Datenbank: Times Digital Archive

Provider: Gale Cengage

		Times Digital Archive
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/TTDA?u=unibern
Distribution		• continously • one volume per year
Scope	Content Purpose Field of use	• Times 1785-2014, newspaper archive plus precursors • The Daily Universal register (1785-1787) • The Times, or, Daily Universal Register (1788)
Time, Place, Language	temporal, local reference	• 1785-2014 • UK, universal • English
Data type	What are the basic data types?	• Facsimiles: TIFF • 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	Each XML file contains bibliographic information for the entire issue, automatically zoned during the OCR process, with individual pages and articles are represented as child elements. At the article level, each individual word is encoded with spatial coordinates of its location on the corresponding image, as well as marker elements indicating new pages or columns. • to 2017: content + metadata (XML): machine-readable text appears within a single XML file per issue, surrounded by layered metadata that

		describes the features of the issue, pages, articles • from 2018: separate issue-level content data (XML)
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)	• to 2007: 300 PPI bitonal TIFFs • after 2007: 400 PPI • no compression
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. About a quarter of the corpus does not have an OCR confidence value associated with it.
Administration, cleanups,	e.g. handling of missing data, cutting, rescaling, NLP preprocessing, used software	Facsimiles: digital restoration was undertaken to reduce the appearance or impact of damaged pages, including manually cropping and cleaning and the insertion of digital titles or page numbers where needed.
Scope /Size	size of data records	1.6 mio pages, 11.8 articles
Metadata	Format/ Standards,	• bespoke metadata schema developed by Gale • hand-keyed issue and article-level metadata • until 2017: content + metadata (XML): machine-readable text appears within a single XML file per issue, surrounded by layered metadata that describes the features of the issue, pages, articles • metadata fields: article title, article subheadings, attribution information, illustration captions • from 2018: separate metadata files: 1. title or publication-level metadata (XML), 2. issue-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 dating back to 1785 may contain language and themes that today's users may find offensive.
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 Exact prices to be quoted in an offer in April 2021

Stand 30.3.2022