

# Data for MT

PV061

**Pavel Rychlý**

NLP Centre, FI MU

10 Dec 2024

## Common Crawl

- <https://commoncrawl.org/>
- around 10 crawls per year (~85 in total from 2013)
- 90 TB of compressed data each
  - 60-80 TB 2020-2022
  - 40-60 TB 2017-2019
  - 30-50 TB 2014-2016
- each file ~ 1GB
- raw data - WARC format (also WET - text)
- textual content only: HTML, PDF, XML, ...
- data accessible from  
[https://data.commoncrawl.org/\[...\]](https://data.commoncrawl.org/[...])

# ClueWeb09

- <http://lemurproject.org/clueweb09/>
- crawling in January and February 2009
- 1 billion web pages, in 10 languages
- 5 TB, compressed. (25 TB, uncompressed.)
- distributed on on 8TB hard disk (\$380)

# WARC format

- raw data from from web servers
- includes response headers, HTML, JavaScript, ...

```
WARC/1.0
WARC-Type: response
WARC-Date: 2014-08-02T09:52:13Z
WARC-Record-ID: <urn:uuid:ffbfb0c0-6456-42b0-af03-3867be6fc09f>
Content-Length: 43428
Content-Type: application/http; msgtype=response
WARC-Warcinfo-ID: <urn:uuid:3169ca8e-39a6-42e9-a4e3-9f001f067bdf>
WARC-Concurrent-To: <urn:uuid:d99f2a24-158a-4c77-bb0a-3cccd40aad56>
WARC-IP-Address: 212.58.244.61
WARC-Target-URI: http://news.bbc.co.uk/2/hi/africa/3414345.stm
WARC-Truncated: length
```

```
HTTP/1.1 200 OK
Content-Type: text/html
Date: Sat, 02 Aug 2014 09:52:13 GMT
Set-Cookie: BBC-UID=15730d9c1b741c0d3942e2aca1317fbf39e57b90be68a32
```

```
<!doctype html public "-//W3C//DTD HTML 4.0 Transitional//EN" "http://www.w3.org/TR/REC-html40/loose.dtd">
```

```
<html>
<head>
<title>
```

```
    BBC NEWS | Africa | Namibia braces for Nujoma exit
```

```
</title>
```

```
<meta name="keywords" content="BBC, News, BBC News, news online, world, uk, international, forei
```

```
<meta name="OriginalPublicationDate" content="2004/01/22 00:48:49" />
```

```
Pavel Rychly • Data for MT • 10 Dec 2024
```

# Original web page

**BBC** Home **News** Sport Radio TV Weather Languages  Search

[an error occurred while processing this directive]

Low graphics | Accessibility help

**BBC NEWS** [Watch](#) One-Minute World News 

News services  
Your news when you want it 

News Front Page 

Last Updated: Thursday, 22 January, 2004, 00:48 GMT  
[E-mail this to a friend](#) [Printable version](#)

## Namibia braces for Nujoma exit

By Robin White  
BBC's former Focus on Africa editor

**President Sam Nujoma works in very pleasant surroundings in the small but beautiful old State House slap bang in the middle of Windhoek just a stone throw away from Namibia's equally beautiful parliament**

The question on everyone's lips is why on earth does he want to build a new State House.

Currently a massive concrete presidential monstrosity is being erected on the outskirts of the capital costing millions of Namibian dollars at a time when the Namibian Government is appealing for relief aid for drought victims.



President Nujoma may be a hard act to follow in Namibia

**SEE ALSO:**

- Deal ends Namibian land invasions 07 Nov 03 | Africa
- Surprise reshuffle in Namibia 27 Aug 02 | Africa
- Nujoma 'will not seek fourth term' 26 Nov 01 | Africa
- Nujoma's war on waste 09 Feb 01 | Africa
- Country profile: Namibia 07 Nov 03 | Country profiles
- Timeline: Namibia 07 Nov 03 | Country profiles

**RELATED INTERNET LINKS:**

- Namibian Government

The BBC is not responsible for the content of external internet sites

**TOP AFRICA STORIES**

- Nigeria state oil firm 'insolvent'
- France to help Africa veterans
- Churches call for Sudan to split

 | News feeds

# WET format

WARC/1.0

WARC-Type: conversion

WARC-Target-URI: http://news.bbc.co.uk/2/hi/africa/3414345.stm

WARC-Date: 2014-08-02T09:52:13Z

WARC-Block-Digest: sha1:JROHLC55SKMBR6XY46WXREW7RXM64EJC

Content-Type: text/plain

Content-Length: 6724

BBC NEWS | Africa | Namibia braces for Nujoma exit

[an error occurred while processing this directive]

Low graphics|Accessibility help

One-Minute World News

News services

Your news when you want it

News Front Page

Africa

Americas

Asia-Pacific

Europe

Middle East

South Asia

UK

Business

Health

Science & Environment

Technology

Entertainment

Also in the news

-----

Video and Audio

-----

Programmes

# Corpus Tools

- <https://corpus.tools/>
- JusText - removing boilerplate
- Chared - detecting character encoding
- onion - removing duplicate parts
- SpiderLing - web spider for linguistics

# OPUS corpus

- <https://opus.nlpl.eu/>
- ~70 sources (subcorpora)
- 700+ languages
  - dng (Dungan): 6 sentences, 25 tokens
- Tatoeba
  - (short) translated sentences for language learning
  - ~400 languages, 12M segments
- WikiMatrix (Facebook)
  - from wikipedia
  - 85 languages, 135M segments



# Corpora from Common Crawl

- CCNet:
  - 130 languages, monolingual
  - fastText for language identification
  - (statistical) language modeling for filtering
- CCAIghned
  - 113 languages, 2.3G segments
  - aligned to English only
  - similarity of sentence embeddings (LASER)
- CCMatrix
  - from CCNet
  - 90 languages (1200 pairs), 7.4G segments
- MultiCCAighned
  - 112 languages (6075 pairs), 840M segments
  - CS-UK: Czech mostly bad MT

**MUNI**

FACULTY

OF INFORMATICS