**PDF Generator Analysis Final Report**

August 2, 2011

Attendees: Gilbert Borrego, Grace Costantino, Bianca Crowley, Sue Graves, Trish Rose-Sandler

**Stats**

38% of all pdfs created from the BHL portal have some user-generated metadata (i.e. at least one of title, author, keyword)

From Jan 2010-Apr 2011 there were 24,450pdfs that went into Citebank (actually we identified a bug in the harvestor that wasn’t indexing approx 7000 pdfs but that has been fixed)

We have 5 reviewers who looked at 945 out of 24,450pdfs (3.9%)

We rated title, author and true article with a rating of 1-3 (1 being the best)

Titles Average 1.68

Author Average 1.33

True article 1.41

Overall Average rate (combining title, author, and true article) 1.47

Average rate when considering just Title and Author 1.50

**Goals**

* Get more users to generate articles with quality metadata; we’d rather have users enter no metadata than enter poor quality metadata
* Encourage users to share articles within a common repository, i.e. CiteBank
* Refine PDF generator UI to satisfy this goal
* Leave PDF generator UI flexible enough to allow users to generate non-traditional articles\*\* as well as “true articles”\*

\*True Article: A self-contained nonfiction prose composition on a fairly narrow topic or subject, written by one or more authors and published under a separate title in a collection or periodical containing other works of the same form, i.e. the cite-able unit.

VS.

\*\*Article: A generic term used to describe a complete, self-contained composition on a topic written by one or more authors. Whether it’s an article or sub-article it should contain the complete intellectual content of that unit. Species descriptions and taxonomic treatments fall within this category.

**Discussion Points**

* As of 6/15 there are 26,548 articles w/ metadata of 69,293 = 38% of articles created have some kind of metadata associated with them
* At some point in the UI, ask users if they would like to share their article to a common repository OR receive a permanent link to their article in CB…
* Could we embed the user generated article metadata on the coversheet of the PDF they receive or in the metadata of the file itself? (Now that I think about this more, this may create a prohibitive delay in the delivery PDF but it’s worth running by Mike anyway).
* Useful to embed copyright status information on the PDF coversheet
* Would be useful to have a stricter definition of what CiteBank is to make determinations about the kinds of user generated content we want in the CiteBank collection
* How many people click on the PDF tips guide link? Are these instructions even useful?
* Interesting to do analysis to see if there is a corollary of between email extensions and the accuracy of metadata OR getting a country breakdown…
* Could we give users stats on how many articles they generated within the CB UI?
* Should we allow users to edit their metadata in citebank? How would we enable this so that Citebank knows they are the “owner” of that record?

**Preliminary Recommendations to change in UI**

* Make the article title metadata a required field
* Send users an automated email with a link to their content in CiteBank when it becomes available; understand there will be a delay between the pdf email and the CB link email but this is OK

Embed the title and author into the pdfs metadata header (according to Mike this is already being done)

* + - * Add author and rights md to cover page of pdf
* Throw out current BHL PDF articles in CiteBank that do not have title metadata OR where title metadata consists of a single character; 618 pdfs in the spreadsheet of 4/20/2011 had no title metadata provided
* Have better instructions for users re: entering author metadata in the way that corresponds to CiteBank’s requirements, i.e. Lastname, Firstname; Lastname, Firstname
* The number of articles generated with missing page images is not significant enough to warrant a rejection of these articles from entry into CiteBank; currently only 2.4% of articles generated have missing page images

**Working proposal for new PDF generator UI workflow:**

1. User asked to enter email, share with, PDF w/ OCR or not
2. Would you like to contribute your article to a common repository / receive a permanent URL to link to your article? YES or NO
3. If Yes, then enter metadata screen
4. Then go to select pages screen
5. If No, then skip to directly to select pages screen

**Next Steps:**

* + Do mockups of new UI
  + Get feedback from heavy pdf users
  + Research other examples of user generated metadata creation – Is it realistic to turn our users into catalogers to see if this is really useful; “social cataloging”, Open Library?