DETAILED INSTRUCTIONS 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.

GUIDELINES FOR FIXED FIELDS

(Required fields are marked in bold. Instructions in orange are local Yale practices)

- **Leader:**
  - Type of record: Leader/06:
    - a=Language material
  - Bibliographic Level
  - Encoding level
  - Cataloging Form: i=ISBD

Leader

Record Status	c : Corrected or revised
Type of Record	a : Language material
Bibliographic Level	m : Monograph/item
Type of Control	_ : No specific type of control
Encoding Level	_ : Full level
Cataloging Form	i : ISBD punctuation included
Linked Record Requirement	_ : Not specified or not applicable

- **006 Additional Material Characteristics:**

  m=Computer File
  - Form of Item:
    - o=Online
    - q=Direct electronic *(for CD-ROMs, DVD-ROMs, etc.)*
  - Type of File:
    - d=Document
    - e=Bibliographic data
    - Ex. Online document dataset:

006 - Additional Material Characteristics (m - Computer File)

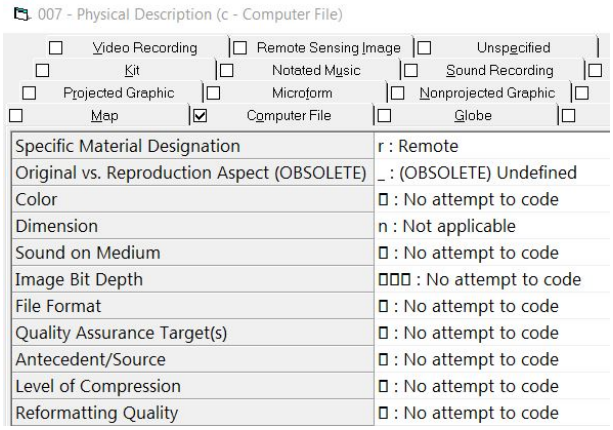
☐ Projected Medium	☐ Nonmusical Sound	☐ Musical Sound	
☐ Notated Music	☐ Manuscript Notated Mu...	☐ Printed Map	☐ Manuscript Map
☐ 3-D Artifact	☐ Continuing Resource	☐ Manuscript Lang.	☐ Books
☐ 2-D Nonprojectable	☑ Computer File	☐ Kit	☐ Mixed Material

Target Audience	_ : Unknown or not specified
Form of Item	o : Online
Type Of File	d : Document
Govt. Publication	_ : Not a government publication

- **007 Physical Description:**

  c=Electronic Resource

  - Specific Material Designation:
    - b=Chip cartridge *(for flash drives)*
    - e=Disc cartridge, type unspecified *(for external hard drives)*
    - o=Optical disc *(for CD-ROMs, etc.)*
    - r=Remote *(for online format)*
  - Dimension:
    - g=4 ¾. or 12 cm *(for CD-ROMs, etc.)*
    - n=Not applicable *(for online format)*
    - □=No attempt to code *(for external hard drives, USB flash drives, etc.*
  - Ex. 1: Online format

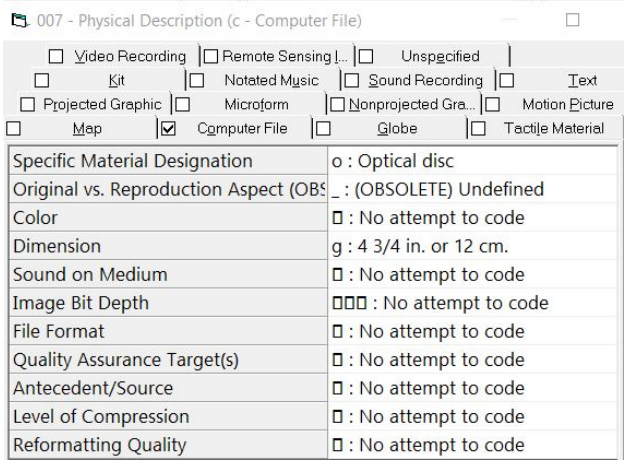


007 - Physical Description (c - Computer File)

☐ Video Recording	☐ Remote Sensing Image	☐ Unspecified
☐ Kit	☐ Notated Music	☐ Sound Recording
☐ Projected Graphic	☐ Microform	☐ Nonprojected Graphic
☐ Map	☑ Computer File	☐ Globe

Specific Material Designation	r : Remote
Original vs. Reproduction Aspect (OBSOLETE)	_ : (OBSOLETE) Undefined
Color	□ : No attempt to code
Dimension	n : Not applicable
Sound on Medium	□ : No attempt to code
Image Bit Depth	□□□ : No attempt to code
File Format	□ : No attempt to code
Quality Assurance Target(s)	□ : No attempt to code
Antecedent/Source	□ : No attempt to code
Level of Compression	□ : No attempt to code
Reformatting Quality	□ : No attempt to code

  - Ex. 2. CD-ROM or DVD-ROM format



007 - Physical Description (c - Computer File)

☐ Video Recording	☐ Remote Sensing I...	☐ Unspecified	
☐ Kit	☐ Notated Music	☐ Sound Recording	☐ Text
☐ Projected Graphic	☐ Microform	☐ Nonprojected Gra...	☐ Motion Picture
☐ Map	☑ Computer File	☐ Globe	☐ Tactile Material

Specific Material Designation	o : Optical disc
Original vs. Reproduction Aspect (OBS	_ : (OBSOLETE) Undefined
Color	□ : No attempt to code
Dimension	g : 4 3/4 in. or 12 cm.
Sound on Medium	□ : No attempt to code
Image Bit Depth	□□□ : No attempt to code
File Format	□ : No attempt to code
Quality Assurance Target(s)	□ : No attempt to code
Antecedent/Source	□ : No attempt to code
Level of Compression	□ : No attempt to code
Reformatting Quality	□ : No attempt to code

- **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
        - m=Theses
        - o=Reviews
        - r=Directories
        - □=No attempt to code
    - Literary Form:
        - 0=Non fiction (not further specified)
        - 1=Fiction (not further specified)
        - d=Drama
        - e=Essays
        - f=Novels
        - h=Humor, satires, etc.
        - i=Letters
        - j=Short stories
        - m=Mixed forms
        - p=Poetry
        - s=Speeches
        - u=Unknown
        - □=No attempt to code
    - Biography
    - Language
    - Cataloging source: d=Other
    - Ex.: Online dataset of essays

008 - General Description (Book)

Publication Status	m : Multiple dates
Date 1 (yyyy)	2010
Date 2 (yyyy)	2018
Place of Publication	miu : Michigan
Illustrations 1	_ : No illustrations
Illustrations 2	_ : No illustrations
Illustrations 3	_ : No illustrations
Illustrations 4	_ : No illustrations
Audience	_ : Unknown or not specified
Form of Item	o : Online
Contents 1	_ : No specified nature of contents
Contents 2	_ : No specified nature of contents
Contents 3	_ : No specified nature of contents
Contents 4	_ : No specified nature of contents
Govt. Publication	_ : Not a government publication
Conf. Publication	0 : Not a conference publication
Festschrift	0 : Not a festschrift
Index	0 : No index
Literary Form	e : Essays
Biography	_ : No biographical material
Language	eng : English
Modified Record	_ : Not modified
Cataloging Source	d : Other

GUIDELINES FOR VARIABLE FIELDS *(Required fields are marked in bold. Instructions in orange are local Yale practices)*

- **040 Cataloging Source:**
  - 040 // |a CtY |b eng |e rda |c CtY *(original cataloging)*
  - 040 // |a CtY |b eng |e rda |e pn |c CtY *(for p-n original cataloging)*
  - 040 // |d CtY *(copy cataloging)*
- 041 Language Code:
  - 041 x/ |a [Language code(s)] |h [Language code of original]
- 043 Geographic Area Code:
  - 043 // |a [Geographic code(s)]
- 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)*

090			ǂa yuldset
090			ǂa yuldsetmediated
090			ǂa yuldsettxt

- **1xx Main Entry** *(required when applicable)*
  - 1xx xx |a [Author or uniform title].
- **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

245	0	0	ǂa Psychological warfare and propaganda in World War II dataset : ǂb air dropped and shelled leaflets and periodicals, ǂf 1939-1945 ǂg 1942-1945.
246	3		ǂa Psychological warfare and propaganda in World War Two dataset
246	3		ǂa Air dropped and shelled leaflets and periodicals dataset

- **264 Publication Information:**
  - 264 /1 |a [Location of publisher] : |b [Publisher], |c [Date].

264		1	ǂa [Ann Arbor, Mich.] : ǂb [ProQuest Information and Learning Co.], ǂc [between 2002 and 2017?]

(Note: When deriving a record for an unseen dataset from the database record, bracket the database publication information and assign date range between database publication date and date the dataset was received.)

- **300 Physical Description**:
  - 300 // |a 1 computer disc (5 text files) ; |c 4 ¾ cm
  - 300 // |a 1 external hard drive (10 text files)
  - 300 // |a 1 USB flash drive (15 text files)
  - 300 // |a 1 online resource (approximately 1 million text files)

  (Note: Include number of files for closed sets if known but NOT for continuing resources.)
- **336 Content Types:**
  - 336 // |a computer dataset |b cod |2 rdacontent *(for all datasets)*
  - 336 // |a text |b txt |2 rdacontent *(for text datasets)*
- **337 Media Types:**
  - 337 // |a computer |b c |2 rdamedia
- **338 Carrier Type:**
  - 338 // |a computer disc |b cd |2 rdacarrier *(for CD-ROMs, etc.)*

- 338 // |a online resource |b cr |2 rdacarrier *(for online format)*
- 338 // |a other |b cz |2 rdacarrier *(for external hard drives, USB flash drives, etc.)*

- 347 Digital File Characteristics:

  (Note: 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 // |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]

    (Note: Subfield 3 is not repeatable; use separate fields if necessary; do not include for continuing resources)
  - Ex.: Online text dataset:

300			ǂa 1 online resource (approximately 430,000 files)
336			ǂa computer dataset ǂb cod ǂ2 rdacontent
336			ǂa text ǂb txt ǂ2 rdacontent
337			ǂa computer ǂb c ǂ2 rdamedia
338			ǂa online resource ǂb cr ǂ2 rdacarrier
347			ǂa text file ǂ2 rdaft
347			ǂb PDF ǂb XML
347			ǂ3 Compressed ǂc 14 GB
347			ǂ3 Uncompressed ǂc 16 GB

- 500 General Note:
  - 500 // |a Accompanied by ….
  - 500 // |a Documentation in README file.
  - 500 // |a 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).
  - Ex.: Collection of text data from multiple periodicals:

505	0	0	ǂt American Craft dataset (1965-2005) -- ǂt Apollo dataset (1925-2005) -- ǂt Architectural Review dataset (1896-2005) -- ǂt Architects' Journal dataset (1895-2005) -- ǂt ArtAsiaPacific dataset (1993-2005) -- ǂt Art Monthly dataset (1976-2005) -- ǂt British Journal of Photography dataset (1854-2005) -- ǂt C Magazine dataset (1983-2005) -- ǂt Canadian Architect dataset (1955-2005) -- ǂt Ceramics Technical dataset (1995-2005) --

  (Note: Add enhanced contents note when the dataset includes data from specific titles, using catalogers' judgement to decide whether the information would be useful.)
- **506 Restrictions on Access Note:** *(if applicable)*
  - 506 // |a Access restricted by licensing agreement and agreement to terms of use. *(for mediated resources)*
  - 506 // |a Access restricted by licensing agreement. *(for all other licensed datasets)*
- 520 Summary Note:
  - 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.)

520			ǂa Dataset of articles for text data mining (TDM) from English-language magazines in the fields of art and architecture dating from 1860-2005. Subjects covered include: fine art, decorative arts, architecture, interior design, industrial design, and photography. The set contains digital reproductions in JPEG and XML format.

- 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 Headings and Genre Terms:**
  - When deriving a dataset record from a database record, retain the original subject headings but delete the form subdivision ~~"|v Databases."~~
  - Provide additional subject access. These sample headings are not exhaustive:
    - 600 x0 |a [Individual person] |x Language.
    - 610 x0 |a [Corporate body] |x Language.
    - 630 /0 |a [Uniform titles of sacred works] |x Language, style.
    - 630 /0 |a [Uniform titles of secular works] |x Language.
    - 650 /0 |a [Individual war] |x Language.
    - 650 /0 |a [Class of person] |z [Geographic location] |x Language.
    - 650 /0 |a [Discipline] |x Language.
    - 650 /0 |a [Ethnic group] |x Language.
    - 650 /0 |a [Type of newspaper, e.g. Chinese American newspapers, etc.] |x Language].
    - 650 /0 |a [Type of periodical, e.g. Children's periodicals, etc.] |x Language.
    - 650 /0 |a [Type of publication, e.g. Business records, etc.] |z [Geographic location] |x Language.
    - 650 /0 |a [Language, e.g. English language] |x Written [language, e.g. "Written English"] |z [Geographic location].
    - 650 /0 |a [Language, e.g. English language] |x Government jargon.
    - 651 /0 |a [Geographic location, e.g. country, city, etc.] |x Language.
  - Genre headings:
    - 655 /7 |a Text corpora. |2 lcgft *(for all text datasets)*
    - 655 /7 |a Data sets. |2 lcgft *(for all datasets)*
    - 655 /7 |a [Additional genre heading(s) to reflect the type of text] |2 lcgft
      (E.g. Biographies, Business correspondence, Diaries, Interviews, Legislative materials, Newspapers, Periodicals, Personal correspondence, Promotional materials, Records (Documents), Travel writing, etc.)

- 7xx Added Entries:
  - 7xx x/ |a [Corporate or personal name].
- **740 Dataset Collection Name** *(required if applicable)*
  - 740 x/ |a [Cataloger supplied title, usually based on the collection title of the data source].

740	0		‡a ProQuest historical newspapers dataset collection.

- 76x-78x Linking Entry Fields:
  - 776 Additional Physical Form Entry:
    - 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]
  - 786 Data Source Entry:
    - 786 08 |i Based on (work): |s [Uniform title]. |t [Title]. |k [Series] |w [Record control number]

786	0	8	‡i Based on (work): ‡s Israelite (Cincinnati, Ohio : 1854). ‡t Israelite. ‡d 1854-1874 ‡w (OCoLC)ocm11975053

  - 787 Other Relationship Entry:
    - 787 08 |i Related work: |s [Uniform title]. |t [Title]. |k [Series] |w [Record control number]

- Ex.: One newspaper title is split into more than one dataset:

245	0	4	ǂa The Indianapolis star dataset, 1903-1922.
787	0	8	ǂi Related work: ǂt Indianapolis star dataset, 1923-2004 ǂw 14423775

- **856 Electronic Locations and Access:**
  - For unmediated resources:
    - 856 40 |y Online dataset |u [URL]
  - For resources that require staff mediation:
    - 856 40 |y For data access contact researchdata@yale.edu |u mailto:researchdata@yale.edu?subject=data%20access%20inquiry%20from%20Quicksearch
  - Additional link for related resource:
    - 856 42 |3 [Link text of related resource, e.g. Documentation, Manifest, etc.] |u [URL]

856	4	0	ǂy For data access contact researchdata@yale.edu ǂu mailto:researchdata@yale.edu?subject=data%20access%20inquiry%20from%20Quicksearch

- **946: Local Data:**
  - 946 // |a DO NOT EDIT. DO NOT EXPORT. *(required for licensed materials)*

MFHD

- 852 80 |b yulint |h None |z Online resource *(for online format)*

852	8	0	ǂb yulint ǂh None ǂz Online resource

- 852 80 |b yulintx |h None |z Online resource *(for online format hosted locally)*

852	8	0	ǂb yulintx ǂh None ǂz Online resource

- 852 00 |b [Location code] |h [Classification part] |i [Item part] *(for physical format)*

852	0	0	ǂb sml ǂh PJ6715 ǂi .B658 2018 DVD (LC)

- 583 Action Note *(for materials with digital copies added to Preservica)*
  - 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]

583	0		ǂa transformed digitally ǂc 2019 ǂk Preservica ǂ2 pda ǂ5 CtY