REQUIREMENTS – PAGE TEXT UPLOADS

This document records the requirements for allowing uploads of new page text into BHL. Initially, the supported sources of new page text will be several different transcription tools, but this could be expanded in the future to include additional options.

File Uploads

Supported Formats

The supported file formats are:

- From The Page transcription tool output
- DigiVol transcription tool output
- Smithsonian Transcription Center output

*NOTE: Initially, consideration was given to defining a BHL file format for submissions that do not fit (or come from) one of the predefined transcription formats. File submitters would be responsible for putting their data into the BHL file format. This option is **NOT** being considered for this phase.*

Required Fields

File uploads must meet the following requirements:

- Files must include the sequence/leaf number of each page, so that the transcriptions for each page can be accurately mapped to the correct BHL page records
- Files can contain a subset of pages in a book, as long at the sequence/leaf numbers are present and accurate
- Each file should contain transcription data for a single BHL item.

NOTE: To support multiple items in a single file, an additional required field for BHL Item Identifier would be needed.

Text Markup

Uploaded files should not contain markup within the text. No processes will be created to handle the display or indexing of markup.

Public User Interface

The newly uploaded text files will be presented in the public user interface as follows:

- Replace the existing text files (generated from OCR) with the uploaded text files

NOTE: Changes to text files in BHL are not reflected in the OCR files stored at Internet Archive.

- Use the current OCR area of the book viewer for transcription display (no changes needed)
- Update labels in book viewer (and elsewhere?) to reflect all possible sources of data

NOTE: Label to indicate the displayed text may be either uncorrected OCR or manually transcribed text:

This text is either generated from uncorrected OCR or is a manually produced transcription. As such, it may include inconsistencies with the content of the original page.

- All exports and API responses that include item text will return text from the files produced by the transcription tools (no changes needed)

Administrative User Interface

Dashboard

Add the following to the Data Import box in the lower left corner of the Administrative Dashboard:

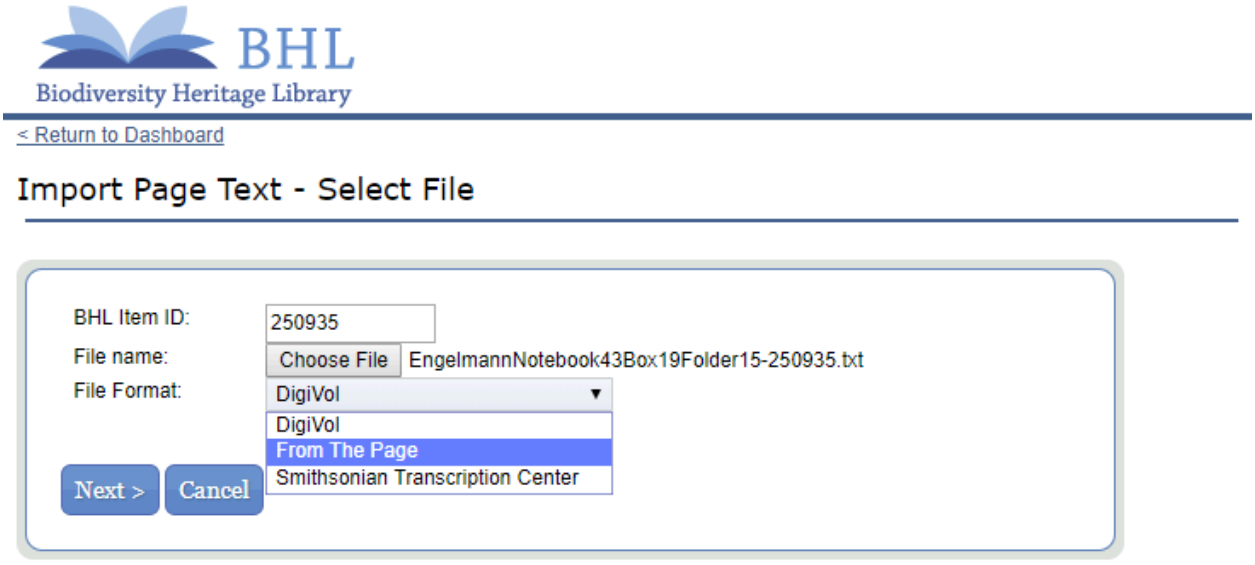
- Import Item Text
- Text Import History

Using "Text" in the labels rather than "Transcriptions" allows room for expansion to other data sources in the future (alternate OCR outputs, gaming outputs, etc).

Import Item Text

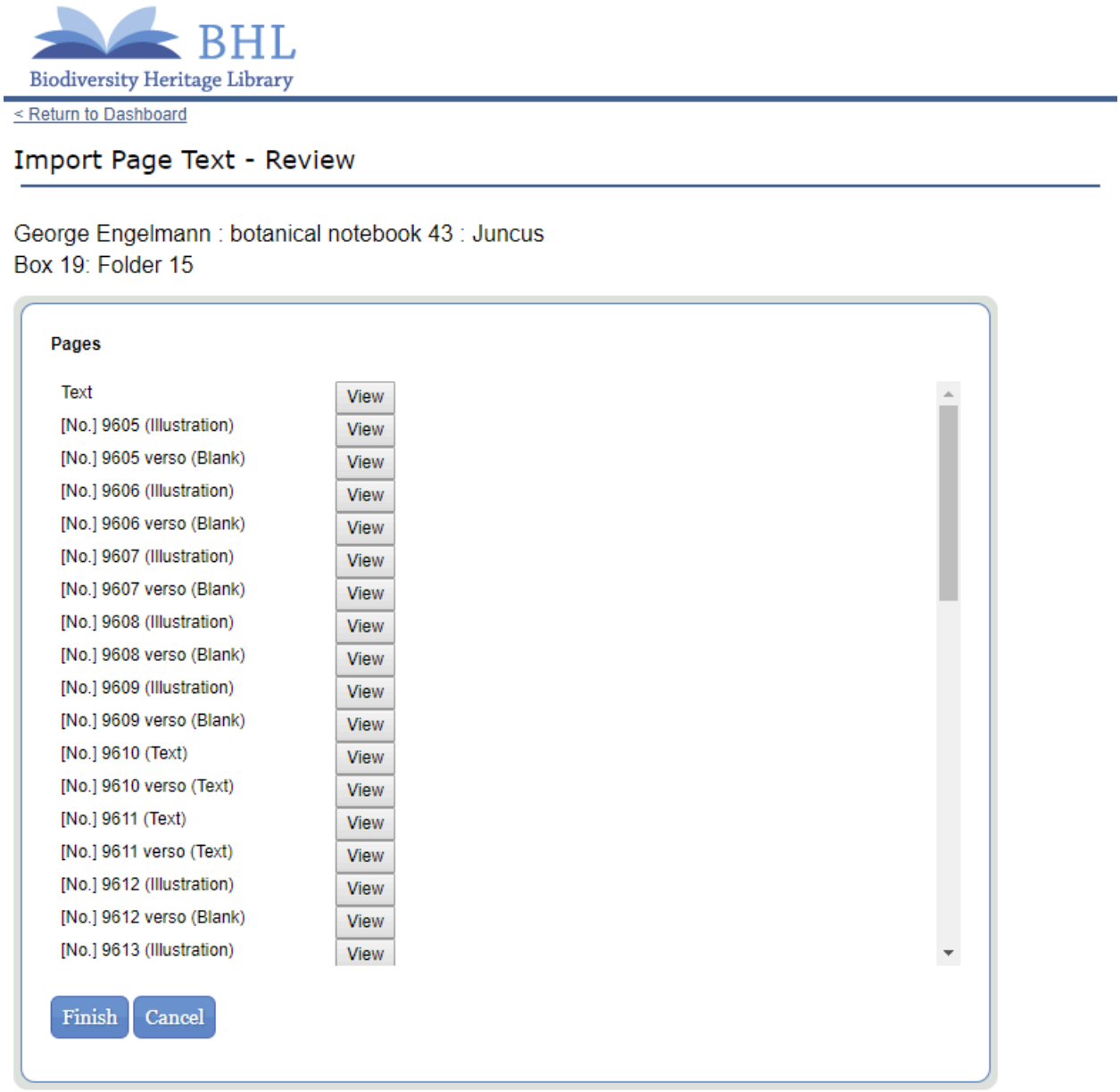
Here is the process that will be initiated by clicking the "Import Item Text" menu option.

1. Specify BHL item, filename, and file format.

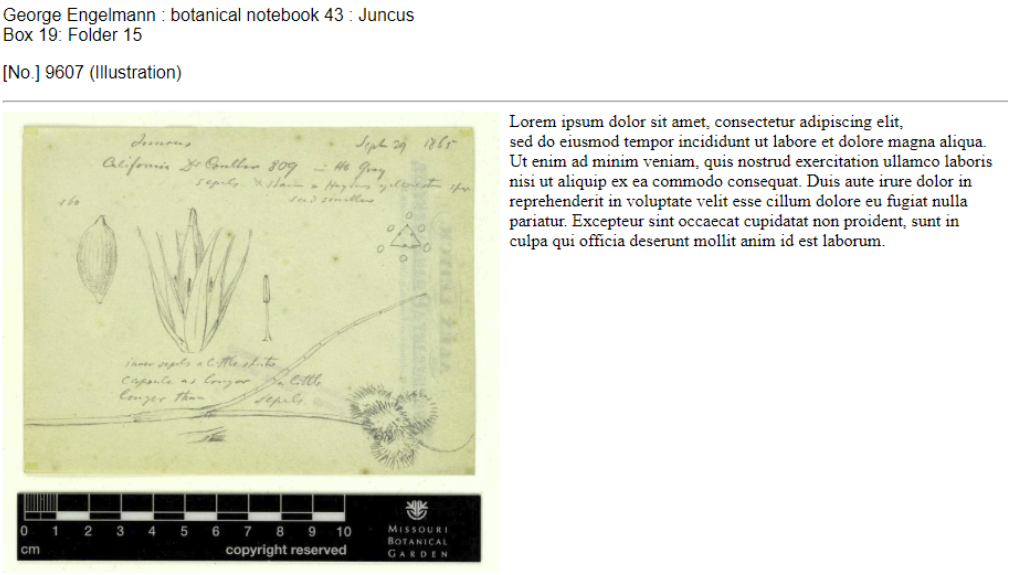


2. Click "Next >" (or "Cancel")
    a. Upload file to temporary location

2. Review list of pages for which text was imported.



a. Click Page to view text side-by-side with a page image (ensure mappings of text-to-page are correct)



3. Click “Next >” (or “Cancel”)
    a. Perform the import operation, replacing the existing text files with the new ones
    b. Create log entries to track the changes

NOTE: There is no “Undo” option. Once the files have been replaced, the action is final.

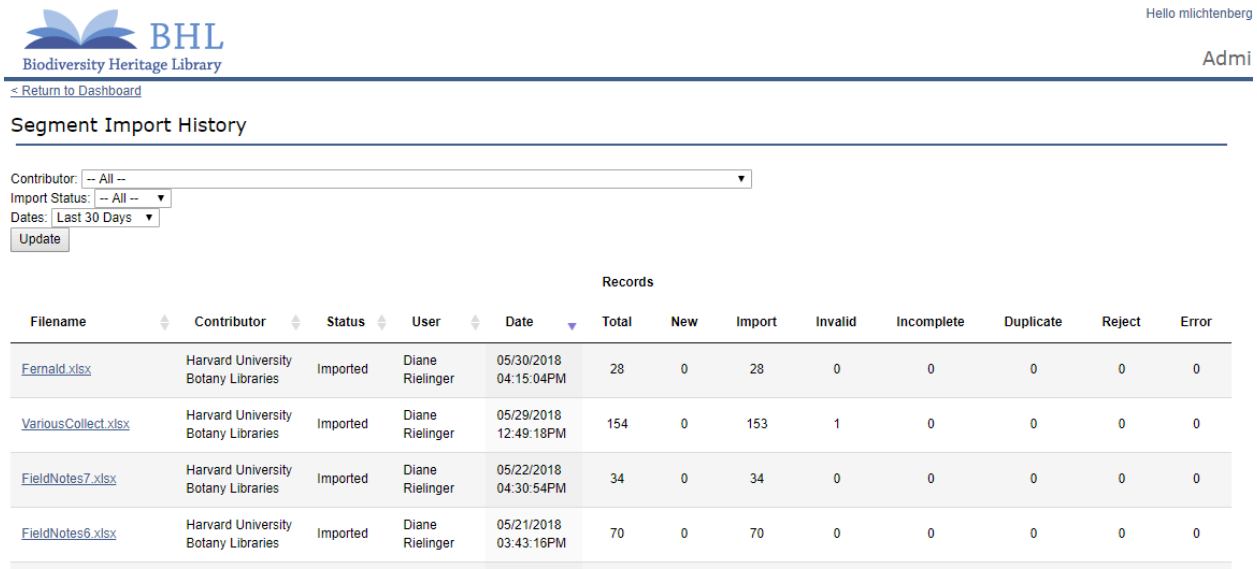
4. Display the Text Import History page with the newly imported file at the top of the list.

Text Import History

Clicking the new “Text Import History” menu item will open a new page that displays a list of the following information about text file uploads:

- Filename
- User/Uploader
- Upload Status (Success or Failure)
- Date
- BHL Item ID
- File Format/Type
- Number of pages of text uploaded

Sorting and filtering of the list will be possible. The functionality will be similar to the functionality of the “Segment Import History” page (shown here for illustration).



Paginator (Editing Page Text)

There was consideration for adding the capability to edit the text of a page within the Paginator interface. This was to be a simple text editor with no support for markup.

It has been decided that this functionality **will not be added in this phase**.

The source platforms (transcription tools, etc) for the text uploads are considered to hold the "master" copies of the text. Therefore, all editing should be done in the source platforms, the data re-exported, and then re-imported into BHL.

Application Programming Interface (API)

No changes to the BHL Application Programming Interfaces (APIs) are needed.

Logging

The following history will be maintained at Page-level:

- Filename
- Date/Time
- User ID
    - Based on the logged in user performing the import
- Source of text
    - OCR (Internet Archive)
    - FromThePage
    - DigiVol
    - Smithsonian Transcription Center

Logging will be performed at the following times:

- When a import is performed on the admin site
- When OCR is re-imported from Internet Archive
    - Currently, this is possible via a “hidden” interface that is part of the Page Insert cleanup tool.
- When a data ingest of new items from Internet Archive occurs (weekly)

Version Control

There will be no version control, or maintaining of prior versions, of the text files in this phase. This may be added at a later date.

Outstanding Questions

Answers shown below in BLUE were decided upon during the June 4, 2018 weekly Tech Conference Call.

- The sample files delivered by Katie Mika contain “[margin]” and “[delete]” markup tags. Should markup be removed before files are uploaded, or should the upload process handle it?
    - If the upload process is handling markup, should it ignore it, remove it, or reject files with markup?
    - ANSWER: Ignore markup. Leave it visible in UI. We can revisit later if we can need to.
- Should "pageocr" URLs (for example: biodiversitylibrary.org/pageocr/1000) be deprecated in favor of "pagetext" URLs? If so, all requests for "pageocr" would be redirected to "pagetext".
    - ANSWER: Change pageocr to pagetext and provide a redirect. Notify Rod Page and others who might be using it.
- Does there need to be a way to view logs/history **by individual page** on the admin site, or is it enough that the data can be retrieved from the database if needed?
- How will the output of the Purposeful Gaming project be accommodated? The outputs from that project are simple text files named with a value that maps to the "FileNamePrefix" value in the BHL Page table (example: americanbirdmaga51905worc_0129.txt).
    - Should this be a separate one-time import, that both loads the files and records all necessary log values?
    - ANSWER: Create a one-time import to replace the OCR files with the game output and create the necessary log entries.

- What if a book's OCR is regenerated due to "page inserts" after a transcription has been contributed?
    - How to prevent overwriting transcriptions?
    - Do we need a flag to prevent ingest of OCR from IA, or at least to alert user that ingesting from IA might be a bad idea?
    - Should OCR ingests from IA only be allowed if the source of the existing text is OCR/IA? (Never overwriting text from any other source?)
    - ANSWER: If the source of the existing text is **not** IA OCR, then warn the user before allowing the text to be replaced with IA OCR.

Timeline

Dates	**Activity**
Mid-June	Mike will contact Ricc, Susan, and Joe to request transcription test sets
Mid-June – Early July	Implement transcription import functionality
Early July	Delivery of transcription test sets by Ricc, Susan, and Joe, as well as documentation of any data cleaning that was done
Early July – Late July	Beta testing: Ricc, Susan, Joe, and Carolyn
Late July or Early Aug	Release to Production

Additional Documentation

- https://bhl.wikispaces.com/Transcriptions+Planning+2017
- https://bhl.wikispaces.com/Transcriptions%20Task%20Group