CHECKLIST FOR CATALOGING TEXT DATASETS

INTRODUCTION

This document outlines procedures for cataloging text datasets, including remote and direct electronic formats. A text dataset is a collection of primarily text data used for natural language processing and analysis. Data sets for single publications (e.g. a newspaper or journal) can be cataloged following provider-neutral (p-n) guidelines. The publication information of the original source publication should be retained while information describing a specific instance (e.g. aggregator, format etc.) is omitted.

FIXED FIELDS	*Required fields are marked in bold. Instructions in orange are local Yale practices)*
Field	**Coded Data**
Leader	Type of record: Leader/06: • a=Language material Bibliographic Level Encoding level: Cataloging Form: i=ISBD
006 Additional Material Characteristics	m=Computer File • Form of Item: o o=Online o q=Direct electronic *(for CD-ROMs, etc.)* • Type of File: o d=Document o e=Bibliographic data
007 Physical Description	c=Electronic Resource • Specific Material Designation: o b=Chip cartridge *(for flash drives)* o e=Disc cartridge, type unspecified *(for external hard drives)* o o=Optical disc *(for CD-ROMs, etc.)* o r=Remote *(for online format)* • Dimension: o g=4 3/4 of 12 cm *(for CD-ROMs, etc.)* o n=Not applicable *(for online format)* o □=No attempt to code *(for external hard drives, USB flash drives, etc.)*
008 General Description	Publication Status Date 1/Date 2 Place of publication Form of Item: • o=online *(for online format)* • q=direct electronic *(for CD-ROMs, DVD-ROMs, external hard drives, USB flash drives, etc.)* Contents (selected): • b=Bibliographies • c=Catalogs • d=Dictionaries

	• e=Encyclopedias • f=Handbooks • □=No attempt to code Literary Form (selected): • 0=Non fiction (not further specified) • 1=Fiction (not further specified) Biography Language Cataloging source: d=Other

VARIABLE FIELDS	*Required fields are marked in bold. Instructions in orange are local Yale practices)*	
Field	**Field Text**	**Notes**
040 Cataloging Source	040 // |a CtY |b eng |e rda |c CtY	*(for original cataloging)*
	040 // |a CtY |b eng |e rda |e pn |c CtY	*(for p-n original cataloging)*
	040 // |d CtY	*(for copy cataloging)*
041 Language Code	041 x/ |a [Language code] |h [Language code of original]	
043 Geographic Area Code	043 // |a [Geographic code]	
050 LC Call Number	050 /4 |a [LC call number]	
090 Local Call Number	090 // |a yuldset	*(add to all datasets)*
	090 // |a yuldsetmediated	*(add to all mediated datasets)*
	090 // |a yuldsettxt	*(add to all text datasets)*
1xx Main Entry	1xx xx |a [Author or uniform title].	*(required when applicable)*
245 Title Statement	245 xx |a [Title] dataset : |b [subtitle], |f [inclusive dates of contents] |g [bulk dates of contents if applicable].	
246 Varying Form of Title	246 x/ |a [Variant title] dataset	
264 Publication Information	264 /1 |a [Location of publisher] : |b [Publisher], |c [Date].	
300 Physical Description	• 300 // |a 1 computer disc ; |c 4 3/4 cm + |e documentation • 300 // |a 1 external hard drive • 300 // |a 1 online resource • 300 // |a 1 USB flash drive	
336 Content Type	336 // |a computer dataset |b cod |2 rdacontent	*(for all datasets)*
	336 // |a text |b txt |2 rdacontent	*(for all text datasets)*
337 Media Type	337 // |a computer |b c |2 rdamedia	
338 Carrier Type	• 338 // |a computer disc |b cd |2 rdacarrier • 338 // |a online resource |b cr |2 rdacarrier • 338 // |a other |b cz |2 rdacarrier	*(for CD-ROMs, etc.)* *(for online format)* *(external hard drives, USBs, etc.)*
347 Digital File Characteristics *(include each type of subfield (|a, |b, |c) in a separate field)*	347 // |a text file |2 rdaft	
	• 347 // |b [Encoding format, e.g. PDF] |b [additional encoding format, e.g. XML] • 347 // |b [Encoding format, e.g. CD-ROM, etc.] |b [additional encoding format]	

	347 // |3 [Materials specified if available, e.g. compressed [specify format if more than one is present], uncompressed [specify format if more than one is present] |c [File size]	*(subfield 3 is not repeatable; use separate fields if necessary do not include for continuing resources)*
500 General Note	500 // |a [Additional notes, e.g. Accompanied by …, Documentation in README file, Title devised by cataloger].	
505 Formatted Contents Note	505 00 |t [Title 1] dataset (inclusive dates of contents for Title 1) -- |t [Title 2] dataset (inclusive dates of contents for Title 2) |t [Title 3] dataset (inclusive dates of contents for Title 3).	*(use catalogers' judgement to decide whether an enhanced contents note would be useful)*
506 Restrictions on Access Note *(if applicable)*	506 // |a Access restricted by licensing agreement and agreement to terms of use.	*(for resources that require staff mediation)*
	506 // |a Access restricted by licensing agreement.	*(for all other licensed datasets)*
520 Summary, etc.	520 // |a [Summary including keywords, e.g. "Dataset of materials for text data mining (TDM) …," etc.; include information on geographic, topical, and chronological coverage, record segmentation/granularity (e.g. by month, issue, and article), and information on digital formats, if available]	
538 System Details Note	538 // |a [System requirements if special software is required; do not make note of standard information or formats].	
588 Source of Description Note	• 588 0/ |a Description based on [source record, e.g. database, print] record. • 588 0/ |a Description based on online resource (viewed [Date]). • 588 0/ |a Title from [Location, e.g. file header, README file, etc.] (viewed [Date]).	
590 Local Note	590 // |a Access is available to the Yale community.	
6xx Subject Access	6xx x0 |a [Subject heading] ~~|v Databases~~.	*(retain subject headings of source database, deleting the form subdivision "Databases")*
		(see the Detailed Guidelines for additional subject access notes)
655 Genre/Form	655 /7 |a Text corpora. |2 lcgft	*(add to text datasets)*
	655 /7 |a Data sets. |2 lcgft	*(add to all datasets)*
	655 /7 |a [Additional genre heading(s) to reflect type of text] |2 lcgft	*(see the detailed instructions for additional information)*
7xx Added Entries	7xx xx |a [Corporate or personal name.]	
740 Dataset Collection Name	740 x/ |a [Cataloger supplied title, usually based on the collection title of the data source, e.g. Archives unbound (a database collection) → Archives unbound dataset collection].	*(required if applicable)*
76x-78x Linking Fields *(use catalogers' judgement to decide whether the*	776 08 |i [Relationship information, e.g. Also issued as, Online version, etc.]: |t [Title]. |h [Physical description, e.g. 1 computer optical disc] |w [Record control number]	*(additional physical format, especially if we retain it)*

information would be useful)	786 08 |i Based on (work): |t [Title] |w [Record control number]	*(link to source title, e.g. newspaper)*
	787 08 |i Related work: |t [Title] |w [Record control number]	
856 Electronic Locations and Access	856 40 |y Online dataset |u [URL]	*(for unmediated resources)*
	856 40 |y For data access contact researchdata@yale.edu |u mailto:researchdata@yale.edu?subject=data%20access%20inquiry%20from%20Quicksearch	*(for resources that require staff mediation)*
	856 42 |3 [Link text of related resource, e.g. Documentation, Manifest, etc.] |u [URL]	*(additional link for related resource)*
946 Local Data	946 // |a DO NOT EDIT. DO NOT EXPORT.	

MFHD	Field Text	Notes
852 Location	852 80 |b yulint |h None |z Online resource	*(for online format)*
	852 80 |b yulintx |h None |z Online resource	*(for online format hosted locally)*
	852 00 |b [Location code] |h [Classification part] |i [Item part]	*(for physical format)*
583 Action Note	583 0/ |a [Action, e.g. transformed digitally, transferred to optimal storage] |c [Time/date of action] |k [Action agent, e.g. Preservica] |2 [Source of term, e.g. pda] |5 [Institution to which field applies, e.g. CtY]	*(for materials with digital copies added to Preservica)*