Datenbank: The Economist Historical Archive 1843-2015

Provider: Gale Cengage

		The Economist Historical 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/ECON?u=unibern
Distribution		- Only as an entire archive until 2015, newer issues will be digitized in batches of 5-10 years and sold as supplements to the original archive
Scope	Content Purpose Field of use	- The Economist 1843-2015
Time, Place, Language	temporal, local reference	- 1843-2015 - Published in London - English
Data type	What are the basic data types?	- Facsimiles: JPEG - 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	The TDM files are three separate xml; - a publication xml - The publication xml includes publication title metadata - an issue xml - includes the issue and article metadata - and a text xml - includes the full text OCR for each article.
Description of databases, tabular	data tables, existing / recommended data	n/a

data	splits (e.g. training / test set)	
Description of image formats	as precisely as possible (e.g. resolution, greyscale / bitonal)	• 300 PPI grayscale and colour jpg to 1998 • 300 PPI colour jpg 1999-2008 • 400 PPI colour jpg from 2009 onwards • 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. The material was digitised from physical copies held by The Economist and national libraries in the UK
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	660k pages
Metadata	Format/ Standards,	• bespoke metadata schema developed by Gale • hand-keyed issue and article-level metadata • metadata fields: article title, article subheadings, attribution information, illustration captions • 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 1843 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

Stand 30.3.2022