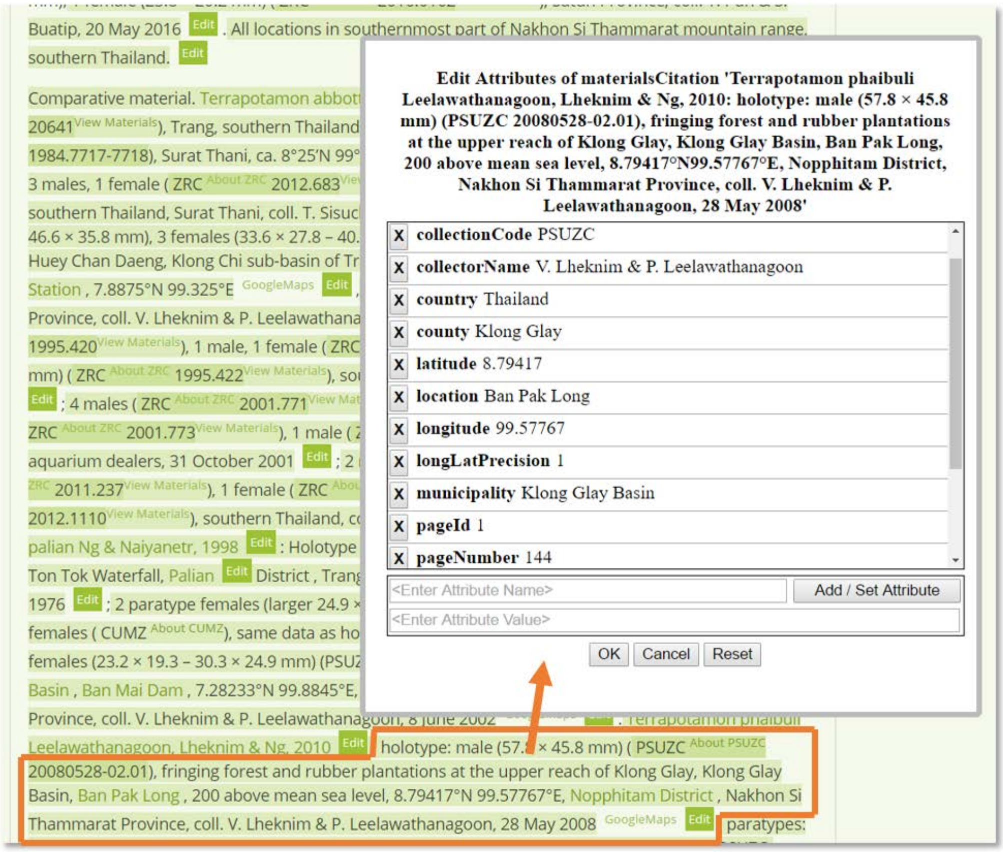
What to do in this dialog?
Please make sure that all the details of this materials citation are marked.
If the materials citation refers to the data from the previous one, please do the following:
- mark the details given here, eg the type status
- mark the remainder of the materials citation as **backReference**
This will import all detail data from the previous materials citation.

collectionCode	specimenCount	specimenCode	typeStatus
collectingCountry	collectingRegion	collectingMunicipality	collectingCounty
location	locationDeviation	determinerName	collectorName
collectingDate	collectedFrom	collectingMethod	geoCoordinate
elevation	backReference		

Material examined. Types:Brazil: MZUSP102268, holotype, 36.5 mm SL, Amapá, Laranjal do Jari, Rio Jari, Igarapé Iratapuru, 0°33’30”S52°34’45”W, M. Carvalho, A. Akama, C. Oliveira & F. Marques

Material examined. Types: Brazil: MZUSP About MZUSP 102268View Materials, holotype, 36.5 mm SL, Amapá, Laranjal do Jari , Rio Jari , Igarapé Iratapuru , 0°33’30”S 52°34’45”W, M. Carvalho, A. Akama, C. Oliveira & F. Marques GoogleMaps Edit , 14 Oct 2007 LBP 5306, 6 paratypes, 36.7–42.4 mm SL, Amapá, Laranjal do Jari, Amazon basin, Rio Jari , Igarapé Iratapuru , 0°34’03”S 52°34’41”W, M. Carvalho et al. GoogleMaps Edit , 10

Figure 17: Interface of the online annotation of observation records.



Buatip, 20 May 2016 Edit . All locations in southernmost part of Nakhon Si Thammarat mountain range, southern Thailand. Edit

Edit Attributes of materialsCitation 'Terrapotamon phaibuli Leelawathanagoon, Lheknim & Ng, 2010: holotype: male (57.8 × 45.8 mm) (PSUZC 20080528-02.01), fringing forest and rubber plantations at the upper reach of Klong Glay, Klong Glay Basin, Ban Pak Long, 200 above mean sea level, 8.79417°N99.57767°E, Nopphitam District, Nakhon Si Thammarat Province, coll. V. Lheknim & P. Leelawathanagoon, 28 May 2008'

X **collectionCode** PSUZC
X **collectorName** V. Lheknim & P. Leelawathanagoon
X **country** Thailand
X **county** Klong Glay
X **latitude** 8.79417
X **location** Ban Pak Long
X **longitude** 99.57767
X **longLatPrecision** 1
X **municipality** Klong Glay Basin
X **pageId** 1
X **pageNumber** 144

Leelawathanagoon, Lheknim & Ng, 2010 Edit holotype: male (57.8 × 45.8 mm) (PSUZC About PSUZC 20080528-02.01), fringing forest and rubber plantations at the upper reach of Klong Glay, Klong Glay Basin, Ban Pak Long , 200 above mean sea level, 8.79417°N 99.57767°E, Nopphitam District , Nakhon Si Thammarat Province, coll. V. Lheknim & P. Leelawathanagoon, 28 May 2008 GoogleMaps Edit paratypes:

Figure 18: Online editing pop-up window

2.3.8 Web Services

Independent of GGI, many of the taggers and parsers are available individually as Web Services for public use, based on the GoldenGATE Web Services infrastructure built in the scope of the ViBRANT project (also FP7, RI-261532).

2.4 Ongoing data processing using GoldenGate Imagine

Two recent publications based on GGI illustrate that the markup is functional. Miller et al (2015) focused on a comparison of Open Access articles to extract and understand key data, such as materials observation and details therein (**Fig. 19**).

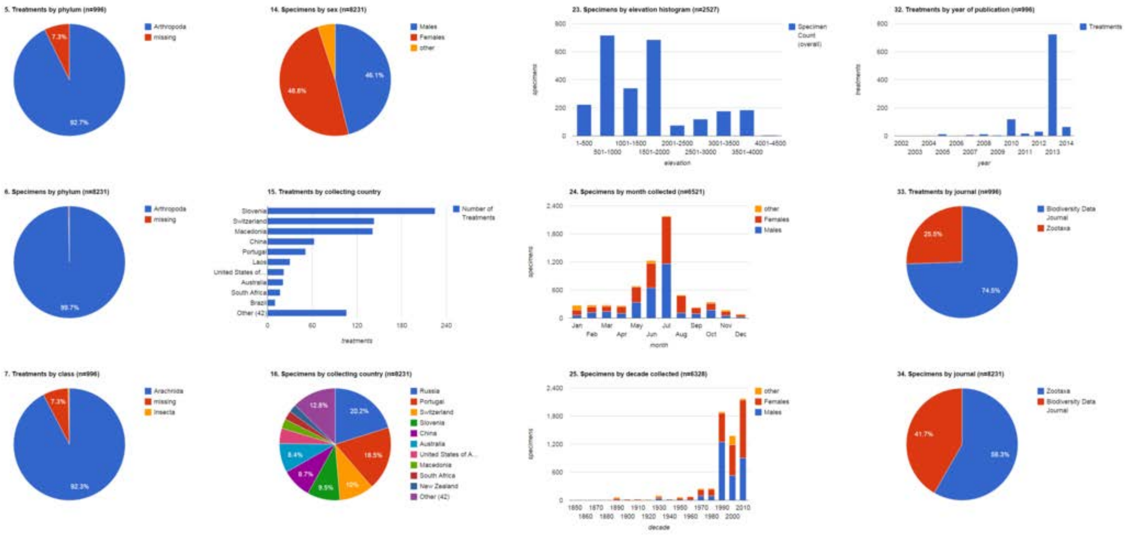
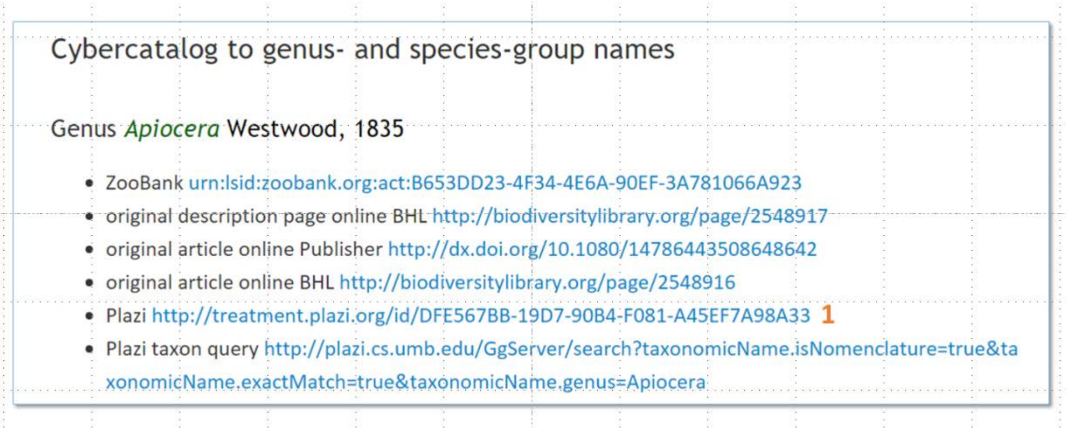


Figure 19: Selection of dashboard charts summarizing content from 37 open access articles published in Zootaxa and five articles published in Biodiversity Data Journal containing treatments on spiders. These charts illustrate interoperability of data from XML-based publishing and subsequently marked up legacy literature (Miller et al., 2015).

Torsten & Agosti (2015; **Fig. 20**) extracted treatments of a fly taxon to demonstrate how treatments can be cited and thus a link from a taxonomic name usage to the underlying treatment and its data can be established.



Cybercatalog to genus- and species-group names

Genus *Apiocera* Westwood, 1835

- ZooBank urn:lsid:zoobank.org:act:B653DD23-4F34-4E6A-90EF-3A781066A923
- original description page online BHL http://biodiversitylibrary.org/page/2548917
- original article online Publisher http://dx.doi.org/10.1080/14786443508648642
- original article online BHL http://biodiversitylibrary.org/page/2548916
- Plazi http://treatment.plazi.org/id/DFE567BB-19D7-90B4-F081-A45EF7A98A33 1
- Plazi taxon query http://plazi.cs.umb.edu/GgServer/search?taxonomicName.isNomenclature=true&taxonomicName.exactMatch=true&taxonomicName.genus=Apiocera

Figure 20: Cyber catalog view with links to external resources. 1: persistent HTTP URI for treatments extracted via GGI and stored in Plazi.

2.4.1 Contribution to GBIF

Within EU BON, GBIF is the aggregator of observation records and as such facilitates data sharing. Plazi's contributions are ongoing at a daily base. This includes by November 28, 2016 38,734 observation records and 32,543 taxonomic names using the DWCA http://www.gbif.org/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c. The difference between the available names in Plazi and those integrated in GBIF is due to the re-indexing of GBIF's re-indexing of their taxonomic backbone every couple of months.

2.4.2 Contribution to BLR

The Biodiversity Literature Repository (BLR, http://biolitrepo.org) is the permanent repository for biodiversity literature and figures, built in collaboration with Zenodo at CERN and Pensoft. It enables minting DataCite Digital Object Identifiers (DOI) for each digital object and thus making them citable and discoverable. This allows linking each observation record to its the source treatment and publication. Currently 14,296 processed articles and 140,000 figures are being added from TreatmentBank.

Another source of daily update of the BLR content is a workflow developed by Pensoft for the biodiversity-related journals published on its' ARPHA Journal Publishing Platform. The workflow automatically uploads both PDF and XML versions of each newly published article into BLR on the day of publication.

2.4.3 Contribution to RefBank

RefBank is a distributed storage facility for bibliographic references, with each node holding the entire dataset, as well as providing a website with search, editing, and upload functionality. Through the latter, users can find the reference data they need, correct errors, and contribute their own collections of references. Currently, RefBank holds about 870,000 references, with Plazi adding an average of 30 to 35 references per article processed. RefBank (Sautter & Biserkov, 2013; King et al., 2013) is a product developed by the ViBRANT project and now maintained by Plazi.

2.4.4 Contribution to UTIS

The Unified Taxonomic Information Service (UTIS) is the taxonomic backbone for the EU BON project. Plazi is one of the taxonomic name contributors (D1.4 10.5281/zenodo.182122)

2.5 Training

Two training sessions were held during the EU BON project. 21 participants attended the workshop in Crete June 8-9, 2015 resulting in a detailed analysis and comparison of GGI with other markup workflows (Faulwetter et al., 2016). A second event was held as part of the EU BON biodiversity data sharing and data publishing workshop in Sofia, March 22-23, 2016 (http://tinyurl.com/zu8kbgw).

3 Challenges and further/future developments

As mentioned above, OCR quality, or error rate, still poses a significant obstacle to the extent that the optimal solution for consistently high accuracy text capture from many legacy documents still is double-keying. Invaluable occurrence data recorded in documents from the ages of discovery and colonization in particular challenge accurate automated text capture due to the quality of the original documents, respective scanned images, and typefaces and formatting for which OCR performs poorly. Building upon previously conducted experiments, we plan on researching and creating an alternative solution that uses data clustering techniques on word images, takes into account font properties and peculiarities, and relies on a crowd based approach for correction akin to reCAPTCHA.

To further automate data extraction, GGI can use document style templates, which allow fully automated processing of PDF documents up to the extraction of bibliographic metadata and references, tables and figures, and taxonomic names and treatments. Style templates model the layout and styling characteristics of individual journals and can vastly speed up the processing of documents from those journals. Users can create style templates from very few documents that follow the same layout, e.g. several articles from a single journal (safe for layout changes), and then rely on their assistance for many more such documents. Furthermore, GGI can fully automatically batch process large numbers of PDF documents once it has a respective document style template available.

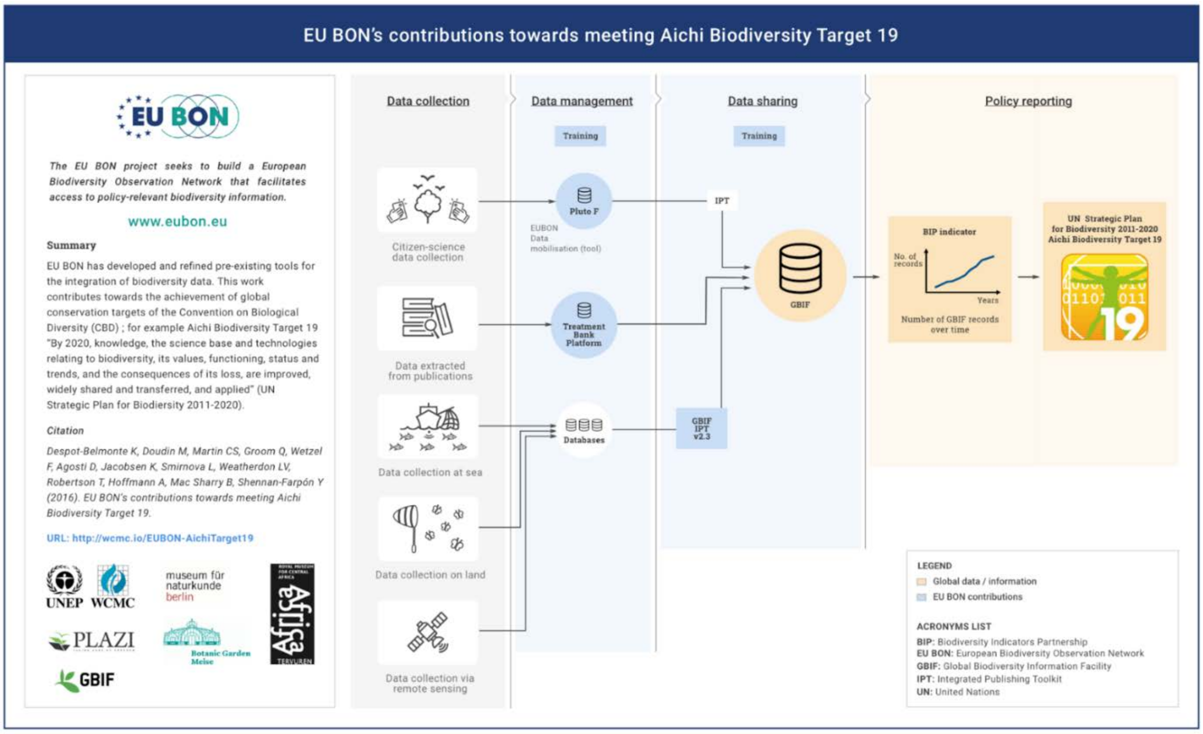
Document style templates thus far exist only for a small number of document layouts. We plan on creating templates for many more document styles, as well as providing an infrastructure that will allow users of GGI to share document style templates of their own creation. This infrastructure will increase the availability of style templates to many more journals, as well as their accuracy and generality, so data extraction effort can be considerably reduced for a wide variety of journals, both legacy and contemporary.

To further increase openness, we plan to more tightly integrate the Plazi treatment repository with CERN’s Zenodo open data repository. All non-copyrightable elements (e.g., figures) of PDF documents contributed to Zenodo will become available to the public and citable individually by means of DOIs.

Currently, GGI is a pure desktop application, which requires a Java Virtual Machine. To further lower the bar for contribution, we plan on building a web based version that allows users to upload PDF documents, or have them downloaded from a URL they provide, and then extract data within their browsers. As a by-product, extraction results will become available publicly by means of the aforementioned collaboration with CERN’s Zenodo platform. An additional benefit is that the browser-based version of GGI will also, enable users to refine, improve, amend, and extend data extracted from taxonomic treatments hosted in the Plazi repository. This will allow marking and parsing further materials citations, or linking specimen citations to other taxonomic databases such as MorphBank by means of simple drag and drop actions.

To increase the number of PDF document that can be automatically loaded into Plazi’s analysis, data extraction, and data sharing systems, we plan on expanding our collaboration with the ContentMine project (http://contentmine.org/), which focuses, apart from data extraction, on the discovery of newly published PDF documents. This will enable the Plazi systems to ingest more PDF documents, and, once respective document style templates come into existence, to extract more treatments fully automatically, thus fostering the growth of the treatment repository and the amount of taxonomic data available to the public.

Participation in WP6 and in Globis-B (http://www.globis-b.eu/), with a main focus on Essential Biodiversity Variables is ongoing to study how the potential of materials citations from the published record can be used, especially to cover the long tail of rare taxa for which hardly other data exists (Miller et al., 2015). TreatmentBank has been integrated in the EU BON workflows such as the contribution towards meeting the Aichi Biodiversity Target 19 (**Fig. 21**).



EU BON's contributions towards meeting Aichi Biodiversity Target 19

The EU BON project seeks to build a European Biodiversity Observation Network that facilitates access to policy-relevant biodiversity information.

www.eubon.eu

Summary

EU BON has developed and refined pre-existing tools for the integration of biodiversity data. This work contributes towards the achievement of global conservation targets of the Convention on Biological Diversity (CBD) ; for example Aichi Biodiversity Target 19 "By 2020, knowledge, the science base and technologies relating to biodiversity, its values, functioning, status and trends, and the consequences of its loss, are improved, widely shared and transferred, and applied" (UN Strategic Plan for Biodiersity 2011-2020).

Citation

Despot-Belmonte K, Doudin M, Martin CS, Groom Q, Wetzel F, Agosti D, Jacobsen K, Smirnova L, Weatherdon LV, Robertson T, Hoffmann A, Mac Sharry B, Shennan-Farpón Y (2016). EU BON's contributions towards meeting Aichi Biodiversity Target 19.

URL: http://wcmc.io/EUBON-AichiTarget19

Data collection
Data management
Data sharing
Policy reporting
Training
Training
Citizen-science data collection
Data extracted from publications
Data collection at sea
Data collection on land
Data collection via remote sensing
Pluto F
EUBON Data mobilisation (tool)
Treatment Bank Platform
Databases
IPT
GBIF IPT v2.3
GBIF
BIP indicator
No. of records
Years
Number of GBIF records over time
UN Strategic Plan for Biodiversity 2011-2020 Aichi Biodiversity Target 19
19

LEGEND

- 🟧 Global data / information
- 🟦 EU BON contributions

ACRONYMS LIST

BIP: Biodiversity Indicators Partnership
EU BON: European Biodiversity Observation Network
GBIF: Global Biodiversity Information Facility
IPT: Integrated Publishing Toolkit
UN: United Nations

Figure 21: TreatmentBank's role in data management in EU BON

4 References

- Agosti D, Egloff W 2009. Taxonomic information exchange and copyright: the Plazi approach. BMC Research Notes 2009, [2:53]. doi: 10.1186/1756-0500-2-53
- Catapano T 2010. TaxPub: An Extension of the NLM/NCBI Journal Publishing DTD for Taxonomic Descriptions. Proceedings of the Journal Article Tag Suite Conference 2010 Article http://www.ncbi.nlm.nih.gov/books/NBK47081/
- Dikow T, Agosti D. 2015. Utilizing online resources for taxonomy: a cybercatalog of Afrotropical apiocerid flies (Insecta: Diptera: Apioceridae). Biodiversity Data Journal 3: e5707 (06 Oct 2015. doi: 10.3897/BDJ.3.e5707
- Egloff W, Agosti D, Kishor P, Patterson D, Miller JA 2016. Copyright and the Use of Images as Biodiversity Data. bioRxiv doi: 10.1101/087015 (preprint submitted Nov 12; article submitted to PLoS Biology Nov 12)
- Faulwetter S, Pafilis E, Fanini L, Bailly N, Agosti D, Arvantidis C, Boicenco L, Catapano T, Claus S, Dekeyzer S, Georgiev T, Legaki A, Mavraki D, Oulas A, Papastefanou G, Penev L, Sautter G, Schigel, Senderov V, Teaca A, Tsompanou M (2016) EMODnet Workshop on mechanisms and guidelines to mobilise historical data into biogeographic databases. Research Ideas and Outcomes 2: e9774. doi: 10.3897/rio.2.e9774
- King D, Sautter G, Morse D, Penev L, Biserkov J, Georgiev T, Robert D, Smith V 2013. Bibliography of Life a freely accessible bibliography of every taxonomic paper ever published. Poster, TDWG 2013, Florence, Italy, 2013. doi: 10.5281/zenodo.181364
- Miller JA, Agosti D, Penev L, Sautter G, Georgiev T, Catapano T, Patterson D, King D, Pereira S, Vos RA, Sierra S 2015. Integrating and visualizing primary data from prospective and legacy taxonomic literature. Biodiversity Data Journal 3: e5063 (12 May 2015) doi: 10.3897/BDJ.3.e5063
- Patterson DJ, Egloff W, Agosti D, Eades D, Franz N, Hagedorn G, Rees J, Remsen DP 2014. Scientific names of organisms: attribution, rights, and licensing . BMC Research Notes 2014, 7:79 doi: 10.1186/1756-0500-7-79
- Sautter G, Biserkov J 2013. Bibliography of Life (RefBank and ReFinder): A tool to discover, download, correct, and format bibliographic references. Proceedings of TDWG 2013, Florence, Italy, 2013.
- Sautter G, Böhm K 2014. Improved bibliographic reference parsing based on repeated patterns. International Journal on Digital Libraries 14.1-2: 59-80. doi: 10.1007/s00799-014-0110-6 Sources available from http://github.com/VBRANT/refparse

5 Acronyms

BHL	Biodiversity Heritage Library	http://www.biodiversitylibrary.org/
BLR	Biodiversity Literature Repository	http://biolitrepo.org
COL	Catalogue of Life	http://www.catalogueoflife.org/
CSV	Comma Separated Values file	https://en.wikipedia.org/wiki/Comma-separated_values
DOI	Digital Object Identifier	http://tinyurl.com/ow7wad3
DST	Document Style Template	
DTD	Document Type Definition	http://www.w3schools.com/xml/xml_dtd_intro.asp
DWC	Darwin Core	http://rs.tdwg.org/dwc/
DWCA	Darwin Core Archive	http://tools.gbif.org/dwca-assistant/
EML	Ecological Metadata Language	https://knb.ecoinformatics.org/#external//emlparser/docs/index.html
ENVO	Environment Ontology	http://bioportal.bioontology.org/ontologies/ENVO
GBIF	Global Biodiversity Information Facility	http://gbif.org
GGI	GoldenGate Imagine	http://plazi.org/resources/treatmentbank/goldengate-editor/
GUI	Graphical User	

	Interface	
HAO	Hymenoptera Anatomy Ontology	http://portal.hymao.org/projects/32/public/ontology/
HTML	Hyper Text Markup Language	http://www.w3.org/TR/html/
IMF	Image Markup File	http://plazi.org/api-tools/image-markup-file-imf/
IPNI	International Plant Names Index	http://www.ipni.org/
ION	Index of Organism Names	http://www.organismnames.com/metrics.htm?page=tsj
JATS	Journal Article Tag Suite	https://jats.nlm.nih.gov/
NLM DTD	National Library of Medicine Data Type Definition	http://dtd.nlm.nih.gov/
OCR	Optical Character Recognition	
PDF	Portable Document Format	https://acrobat.adobe.com/us/en/products/about-adobe-pdf.html
PNG	Portable Network Graphics	http://www.w3.org/TR/PNG/
RefBank		http://refbank.org/
RefParse	RefParse	https://github.com/VBRANT/refparse
tab	Tabular key on a keyboard	

TaxPub		https://sourceforge.net/projects/taxpub/
URI	Universal Resource Identifier	https://www.w3.org/Addressing/
ViBRANT	Virtual Biodiversity Research and Access Network for Taxonomy	http://vbrant.eu/
XML	eXtensible Markup Language	http://www.w3.org/TR/xml/
ZIP	“Zipped” or compressed file	