Meeting with Tiltfactor

2014/8/7

Present: Trish, William, Mike L, Suki, Max, designer, intern developer

**Agenda**

* Tiltfactor team will research the feasibility of automating the gathering of locative data for words in manually transcribed manuscripts (or horticultural catalogs with nonstandard strucutres - e.g. two-column layout) and come up with a recommendation for proceeding.

Max Investigated some options

Text image linking tool – automated tool in development. Could use to translate manuscripts similar to OCR. Not sure if ready in time.

Text scribe

Tile

Once MOBOT reviews those tools we will need to make a decision on which games

Max will send a list of options

Trish asked is there an Absolute date to make decisions on which games to go with? Not sure yet. Could start with game you definitely know you want (e.g. for BHL audience) then decide on the other later but it would good not to wait too long because it will determine what data gets stored in the backend database.

* Tiltfactor will continue to work with the Dartmouth contracts office to finalize the contract.

Max will send an update once he talks with Mary and Christine

* BHL will provide TIltfactor example scanned manuscript pages and data files with the manual transcriptions.

Mike Provided sample of 3 transcribed pages from ALA and FromThePage

* BHL will send to Tiltfactor input format and desired backend output data format.

Mike provided 3 sample JSON files

Attached are three files containing \***very preliminary**\* possibilities for game inputs and outputs.

GameInput.json is fairly straightforward; it is basically what we discussed and drew up on the white board.  One thing to note (and to point out to TiltFactor) is that both OCR outputs for the first and last differences contained in that file are wrong.  Depending on how the game works, the output should either simply indicate that neither OCR output was correct or it should contain a gameplayer-entered correct value.

Both of the proposed game outputs assume that we will only want to get the corrected values back, and not any of the other game-generated metadata (number of votes for each value, etc).  Both are similar to the input, except that image URLs and word coordinates do not exist in the output.

GameOutputOption1.json is a proposed output that includes gameplayer-entered values for the first and last difference in the input.

GameOutputOption2.json is a proposed output that contains a true/false field named “other” for the first and last difference in the input, which indicates that some value other than the OCR outputs is correct.

Tiltfactor asked how likely is it that the 2 OCR output coordinates are the same?

Mike is not absolutely sure but should agree within a pixel or two so not a huge issue.

Mike L looked at Tesssaract and their coordinates match those from IA but coordinates from JP2 not the same. William wondered do we need 2 sets of coordinates for each word? He thinks One set of coordinates should work fine. Trish clarified the use of the coordinates 1) will allow Tiltfactor to create images of words from the page to display to users in the game 2) will allow MOBOT to recompile the page. Is it necessary for them to agree to be able to do the later? Mike L said h will use ids to put page back together so no needed for that. *Conclusion – they will mostly agree but when they don’t not really an issue.*

Tiltfactor asked about scan quality and how slanting affects OCR? William says it will Mess with OCR. Does OCR it only use rectangles (i.e. 4 coordinates)? Yes we believe so. How many points do we need? If we have 5 could do polygons. Mike said Prime OCR has 5th coordinate but it it is Avg height of characters, not sure for ABBYY

How do we want to deal with capital differences? Capitalization doesn’t matter for searching and mining names. For readability could be useful. *William said Caps is like italics, bold so Tiltfactor can ignore those differences.*

Output (data passed from Tiltfactor back to MOBOT)

They asked is Passing back what is more correct better than passing back neither are correct?

How does Tiltfactor determine levels of confidence? Based on Number of users that vote for an option.

William things getting number of votes back would be useful info. Tiltfactor cautions against giving us raw number of votes but better to return to us confidence scores. These would be based on number of users that agreed on a word (we would have to determine that threshold)

From a design perspective – 2 options plus 4 we know are bad. Good to give user feedback as to how they are doing

*Decision: TIltfactor Will give us best answer they have and a confidence level and then we decide how to incorporate it into BHL.*

Can allow user to skip giving an answer. Is something better than nothing?Allowing users to type their own answer is not a good option for gamers – Tiltfactor won’t know if their answer is correct or not so can’t reward them by points. But for the non gaming audience who is not motivated by points or rewards typing in an answer can be effective.

For gamer game we would prefer to force them to choose one of the options even if neither is very good (i.e. if none of the above are correct choose the best option). Mike would prefer to get feedback from a person than rely on OCR interpretation.

Immediate reward in gaming is usual better than delayed reward – technically and psychologically difficult

William explains the 20% of alpha numeric characters –

What to do with periods or semicolons? Tiltfactor has some design solutions for this through known bad insertations - it can catch some of these

Only having seen sample input for the game Tiltfactor wants to know How many differences will we they see with actual OCR output? Trish explained that a lot of the worst stuff will not get submitted to the game because we will first filter out those with more than 20% non alpha-numeric characters. Tiltfactor would like to see Examples of worst stuff they will have to deal with. MOBOT can look for those.