DETAILED INSTRUCTIONS FOR CATALOGING TEXT & IMAGE DATASETS

INTRODUCTION

This document outlines procedures for cataloging text and image datasets, including remote and direct electronic formats. A text dataset is a collection of primarily text data used for natural language processing and analysis. An image dataset is a collection consisting primarily of images or videos used for tasks such as facial, action, character, handwriting, and object detection and recognition. Projected media is included here for when it is evaluated for its visual characteristics; when evaluating for sound characteristics, see the document on speech/sound datasets. Bibliographic records for datasets based on published databases should be derived from the bibliographic record for the source database.

GUIDELINES FOR FIXED FIELDS

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

- **Leader:**
  - Type of Record: m=Computer file
  - Bibliographic Level:
    - m=Monograph/item *(treat as monograph unless it is specifically noted to be a continuing resource)*
    - i=Integrating resource
    - s=Serial
  - Encoding level: _=Full level
  - Cataloging Form: i=ISBD

Record Status	c : Corrected or revised
Type of Record	m : Computer file
Bibliographic Level	m : Monograph/item
Type of Control	_ : No specific type of control
Encoding Level	_ : Full level
Cataloging Form	i : ISBD punctuation included
Multipart resource record level	_ : Not specified or not applicable
Length of the length-of-field portion	4 : Number of characters in the length-of-field portion of a Directory entry
Length of the starting-character-position portion	5 : Number of characters in the starting-character-position portion of a Directory entry
Length of the implementation-defined portion	0 : Number of characters in the implementation-defined portion of a Directory entry
Undefined	0 : Undefined

- **006 Additional Material Characteristics:** *(add 006 for type(s) of resource)*
  - a=Books:
    - Form of Item:
      - o=Online
      - q=Direct electronic *(for CD-ROMs, DVD-ROMs, etc.)*
    - Literary Form (selected):
      - 0=Non-Fiction
      - 1=Fiction
      - d=Drama
      - e=Essays
      - i=Letters
      - p=Poetry
      - s=Speeches

- g=Projected medium:
  - Form of Item:
    - o=Online
    - q=Direct electronic *(for CD-ROMs, DVD-ROMs, etc.)*
  - Type of Material:
    - f=Filmstrip
    - m=Moving image
    - v=Videorecording
- k=Two-dimensional non-projectable graphic: *(images)*
  - Form of Item:
    - o=Online
    - q=Direct electronic *(for CD-ROMs, DVD-ROMs, etc.)*
  - Type of Material: (selected)
    - c=Art reproduction
    - i=Picture
- s=Continuing resource
  - Type of Continuing Resource: (selected)
    - n=Newspaper
    - p=Periodical
  - Form of Item:
    - o=Online
    - q=Direct electronic *(for CD-ROMs, DVD-ROMs, etc.)*
  - Nature of Entire Work (selected):
    - f=Handbooks
    - o=Reviews

- 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, DVD-ROMs, etc.)*
    - r=Remote *(for online format)*
  - Dimension:
    - g=4 ¾. or 12 cm *(for CD-ROMs, DVD-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

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:**
  - Publication Status
  - Date 1/Date 2
  - Place of publication
  - Form of Item:
    - o=Online
    - q=direct electronic *(for CD-ROMs, DVD-ROMs, external hard drives, USB flash drives, etc.)*
  - Type of File:
    - c=(chiefly) Representational *(image datasets, including moving images)*
    - d=(chiefly) Document *(text datasets)*
    - e=(chiefly) Bibliographic data
    - (do not use ~~m=Combination~~)
  - Language
  - Cataloging source: d=Other

- Ex. 1: Online text dataset

008 - General Description (Computer)

Publication Status	m : Multiple dates
Date 1 (yyyy)	2007
Date 2 (yyyy)	2018
Place of Publication	miu : Michigan
Frequency (OBSOLETE)	_ : (OBSOLETE) No determinable frequency
Regularity (OBSOLETE)	_ : (OBSOLETE) Not applicable
Audience	_ : Unknown or not specified
Form of item	o : Online
Type of File	d : Document
Govt. Publication	_ : Not a government publication
Language	eng : English
Modified Record	_ : Not modified
Cataloging Source	d : Other

- Ex. 2: Direct electronic resource image dataset

008 - General Description (Computer)

Publication Status	m : Multiple dates
Date 1 (yyyy)	2013
Date 2 (yyyy)	2018
Place of Publication	miu : Michigan
Frequency (OBSOLETE)	_ : (OBSOLETE) No determinable frequency
Regularity (OBSOLETE)	_ : (OBSOLETE) Not applicable
Audience	_ : Unknown or not specified
Form of item	q : Direct electronic
Type of File	c : Representational
Govt. Publication	_ : Not a government publication
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 // |d CtY *(copy cataloging)*
- 041 Language Code:
  - 041 X/ |a [Language code(s)]
- 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 yuldsetimg *(add to all image datasets)*

- 090 // |a yuldsettxt *(add to all text datasets)*

090			ǂa yuldset
090			ǂa yuldsetimg
090			ǂa yuldsettxt

- 1XX Main Entry
- **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 3/ |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:**
  - *(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.)*
  - 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?]

- **300 Physical Description**: ****(include number of files for closed sets if known but NOT for continuing resources)****
  - 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 (1 million text files)
  - 300 // |a 1 online resource *(if number of files is unknown or resource is incomplete)*
- **336 Content Types:**
  - 336 // |a computer dataset |b cod |2 rdacontent *(for all datasets)*
  - 336 // |a text |b txt |2 rdacontent *(for text datasets)*
  - 336 // |a still image |b sti |2 rdacontent *(for image-heavy datasets)*
  - 336 // |a three-dimensional moving image |b tdm |2 rdacontent *(for 3D film)*
  - 336 // |a two-dimensional moving image |b tdi |2 rdacontent *(for standard film)*
- **337 Media Types:**
  - 337 // |a computer |b c |2 rdamedia
- **338 Carrier Type:**
  - 338 // |a computer disc |b cd |2 rdacarrier *(for CD-ROMs, DVD-ROMs, etc.)*
  - 338 // |a online resource |b cr |2 rdacarrier *(for online versions)*
  - 338 // |a other |b cz |2 rdacarrier *(for external hard drives, USB flash drives, etc.)*
- 347 Digital File Characteristics: *(include each type of subfield (|a, |b, |c) in a separate field)*
  - 347 // |a image file |2 rdaft
  - 347 // |a text file |2 rdaft
  - 347 // |b [Encoding format] |b [additional encoding format]
  - 347 // |3 [Materials specified if available, e.g. compressed [format], uncompressed [format]] |c [File size] *(do not include for continuing resources)*

- Ex. 1: Online text dataset:

300			ǂa 1 online resource (approximately 6 million text 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			ǂc 215.25 GB

- Ex. 2: Online image dataset with file size:

300			‡a 1 online resource (approximately 6 million image files)
336			‡a computer dataset ‡b cod ‡2 rdacontent
336			‡a still image ‡b sti ‡2 rdacontent
337			‡a computer ‡b c ‡2 rdamedia
338			‡a online resource ‡b cr ‡2 rdacarrier
347			‡3 uncompressed thumbnail ‡a image file ‡2 rdaft
347			‡3 uncompressed thumbnail ‡b TIFF

- Ex. 3: Image dataset for moving images on DVD-ROM:

300			ǂa 6 DVD-ROMS ; ǂc 4 3/4 in
336			ǂa computer dataset ǂb cod ǂ2 rdacontent
336			ǂa two-dimensional moving image ǂb tdi ǂ2 rdacontent
337			ǂa computer ǂb c ǂ2 rdamedia
338			ǂa computer disc ǂb cd ǂ2 rdacarrier
347			ǂa video file ǂ2 rdaft
347			ǂb dvd video

- 500 General Notes:
  - 500 // |a Accompanied by ….
  - 500 // |a Documentation in README file.
- 505 Formatted Contents Note:
  - *(add enhanced contents note when the dataset includes data from specific titles)*
  - 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. 1: Newspaper title that underwent several title changes:

505	0	0	ǂt New-York tribune dataset (1841-1842; 1866-1924) -- ǂt New-York daily tribune dataset (1842-1866) -- ǂt New York herald, New York tribune dataset (1924-1926) -- ǂt New York herald tribune dataset (1926-1962).

  - Ex. 2: Collection of text data from specific 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) --

- **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:
    - E.g.: "Dataset of materials for text data mining (TDM) …", "Dataset of articles for text data mining (TDM) …., " "Dataset of materials for optical pattern recognition…", etc.
    - *(Note: 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:
    - *(Note: use only if special software is required; do not make note of standard information.)*

538			ǂa System requirements: DjView software.

- 583 Action Note:
    - *(Note: use 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

- **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]); title from [Location]
    - 588 0/ |a Description based on [source of description]; title devised by cataloger.
- **590 Local Note:**
    - 590 // |a Access is available to the Yale community.
- **6XX Subject Access Headings and Genre Terms:**
    - *(Note: when deriving a dataset record from a database record, retain the original subject headings but delete ~~ǂv Databases.")~~*
    - Text Datasets:
        - Provide subject access. These are sample headings; they may not all apply:
            - 600 X0 |a [Individual person] |x Language.
            - 610 X0 |a [Corporate body] |x Language.
            - 630 /0 |a [Uniform title] |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 [Type of periodical, e.g. English newspapers] |x Language].
            - 650 /0 |a [Type of publication, e.g. Government publications] |z Location |x Language].
            - 650 /0 |a [Language, e.g. English language, Indo-European languages] |x Government jargon.
            - 650 /0 |a [Language, e.g. English language] |x [Form of communication, e.g. "Written English" or "Spoken English"] |z [Geographic location].
        - 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 type of text] |2 lcgft
  (e.g. Biographies, Business correspondence, Diaries, Interviews, Legislative materials, Literature, Newsletters, Newspapers, Oral histories, Periodicals, Personal correspondence, Personal narratives, Promotional materials, Records (Documents), Religious materials, Textbooks, Tracts (Ephemera), Travel writing, Trial and arbitral proceedings, etc.)

o Image Datasets:

- Provide subject access. These are sample headings; they may not all apply:
  - 650 /0 |a [Class of person, e.g. Older people or Women] |x Identification.
  - 650 /0 |a [Topical heading, e.g. Street names or Shop signs] |x Identification.
  - 650 /0 |a Optical pattern recognition.
  - 650 /0 |a Human face recognition (Computer science)
  - 650 /0 |a Emotion recognition.
  - 650 /0 |a Human activity recognition.
- Add the following genre headings to all image datasets:
  - 655 /7 |a Pictures. |2 lcgft
  - 655 /7 |a Data sets. |2 lcgft
  - 655 /7 |a [Additional genre heading(s) to reflect type of image] |2 lcgft
    (e.g. Cartoons (Humor), Illustrated works, Motion pictures (and specific types of motion pictures), Pattern books, Theater announcements (Motion pictures), etc.)

- 710 Corporate Name:
  - 710 X/ |a [Corporate body, including author or publisher. Do NOT include field for a specific Yale Library *location*].
- **740 Dataset Collection Name:** *(required if applicable)*
  - 740 X/ |a [Dataset collection title based on vendor's database collection not YUL location].

740	0		ǂa ProQuest historical newspapers dataset collection.

- 76x-78x Linking Entry Fields: *(use catalogers' judgement to add when applicable and useful*):
  - 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	ǂi Based on (work): ǂs Newsday (Nassau edition). ǂt Newsday. ǂb Nassau edition ǂw (OCoLC)ocm05371847

  - 787 Other Relationship Entry:
    - 787 08 |i Related work: |s [Uniform title]. |t [Title]. |k [Series] |w [Record control number]
    - Ex. 1: Dataset record points to database it is related to:

245	0	4	ǂa The Louisville courier journal dataset, 1923-2000.
787	0	8	ǂi Related work: ǂt Proquest historical newspapers, Louisville courier journal ǂw (OCoLC)ocm08784449

    - Ex. 2: One newspaper title is split into more than one part:

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 with one link:
    - 856 40 **|y** Online dataset |u [URL]

856	4	0	ǂy Online dataset ǂu http://ssrs.yale.edu/data/SSDA/indiamaps/1991/

- For unmediated resources with multiple links:
  - 856 40 |y Online dataset |u [URL]
  - 856 42 |3 [Link text of related resource, e.g. Documentation, Manifest, etc.] |u [URL]

856	4	0	ǂy Online dataset ǂu http://ssrs.yale.edu/data/SSDA/LDC/LDC2017S06/
856	4	2	ǂ3 Documentation ǂu https://catalog.ldc.upenn.edu/docs/LDC2017S06/

- 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

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 all online datasets)*

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 items)*

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