Datenbank: WBIS Online

Provider: De Gruyter

		WBIS Online
Access	Web address, API, Dumps, offline back up copy	• https://wbis.degruyter.com • API access is not available
Documentation	Web address	• n/a
Distribution		• Continuously
Scope	Content Purpose Field of use	• The most comprehensive biographical database available • Biographical information on over 6 million people from the 8th century B.C. to the present • Included are 8.5 Million digital facsimile articles from biographical reference works
Time, Place, Language	temporal, local reference	• From 8th century B.C. to the present • International • Multiple languages
Data type	What are the basic data types?	• XML files • TIF Images
Provenance, dependencies, accompanying material	original data source, manufacturer, data collection procedure, dependencies / links to other data sets / online resources, old versions	• The definitive dataset for biographical information is kept in a proprietary XML format, and from this its online biographical datasets are derived • The digital facsimile articles are provided in TIF format and accompanied by metadata in the related biographical XML and a source XML entry
Description Structured text data	Text markup or data structure e.g. TXT, XML, ALTO, TEI, versions	• Each XML entry contains biographic information for a specified person and bibliographical information for the specified sources • Also included are XML files for the sources themselves
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)	• TIF • 400 dpi, greyscale
Standards, vocabularies	as precisely as possible: standards and vocabularies used	• No standard vocabularies where used but WBIS Online has its own classification system, e.g. for occupations, that accompanies each entry and can be provided separately and standardized keywords are also included

Data quality: OCR; missing, incorrect, redundant data, noise	For example. OCR error rate, OCR process; different raw data available? Used software?	• n/a
Administration, cleanups,	e.g. handling of missing data, cutting, rescaling, NLP preprocessing, used software	• Missing data is added when identified, happens very rarely
Scope /Size	size of data records	• Over 6 million biographical entries with 8.5 million facsimiles
Metadata	Format/ Standards,	• Proprietary, see above
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	• n/a
Use	Recommendations for use/ not recommended use	• All purposes of TDM
Text and Data Mining	Additional costs? If so, how much? Trial possible?	• Small cost for delivering the data on hard drives.

Stand 30.3.2022