On September 8th, 2015 a group of people involved in Transcriptions (William UIate, Trish Rose-Sandler, Mike Lichtenberg, Julia Blase, Carolyn Sheffield, Joe DeVeer. Excused: Martin Kalfatovic) had a Conference call to talk about transcription in BHL, particularly:

1) How to incorporate current transcriptions
2) How to proceed in the future with new transcriptions

We have several sources of transcriptions – FromThePage (Purposeful Gaming), DigiVol (BHL Australia & Purposeful Gaming) & Smithsonian Transcription Center.

One of the topics discussed was the usage of Markup. Australia, for example, has been marking up taxonomic and geographic names. The Purposeful Gaming project is not making use of markup tags but they are including them, so exports are just plain text .csv files.

For handwritten content there are OCR files in BHL but they are mostly empty files.

Currently, Mike has a workflow for transcriptions from PG but not the other sources.

Users would like to upload the transcriptions they have. They are mainly asking where to put the export files so it shows up in the BHL portal where OCR currently displays.

For full text Search, we would have to exclude words in marked up content from indexing. To deal with current markup would require development resources to be committed to it. We can deal with the non-marked up text now w/o too much.

Future: Standarize Markup.
Need to define requirements for the future: how do we want folks to markup documents? Mining Biodiversity is currently marking up entities and relationships on the texts and expect to provide this to BHL.

For now, focus on the OCR we have right now and bring in plain .csv data. For the future, consider what type of transcription tool do we want? Again, our colleagues from Mining Biodiversity could contribute to this with their tool: Argo, as an installation or as a service.

We don’t currently have a version control for OCR either in IA or in BHL. We can’t just go into the Admin dashboard and replace an existing file. But we could add this functionality.

Any changes in IA’s OCR files will not be reflected back to BHL and any changes to BHL’s OCR files will not be reflected back to BHL.

We need to keep Joel in the loop about our decisions. We should also include Nicole Kearney at BHL Australia in the conversations since she manages their transcriptions in DigiVol

Next steps

– Mike will build the functionality into the Adm portal so anyone can replace the existing OCR. He will need to review existing transcriptions and come up with list of questions (about 2 weeks)

-We will need to have future discussions about which transcription tools we want to continue to support beyond the projects and what markup guidelines we want to develop so there is consistency.