Datenbank: Eighteenth Century Collections Online (ECCO)

Provider: Gale Cengage

		Eighteenth Century Collections Online
Access	Web address, API, Dumps, offline back up copy	• text-mining drives (includes directories, title manifests, XML files and image files, containing metadata and page facsimiles (fee, available only for content the UB subscribes to or has purchased) • User can create batches of specific 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/ECCO?u=unibern
Distribution		
Scope	Content Purpose Field of use	• 180k books and pamphlets published in the English-speaking world between 1700-1799, based on the English Short Title Catalogue (ESTC).
Time, Place, Language	temporal, local reference	• 1700-1799 • Mostly in English • Also: French, Latin, Ancient Greek, German, Italian, Scots Gaelic, Spanish, Welsh and a few more • Works published in the UK and colonial territories.
Data type	What are the basic data types?	• Facsimiles: TIFF • document text files with structural mark up (pages, subdivided): XML • bibliographic information: XML, partly within document 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	The TDM files are two seperate xml; • a document xml - The document xml includes bibliographic 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.	• 300 PPI bitonal tiff

	resolution, greyscale / bitonal)	
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 microfilm.
Administration, cleanups,	e.g. handling of missing data, cutting, rescaling, NLP preprocessing, used software	PrimeOCR engine used to create ECCO I OCR text. ABBYY OCR engine used to create ECCO II OCR text.
Scope /Size	size of data records	32.9M pages 184k monographs 207k volumes
Metadata	Format/ Standards,	• bespoke metadata schema developed by Gale • MARC record metadata and hand-keyed page-level metadata. Keyed metadata fields: source page number, section headers and graphic caption. • separate metadata files: 1. Document level metadata (XML), 2. 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 18th century will contain views and material that may cause offense.
Use	Recommendations for use/ not recommended use	
Text and Data Mining	Additional costs? If so, how much? Trial possible?	

Stand 30.3.2022