**11/4/15 Transcription discussion**

Mike L, William, Trish, Bianca, Joel, Martin, Joe, Carolyn, Lesley

Nicole and BHL Australia’s approach- approach of loading in BHL so far has been loading one Volume with image and separate volume with transcription. Joe Coleman at Australia suggested having them side by side but Australia doesn’t want to impose further development on the BHL book viewer

Martin says Creating books from parts of books is not desirable by BHL community for standard published books but maybe different for unpublished texts. Martin wants a consistent workflow and display of transcriptions. Would like transcriptions to appear where OCR currently displays.

Bianca also would like transcriptions to display where OCR is and knows that is what our users want.

Carolyn agrees because added features could be applied to the transcriptions.

What about the layout of transcriptions ie. Tagging? Not in terms of blocks but line by line. Actual text is not in blocks.

Martin says Without a central transcription tool and process we will always have multiple sources of transcriptions and different rules for creating them.

Lesley said she likes being able to put them through the IA workflow because can do a text search in IA and that functionality currently isn’t available in BHL

Joe’s experience with transcribing indicates some types of markup is important for things like tables/charts that happen in hort catalogs.

Martin’s concern is scale – markup whether tagging entities or indicating layout is maybe not realistic to do for everything we will have. Its More important to have basic text transcriptions for full text searching. That would align with Our current OCR which doesn’t have semantic markup.

Wiiliam says Whatever decision we make it will be hard to roll back and change later.

Martin likes model that Lesley has done – sep volumes of page images and transcription. He does’t care for Australia model which creates Frankenbooks. What about Hybrid option - including pdf of transcript and putting in transcription into OCR field?

Mike what are immediate problems with ingesting page level transcriptions? Replacing OCR with transcriptions? Mikes said that could be done very easily. We could do a test with 1 volume.

Why do we need the pdf version? Sometimes can Help with viewing of layout. IF downloading OCR for whole book that comes from IA.

Mike wonders about Duplicate of a volume thats already there - how do we handle duplicates in APIs and searches.? Introduces some interesting challenges.

Back to Mike’s question – transcript as alternate download which is different than how they currently work.

If we download do we store the IA file? Yes could be sep file in same folder as scans.

We agreed on:

Everyone likes idea of transcription appearing in place of OCR for field books. Should we prioritize this first then look into PDF of transcription after that? Yes.

Martin asked What is workflow for that and technical issues with just replacing OCR with transcription? Mike has already started working on this with the PG project and transcriptions from DigiVol and FromThePage. Timeline - he should be able to work on this next week but not sure how long it will take.

Martin suggested talking to Australia to let them know we are working on this process and would prefer not to have the FRankenbooks uploaded right now til we figure this out. William said Nicole is already in process of uploading some new stuff so we should ask them to put on hold for now. Australia mentioned involved in a collection move for next 2 months and not sure how it will impact their transcription work.

Next steps:

We will update progress through tech meetings and keep this group up to date via email