### To: UT System Directors

### From: Ja-Nice Woolaver, Bibliographic Database Manager, University of Texas Libraries, UT Austin

### Date: Monday, April 27, 2015

# RE: Procedures for HathiTrust Extraction

# File creation

Each School should create three types of files; serials, single-part (monograph) and multi-part (monograph). All print holdings regardless of publication date. Our largest review file is 250,000. Each of our files will be in those increments, and then we will concatenate them. If we use email, we can only send 50M at a time. So I think it is preferable to use drop box. All files need to be saved as text files tab delineated with extension tsv. To prevent extraneous characters, please avoid using Excel or Word. I will be creating our files as txt then renaming after the fact to tsv. For example:

Utexas\_single-part\_date.tsv or UTSA\_single-part\_date.tsv for University of Texas at San Antonio.

The submitted files will be named as follows, one file per type:

utsystem\_single-part\_<date>.tsv

utsystem\_multi-part\_<date>.tsv

utsystem\_serials\_<date>.tsv date in format yyyymmdd

For those of you who have a drop box account, I will share a folder called HathiTrust with you and you can place your files there. Please send me an email because this is just my personal account and I will need to keep moving the files out. If you are not able to change your bib numbers to have a prefix, or the holdings status to match HathiTrust’s then email me and we can work out a plan to do that once your files reach UT Austin. Also if you do not have the mechanism to join all your information into just 3 files, we can do that here as well.

Below is how files should be formatted. For reference, the requirements by Hathi on how data should be formatted are located on the following webpage:  
http://www.hathitrust.org/print\_holdings

# For Serial files

Column 1 OCLC number(s) (with or without prefix, UT Austin does not keep that information.)

Column 2 Local bib number (for UT Austin b36499183, we will convert to ut.b36499183 (utsa.local number)

Column 3 ISSN

Column 4 Government Document 0=not govdocs 1=govdocs (we will use 008 byte for older titles at least, you your best judgment to provide this information)

*The first two columns are required*

DATA EXAMPLES:

01039883 ut.b17285860 0031-1723 0  
05252589 ut.b61886269 0 *(NOTE: include third column tab even if it is blank)*  
(OCoLC)02240015 utsa.102 0095-0033 0  
(OCoLC)ocm39370286 utd.b9992997801421 0  
10761886 utep.b10004087 0747-5535 1  
ocm02241027 utmda.232204 0098-311X 0  
(OCoLC)ocm04839041,(OCoLC)4839041 utsod.232210 0019-3992 0 *(NOTE: multiple OCLC numbers should be listed with commas and no spaces separating them in the first column)*

# For Mono-single part files

Column 1 OCLC number (with or without prefix, UT Austin does not keep that information.)

Column 2 Local bib number (for UT Austin b36499183, we will convert to ut.b36499183 (utsa.local number)

Column 3 Holding status (ours will be taken from our Sierra fixed field and converted to CH=current, WD=withdrawn and LM=lost or missing)

Column 4 Condition (UT Austin cannot supply this information but will leave a blank spot for it) BRT= brittle

Column 5 Government Document 0=not govdocs 1=govdocs (we will use 008 byte for older titles at least, use your best judgment to provide this information)

*The first three columns are required*

DATA EXAMPLES:

(OCoLC)03750927 uta.14 CH 0 *(NOTE: include the fourth column tab even if it is blank)*  
(OCoLC)08864899 utd.b993101421 CH 1  
10696482 utep.b10000069 CH 1  
8961927 ut.b1000077x CH 0  
(OCoLC)ocm02645404 utmb.200 CH 0  
ocm02633009 utmb.280 CH 0

# For Mono-multi part files

Column 1 OCLC number(s) (with or with prefix, UT Austin does not keep that information.)

Column 2 Local bib number (for UT Austin b36499183, we will convert to ut.b36499183 (utsa.local number)

Column 3 Holding status (ours will be taken from our Sierra fixed field and converted to CH=current, WD=withdrawn and LM=lost or missing)

Column 4 Condition (UT Austin cannot supply this information but will leave a blank spot for it) BRT= brittle (if available)

Column 5 Item-specific enumeration

Column 6 Government Document 0=not govdocs 1=govdocs (we will use 008 byte for older titles at least, use your best judgment to provide this information)

*The first three columns are required*

DATA EXAMPLES:

(OCoLC)01255964 uta.667 CH v.2 0 *(NOTE: include fourth column tab even if blank)*  
ocm01778915 utmb.81009 CH 0 *(NOTE: include fourth and ffrth tab even if blank)*  
OCoLC)ocm56703204b utmb.286729 CH 0  
1150082,24858839 ut.b27605231 CH 1926 1  
2948516 ut.b27605309 CH PLATES 0  
1159532,3109889,4400198,1678805,15188981,17754596,21881936,20465283 ut.b2760567x CH SER.1 0 *(NOTE: multiple OCLC numbers should be listed with commas and no spaces separating them in the first column)*  
(OCoLC)02674103 utsa.24635 WD 1  
(OCoLC)ocm00026908 utsph.7487 CH no. 31 suppl.,1976 0  
7545948,1877011,16571812,20080266,20079915,14445283 ut.b20755909 CH NO.7 9-10 17-18 22-23 1