SwissText Hackathon Swissdox

On this page, you find an overview of the data for the SwissText Hackathon 2023 (https://www.swisstext.org/swissdox-hackathon/).

Sources

The dataset consists of 100,000 articles each from 9 newspapers from Switzerland, 5 of them published in German, 4 in French. The articles from the German newspapers span the years 1996 - 2023, the French ones 1998 - 2023. The data comes from the collaboration between Swissdox and LiRI in the project Swissdox@LiRI.

	acronym	source name
German	BAZ	Basler Zeitung
	BU	Der Bund
	BZ	Berner Zeitung
	NZZ	Neue Zürcher Zeitung
	TA	Tages-Anzeiger
French	HEU	24 Heures
	TDG	Tribune de Genève
	TLM	Le Matin
	TPS	Le Temps

The following table gives an overview of the articles per year and source included in the dataset:

	German					French			
year	**BAZ**	**BU**	**BZ**	**NZZ**	**TA**	**HEU**	**TDG**	**TLM**	**TPS**
1996	3 213	3 681	0	3 681	3 681	-	-	-	-
1997	3 272	3 681	0	3 681	3 681	-	-	-	-
1998	3 270	3 681	1 333	3 681	3 681	3 968	3 968	3 984	3 968
1999	3 626	3 681	5 300	3 681	3 681	3 968	3 968	3 985	3 968
2000	4 081	3 681	5 299	3 681	3 681	5 000	3 968	3 985	3 968
2001	3 853	3 681	5 299	3 681	3 681	5 000	3 968	3 985	3 968
2002	3 884	3 681	5 299	3 681	3 681	5 001	3 968	3 985	3 968
2003	3 884	3 681	5 299	3 681	3 681	3 852	3 968	2 595	3 968
2004	3 884	3 681	5 299	3 681	3 681	5 001	3 968	5 165	3 968
2005	3 842	3 681	3 681	3 681	3 681	2 729	3 968	5 165	3 968
2006	3 681	3 681	3 681	3 681	3 681	0	3 968	5 165	3 968
2007	3 681	3 681	3 681	3 681	3 681	0	3 968	5 165	3 968
2008	3 681	3 681	3 681	3 681	3 681	5 001	3 968	5 165	3 968
2009	3 681	3 681	3 681	3 681	3 681	5 000	3 968	5 165	3 968
2010	3 681	3 680	3 681	3 680	3 680	5 000	3 968	5 165	3 968
2011	3 681	3 681	3 681	3 681	3 681	5 000	3 968	5 165	3 968
2012	3 681	3 681	3 681	3 681	3 681	5 000	3 968	5 165	3 968
2013	3 681	3 681	3 681	3 681	3 681	3 968	3 968	5 165	3 968
2014	3 681	3 681	3 681	3 681	3 681	3 968	3 968	5 165	3 968
2015	3 681	3 681	3 681	3 681	3 681	3 968	3 968	5 166	3 968

2016	3 681	3 681	3 681	3 681	3 681	3 968	3 968	5 166	3 968
2017	3 681	3 681	3 681	3 681	3 681	3 969	3 969	5 167	3 969
2018	3 681	3 681	3 681	3 681	3 681	3 969	3 969	5 167	3 969
2019	3 681	3 681	3 681	3 681	3 681	3 969	3 969	0	3 969
2020	3 681	3 681	3 681	3 681	3 681	3 969	3 969	0	3 969
2021	3 681	3 681	3 681	3 681	3 681	3 969	3 969	0	3 969
2022	3 681	3 681	3 681	3 681	3 681	3 969	3 969	0	3 969
2023	614	614	614	614	614	794	794	0	794
	100 000	**100 000**	**100 000**	**100 000**	**100 000**	**100 000**	**100 000**	**100 000**	**100 000**

Collecting 100,000 articles evenly distributed over *n* years would ideally result in a fixed number of articles per year per source. However, because this amount of articles is not available for all sources per year, the actual numbers of provided articles differ slightly. The missing articles for a year (or several years) have been collected from adjacent years in which more articles were available: E.g. for BZ 100,000 articles evenly distributed over 27 years and 2 months would result in selecting 3,681/year. However, there are no articles available for 1996/97 and only 1,333 for the year 1998. Therefore, the missing articles were evenly distributed over the following 6 years as additional selected articles.

For your participation at the hackathon, you are free in your choice of sources, you may e.g. only want to use French texts, or only those dealing with your topic of interest, such as migration or climate change, or you are also free to add data of your own and copmare or correlate, such as tweets, Worry Barometer, outcomes of votes, seasons, important events such as summits, earth quakes, financial crises.

Data Structure

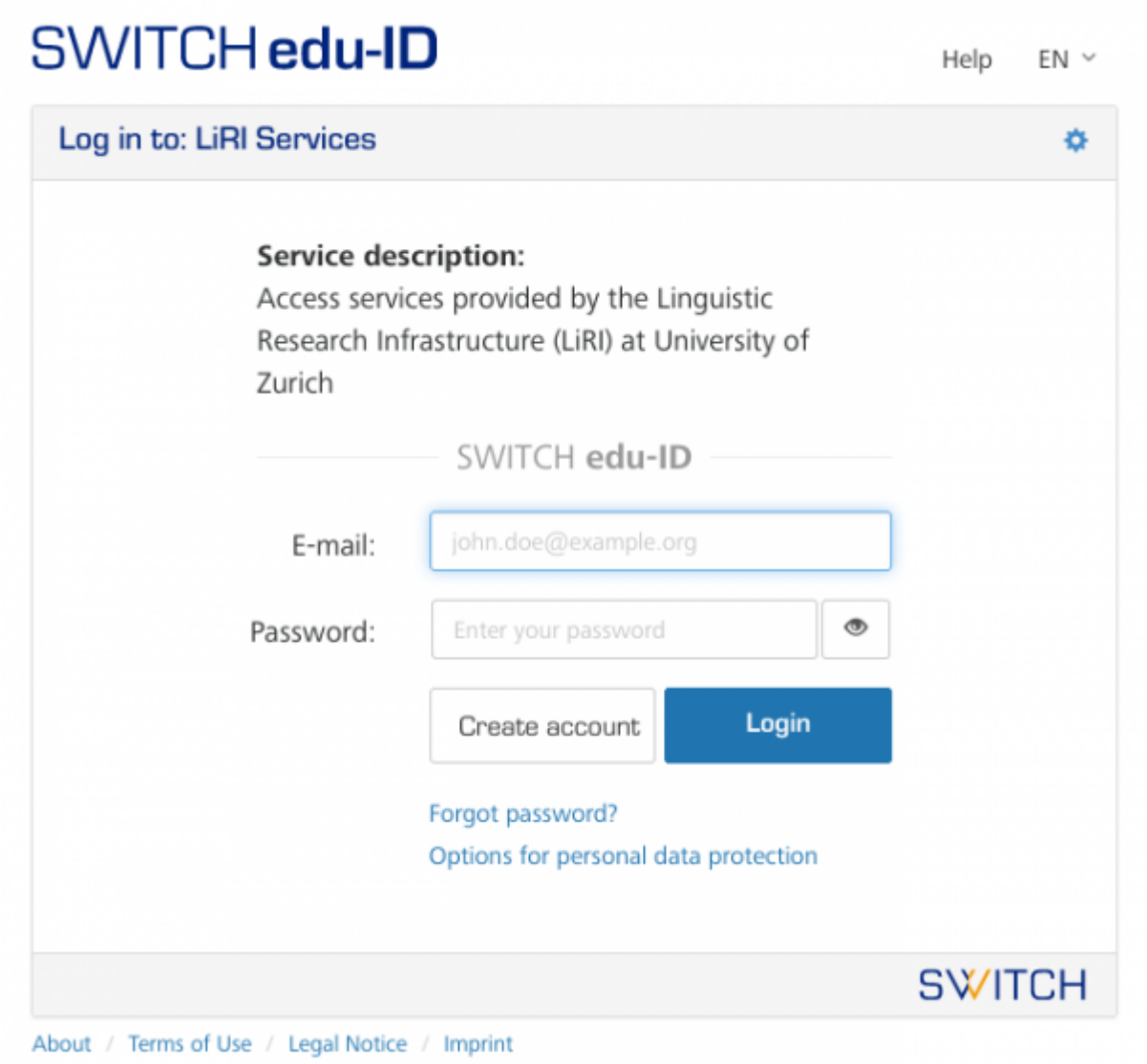
The name of an article is a UUID.

- Data for each source (i.e. newspaper) is in its own file.
- Data for every year is in its own directory
- Based on the first two characters of the name of articles, subdirectories were created.
- In each subdirectory, articles starting with those two characters are found in two formats:
  - `<UUID>.xml` (the original XML that is stored in the database, containing the whole article with all data such as, text, tables, image legends, author names, etc.)
  - `<UUID>.txt` (the verticalized text of the XML-article, in CoNLL-U format)
- The tags present in the XML have the following semantics:
  - `<a>` Link
  - `<au>` Author
  - `<ka>` Table
  - `<ld>` Lead
  - `<lg>` Caption
  - `<p>` Paragraph
  - `<zt>` Subheading

Log in and accept the terms of use

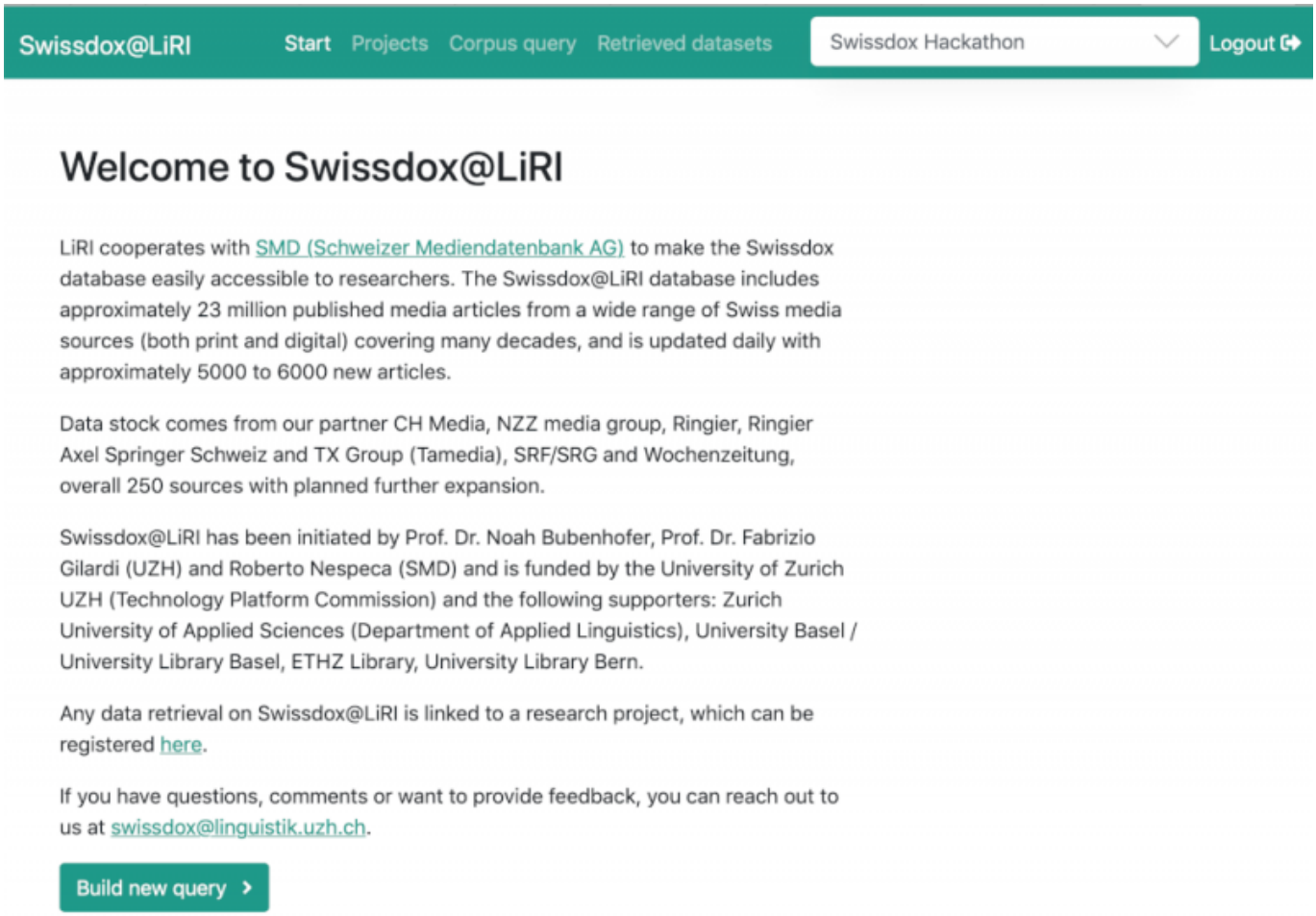
You have been sent an email from noreply@linguistik.uzh.ch with the title “[LiRI account manager] Invitation for Swissdox@LiRI”. When you open the link, you will need to log in using your SWITCH edu-ID, and accept the terms of use for Swissdox@LiRI. Please observe the conditions, for instance that

the raw data must be deleted six months after the end of the project.

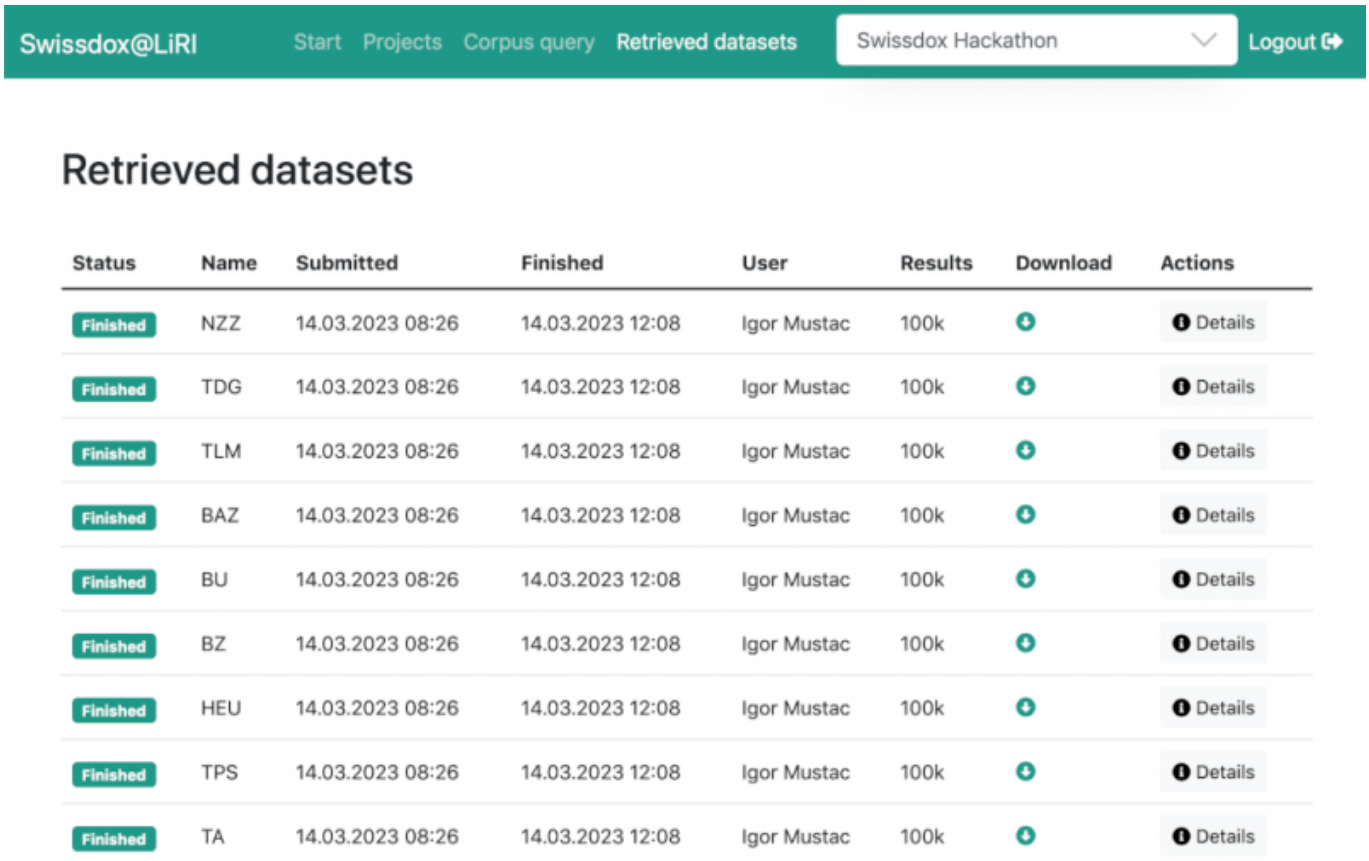


Download the data

After logging in, make sure that the project “Swissdox Hackathon” is selected (top right) as shown in the following screenshot.



Select the tab “Retrieved datasets”. You now see the list of the 9 Swiss newspapers. Download each of them separately using the arrow in the second last column.



Contact

If you have any questions or feedback concerning the data, please write to swissdox@linguistik.uzh.ch

We are looking forward to an exciting workshop.