UNIVERSITÄT BERN

Universitätsbibliothek Bern

Datenbank: Swissdox@LiRI

Provider: LiRI

		Swissdox@LiRI
Access	Web address, API, Dumps, offline back up copy	• https://swissdox.linguistik.uzh.ch/ • Corpus bulk download according to search (filtering for language, time interval, keywords, sources and document types) • API access to come mid 2022
Documentation	Web address	• Provider website • User's Guide • Flyer
Distribution		• Continuously (daily)
Scope	Content Purpose Field of use	• The Swissdox@LiRI database includes approx 23 million published media articles from a wide range of Swiss media sources (both print and digital) covering many decades, and is updated daily with approximately 5000 to 6000 new articles.
Time, Place, Language	temporal, local reference	• Starting 1910 (main: • Switzerland • German, French, (Italian, Romansh, English)
Data type	What are the basic data types?	• TSV file, xy-zipped • Full texts: XML
Provenance, dependencies, accompanying material	original data source, manufacturer, data collection procedure, dependencies / links to other data sets / online resources, old versions	• Data stock comes from our partner CH Media, NZZ media group, Ringier, Ringier Axel Springer Schweiz and TX Group (Tamedia), SRF/SRG and Wochenzeitung, overall 250 sources with planned further expansion.
Description Structured text data	Text markup or data structure e.g. TXT, XML, ALTO, TEI, versions	• Documentation of metadata and article format • Full texts: XML (spec)
Description of databases, tabular data	data tables, existing / recommended data splits (e.g. training / test set)	• TSV (tab separated value) format • zipped (.xy), unzipping with WinZip, 7-Zip (Win) or Unarchiver (Mac)
Description of image formats	as precisely as possible (e.g. resolution, greyscale / bitonal)	• n/a
Standards, vocabularies	as precisely as possible: standards and vocabularies used	• n/a

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	• n/a
Scope /Size	size of data records	• approx. 23 mio news/media articles
Metadata	Format/ Standards	• proprietary, part of the corpus
Rights	licenses for metadata, full texts (TDM), rights / use (e.g. on-site, groups, scientific use)	• Terms of Use • service is for academic use only • data from the corpus may only be used for the declared research project • storage is allowed only in one's own IT infrastructure (at an academic institution) • storage on third-party cloud platforms (e.g. Dropbox, Google Docs, Google Cloud Platform, Amazon Web Services, Microsoft Azure) is not permitted • data may not be shared with third parties • it is not allowed to receive all the data • raw data have to be deleted six months after project completion • details of the project title, project management and duration of the research project will be communicated to Swissdox and may be mentioned on the SMD/Swissdox and LiRI websites
Ethical Issues	Personal and / or Confidential Information; Bias / representation; offensive / insulting / sensitive content	• n/a
Use	Recommendations for use/ not recommended use	Users have to • login (SWITCH edu-ID) • accept the terms and conditions • register their project • enter their corpus query • download the retrieved datasets in TSV (tab separated) format (.xz compressed) • extract the data in order to work with them. More information can be found in the users' guide: https://liri.linguistik.uzh.ch/wiki/langtech/swissdox/start

Text and Data Mining	Additional costs? If so, how much? Trial possible?	• No additional costs • Online analysis/NLP functionality to come end of 2022

Stand 30.3.2022