Datenbank: Nineteenth Century Collections Online: British Theatre, Music and Literature

Provider: Gale Cengage

		Nineteenth Century Collections Online: British Theatre, Music and Literature
Access	Web address, API, Dumps, offline back up copy	• text-mining drives (includes directories, title manifests, XML files and image files, containing metadata, article segmentation, and page facsimiles (fee, available only for content the UB subscribes to or has purchased) • User can create batches of specific issues or titles for bulk download through the Gale Digital Scholar Lab (subscription service) • API access is not available
Documentation	Web address	https://link.gale.com/apps/NCCO?u=unibern
Distribution		• Only as an entire collection, closed
Scope	Content Purpose Field of use	• Rare documents relating to the arts in Victorian Britain. Includes music compositions, playbills, minutes of meetings, financial records and various other content types. • Full list of collections included: Archives of the Royal Literary Fund British Playbills, 1754-1882 Crystal Palace Handel Triennial Crystal Palace Saturday Concerts Drury Lane Receipts Drury Lane Theatre Archive Drury Lane under Sheridan, 1776-1812: Manuscript Plays and Correspondence English Stage after the Restoration, 1733-1822 JW Davison Papers King's Theatre Haymarket Archive Konzert Programm Austausch Lord Chamberlain's Plays Oratorio Concert Programmes Popular Literature in 18th and 19th Century Britain, Parts Three-Ten: The Barry Ono Collection of Bloods and Penny Dreadfuls Popular Literature in 18th and 19th Century Britain, Part Two: The Sabine Baring-Gould and Thomas Crampton Collections Queen's Hall Programmes Royal Albert Hall Royal Philharmonic Society Archive Royal Philharmonic Society Music Manuscripts Sir George Smart Papers Sir George Smart Programmes St James Hall Monday/Saturday Popular Concerts

		Wandering Minstrels Archive
Time, Place, Language	temporal, local reference	«Long» 19th Century i.e. 1789-1914 Britain English
Data type	What are the basic data types?	• Facsimiles: JPEG • Document and issue text files with structural mark up (pages, subdivided or zoned into articles): XML • bibliographic information: XML, partly within issue text files
Provenance, dependencies, accompanying material	original data source, manufacturer, data collection procedure, dependencies / links to other data sets / online resources, old versions	A DTD file is provided on the text-mining drives (not online) and the fields are comparable to those found in Dublin Core, MARC and other standard bibliographic standards The definitive dataset is kept in a proprietary XML format, known as the Gale Interchange Format or GIFT, and from this its text-mining and online datasets are derived.
Description Structured text data	Text markup or data structure e.g. TXT, XML, ALTO, TEI, versions	Periodical/Newpaper content: The TDM files are three seperate xml; • a publication xml - The publication xml includes publication title metadata • an issue xml - inlcudes the issue and article metadata • and a text xml - includes the full text OCR for each article. Manuscript and Monogpraph content: The TDM files are two seperate xml; • a document xml - The document xml includes document level metadata • and a page xml - includes the full text OCR for each page.
Description of databases, tabular data	data tables, existing / recommended data splits (e.g. training / test set)	n/a
Description of image formats	as precisely as possible (e.g. resolution, greyscale / bitonal)	• 400 PPI grayscale and colour jpeg
Standards, vocabularies	as precisely as possible: standards and vocabularies used	
Data quality: OCR; missing, incorrect, redundant data, noise	For example. OCR error rate, OCR process; different raw data available? Used software?	OCR confidence rating varies across the corpus. The corpus was digitised from a mixture of physical copies and microfilm.
Administration, cleanups,	e.g. handling of missing data, cutting, rescaling, NLP preprocessing, used	• ABBYY OCR engine used to create OCR text.

	software	
Scope /Size	size of data records	1.2M pages 566 monographs 4850 manuscript volumes 3152 newspaper issues
Metadata	Format/ Standards,	• bespoke metadata schema developed by Gale • hand-keyed issue, page, and article-level metadata • metadata fields: article title, article subheadings, attribution information, illustration captions, source page number, section headers • separate metadata files: 1. title or publication-level metadata (XML), 2. Issue or page-level metadata (XML)
Rights	licenses for metadata, full texts (TDM), rights / use (e.g. on-site, groups, scientific use)	Institutions have rights for non-commercial use by Authorised Users of the institutions only.
Ethical Issues	Personal and / or Confidential Information; Bias / representation; offensive / insulting / sensitive content	Historical content from the 19th century will contain views and material that may cause offense.
Use	Recommendations for use/ not recommended use	All purposes of TDM
Text and Data Mining	Additional costs? If so, how much? Trial possible?	Option 1: Small cost for delivering the data on hard drives Option 2: Annual subscription cost for access to the Gale Digital Scholar Lab

Stand 30.3.2022