Nasjonalbiblioteket
National Library of Norway

Freely available documents from Norwegian state institutions (Målfrid 2026)

Description

This corpus consists of documents collected from 720 domains belonging to Norwegian state institutions and contains approximately 2.6 billion tokens, making it one of the largest freely available resources for Norwegian Bokmål and Nynorsk. In addition to Norwegian, the corpus also contains texts in Northern Sami, Lule Sami, Southern Sami, and English.

Number of tokens per language:

Norwegian Bokmål	'nob'	1 890 232 881
English	'eng'	570 099 658
Norwegian Nynorsk	'nno'	170 231 965
Northern Sami	'sme'	2 401 814
Southern Sami	'sma'	415 173
Lule Sami	'smj'	329 340

The data was collected as part of the Målfrid project, in which the National Library of Norway, on behalf of the Ministry of Culture and in collaboration with the Language Council of Norway, collects and aggregates data to map the use of Norwegian Bokmål and Norwegian Nynorsk on the websites of Norwegian state institutions.

The corpus is the result of a focused crawl conducted between December 2025 and January 2026. Text documents (HTML, DOC(X)/ODT, and PDF) were downloaded recursively from a set of domains (down to and including level 12), while respecting robots.txt and politeness restrictions. Crawling and processing were carried out using the freely available Python package maalfrid_toolkit (https://github.com/NationalLibraryOfNorway/maalfrid_toolkit). The package extracts natural language from HTML using the boilerplate removal system jusText (http://corpus.tools/wiki/Justext), from Word documents using docx2txt/antiword, and from PDFs using a custom text extractor developed by the AI Lab at the National Library of Norway.

The extracted text was classified at the document level using Gielladetect/pytextcat (https://github.com/NationalLibraryOfNorway/gielladetect), and the detected language is provided as part of the metadata. Please note that language identification, especially for similar languages and noisy web data, is never perfect. Exact duplicates were removed at the domain level.

Format

Each tarball (one per language) contains gzipped JSON Lines (JSONL) files, with one document per line. There is one JSONL file for each combination of domain (e.g. nb.no) and content type (e.g. HTML). The files are encoded in UTF-8. Each document contains the following keys:

- doc_id: a unique ID for the document (use this for reference)
- doc_hash: a sha256 checksum of the full text
- domain: domain of the document at crawl time
- url: the URL of the document at crawl time
- date: crawl date
- mimetype: simplified media type: HTML, DOC, or PDF
- title: title of the document (HTML only; otherwise null)
- lang: language of the document (detected using gielladetect)
- fulltext: an array of strings, where each string represents one paragraph

License

The data is provided under the Norwegian Licence for Open Government Data (NLOD) 2.0: https://data.norge.no/nlod/en/2.0/

Contact

If you have questions about the material, please contact us at sprakbanken@nb.no.