

DH Potsdam EC Fellowship Programme

Winter Semester 2025/2026

Schedule

— — —

23.10.2025: From Texts to Data/Metadata (L. Giovannini, D. Skorinkin)

06.11.2025: Cleaning and Managing Data (L. Giovannini, D. Skorinkin)

20.11.2025: Processing and Sharing Data (A. Gerstenberg, H. Sluyter-Gäthje)

04.12.2025: Data visualisation (K. Tikhomirova)

08.01.2026: Working Efficiently with LLMs (L. Giovannini, D. Skorinkin)

29.01.2026: Project consulting

Extra events

— — —

13.11.2025: Potsdam DH Day

05.12.2026: XML-TEI Workshop

17.12.2025: Introduction to Transkribus (Gerstenberg)

Any question?

Introduction round

What's your name?
What do you study?

23.10.2025

#1: From Texts to Data/Metadata

Today

— — —

9:00–9:15 Housekeeping

9:15–9:45 What are “data” in the humanities?

9:45–10:30 Turning sources into data

Hands-on: OCR of textual historical sources

10:30–10:45 Break ☹☹☹

10:45–12:00 Structured metadata, annotation, formats

What are 'data' in the humanities

Following
Drücker 2021: 19–33

- Critical perspective: Data are produced, not given
- All “data are capta”, i.e., information captured through abstraction and interpretation
- Data modeling: process of selecting and structuring information from sources
- Every stage raises ethical issues about privacy, data ownership and usage

— — —

unstructured data

— — —

Ambiguous, context-dependent (think of the word “Berlin”)

Without a “secondary structure”
(data model) imposed upon it

Examples: analog materials, simple
text/image/sound files

structured data

Explicit, discrete, unambiguous

Has an explicit form that has been
deliberately designed, generally
for a specific purpose

Example: encoded texts, tables,
spreadsheets

Checklist for creating a data model

— — —

- Identify information that can be described unambiguously in your sources.
- Create a set of labeled **categories** that are clear and distinct.
- Decide which categories will be specified numerically, by true/false values, or by descriptive terms.
- Assess the difficulty of keeping your entries consistent.
- **Standardise** the format for each data type (dates, places, and names).
- Limit the **vocabulary** for description except in the comments.
- Read the categories you have established to see what the data model will take from the original sources.
- Check for assumptions about categories that may embody bias.
- Check your data model for completeness and accuracy.

Parametrisation

— — —

Determining what can be measured
(according to quantity, scale,
size, direction, frequency...)

e.g. number of words in a text

Tokenisation

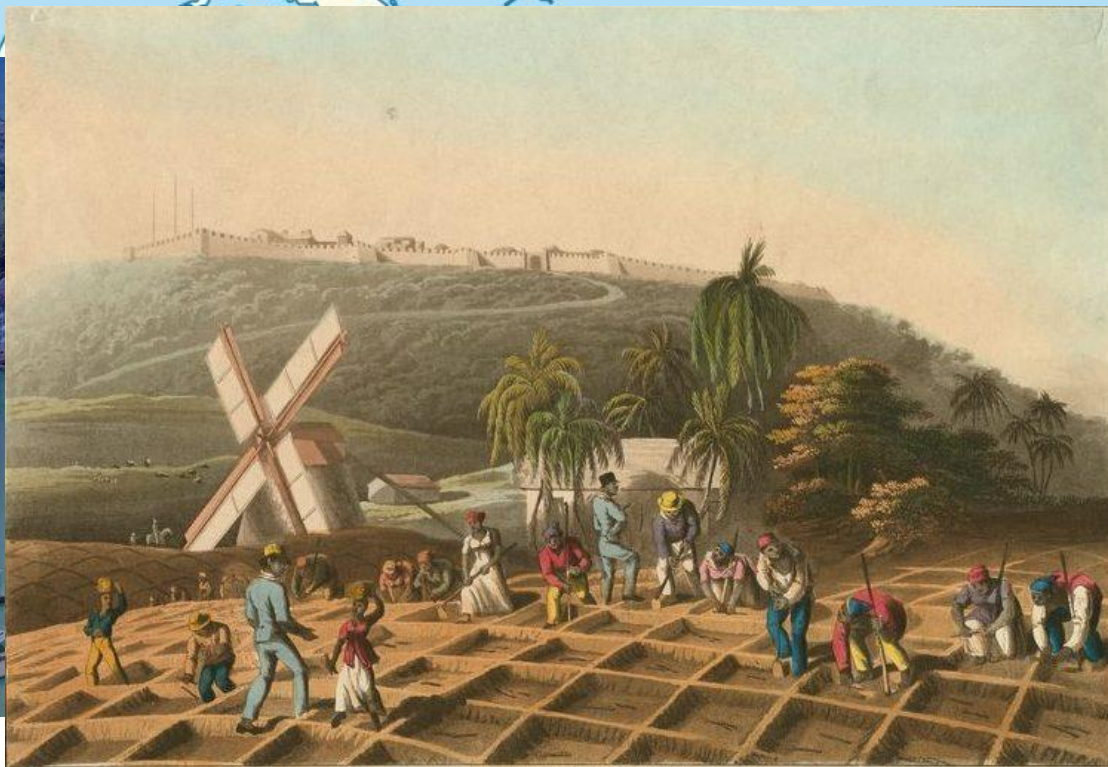
Determining what units of
representation can be identified
in a corpus or artifact

eg. individual words

Both parametrisation and require judgments about significance
and context!

Observer position matters, since metrics change with
cultural/historical perspective

Turning sources into data



Our practice dataset: The Barbados Mercury and Bridge-town Gazette (1783-1839)

**Ever heard of OCR
(Optical Character Recognition)?**

Hands-on

Let's transcribe the
Barbados Mercury with
Transkribus

Open the Moodle link!



Structured (meta)data, annotation, formats

unstructured data

— — —

Ambiguous, context-dependent (think of the word “Berlin”)

Without a “secondary structure”
(data model) imposed upon it

Examples: analog materials, simple
text/image/sound files

structured data

Explicit, discrete, unambiguous

Has an explicit form that has been
deliberately designed, generally
for a specific purpose

Example: encoded texts, tables,
spreadsheets

unstructured data

— — —

Ambiguous, context-dependent (think of the word “Berlin”)

Without a “secondary structure”
(data model) imposed upon it

Examples: analog materials, simple
text/image/sound files

structured data

Club	MP	W	D	L	GF	GA	GD	Pts	Last 5
1  Bayern	7	7	0	0	27	4	23	21	✓✓✓✓✓
2  RB Leipzig	7	5	1	1	10	9	1	16	✓✓✓-✓
3  VfB Stuttgart	7	5	0	2	11	6	5	15	✗✓✓✓✓
4  Dortmund	7	4	2	1	13	6	7	14	✓✓✓-✗
5  Leverkusen	7	4	2	1	16	11	5	14	✓-✓✓✓
6  Köln	7	3	2	2	12	10	2	11	-✗✗✓-
7  Eintracht Fran...	7	3	1	3	19	18	1	10	✗✗✓✗-
8  Hoffenheim	7	3	1	3	12	12	0	10	✓✗-✗✓
9  Union Berlin	7	3	1	3	11	14	-3	10	✗✓-✗✓
10  SC Freiburg	7	2	3	2	11	11	0	9	✓✓- -

unstructured data

— — —

Ambiguous, context-dependent (think of the word “Berlin”)

Without a “secondary structure”
(data model) imposed upon it

Examples: analog materials, simple
text/image/sound files

structured data

Explicit, discrete, unambiguous

Has an explicit form that has been
deliberately designed, generally
for a specific purpose

Example: encoded texts, tables,
spreadsheets

unstructured data

— — —

THE BARBADOS MERCURY,
AND BRIDGE-TOWN GAZETTE.
SIX DOLLARS A JOURNAL
TUESDAY, JANUARY 16, 1816.

structured data

Sent-ID	Token	Lemma	POS-tag	Entity-tag
1	THE	the	DET	
1	BARBADOS	Barbados	PROPN	LOC
1	MERCURY	Mercury	PROPN	
1	,	,	PUNCT	
1	AND	and	CCONJ	
1	BRIDGE	Bridge	PROPN	LOC
1	-	-	PUNCT	LOC
1	TOWN	Town	PROPN	LOC
1	GAZETTE	gazette	PROPN	ORG
1	.	.	PUNCT	
2	SIX	six	NUM	
2	DOLLARS	dollar	NOUN	
2	A	a	DET	
2	JOURNAL	journal	NOUN	
2			SPACE	
2	TUESDAY	Tuesday	NOUN	
2	,	,	PUNCT	
2	JANUARY	January	PROPN	DATE
2	16	16	NUM	DATE
2	,	,	PUNCT	
2	1816	1816	NUM	DATE
2	.	.	PUNCT	

Popular structured data formats in research

location	latitude	longitude
Barbados	13.19	-59.54
Potsdam	52.39	13.06
Spitzbergen	78.9	13.48

— — —

CSV/TSV (modeling tables with commas/tabs)	XML (annotating texts with angled brackets and HTML-like syntax)	JSON (storing data with curly brackets)
location, latitude, longitude Barbados, 13.19, -59.54 Potsdam, 52.39, 13.06 Spitzbergen, 78.9, 13.48	<pre><row> <location>Barbados</location> <latitude>13.19</latitude> <longitude>-59.54</longitude> </row> <row> <location>Potsdam</location> <latitude>52.39</latitude> <longitude>13.06</longitude> </row> <row> <location>Spitzbergen</location> <latitude>78.9</latitude> <longitude>13.48</longitude> </row></pre>	<pre>[{ "location": "Barbados", "latitude": 13.19, "longitude": -59.54 }, { "location": "Potsdam", "latitude": 52.39, "longitude": 13.06 }, { "location": "Spitzbergen", "latitude": 78.9, "longitude": 13.48 }]</pre>

Popular structured data formats in research

location	latitude	longitude
Barbados	13.19	-59.54
Potsdam	52.39	13.06
Spitzbergen	78.9	13.48

CSV/TSV (modeling tables with commas/tabs)	XML (annotating texts with angled brackets and HTML-like syntax)	JSON (storing data with curly brackets)
location, latitude, longitude Barbados, 13.19, -59.54 Potsdam, 52.39, 13.06 Spitzbergen, 78.9, 13.48	<row> <location>Barbados</location> <latitude>13.19</latitude> <longitude>-59.54</longitude> </row> <row> <location>Potsdam</location> <latitude>52.39</latitude> <longitude>13.06</longitude> </row> <row> <location>Spitzbergen</location> <latitude>78.9</latitude> <longitude>13.48</longitude> </row>	[{ "location": "Barbados", "latitude": 13.19, "longitude": -59.54 }, { "location": "Potsdam", "latitude": 52.39, "longitude": 13.06 }, { "location": "Spitzbergen", "latitude": 78.9, "longitude": 13.48 } <\/>

Popular structured data formats in research

location	latitude	longitude
Barbados	13.19	-59.54
Potsdam	52.39	13.06
Spitzbergen	78.9	13.48

CSV/TSV (modeling tables with commas/tabs)	XML (annotating texts with angled brackets and HTML-like syntax)	JSON (storing data with curly brackets)
location, latitude, longitude Barbados, 13.19, -59.54 Potsdam, 52.39, 13.06 Spitzbergen, 78.9, 13.48	<pre><row> <location>Barbados</location> <latitude>13.19</latitude> <longitude>-59.54</longitude> </row> <row> <location>Potsdam</location> <latitude>52.39</latitude> <longitude>13.06</longitude> </row> <row> <location>Spitzbergen</location> <latitude>78.9</latitude> <longitude>13.48</longitude> </row></pre>	<pre>[{ "location": "Barbados", "latitude": 13.19, "longitude": -59.54 }, { "location": "Potsdam", "latitude": 52.39, "longitude": 13.06 }, { "location": "Spitzbergen", "latitude": 78.9, "longitude": 13.48 }]</pre>

Popular structured data formats in research

location	latitude	longitude
Barbados	13.19	-59.54
Potsdam	52.39	13.06
Spitzbergen	78.9	13.48

CSV/TSV (modeling tables with commas/tabs)	XML (annotating texts with angled brackets and HTML-like syntax)	JSON (storing data with curly brackets)
location, latitude, longitude Barbados, 13.19, -59.54 Potsdam, 52.39, 13.06 Spitzbergen, 78.9, 13.48	<pre><row> <location>Barbados</location> <latitude>13.19</latitude> <longitude>-59.54</longitude> </row> <row> <location>Potsdam</location> <latitude>52.39</latitude> <longitude>13.06</longitude> </row> <row> <location>Spitzbergen</location> <latitude>78.9</latitude> <longitude>13.48</longitude> </row></pre>	<pre>[{ "location": "Barbados", "latitude": 13.19, "longitude": -59.54 }, { "location": "Potsdam", "latitude": 52.39, "longitude": 13.06 }, { "location": "Spitzbergen", "latitude": 78.9, "longitude": 13.48 }]</pre>

Motivation for structuring data in a machine-readable way

- Explicit encoding (formalisation) of knowledge
- Makes it machine-actionable, e.g.
- Structured dataset as the middle-ground between your raw data and the output of some processing model

```
import pandas as pd...
```

```
# Read the CSV file...
```

	location	latitude	longitude
0	Barbados	13.19	-59.54
1	Potsdam	52.39	13.06
2	Spitzbergen	78.90	13.48

```
# Create the plot...
```



unstructured data

— — —

THE BARBADOS MERCURY,
AND BRIDGE-TOWN GAZETTE.
SIX DOLLARS A JOURNAL
TUESDAY, JANUARY 16, 1816.

structured data

Sent-ID	Token	Lemma	POS-tag	Entity-tag
1	THE	the	DET	
1	BARBADOS	Barbados	PROPN	LOC
1	MERCURY	Mercury	PROPN	
1	,	,	PUNCT	
1	AND	and	CCONJ	
1	BRIDGE	Bridge	PROPN	LOC
1	-	-	PUNCT	LOC
1	TOWN	Town	PROPN	LOC
1	GAZETTE	gazette	PROPN	ORG
1	.	.	PUNCT	
2	SIX	six	NUM	
2	DOLLARS	dollar	NOUN	
2	A	a	DET	
2	JOURNAL	journal	NOUN	
2			SPACE	
2	TUESDAY	Tuesday	NOUN	
2	,	,	PUNCT	
2	JANUARY	January	PROPN	DATE
2	16	16	NUM	DATE
2	,	,	PUNCT	
2	1816	1816	NUM	DATE
2	.	.	PUNCT	

unstructured data

— — —

THE BARBADOS MERCURY,
AND BRIDGE-TOWN GAZETTE.
SIX DOLLARS A JOURNAL
TUESDAY, JANUARY 16, 1816.

structured data (CSV)

Sent-ID,Token,Lemma,POS-tag,Entity-tag

```
1,THE,the,DET,  
1,BARBADOS,Barbados,PROPN,LOC  
1,MERCURY,Mercury,PROPN,  
1,"","","",PUNCT,  
1,AND,and,CONJ,  
1,BRIDGE,Bridge,PROPN,LOC  
1,-,-,PUNCT,LOC  
1,TOWN,Town,PROPN,LOC  
1,GAZETTE,gazette,PROPN,ORG  
1,...,PUNCT,  
2,SIX,six,NUM,  
2,DOLLARS,dollar,NOUN,  
2,A,a,DET,  
2,JOURNAL,journal,NOUN,  
2, , ,SPACE,  
2,TUESDAY,Tuesday,NOUN,  
2,"","","",PUNCT,  
2,JANUARY,January,PROPN,DATE  
2,16,16,NUM,DATE  
2,"","","",PUNCT,  
2,1816,1816,NUM,DATE  
2,...,PUNCT,
```

unstructured data

— — —
THE BARBADOS MERCURY,
AND BRIDGE-TOWN GAZETTE.
SIX DOLLARS A JOURNAL
TUESDAY, JANUARY 16, 1816.

structured data (XML)

```
<document>
  <tok sent="1" lemma="the" pos="DET" ent="">THE</tok>
  <tok sent="1" lemma="Barbados" pos="PROPN" ent="LOC">BARBADOS</tok>
  <tok sent="1" lemma="Mercury" pos="PROPN" ent="">MERCURY</tok>
  <tok sent="1" lemma="," pos="PUNCT" ent="">,</tok>
  <tok sent="1" lemma="and" pos="CCONJ" ent="">AND</tok>
  <tok sent="1" lemma="Bridge" pos="PROPN" ent="LOC">BRIDGE</tok>
  <tok sent="1" lemma="-" pos="PUNCT" ent="LOC">-</tok>
  <tok sent="1" lemma="Town" pos="PROPN" ent="LOC">TOWN</tok>
  <tok sent="1" lemma="gazette" pos="PROPN" ent="ORG">GAZETTE</tok>
  <tok sent="1" lemma="." pos="PUNCT" ent="">.</tok>

  <tok sent="2" lemma="six" pos="NUM" ent="">SIX</tok>
  <tok sent="2" lemma="dollar" pos="NOUN" ent="">DOLLARS</tok>
  <tok sent="2" lemma="a" pos="DET" ent="">A</tok>
  <tok sent="2" lemma="journal" pos="NOUN" ent="">JOURNAL</tok>
  <tok sent="2" lemma=" " pos="SPACE" ent=""> </tok>
  <tok sent="2" lemma="Tuesday" pos="NOUN" ent="">TUESDAY</tok>
  <tok sent="2" lemma="," pos="PUNCT" ent="">,</tok>
  <tok sent="2" lemma="January" pos="PROPN" ent="DATE">JANUARY</tok>
  <tok sent="2" lemma="16" pos="NUM" ent="DATE">16</tok>
  <tok sent="2" lemma="," pos="PUNCT" ent="">,</tok>
  <tok sent="2" lemma="1816" pos="NUM" ent="DATE">1816</tok>
  <tok sent="2" lemma="." pos="PUNCT" ent="">.</tok>
</document>
```

unstructured data

— — —

THE BARBADOS MERCURY,
AND BRIDGE-TOWN GAZETTE.
SIX DOLLARS A JOURNAL
TUESDAY, JANUARY 16, 1816.

structured data (JSON)

```
"document": {  
  "tokens": [  
    { "sent": "1", "lemma": "the", "pos": "DET", "ent": "", "text": "THE" },  
    { "sent": "1", "lemma": "Barbados", "pos": "PROPN", "ent": "LOC", "text": "BARBADOS" },  
    { "sent": "1", "lemma": "Mercury", "pos": "PROPN", "ent": "", "text": "MERCURY" },  
    { "sent": "1", "lemma": ",", "pos": "PUNCT", "ent": "", "text": "," },  
    { "sent": "1", "lemma": "and", "pos": "CCONJ", "ent": "", "text": "AND" },  
    { "sent": "1", "lemma": "Bridge", "pos": "PROPN", "ent": "LOC", "text": "BRIDGE" },  
    { "sent": "1", "lemma": "-", "pos": "PUNCT", "ent": "LOC", "text": "-" },  
    { "sent": "1", "lemma": "Town", "pos": "PROPN", "ent": "LOC", "text": "TOWN" },  
    { "sent": "1", "lemma": "gazette", "pos": "PROPN", "ent": "ORG", "text": "GAZETTE" },  
    { "sent": "1", "lemma": ".", "pos": "PUNCT", "ent": "", "text": "." },  
  
    { "sent": "2", "lemma": "six", "pos": "NUM", "ent": "", "text": "SIX" },  
    { "sent": "2", "lemma": "dollar", "pos": "NOUN", "ent": "", "text": "DOLLARS" },  
    { "sent": "2", "lemma": "a", "pos": "DET", "ent": "", "text": "A" },  
    { "sent": "2", "lemma": "journal", "pos": "NOUN", "ent": "", "text": "JOURNAL" },  
    { "sent": "2", "lemma": " ", "pos": "SPACE", "ent": "", "text": " " },  
    { "sent": "2", "lemma": "Tuesday", "pos": "NOUN", "ent": "", "text": "TUESDAY" },  
    { "sent": "2", "lemma": ",", "pos": "PUNCT", "ent": "", "text": "," },  
    { "sent": "2", "lemma": "January", "pos": "PROPN", "ent": "DATE", "text": "JANUARY" },  
    { "sent": "2", "lemma": "16", "pos": "NUM", "ent": "DATE", "text": "16" },  
    { "sent": "2", "lemma": ",", "pos": "PUNCT", "ent": "", "text": "," },  
    { "sent": "2", "lemma": "1816", "pos": "NUM", "ent": "DATE", "text": "1816" },  
    { "sent": "2", "lemma": ".", "pos": "PUNCT", "ent": "", "text": "." }  
  ]  
}
```

What are metadata?

— — —

- in short: information about a resource or digital asset
 - Examples: author name, date of creation, place, genre, language, keywords, collection name
- can be descriptive, administrative, operational
- content models/metadata schemes
 - Examples: Dublin Core, CIDOC-CRM, MODS
- metadata is interpretive: choices about terms, categories, and structure affect meaning and accessibility

Structured metadata for our corpus (in CSV):

— — —

filename,URL,date

AA00047511_01182.pdf,https://dloc.com/AA00047511/01182/pdf/0,1816-01-16
AA00047511_01194.pdf,https://dloc.com/AA00047511/01194/pdf/0,1816-02-27
AA00047511_01203.pdf,https://dloc.com/AA00047511/01203/pdf/0,1816-03-30
AA00047511_01206.pdf,https://dloc.com/AA00047511/01206/pdf/0,1816-04-09
AA00047511_01209.pdf,https://dloc.com/AA00047511/01209/pdf/0,1816-05-04
AA00047511_01220.pdf,https://dloc.com/AA00047511/01220/pdf/0,1816-06-11
AA00047511_01229.pdf,https://dloc.com/AA00047511/01229/pdf/0,1816-07-13
AA00047511_01237.pdf,https://dloc.com/AA00047511/01237/pdf/0,1816-08-10
AA00047511_01246.pdf,https://dloc.com/AA00047511/01246/pdf/0,1816-09-10
AA00047511_01256.pdf,https://dloc.com/AA00047511/01256/pdf/0,1816-10-15
AA00047511_01264.pdf,https://dloc.com/AA00047511/01264/pdf/0,1816-11-16
AA00047511_01273.pdf,https://dloc.com/AA00047511/01273/pdf/0,1816-12-17

https://github.com/dh-network/dh-ecf/blob/main/corpus_sources.csv

Can you open a CSV? 🤔

The rightmost column is in w3c date format

— — —



Examples of real-life metadata tables:

— — —

- DraCor: dracor.org/api/v1/corpora/ger/metadata/csv
- (same in JSON): dracor.org/api/v1/corpora/ger/metadata)
- ELTeC:
github.com/COST-ELTeC/ELTeC-eng/blob/master/ELTeC-eng_metadata.tsv
- Digital Damaged Ceramics:
<https://zenodo.org/records/15854369/files/object.csv?download=1>

06.11.2025

#2: Cleaning and Managing Data

Today

— — —

9:15–9:45 Research Data Management (RDM) fundamentals

- FAIR principles, data lifecycle
- Metadata standards and good documentation practices

9:45–10:30 Learning versioning and collaboration: GitHub

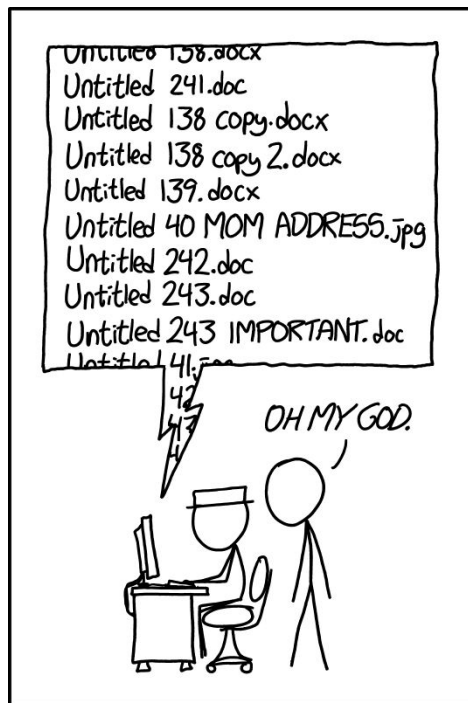
10:30–10:45 Break 🙄🙄🙄

10:45: Data cleaning

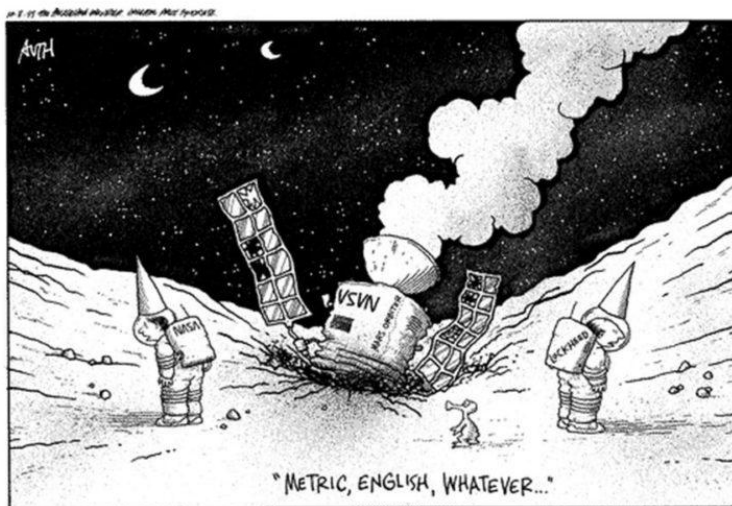
- Opening/editing data in advanced text editors
- Using RegEx for search/replace

11:20–11:45 Demo: OpenRefine

Why research data management?



PRO TIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.



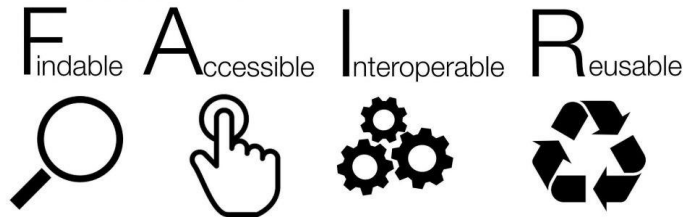
Remember the Mars Climate Orbiter incident from 1999?

Why research data management?

— — —

- Research data is valuable but fragile
 - “By 20 years post-publication, 80% of that data obtained through publicly funded research is inaccessible due to mundane issues, primarily old email addresses and obsolete storage devices” ([Vines et al. 2014](#))
- Poor RDM = Wasted Scholarly Labor
 - You spend years encoding texts, building databases, transcribing archives BUT you miss documentation, use inconsistent file naming, scatter data across devices
 - Consequences:
 - You can't reproduce your own analysis
 - Colleagues can't build on your work
 - Your data won't outlive the project
 - Years of work become invisible and lost

Make it FAIR!



Findable – Use persistent identifiers ([DOIs](#)), rich metadata, descriptive filenames

Accessible – Store data in reliable repositories with clear access conditions

Interoperable – Prefer standard formats (CSV, JSON) and follow community metadata standards

Reusable – Use clear licenses, provide complete documentation and provenance info on your data

👁 FAIR ≠ Open (data can be restricted but still well-managed)

Research data lifecycle

— — —

Plan: What data? What formats? Where stored? (Data Management Plan)

Collect: Consistent naming, document immediately, keep raw data separate

Analyse: Version control, record all transformations, use scripts

Preserve: Deposit in repositories, assign DOI, include README

Share & Reuse: License clearly, make discoverable, enable future research



[Source: UniVie](#)

Practical tips

— — —

- **Effective naming and meaningful folder structure:**
`YYYY-MM-DD_project-name_version.csv` instead of `data.csv`
- **Document everything** in a README (future-you is a stranger!)
- Store important data in multiple copies and on cloud
- When done: **deposit in Zenodo**, OSF, or institutional repository (**get a DOI**)
- For version control and collaboration: try **GitHub**

Data management with GitHub: Hands-on

— — —



GitHub is

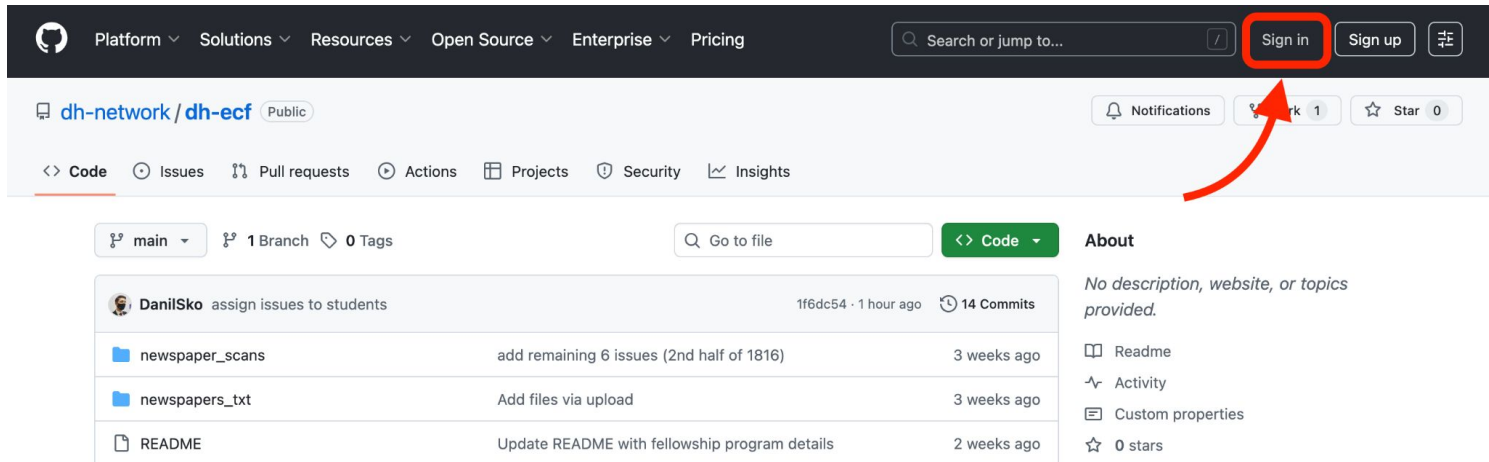
— — —

- used to store backups of your code and/or data
- a ‘track changes on steroids’ (the **Git** part)
- a collaboration platform (the **Hub** part)
- FAIR (kinda)

Let's try it out together

— — —

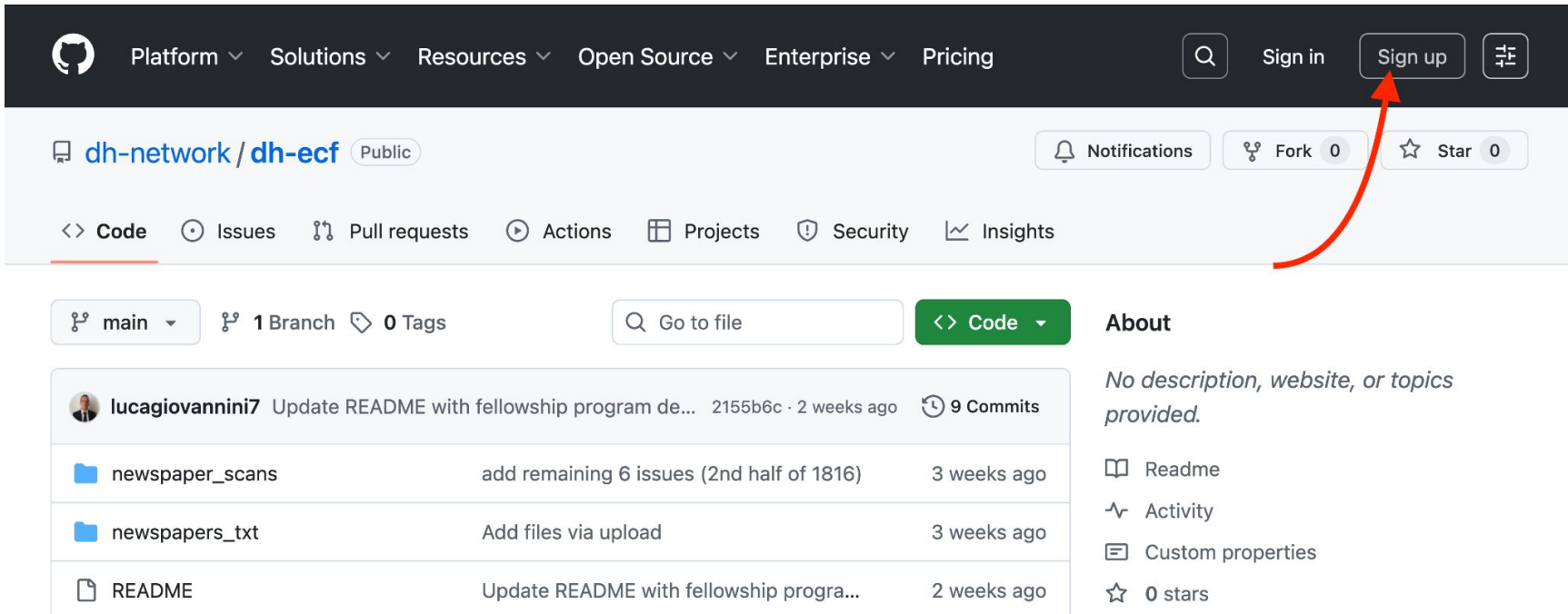
- First, you need to be logged into your GitHub account:



The screenshot shows the GitHub interface for the repository `dh-network / dh-ecf`. The top navigation bar includes links for Platform, Solutions, Resources, Open Source, Enterprise, and Pricing, along with a search bar and buttons for Sign in and Sign up. The repository page shows the file explorer with a table of files and their commit history. The 'Sign in' button is highlighted with a red box and a red arrow pointing to it.

File	Commit Message	Commit Hash	Time Ago
newspaper_scans	add remaining 6 issues (2nd half of 1816)	1f6dc54	1 hour ago
newspapers_txt	Add files via upload		3 weeks ago
README	Update README with fellowship program details		2 weeks ago

Or create a GitHub account if you haven't yet



The screenshot shows the GitHub interface for the repository `dh-network / dh-ecf`. The top navigation bar includes links for Platform, Solutions, Resources, Open Source, Enterprise, and Pricing, along with search, Sign in, and Sign up buttons. A red arrow points to the 'Sign up' button. Below the repository name, there are buttons for Notifications, Fork (0), and Star (0). The repository is public and has a main branch with 1 branch and 0 tags. The file list shows three items: `newspaper_scans`, `newspapers_txt`, and `README`. The right sidebar contains the 'About' section, which states 'No description, website, or topics provided.' and lists links for Readme, Activity, Custom properties, and 0 stars.

Platform Solutions Resources Open Source Enterprise Pricing

Search Sign in Sign up

dh-network / dh-ecf Public

Notifications Fork 0 Star 0

Code Issues Pull requests Actions Projects Security Insights

main 1 Branch 0 Tags

Go to file Code

lucagiovannini7 Update README with fellowship program de... 2155b6c · 2 weeks ago 9 Commits

newspaper_scans	add remaining 6 issues (2nd half of 1816)	3 weeks ago
newspapers_txt	Add files via upload	3 weeks ago
README	Update README with fellowship progra...	2 weeks ago

About

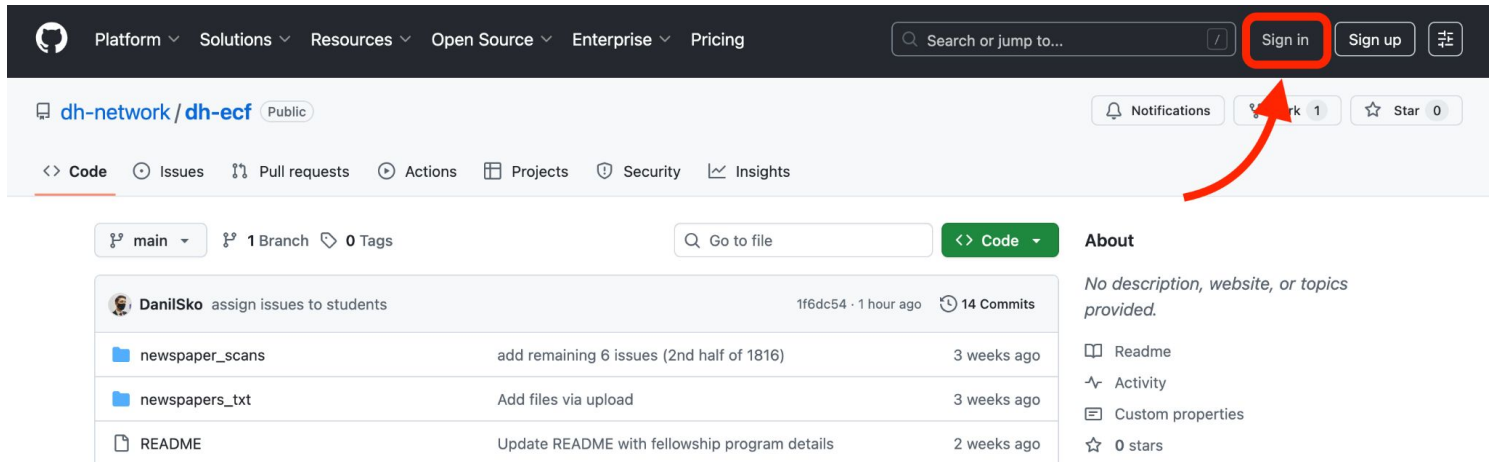
No description, website, or topics provided.

Readme Activity Custom properties 0 stars

Let's try it out together

— — —

- First, you need to be logged into your GitHub account:



- Once everyone is done... 

Let's try it out together

- We will be enriching our metadata table together with some (hopefully) meaningful data
- Here is a [list](#), in which **everyone is assigned one issue** of the Barbados Mercury Gazette
- Take the issue that is assigned to you, skim through the text and **try to count the number of ads for the search of missing slaves**; in the end we'll need just **one number**

Examples

RUNAWAY from the subscriber, on Friday last, a tall, slim, black GIRL, named Nanny, formerly the property of Kitty Cox, in Reed-street, and was in the habit of selling bread and cakes. Whoever will bring the said slave to No, 8, in Jew-street, shall receive the reward of Four Dollars. All persons are cautioned against harbouring her, as the law will be enforced against any so offending.

Jan. 16—2a. **LEVI ELEASER.**

CAUTION.
ABSCONDED from the service of her owner, on the 2d instant, a Negro WOMAN named Thomazin, formerly the property of Charlotte Doldron. Whoever will apprehend the said Slave, and lodge her in the Cage, shall receive, by application at this Office, a Reward of Five Pounds; and all persons are hereby cautioned against harbouring, employing, or taking her off the Island, as the law, in either case, will be most rigidly enforced against any so offending.

April 6—2a.

ABSENTED from the service of the subscriber, since the month of August last, a young Barbadian Negro WOMAN named Charlotte, aged about 22 years, 5 feet 7 inches in height, slender made, parrot-toed, and her right breast larger than the left. She has been seen near the Belle Plantation, the Hole-Town, Porey Spring, and lately at the Place of Mr. Lloyd, near Pilgrim. A reward of Fifty Shillings will be given for apprehending and delivering her to the subscriber, at his Place in Cotton Coast, or lodging her in the Cage of Bridge-Town;—and whoever will give information that will be sufficient to convict any white or free person or persons having harboured said slave, shall receive all the penalty that shall be recovered by virtue of the Detinue Act, against such offender or offenders. All masters of vessels are cautioned carrying said slave from this Island;—and the subscriber forbids all persons employing, or in any manner dealing with any or either of the following slaves, viz—Francis, a mason; Ben Kitt, a carpenter; and King, a taylor; as the law will be enforced against all offending.

Jan. 16—3a. **T. BUSBY.**
 If the Woman, Charlotte, return home of her own accord, she will be pardoned.

TEN POUNDS REWARD.
ABSENTED herself on the 1st January, without provocation, the subscriber's Negro WOMAN named Peggy Phillis. She formerly belonged to Mrs. Eliz. Davis, and lately to Mr. Salom. She is an African, but speaks well. She some time last year was selling goods for Mr. Thomas Atkins. The above reward will be given to any person delivering her to the subscriber; and one-half the sum recovered by the Detinue Act, will be given to any person legally convicting any white or free coloured person of harbouring or employing her, as the Law will be most rigidly enforced against all so offending.

April 13—2a. **CHARLES CARRINGTON.**

A REWARD of Ten Pounds each will be paid by the subscriber, for apprehending and delivering to him, or lodging in the Cage, the following SLAVES, viz. Mary Ann, Betsy Renn, Philly, and Kitty (Women), and Sammy (a man), a mason by trade. These people were formerly the property of Mr. B. Odwin. The Women are supposed to be harboured at Upper Birney's and Friendship Estates; or in Bridge-Town, by some of the servants of Wm. Yard, Taylor; and the Man, by his sister, at Mrs. Sparrock's, in the Bay.—Persons are forbid harbouring or concealing, and Captains of vessels taking them off the Island, as the law will be rigidly enforced against such offenders.

April 6—tf.

W. LICORISH,
 Overseer on Port's Plantation.

RUNAWAY from the subscriber, a Negro Boy named TIM:—One Dollar reward will be given to apprehend him, and a further reward of £5. to any person that will give such information as will prove his being harboured with any free coloured or white person.

Aug. 31—2a.

JOHN WELCH

(each counts as one)

Let's try it out together

- We will be enriching our metadata table together with some (hopefully) meaningful data
- Here is a [list](#), in which **everyone is assigned one issue** of the Barbados Mercury Gazette
- Take the issue that is assigned to you, skim through the text and **try to count the number of ads for the search of missing slaves**; in the end we'll need just **one number**
- Once everyone is done and has the number... 

Let's try it out together

1. Click on [corpus_sources.csv](#) in the repository.
2. Click the  **Edit this file** icon (top right).
3. Find the row corresponding to the [issue assigned to you](#).
4. In the column **n_ads**, enter the number of advertisements you have counted (after the last comma; do not remove the commas).
5. Scroll down to the **Commit changes** section:
 - a. Write a short message like *Added n_ads value for issue 1875-03-10 (Anna)*
 - b. Select **“Commit directly to the main branch.”**
6. Click **Commit changes**.

dh-ecf / corpus_sources.csv

6271ec4 · 2 hours ago

Preview Code Blame 13 lines (13 loc) · 900 Bytes

Search this file

	filename	URL	date	n_ads
1	AA00047511_01182.pdf	https://dloc.com/AA00047511/01182/pdf/0	1816-01-16	
2	AA00047511_01194.pdf	https://dloc.com/AA00047511/01194/pdf/0	1816-02-27	
3	AA00047511_01203.pdf	https://dloc.com/AA00047511/01203/pdf/0	1816-03-30	
4	AA00047511_01206.pdf	https://dloc.com/AA00047511/01206/pdf/0	1816-04-09	
5	AA00047511_01209.pdf	https://dloc.com/AA00047511/01209/pdf/0	1816-05-04	
6	AA00047511_01220.pdf	https://dloc.com/AA00047511/01220/pdf/0	1816-06-11	
7	AA00047511_01229.pdf	https://dloc.com/AA00047511/01229/pdf/0	1816-07-13	
8	AA00047511_01237.pdf	https://dloc.com/AA00047511/01237/pdf/0	1816-08-10	
9	AA00047511_01246.pdf	https://dloc.com/AA00047511/01246/pdf/0	1816-09-10	
10	AA00047511_01256.pdf	https://dloc.com/AA00047511/01256/pdf/0	1816-10-15	
11	AA00047511_01264.pdf	https://dloc.com/AA00047511/01264/pdf/0	1816-11-16	
12	AA00047511_01273.pdf	https://dloc.com/AA00047511/01273/pdf/0	1816-12-17	

dh-ecf / corpus_sources.csv in main

Cancel changes Commit changes...

Spaces 2 No wrap

Edit Preview

```
1 filename, URL, date, n_ads
2 AA00047511_01182.pdf, https://dloc.com/AA00047511/01182/pdf/0, 1816-01-16,
3 AA00047511_01194.pdf, https://dloc.com/AA00047511/01194/pdf/0, 1816-02-27,
4 AA00047511_01203.pdf, https://dloc.com/AA00047511/01203/pdf/0, 1816-03-30,
5 AA00047511_01206.pdf, https://dloc.com/AA00047511/01206/pdf/0, 1816-04-09,
6 AA00047511_01209.pdf, https://dloc.com/AA00047511/01209/pdf/0, 1816-05-04,
7 AA00047511_01220.pdf, https://dloc.com/AA00047511/01220/pdf/0, 1816-06-11,
8 AA00047511_01229.pdf, https://dloc.com/AA00047511/01229/pdf/0, 1816-07-13,
9 AA00047511_01237.pdf, https://dloc.com/AA00047511/01237/pdf/0, 1816-08-10,
10 AA00047511_01246.pdf, https://dloc.com/AA00047511/01246/pdf/0, 1816-09-10,
11 AA00047511_01256.pdf, https://dloc.com/AA00047511/01256/pdf/0, 1816-10-15,
12 AA00047511_01264.pdf, https://dloc.com/AA00047511/01264/pdf/0, 1816-11-16,
13 AA00047511_01273.pdf, https://dloc.com/AA00047511/01273/pdf/0, 1816-12-17,
14
```

dh-ecf / corpus_sources.csv in main

Cancel changes Commit changes...

Edit Preview

Commit changes

Commit message

added n_ads for issue AA00047511_01229 (Danil)

Extended description

Add an optional extended description...

☒ Commit directly to the main branch

☐ Create a new branch for this commit and start a pull request

[Learn more about pull requests](#)

Cancel Commit changes

Preview Code Blame 13 lines (13 loc) · 901 Bytes

Raw

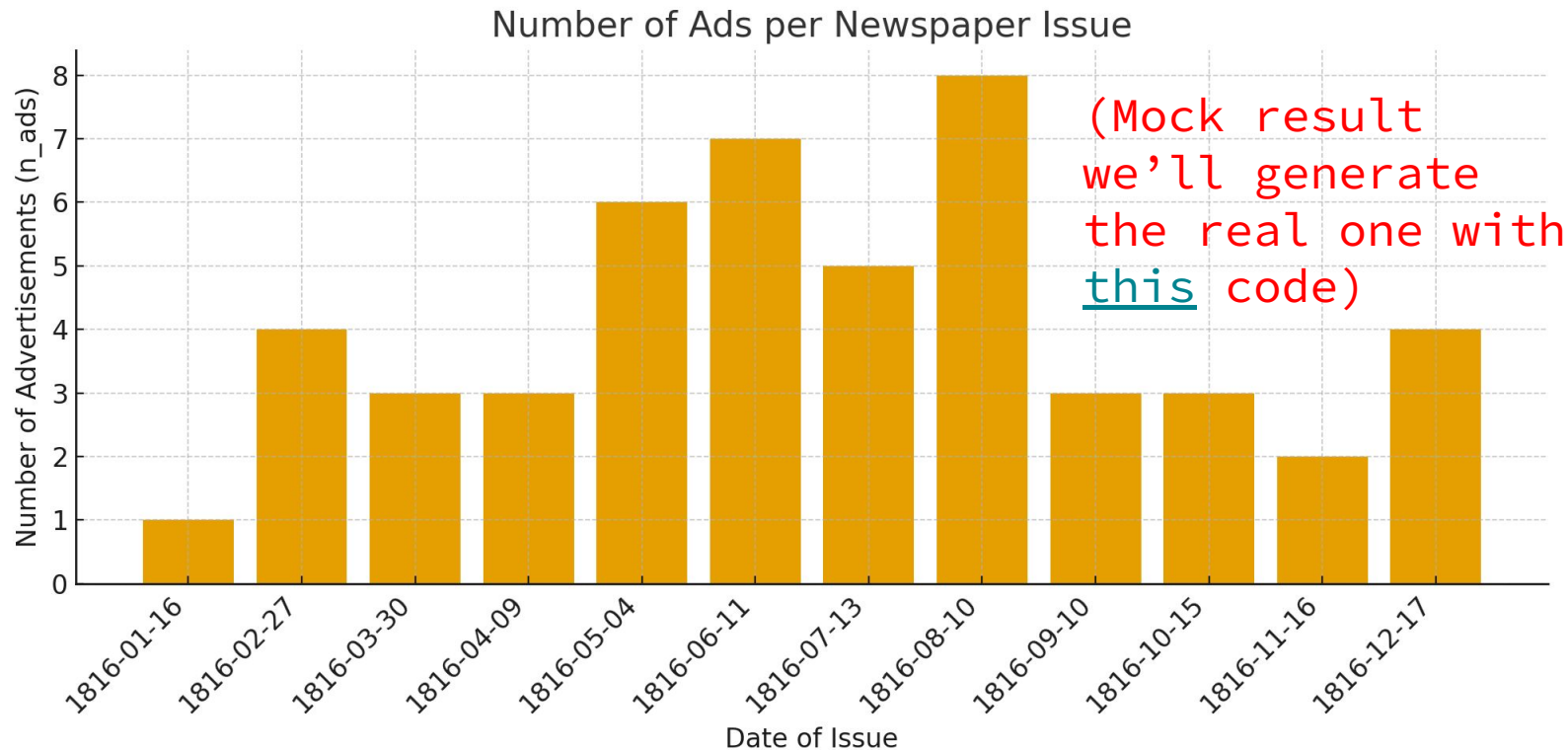
Search this file

	filename	URL	date	n_ads
1	AA00047511_01182.pdf	https://dloc.com/AA00047511/01182/pdf/0	1816-01-16	
2	AA00047511_01194.pdf	https://dloc.com/AA00047511/01194/pdf/0	1816-02-27	
3	AA00047511_01203.pdf	https://dloc.com/AA00047511/01203/pdf/0	1816-03-30	
4	AA00047511_01206.pdf	https://dloc.com/AA00047511/01206/pdf/0	1816-04-09	
5	AA00047511_01209.pdf	https://dloc.com/AA00047511/01209/pdf/0	1816-05-04	
6	AA00047511_01220.pdf	https://dloc.com/AA00047511/01220/pdf/0	1816-06-11	
7	AA00047511_01229.pdf	https://dloc.com/AA00047511/01229/pdf/0	1816-07-13	5
8	AA00047511_01237.pdf	https://dloc.com/AA00047511/01237/pdf/0	1816-08-10	
9	AA00047511_01246.pdf	https://dloc.com/AA00047511/01246/pdf/0	1816-09-10	
10	AA00047511_01256.pdf	https://dloc.com/AA00047511/01256/pdf/0	1816-10-15	
11	AA00047511_01264.pdf	https://dloc.com/AA00047511/01264/pdf/0	1816-11-16	
12	AA00047511_01273.pdf	https://dloc.com/AA00047511/01273/pdf/0	1816-12-17	

Done! 🎉

Let's look at the results

--



Now, if we trusted you less...

- we wouldn't give you the write-access to the repo
- but you'd still be able to introduce changes
- you'd do it by **forking** (i.e. copying) the whole repository...
- ...applying changes in your 'forked' copy...
- ...and making a **pull request** (i.e. asking us to pull your changes upstream into the main repository)
- **pull requests** are a basic 'pls review my changes' mechanism in the software world

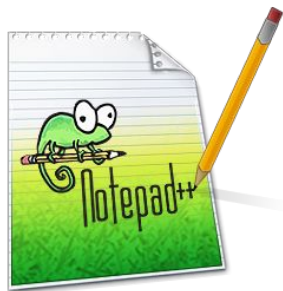
How collaboration on GitHub works IRL: a 3-min video

— — —



Advanced notepad-like editors

- highlight syntax for data formats (JSON, XML etc) and code
- allow encoding switch (solves the `?????????`-like gibberish problem)
- support advanced search options (e.g. with regular expressions) ➡



Brackets

Regular Expressions (regex) are

- a "super-powered search/replace"
- a "language of patterns" that allows to search for things like:
 - all words of any length starting with "pre" and ending with "ment"
 - all 4-digit numbers in the file
 - all words that repeat twice in a row
 - all lines in the file that contain 7 to 9 commas
- handy when dealing with messy/broken data

Regular Expressions are ‘expressed’ with metacharacters:

- Metacharacter Dot(**.**) means any symbol:

REGULAR EXPRESSION

83 matches

:/.

TEST STRING

Today•at•DH•Netzwerk•Potsdam↵
We•are•learning•regular•expressions↵
with•Daniil•and•Luca|

Regular Expressions are ‘expressed’ with metacharacters:

- Metacharacter Dot(**.**) means any symbol:

The screenshot shows a web-based regular expression testing tool. At the top, the 'REGULAR EXPRESSION' field contains the pattern `u.a`, with the dot character highlighted in blue. To the right of this field, a green badge indicates '2 matches'. Below the expression field is the 'TEST STRING' section, which contains the text: 'Today at DH Netzwerk Potsdam', 'We are learning regular expressions', and 'with Daniil and Luca'. The words 'regular' and 'Luca' are highlighted in blue, indicating they are the matches for the pattern `u.a`.

REGULAR EXPRESSION	TEST STRING
<code>u.a</code>	Today at DH Netzwerk Potsdam We are learning regular expressions with Daniil and Luca

Regular Expressions are ‘expressed’ with metacharacters:

- Metacharacter Plus(+) means any positive number of previous symbol:

REGULAR EXPRESSION

5 matches

:

/

i

+

TEST STRING

Today•at•DH•Netzwerk•Potsdam↵
We•are•learning•regular•expressions↵
with•Daniil•and•Luca↵
ii|

Regular Expressions are 'expressed' with metacharacters:

- Metacharacter Plus(+) means any positive number of previous symbol:

REGULAR EXPRESSION

4 matches

:/ .+

this looks for any number of any symbols

TEST STRING

Today•at•DH•Netzwerk•Potsdam↵
We•are•learning•regular•expressions↵
with•Daniil•and•Luca↵
iiiiiiiiiiiiiiiiiiiiiiiiiiiiii

Regular Expressions are ‘expressed’ with metacharacters:

- Group of metacharacters (`[ae]`) means any one of the symbols `[` inside `]` the brackets:

REGULAR EXPRESSION

2 matches

`/gr[ae]y`

TEST STRING

the•clouds•in•Potsdam•are•gray↵
Today•at•DH•Netzwerk•Potsdam↵
We•are•learning•regular•expressions•|↵
the•weather•is•grey

Regular Expressions are 'expressed' with metacharacters:

- Group of metacharacters (`[A-Z]`) means any capital latin letter:

REGULAR EXPRESSION

8 matches

⋮ / `[A-Z]`

TEST STRING

Today•at•DH•Netzwerk•Potsdam↵
We•are•learning•regular•expressions↵
with•Daniil•and•Luca↵
ii

Regular Expressions are 'expressed' with metacharacters:

- Group of metacharacters (`\w`) means any alphanumeric character or digit

SAVE & SHARE

 Save new Regex ⌘+s

 Add to Community Libr...

FLAVOR

</> PCRE2 (PHP >=7.3)

</> PCRE (PHP <7.3)

</> ECMAScript (JavaScript)

</> Python ✓

REGULAR EXPRESSION 97 matches

`: r" \w`

TEST STRING

Today • on • 6.11.2025 • at • DH • Netzwerk • Potsdam ↵
We • are • learning • regular • expressions • ↵
With • Daniil • and • Luca ↵
Привет, • Лука ↵
义勇军进行曲

Regular Expressions are 'expressed' with metacharacters:

- If you combine(`\w`) with (`+`), it gives an approximation of word count

SAVE & SHARE

 Save new Regex 

 Add to Community Lib...

FLAVOR

</> PCRE2 (PHP >=7.3)

</> PCRE (PHP <7.3)

</> ECMAScript (JavaScript)

</> **Python** 

REGULAR EXPRESSION

21 matches

`:\w+`

TEST STRING

Today on 6.11.2025 at DH Netzwerk Potsdam
We are learning regular expressions
With Daniil and Luca
Привет, Лука
义勇军进行曲

Regular Expressions are 'expressed' with metacharacters:

- Group of metacharacters (`\d`) means any arabic digit

```
REGULAR EXPRESSION7 matches  
  
::/\d  
  
TEST STRING  
  
Today, on 6.11.2025 at DH Netzwerk Potsdam  
We are learning regular expressions  
with Daniil and Luca  
iiiiiiiiiiiiiiiiiiiiiiiii
```

Regular Expressions are 'expressed' with metacharacters:

- Group of metacharacters (`{n}`) means the previous character (or metacharacter) repeats n times

REGULAR EXPRESSION

1 match

`:/ \d{4}` this looks for any four digits

TEST STRING

Today, on 6.11.2025, at DH Netzwerk Potsdam

We are learning regular expressions

with Daniil and Luca

iiiiiiiiiiiiiiiiiiiiiiiiii

A small hands-on task

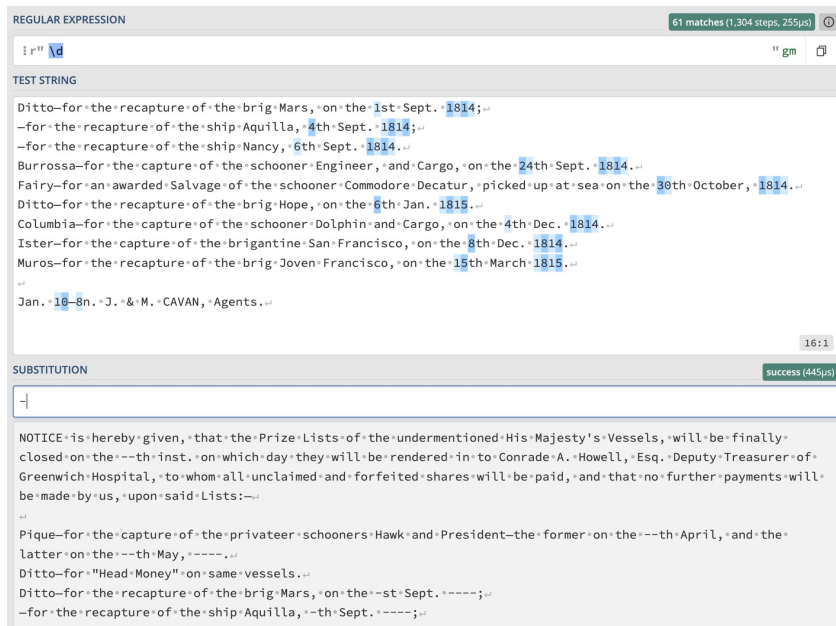
Try to find all mentions of **full years** in this issue of the Gazette:

github.com/dh-network/dh-ecf/blob/main/newspapers_txt/AA00047511_01194.txt

(hint: a digit is found with `\d`
or `[0-9]`)

Regular Expressions can be used to modify text with replace:

- E.g. you might want to remove all `\d` to remove sensitive data



Regular Expressions can be used to modify text with replace:

— — —

REGULAR EXPRESSION

12 matches (72 steps, 200µs)

REGULAR EXPRESSION

`/(.+)`

" gm

TEST STRING

Taisija
Tim
Luca
Damla
Anna
Ana-Maria
Anna
Alberto
Peer
Samira
Sophie
Nina

12:5

SUBSTITUTION

success (65µs)

SUBSTITUTION

`<name>\1</name>`

`<name>Taisija</name>`
`<name>Tim</name>`
`<name>Luca</name>`
`<name>Damla</name>`
`<name>Anna</name>`
`<name>Ana-Maria</name>`
`<name>Anna</name>`
`<name>Alberto</name>`
`<name>Peer</name>`
`<name>Samira</name>`

A small hands-on task

— — —

Try to estimate the number of words in this text with regex:

J. LAW

HAS imported in the vessels just arrived from England and Ireland, the following articles, which he offers for sale at reduced prices:—

Superfine broad cloths and kerseymeers of fashionable colours; second quality cloths, coarse green cloth, superfine white Welsh flannels and swansdowns; 4-4 Irish linen of all qualities; 5-4 and 6-4 Irish sheetings; long lawns, diaper tablecloths 8-4 and 10-4 huckaback, damask napkins, napkin and hiping diaper, French cambrics, Russia sheetings, white and coloured Marseilles waisting, black silk florentine, India nankeens and twills; white Manchester nankeen, and calicoes, plain and twilled; York stripe, dimities; India calicoes, blue and white; romal handkerchiefs, shirting calico, denims, rich furniture chintz, fine printed calicoes, Marseilles quilts, cotton counterpanes; mattresses, bolsters, feather-pillows, and feathers in bags; Irish and Flanders bed-ticking, and ready-made beds; Nun's thread and sewing silk; bombazeen, bombazeette, black and coloured crape, 6-4 black cambric muslin, plain sarcenet, shalloons, fashionable parasols, silk and gingham umbrellas, men's and boys' beaver and silk hats, broad-brim drab hats; silk and cotton hose, children's do. men's cotton gloves, white and coloured unbleached cotton stockings and socks, 4-4 indigo blue pennistone, Drogheda linens, Monmouth hats, Kilmarnock caps, linen and cotton checks and stripes, coarse threads, London-made hoes and cane bills, nails of all sorts, paints, oil; sheathing copper, and patent nails; bar iron, blistered steel, puncheon hoops, sheet lead, window-glass, East Coker canvass, Dry-Books, Invoice Books, &c.; common writing paper by the ream; table knives and forks, mahogany knife-cases, portable writing desks, looking glasses, Japan ware, saddles. Also for sale, 100 new sugar tierces; a lot of soap and candles now landing; tobacco in barrels and hogsheads; and gin, in pipes.

Jan. 13—tf.

Cleaning data with OpenRefine: Demo



OpenRefine

20.11.2025

#3: Processing and Sharing Data

Today

9:15-10:00 Sharing Data: Zenodo, DOI, Licenses (Annette Gerstenberg)

10:00-11:45 Processing (textual) Data: Basic NLP routines (Henny Sluyter-Gäthje)

Topics

- CARE
- FAIR, F – A – I – R
- Licenses: Creative Commons

CARE principles



Stephanie Russo Carroll u. a. 2022. „Using Indigenous Standards to Implement the CARE Principles: Setting Expectations through Tribal Research Codes“. *Frontiers in genetics* 13:823309. <https://doi.org/10.3389/fgene.2022.823309>

Research Data Management Plans based on FAIR principles

- In today's research, Research Data Management Plans (RDMP) have become an important tool for professionally organising your research data, and demonstrating good practice.
- Many organizations in your field of research provide RDMP models, e. g. Text+. 2025. *Research Data Management* (v. 2025-11-04. Online: Text+ Consortium. <https://text-plus.org/en/themen-dokumentation/forschungsdatenmanagement/>

FAIR principles – choice and obligation

- Choice: You are a fair person.
 - FAIR principles reflect the spirit of Open Science.
 - The FAIR principles grant ethical standards in the **humanities**.
 - FAIR principles promote sustainability.
- Obligation: FAIR principles in your Research Data Management Plan (RDMP)
 - The RDMP can be crucial for data management in your master's thesis.
 - The RDMP is crucial in the application of third party funding.
 - The RDMP can be crucial for publishing research paper based on your data.

What do FAIR principles apply to?

- „The FAIR principles apply to both data and metadata“ (Harms Buss u. a. 2023)
- What else is covered by FAIR principles?
 - Methodology descriptions
 - Codebooks
 - Questionnaires
 - Scripts like editor- and do-files (STATA)
 - Laboratory notebooks and experimental protocols
 - Software syntax and output files
 - Database schemes
 - Provenance information about secondary data
 - The finalised data management plan

Mark D. Wilkinson u. a. 2016. „The FAIR Guiding Principles for scientific data management and stewardship“. *Scientific data* 3:1–9. <https://doi.org/10.1038/sdata.2016.18>

FAIR, F ‘findable’: the crucial role of (meta-)data

- Administrative metadata are data about a project or resource that are relevant for managing it; for example, project/ resource owner, principal investigator, project collaborators, funder, project period, etc. They are usually assigned to the data, before you collect or create them.
- Descriptive or citation metadata are data about a dataset or resource that allow people to discover and identify it; for example, authors, title, abstract, keywords, persistent identifier, related publications, etc.

- Structural metadata
 - . . . are data about how a dataset or resource came about, but also how it is internally structured.
 - Structural metadata describe, for example, the unit of analysis, collection method, sampling procedure, sample size, categories, variables, etc.
 - Structural metadata have to be gathered by the researchers according to best practice in their research community and will be published together with the data.
 - Descriptive and structural metadata should be added continuously throughout the project.

Metadata standards: Resource Description Framework (RDF)

- Graph model
- Using triples of subject, predicate, and object
- RDF can be written in XML: RDF/XML (but other options are available)
- RDF is used in the Semantic Web, it enables linking and integrating data from various sources

Metadata standards

- Resource Description Framework (RDF) for the Semantic Web: is a model
- Extensible Markup Language (XML): is a syntax
- RDF can be written in XML: RDF/XML (but other options are available)
- RDF can rely on an Ontology, a formal representation of knowledge, representing hierarchies and networks

Metadata standards

- Dublin Core Metadata Initiative (DCMI) = Dublin Core Metadata Initiative. 2025.
DCMI Schemas (v. 0.89.3 on Mon, 03 Jan 2022). Leesburg (VA): asis&t. <https://dublincore.org/schemas/>
- Data Documentation Initiative (DDI)
- Text Encoding Initiative (TEI) = TEI Consortium. 2021. „P5 Guidelines 31st August 2021 (Version 4.3.0, revision b4f72b1ff): P5/TEI“. In *Text Encoding Initiative: TEI*, herausgegeben von TEI Consortium. Institute for Advanced Technology in the Humanities Charlottesville (Vi).
<http://www.tei-c.org/Guidelines/P5/index.xml>

Metadata standards: DCMI

- title
- Subject
- Description
- Type
- Source
- Relation
- Coverage
- Creator
- Publisher
- Contributor
- Rights
- Date
- Format
- Identifier
- Language
- Audience
- RightsHolder

- **InstructionalMethod:** A process, used to engender knowledge, attitudes and skills, that the resource is designed to support. Instructional Method will typically include ways of presenting instructional materials or conducting instructional activities, patterns of learner-to-learner and learner-to-instructor interactions, and mechanisms by which group and individual levels of learning are measured. Instructional methods include all aspects of the instruction and learning processes from planning and implementation through evaluation and feedback.
- **AccrualMethod:** The method by which items are added to a collection. Recommended best practice is to use a value from a controlled vocabulary.
- **AccrualPeriodicity:** The frequency with which items are added to a collection. Recommended best practice is to use a value from a controlled vocabulary.
- **AccrualPolicy:** The policy governing the addition of items to a collection. Recommended best practice is to use a value from a controlled vocabulary.

Dublin Core Metadata Initiative 2025 DCMI
elements

DublinCore: applications

- DublinCore metadata are standards, need to be applied and specified
- Applications are also standardized, e. g. by the Open Languages Archiving Community (OLAC)

Example of metadata using Dublin Core, RDF/XML

```
<dc:title xml:lang ="fr">ESL02: entretien 1053
</dc:title>
<dc:description xml:lang ="fr">Extrait de la
seconde Enquête Sociolinguistique à Orléans
réalisée autour des années 2010. </dc:description>
<dc:source xml:lang ="fr">Bande magnétique 1053
</dc:source>
<dc:publisher>Laboratoire Ligérien de Linguistique
</dc:publisher>
<dc:contributor olac:code ="depositor" xsi:type
="olac:role">Laboratoire Ligérien de
Linguistique </dc:contributor>
```

Metadata for
ESL02 ENT 1053

To make your data accessible and easy to find, you must provide your data and metadata with a persistent identifier (PID).

A PID is a long-lasting reference to a digital resource and provides the information required to reliably identify, verify and locate your research data.

Harms Buss u. a. 2023

Persistent identifiers (PID) grant findability

- Persistent URL (PURL)
- Digital Object Identifiers (DOI)
- Uniform Resource Name (URN)
- Other Handles, such as Archival Resource Key (ARK)

Good practice: complete references to internet resources

- Uniform Resource locators (URL) don't identify, they (temporarily) locate, and thus don't grant a persistent reference.
- Neither DOI nor URL are stand-alone references.
- Always provide author(s), title, place, institution, date when referencing any internet source.

Persistent Identifier: Digital Object Identifier (DOI)



The DOI Foundation is a not-for-profit organization. We govern the Digital Object Identifier (DOI) system on behalf of the agencies who manage DOI registries and provide services to their respective communities. We are the registration authority for the ISO standard (ISO 26324) for the DOI system and we are governed by our Registration Agencies.

Persistent Identifier: Archival Resource Key (ARK)

NOTICE

Titre : Les Discours de Guillaume II pendant la guerre. Recueillis par Mme Marie Méring. Traduits par Mme A. H.,...

Auteur : Guillaume II (empereur d'Allemagne ; 1859-1941). Auteur du texte

notice.date : 1918

Contributeur : Méring, Marie. Éditeur scientifique

Type : monographie imprimée

Langue : français

Format : Nombre total de vues : 64

Droits : domaine public

Identifiant : ark:/12148/bpt6k372873v

Source : Bibliothèque nationale de France, département Philosophie, histoire, sciences de l'homme, 8-M-20634

Notice du catalogue : <http://catalogue.bnf.fr/ark:/12148/cb322065882>

Provenance : Bibliothèque nationale de France

Date de mise en ligne : 09/05/2011

Link to the catalogue entry at Bibliothèque Nationale de France (BNF)/Gallica

FAIR, A ‘accessible’: repositories

- Darjah, Clarin, University repositories, FID Romanistik
- OSF = Center for Open Science. 2025. *Open Science Framework (OSF)*. Charlottesville: Center for Open Science. <https://osf.io/>
- Zenodo = European Organization for Nuclear Research und OpenAIRE. 2013. *Zenodo*. Geneva: CERN Data Centre / Invenio / CERN. <https://doi.org/10.25495/7gxx-rd71>

Zenodo

- Allows for publishing all sorts of data types.
- Allows for creating a proper reference with DOI.
- Allows groups to link their projects.

Example: RCFC – réseau de corpus du français préclassique et classique

Zenodo: Bingo, and review

- What different kinds of uploads do you find on Zenodo?
- What do you learn about the related metadata?
- What is missing in the available metadata?

FAIR, I ‘interoperable’: data can be used in different programs

→ as stated by Ferdinand de Saussure (1857–1913): *le signe (linguistique) est arbitraire.*

→ a simple proof of interoperability: you can open it in an editor!

FAIR, R ‘reusable’: ethics, legal, copyright issues

- Ethics: BAAL = The British Association for Applied Linguistics. 2021. *Recommendations on Good Practice in Applied Linguistics, 4th Edition*. BAAL. <https://www.baal.org.uk/wp-content/uploads/2021/03/BAAL-Good-Practice-Guidelines-2021.pdf>
 - DFG Good practice = Deutsche Forschungsgemeinschaft. 2019. *Leitlinien zur Sicherung guter wissenschaftlicher Praxis. Kodex, korrigierte Version 1.2 September 2024*. Bonn: DFG. <https://doi.org/10.5281/zenodo.14281892>
 - General Data Protection Regulation (GDPR), Datenschutz Grundverordnung (DSGVO), see Universität Potsdam. 2024. *Datenschutzerklärung zu den Webseiten der Universität Potsdam (28.03.2024)*. Potsdam: Universität Potsdam. <https://www.uni-potsdam.de/de/datenschutzerklaerung>
- Social Media Data: check for terms of service, policies . . .
- Interviews, questionnaires: informed consent, check for standards

The spirit of GDPR

Key principles (Applied Linguistics 2021)

- Lawfulness, fairness and transparency
- Purpose limitation
- Data minimisation
- Accuracy
- Storage limitation
- Integrity and confidentiality (security)
- Accountability

Who owns research data?

In der deutschen Literatur wird mit überzeugenden Argumenten für das grundsätzliche Erfordernis eines Dateneigentums argumentiert, das einem analogen Sacheigentum entspricht. Eine endgültige Beurteilung, ob auch für Forschungsdaten eine derartige Regelungslücke besteht, kann jedoch erst erfolgen, wenn letztlich alle Schutzbedürfnisse und bereits bestehenden Schutzrechte verglichen worden sind.²⁹ Die Rechtssetzung auf europäischer Ebene vermag jedenfalls kein dingliches Eigentum für Forschungsdaten zu begründen.

Paul Baumann. 2023. *Rechte an Forschungsdaten*. Trier: Universität Trier / Institut für Recht und Digitalisierung Trier.
<https://doi.org/10.25353/ubtr-81e9-927d-e6dd>

Creative Commons

- CC 'Creative Commons': „Creative Commons licenses give everyone from individual creators to large institutions a standardized way to grant the public permission to use their creative work under copyright law“
- BY 'credit must be given to the creator'
- SA 'Share Alike': „This license enables reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use. If you remix, adapt, or build upon the material, you must license the modified material under identical terms.“
- NC Only 'noncommercial' uses are permitted
- ND 'No derivatives' or adaptations of the work are permitted

Creative Commons. 2019. *About CC Licences*. Mountain View (CA): CC. <https://creativecommons.org/share-your-work/cclicenses/>

Confused?

- Forschungsdatenmanagement Uni Potsdam
- NFDI-Konsortium Text+
- NFDI-Konsortium 4memory
- Fachinformationsdienst Romanistik:
Forschungsdaten

Bibliography

- Applied Linguistics, The British Association for. 2021. *Recommendations on Good Practice in Applied Linguistics, 4th Edition*. BAAL. <https://www.baal.org.uk/wp-content/uploads/2021/03/BAAL-Good-Practice-Guidelines-2021.pdf>.
- Baumann, Paul. 2023. *Rechte an Forschungsdaten*. Trier: Universität Trier / Institut für Recht und Digitalisierung Trier. <https://doi.org/10.25353/ubtr-81e9-927d-e6dd>.
- Carroll, Stephanie Russo, Ibrahim Garba, Rebecca Plevel, Desi Small-Rodriguez, Vanessa Y. Hiratsuka, Maui Hudson und Nanibaa' A. Garrison. 2022. „Using Indigenous Standards to Implement the CARE Principles: Setting Expectations through Tribal Research Codes“. *Frontiers in genetics* 13:823309. <https://doi.org/10.3389/fgene.2022.823309>.
- Center for Open Science. 2025. *Open Science Framework (OSF)*. Charlottesville: Center for Open Science. <https://osf.io/>.
- Consortium, TEI. 2021. „P5 Guidelines 31st August 2021 (Version 4.3.0, revision b4f72b1ff): P5/TEI“. In *Text Encoding Initiative: TEI*, herausgegeben von TEI Consortium. Institute for Advanced Technology in the Humanities Charlottesville (Vi). <http://www.tei-c.org/Guidelines/P5/index.xml>.
- Creative Commons. 2019. *About CC Licences*. Mountain View (CA): CC. <https://creativecommons.org/share-your-work/cclicenses/>.
- Deutsche Forschungsgemeinschaft. 2019. *Leitlinien zur Sicherung guter wissenschaftlicher Praxis. Kodex, korrigierte Version 1.2 September 2024*. Bonn: DFG. <https://doi.org/10.5281/zenodo.14281892>.

Dublin Core Metadata Initiative. 2025. *DCMI Schemas (v. 0.89.3 on Mon, 03 Jan 2022)*. Leesburg (VA): asis&t. <https://dublincore.org/schemas/>.

European Organization for Nuclear Research und OpenAIRE. 2013. *Zenodo*. Geneva: CERN Data Centre / Invenio / CERN. <https://doi.org/10.25495/7gxx-rd71>.

Harms Buss, Mareike Christina, Daniella Bayle Deutz, Katrine Flindt Holmstrand, Asger Vaering Larsen und Evgenios Vlachos. 2023. *How to Fair*. University of Southern Denmark: Danish National Forum for Research Data Management. <https://www.howtofair.dk/>.

Text+. 2025. *Research Data Management (v. 2025-11-04)*. Online: Text+ Consortium. <https://text-plus.org/en/themen-dokumentation/forschungsdatenmanagement/>.

Universität Potsdam. 2024. *Datenschutzerklärung zu den Webseiten der Universität Potsdam (28.03.2024)*.

Potsdam: Universität Potsdam. <https://www.uni-potsdam.de/de/datenschutzerklaerung>.

Wilkinson, Mark D., Michel Dumontier, I. Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg u. a. 2016. „The FAIR Guiding Principles for scientific data management and stewardship“. *Scientific data* 3:1–9. <https://doi.org/10.1038/sdata.2016.18>.

Basic NLP routines: Text for a Computer

— — —

ANY person willing to undertake the removal of the Copper Gutters at King's-House, and lay them anew, with lead, agreeable to a description to be seen at this Office, or that of the Deputy Assistant Quarter-Master-General,

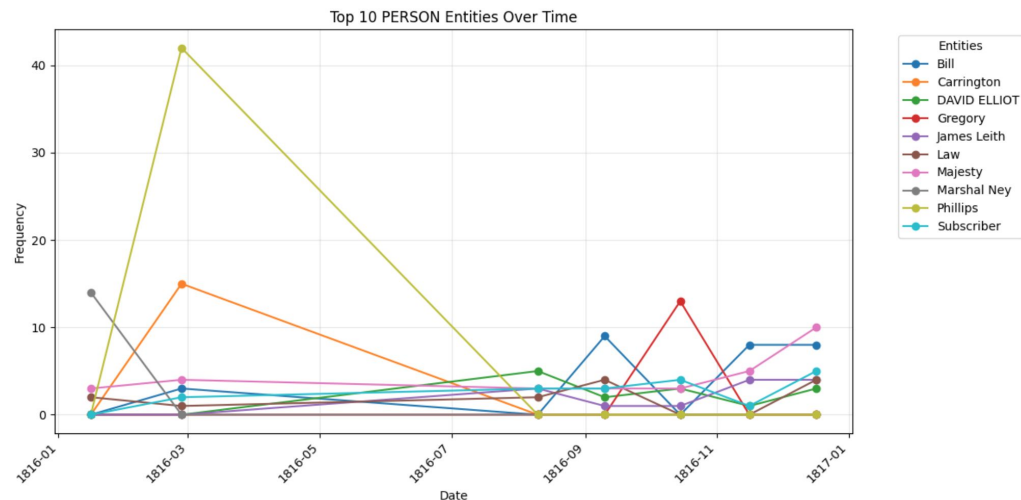
Will please send sealed Tenders to this Office, in Triplicate, expressing, in words at length, the rate at which they will undertake the same, on or before Thursday the 18th inst. at twelve o'clock, when they will be opened, and the one most favourable for the Public approved of.



jwioJ XCJSijs CISO cijos CIOj
sciOJC Diocdj dv dv reoigo
ofofieowhIO diowejifa er aio arjio
avoa avjio reig vajio vrjiov ro
vijearo vao vrjio vaero brio gtri
aeio verb ibar vbir vur ivahuirhv
urhiorn grior re i og gio giog
hiogo ernht iojiaore iorajojg aro
ggriao ioioeger njkernkaj reaio
ioriaioia ioisoifs iofdbdfo
ofdoif igueri ae aoira ioirai
haioai ihiovif orpwnbjp iro igrohi
aruhr tjiotis i i ij regio
agriaojiji jior ajair ojio
aeriaojio graiori aewvjfiop

Basic NLP routines: What if ...

... we want to analyse which **persons** are the most prominent in the *Barbados Mercury and Bridge-Town Gazette* and how the count develops over time



Basic NLP routines: Methods

NLP = Natural Language Processing

One part of NLP: Linguistic annotation of text data

- Tokenisation
- Lemmatisation
- Part-of-Speech-Tagging (PoS-Tagging)
- Dependency Parsing
- Named Entity Recognition

Basic NLP routines: Tokenisation

— — —

ANY person willing to undertake the removal of the Copper Gutters at King's-House, and lay them anew, with lead, agreeable to a description to be seen at this Office, or that of the Deputy Assistant Quarter-Master-General,

Will please send sealed Tenders to this Office, in Triplicate, expressing, in words at length, the rate at which they will undertake the same, on or before Thursday the 18th inst. at twelve o'clock, when they will be opened, and the one most favourable for the Public approved of.



```
['ANY', 'person', 'willing', 'to',  
'undertake', 'the', 'removal', 'of', 'the',  
'Copper', 'Gutters', 'at', 'King's', '-',  
'House', ',', 'and', 'lay', 'them', 'anew',  
',', 'with', 'lead', ',', 'agreeable', 'to',  
'a', 'description', 'to', 'be', 'seen', 'at',  
'this', 'Office', ',', 'or', 'that', 'of',  
'the', 'Deputy', 'Assistant', 'Quarter', '-',  
'Master', '-', 'General', ',', '\n\n',  
'Will', 'please', 'send', 'sealed',  
'Tenders', 'to', 'this', 'Office', ',', 'in',  
'Triplicate', ',', 'expressing', ',', 'in',  
'words', 'at', 'length', ',', 'the', 'rate',  
'at', 'which', 'they', 'will', 'undertake',  
'the', 'same', ',', 'on', 'or', 'before',  
'Thursday', 'the', '18th', 'inst', '.', 'at',  
'twelve', 'o'clock', ',', 'when', 'they',  
'will', 'be', 'opened', ',', 'and', 'the',  
'one', 'most', 'favourable', 'for', 'the',  
'Public', 'approved', 'of', '.']
```

Basic NLP routines: Lemmatisation

— — —

ANY person willing to undertake the removal of the Copper Gutters at King's-House, and lay them anew, with lead, agreeable to a description to be seen at this Office, or that of the Deputy Assistant Quarter-Master-General,

Will please send sealed Tenders to this Office, in Triplicate, expressing, in words at length, the rate at which they will undertake the same, on or before Thursday the 18th inst. at twelve o'clock, when they will be opened, and the one most favourable for the Public approved of.



```
['any', 'person', 'willing', 'to',  
'undertake', 'the', 'removal', 'of', 'the',  
'Copper', 'Gutters', 'at', 'King's', '-',  
'House', ',', 'and', 'lie', 'they', 'anew',  
',', 'with', 'lead', ',', 'agreeable', 'to',  
'a', 'description', 'to', 'be', 'see', 'at',  
'this', 'Office', ',', 'or', 'that', 'of',  
'the', 'Deputy', 'Assistant', 'Quarter', '-',  
'Master', '-', 'General', ',', '\n\n',  
'will', 'please', 'send', 'seal', 'tender',  
'to', 'this', 'Office', ',', 'in',  
'Triplicate', ',', 'express', ',', 'in',  
'word', 'at', 'length', ',', 'the', 'rate',  
'at', 'which', 'they', 'will', 'undertake',  
'the', 'same', ',', 'on', 'or', 'before',  
'Thursday', 'the', '18th', 'inst', '.', 'at',  
'twelve', 'o'clock', ',', 'when', 'they',  
'will', 'be', 'open', ',', 'and', 'the',  
'one', 'most', 'favourable', 'for', 'the',  
'public', 'approve', 'of', '.']
```

Basic NLP routines: PoS-Tagging

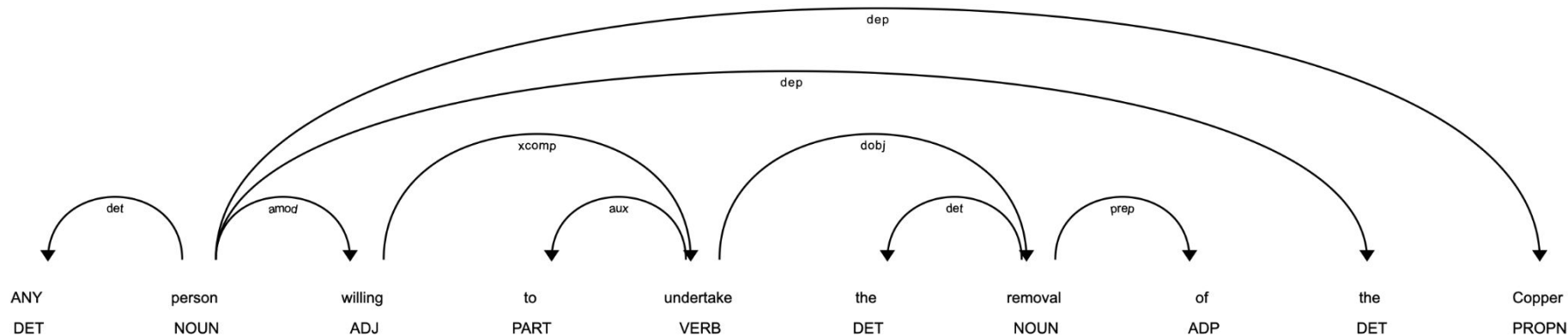
— — —

Tagset: Universal POS Tags [[LINK](#)]

- [ADJ](#): adjective
- [ADP](#): adposition
- [ADV](#): adverb
- [AUX](#): auxiliary
- [CCONJ](#): coordinating conjunction
- [DET](#): determiner
- [INTJ](#): interjection
- [NOUN](#): noun
- [NUM](#): numeral
- [PART](#): particle
- [PRON](#): pronoun
- [PROPN](#): proper noun
- [PUNCT](#): punctuation
- [SCONJ](#): subordinating conjunction
- [SYM](#): symbol
- [VERB](#): verb
- [X](#): other

Token	PoS	Lemma
ANY	DET	any
person	NOUN	person
willing	ADJ	willing
to	PART	to
undertake	VERB	undertake
the	DET	the
removal	NOUN	removal
of	ADP	of
the	DET	the
Copper	PROPN	Copper
Gutters	PROPN	Gutters
at	ADP	at
King's	PROPN	King's
–	PUNCT	–
House	PROPN	House

Basic NLP routines: Dependency Parsing



det: determiner
amod: adjectival modifier
dep: unclassified dependent
aux: auxiliary
dobj: direct object
prep: prepositional modifier

Basic NLP routines: Named Entity Recognition

— — —

ANY person willing to undertake the removal of **the Copper Gutters** **ORG** at **King's-House** **ORG** , and lay them anew, with lead, agreeable to a description to be seen at this Office, or that of the Deputy Assistant Quarter-Master-General,

Will please send sealed Tenders to this Office, in **Triplicate** **GPE** , expressing, in words at length, the rate at which they will undertake the same, on or before **Thursday** **DATE** the **18th** **ORDINAL** inst. at **twelve o'clock** **TIME** , when they will be opened, and the **one** **CARDINAL** most favourable for the **Public** **ORG** approved

- **CARDINAL**: Numerals that do not fall under another type
- **DATE**: Absolute or relative dates or periods
- **EVENT**: Named hurricanes, battles, wars, sports events, etc.
- **FAC**: Buildings, airports, highways, bridges, etc.
- **GPE**: Countries, cities, states
- **LANGUAGE**: Any named language
- **LAW**: Named documents made into laws
- **LOC**: Non-GPE locations, mountain ranges, bodies of water
- **MONEY**: Monetary values, including unit
- **NORP**: Nationalities or religious or political groups
- **ORDINAL**: "first", "second", etc.
- **ORG**: Companies, agencies, institutions, etc.
- **PERCENT**: Percentage, including "%"
- **PERSON**: People, including fictional
- **PRODUCT**: Objects, vehicles, foods, etc. (not services)
- **QUANTITY**: Measurements, as of weight or distance
- **TIME**: Times smaller than a day
- **WORK_OF_ART**: Titles of books, songs, etc

Basic NLP routines with Python 🐍 🪐

— — —

We'll be working with a Jupyter Notebook:

[https://github.com/dh-network/dh-ecf/blob/main/data-processing/DH-ECF-Data Processing.ipynb](https://github.com/dh-network/dh-ecf/blob/main/data-processing/DH-ECF-Data%20Processing.ipynb)

We'll execute it on a JupyterHub: <https://test.jupyterhub.uni-potsdam.de/>

⚠ If you don't have an Uni Potsdam account, you can execute the notebook on Google Colab:

<https://colab.research.google.com/drive/1Bf4asZzIWJX61phXXT5oQXdZUURJQcfZ?usp=sharing>

Working in the JupyterHub

— — —

1. Navigate to <https://test.jupyterhub.uni-potsdam.de/>
2. Log in with your UP credentials
3. Once logged in, choose the first option “minimal notebook”

Server Options

Minimal Notebook

Command-line tools useful when working in Jupyter applications: Common useful utilities like curl, git, nano (actually nano-tiny), tzdata, unzip, and vi (actually vim-tiny), TeX Live for notebook document conversion.

R Notebook

Popular packages from the R ecosystem: The R interpreter and base environment, IRKernel to support R code in Jupyter notebooks, tidyverse packages from conda-forge, caret, crayon, devtools, forecast, hexbin, htmltools, htmlwidgets, nycflights13, randomforest, rcurl, rmarkdown, rdbc, rsqLite, shiny, tidymodels, unixodbc packages from conda-forge.

Julia Notebook

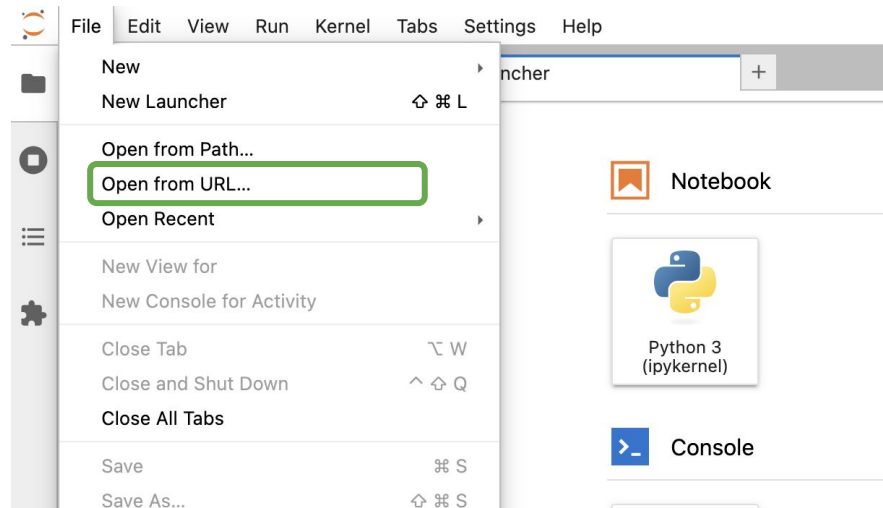
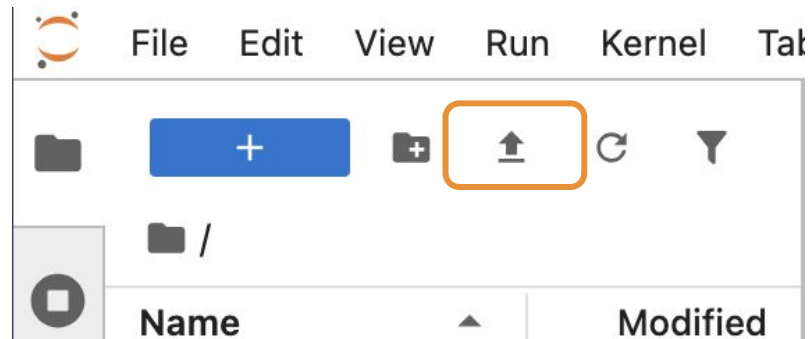
Popular packages from the Julia ecosystem: The Julia compiler and base environment, IJulia to support Julia code in Jupyter notebook, Pluto.jl reactive Julia notebook interface, made accessible with jupyter-pluto-proxy, HDF5 package.

SciPy Notebook

Working in the JupyterHub

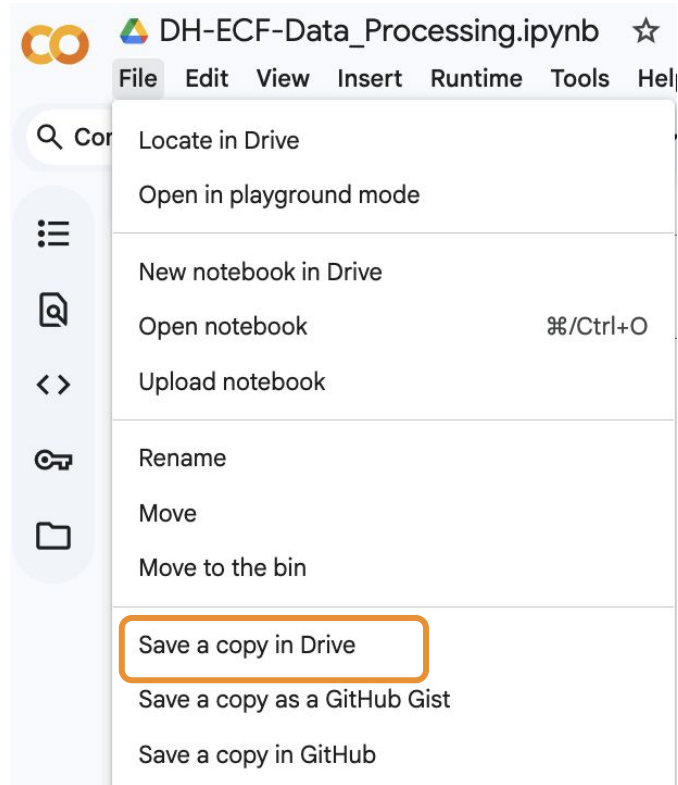
4. Once you see the “launcher”, you

- (a) can upload the notebook from your computer (orange box) or
- (b) load it directly from GitHub (green box), for this you need the **link to raw file**



Working with Colab

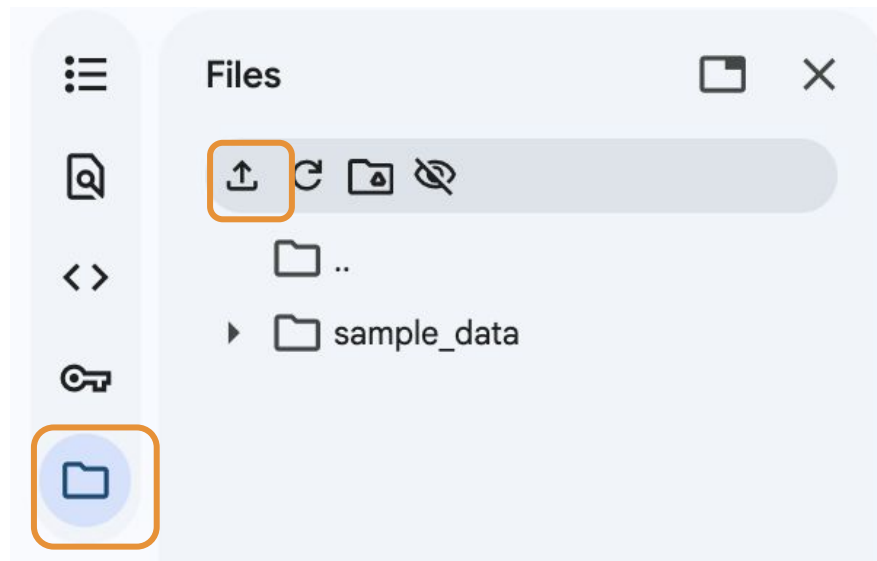
1. Save a copy in your own Google Drive



Working with Colab

— — —

2. Upload data to your session



04.12.2025
#4: Data Visualisation

Today

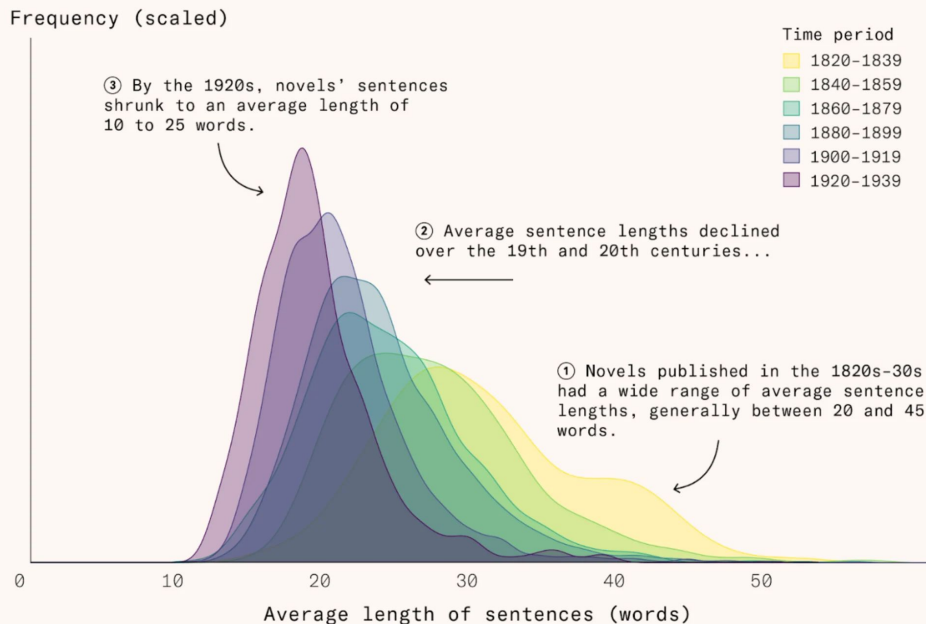
— — —

1. 👁 How to make chart clear
2. 📈 Examples from media
3. Practice using Datawrapper:
 - ☂️ tourism in Germany
 - and 🍓 strawberry yield

Your Experience with Data Visualization

How charts look

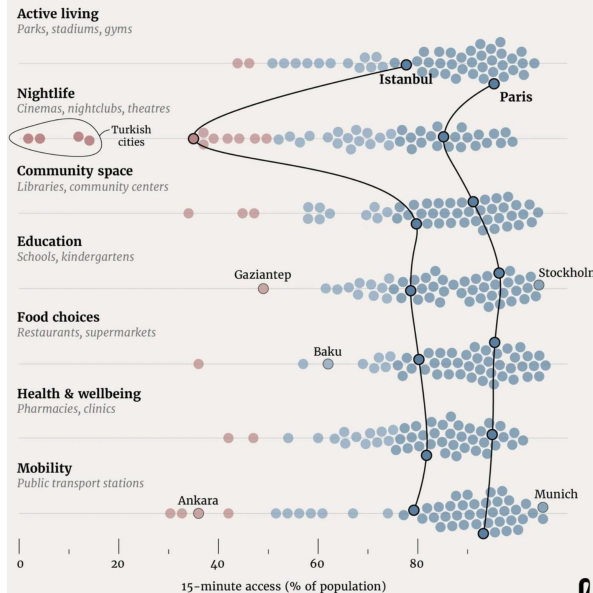
Novels of the 20th century had shorter sentences than those of the 19th century



Source: Adapted from Christoph Schöch; Project Gutenberg (2021)

15-minute cities

Share of residents with 15-minute walk, bike, or transit access to key services in Europe's 50 largest cities



Source: Cityaccessmap.com (2025)

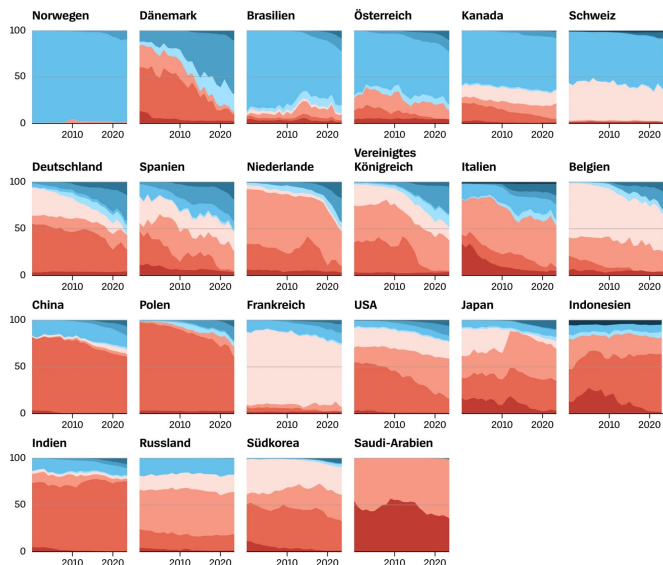
Datawrapper. Data Vis Dispatch

How charts look

Unter Strom

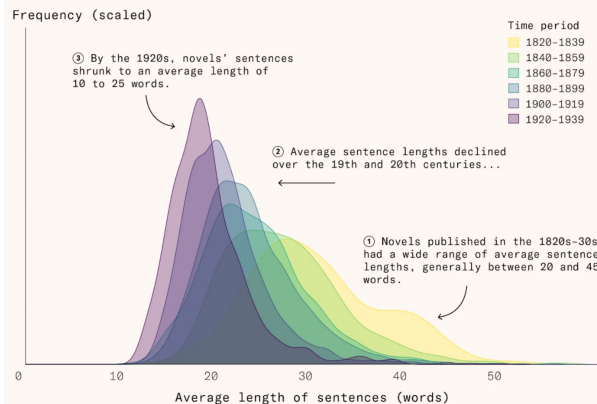
Anteil verschiedener Energieträger an der Stromerzeugung zwischen 2000 und 2024 in ausgewählten Ländern, Prozent

■ Andere ■ Sonstige Fossile ■ Kohle ■ Gas ■ Kernkraft ■ Biomasse ■ Wasser
■ Wind ■ Solar ■ Sonstige Erneuerbare



Quelle: Ember

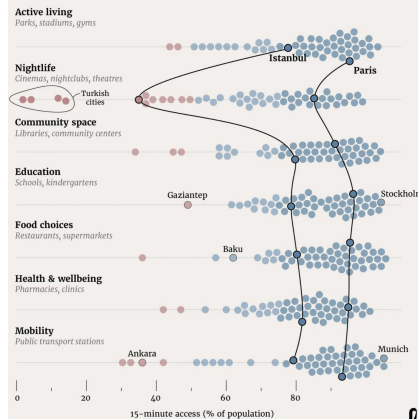
Novels of the 20th century had shorter sentences than those of the 19th century



Source: Adapted from Christoph Schöch; Project Gutenberg (2021)

15-minute cities

Share of residents with 15-minute walk, bike, or transit access to key services in Europe's 50 largest cities



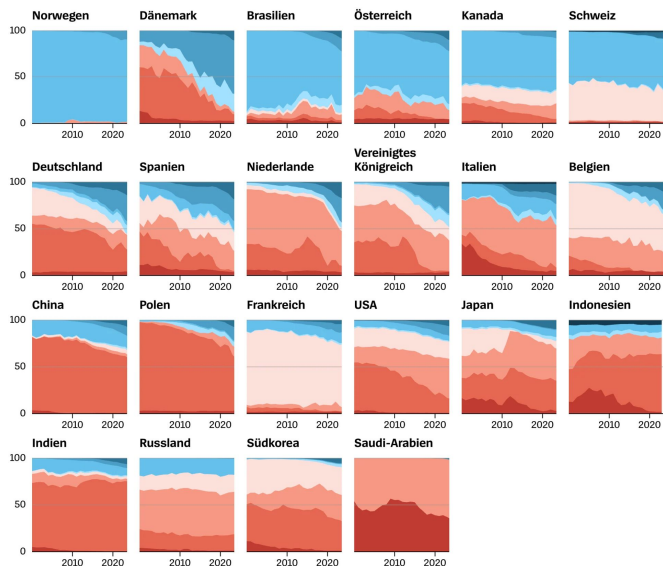
Source: Cityaccessmap.com (2023)

How charts look

Unter Strom

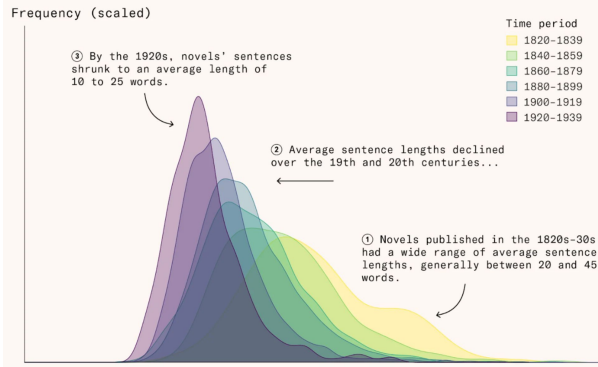
Anteil verschiedener Energieträger an der Stromerzeugung zwischen 2000 und 2024 in ausgewählten Ländern, Prozent

Andere Sonstige Fossile Kohle Gas Kernkraft Biomasse Wasser
Wind Solar Sonstige Erneuerbare



Quelle: Ember

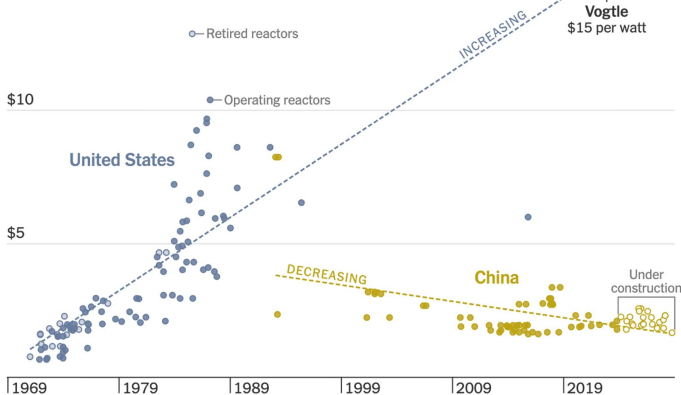
Novels of the 20th century had shorter sentences than those of the 19th century



Construction costs of nuclear reactors

\$15 per watt

Source:

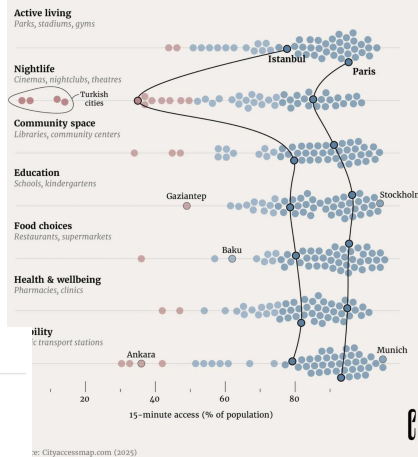


Hover to explore the data

Note: Reactors are placed according to the date they entered or are expected to enter commercial operation. Chart shows inflation-adjusted overnight costs, which exclude interest payments. Trend lines show linear regression slope. Source: [Liu et al. 2025](#)

15-minute cities

Share of residents with 15-minute walk, bike, or transit access to key services in Europe's 50 largest cities

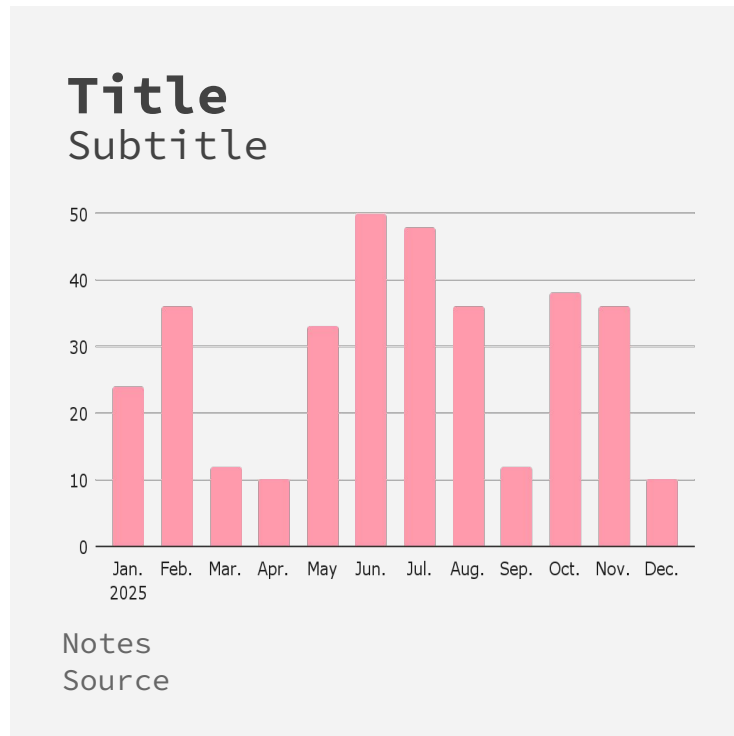


Vis Dispatch

Around the chart

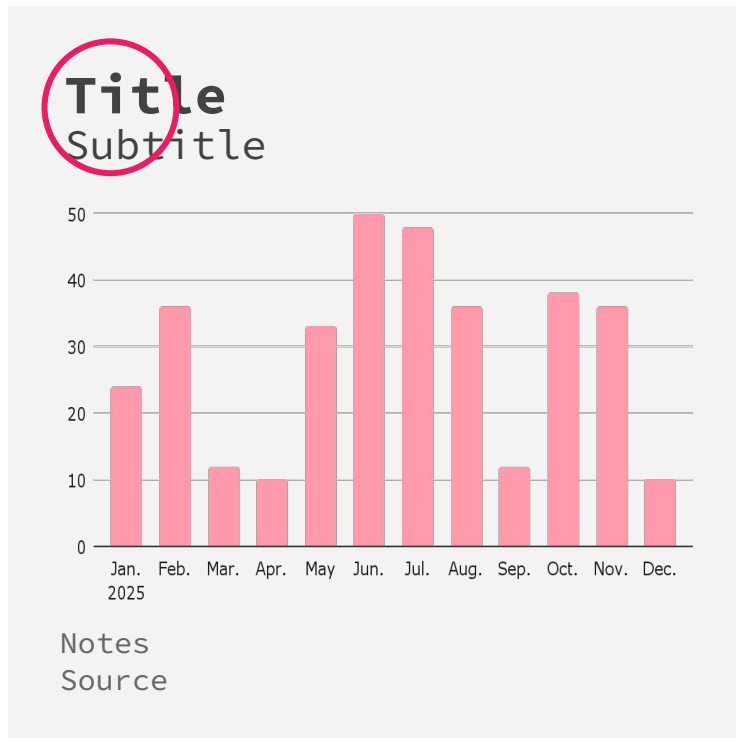
— — —

- Title & subtitle
- Source & Notes
- Axes
- Colors & Legend



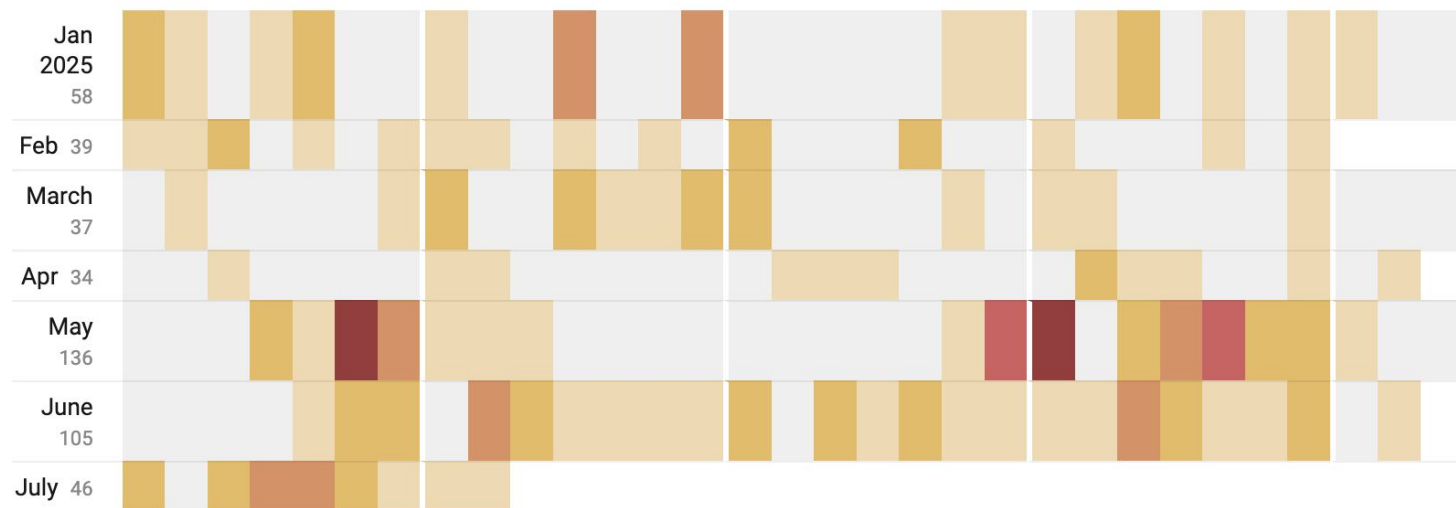
Title & Subtitle

- What is the chart about?
 - Topic
 - Indicator name
 - Insight (narrative title)
- Keep titles within $\frac{1}{3}$ of the image



Russian airports have closed 455 times since the beginning of 2025.

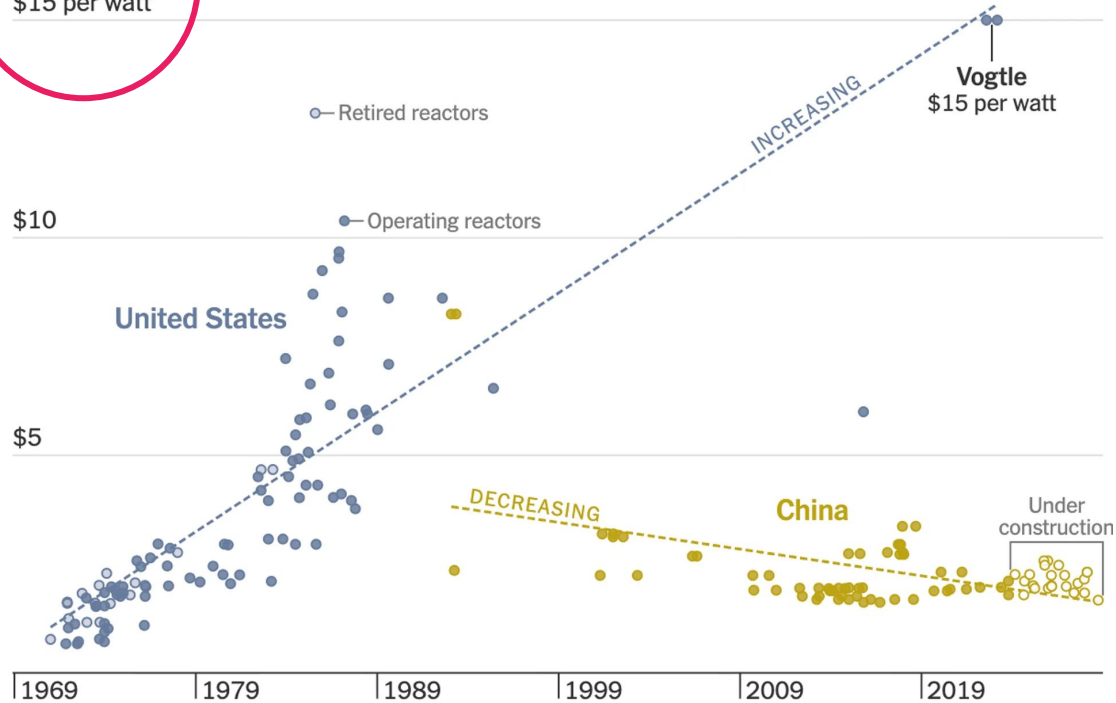
Number of temporary airport closures per day due to air defense operations



Source: Rosaviatsia messages on Telegram, calculations by Novaya Evropa

Construction costs of nuclear reactors

\$15 per watt

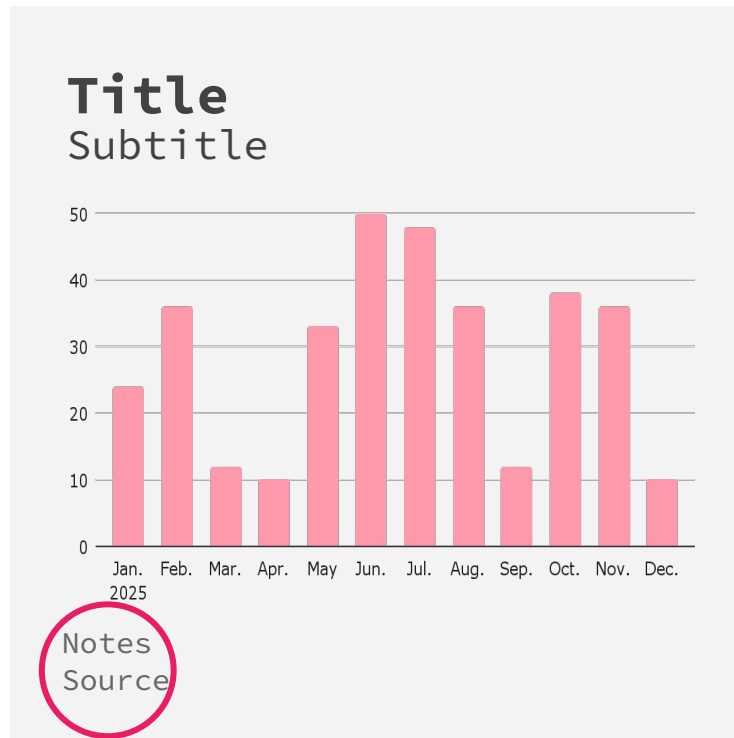


🖱️ Hover to explore the data

Note: Reactors are placed according to the date they entered or are expected to enter commercial operation. Chart shows inflation-adjusted overnight costs, which exclude interest payments. Trend lines show linear regression slope. • Source: [Liu et al. 2025](#)

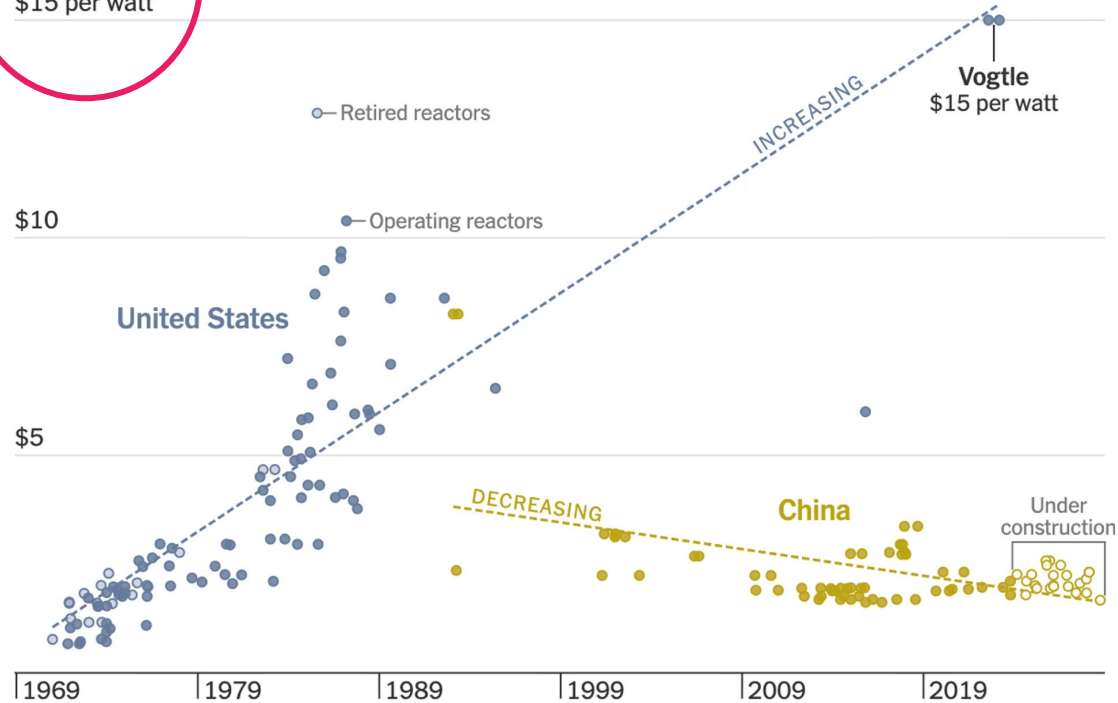
Source & Notes

- Where the data comes from
- Explanation of special terms
- Data restrictions



Construction costs of nuclear reactors

\$15 per watt

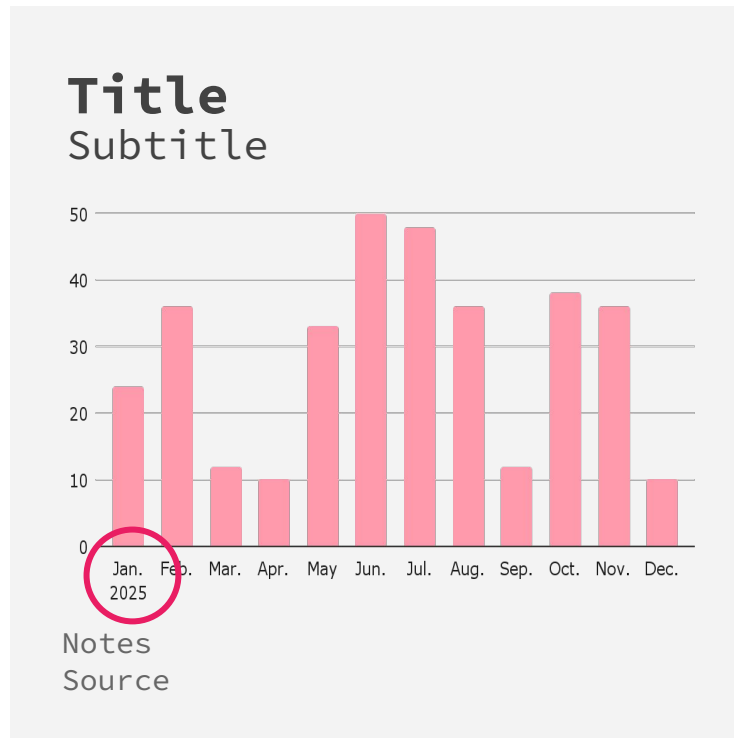


Hover to explore the data

Note: Reactors are placed according to the date they entered or are expected to enter commercial operation. Chart shows inflation adjusted overnight costs, which exclude interest payments. Trend lines show linear regression slope. • Source: [Liu et al. 2025](#)

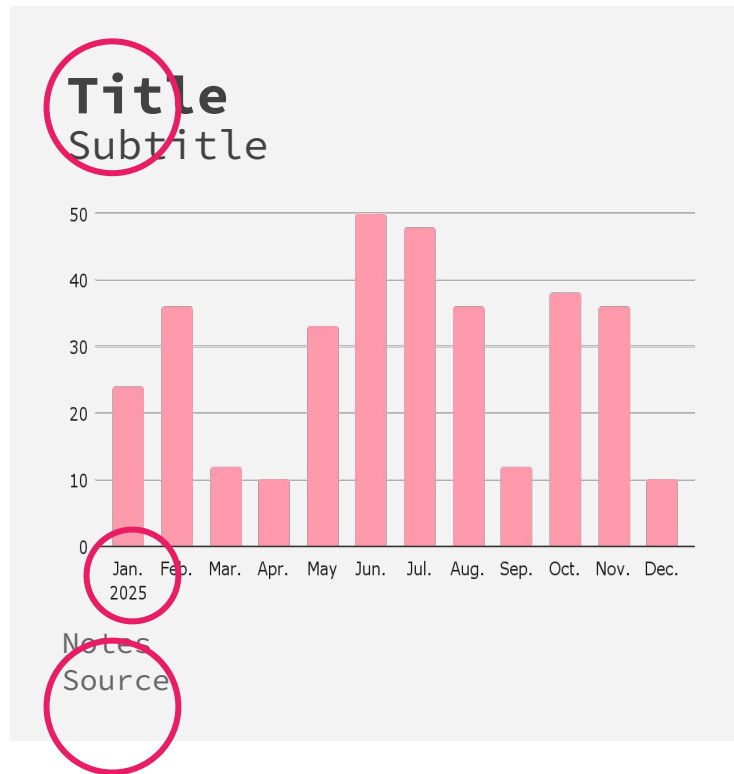
Don't forget the year!

- Time series data
- or



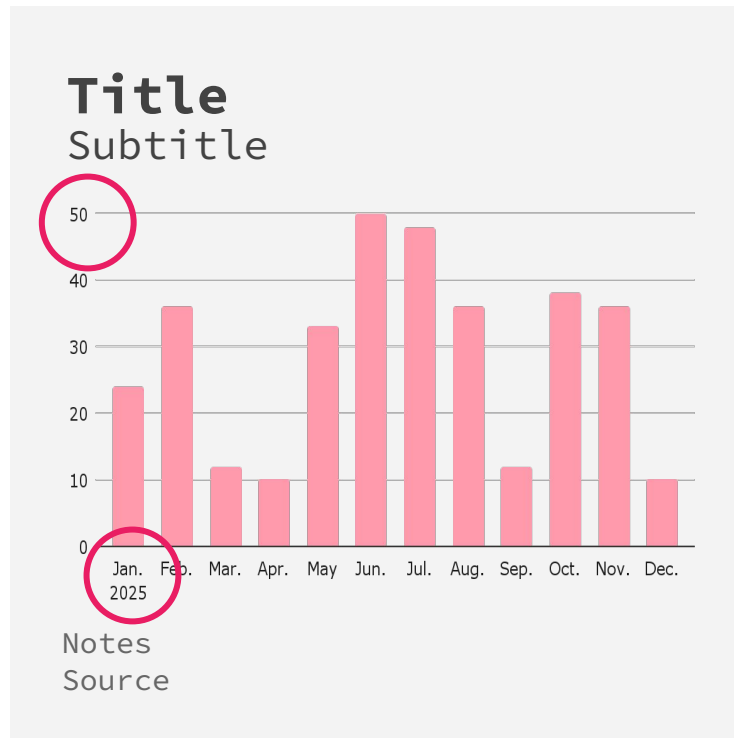
Don't forget the year!

- Time series data
or mention it in
- Source
- Title / Subtitle



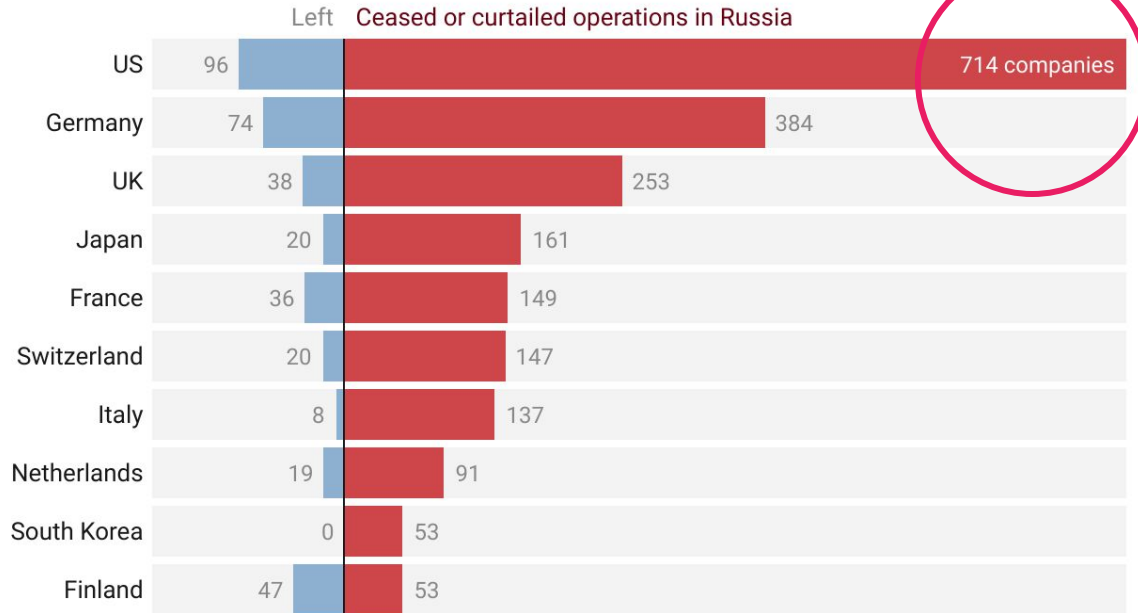
X & Y Axes

- Rounding of values
(100 000 → 100k; 10.0% → 10%)
- Units of measurement
- No rotation
- Axis label → subtitle



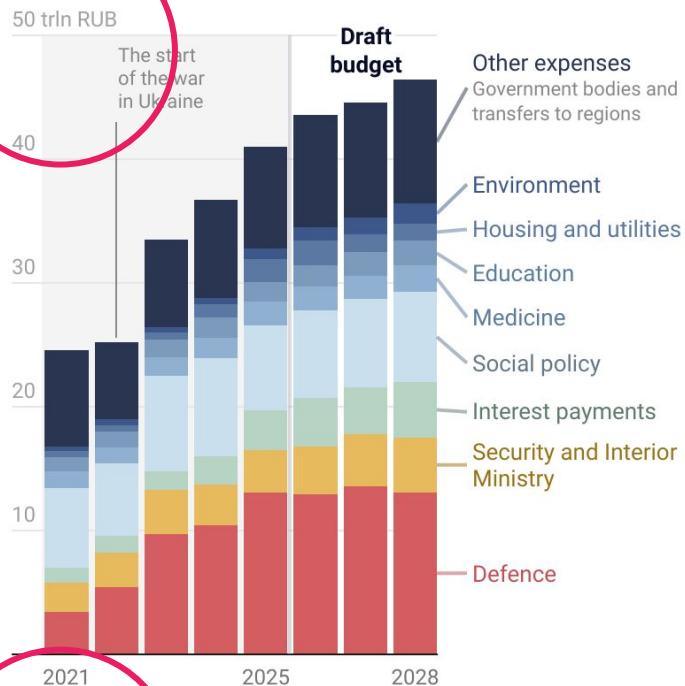
US companies most prevalent among “unfriendly” investors in Russia

Most foreign companies remained in Russia after the outbreak of the war in Ukraine



Source: Leave Russia, August 2025

38% of the Russian budget will go on defence and security from 2026 to 2028



2026–2028 budget: draft as of September 2025.

2025: adjustments to the May 2025 budget.

2022 defence spending: Gaidar Institute estimates; 2023: Reuters data.

Source: Russian Finance Ministry

Effective color usage

— — —

- ≤ 5 colors (or categories)

Use greys palette for the rest

- Check accessibility, try B&W mode

Avoid mixing  reds and  greens

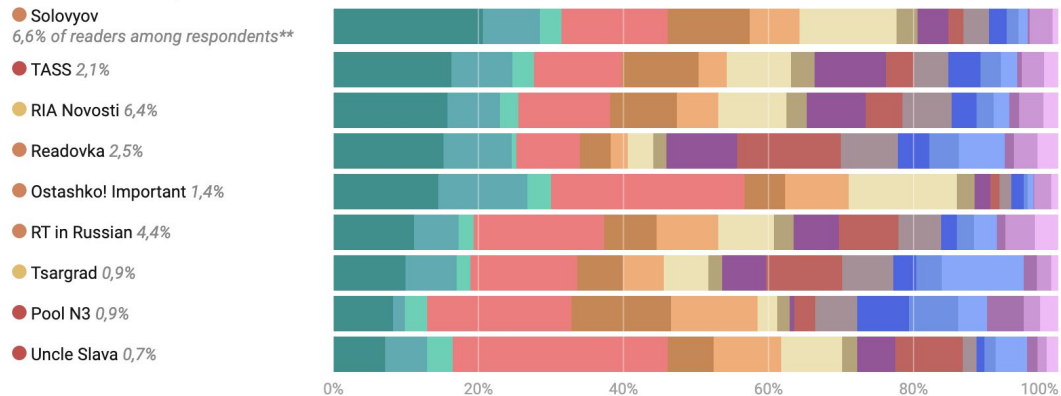
- Include a legend

Clusters of pro-government channels by events topics*

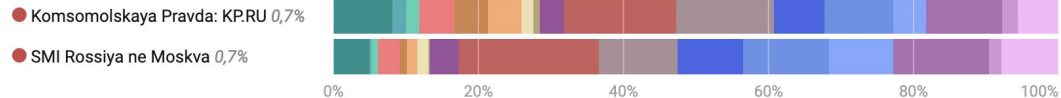
Hover over a topic to see its share

Shelling Situation on the frontlines Weapons and military cooperation International politics related to Ukraine International diplomacy (not related to Ukraine) U.S. politics Israel and the Middle East Foreign policy Arrests and trials (political and economic) Violent crimes and criminal activity Emergencies and incidents Economy Legislative bills and regulations Domestic politics and social issues Food and health Sports and political conflicts Technology and IT

Focus on international politics



Lifestyle and daily news



News from the frontlines



● ● ● Pro-war sentiment of readers from average to maximum

* Event here means an occurrence that has more than one post dedicated to it

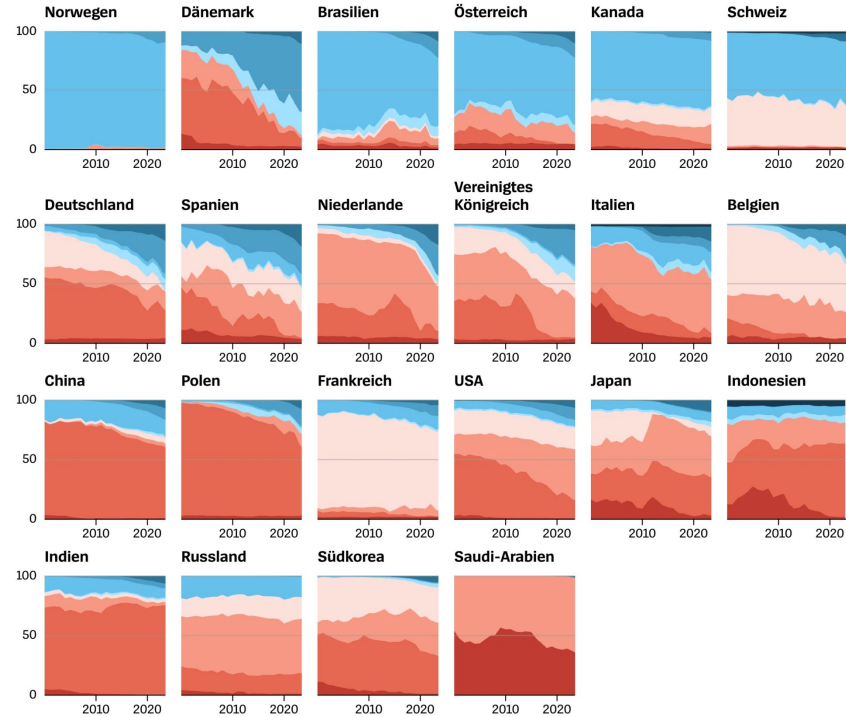
** Percentage of respondents who answered the question about the channels they read on Telegram

Chart: Cedar • Source: Cedar calculations based on Telegram posts, 2025 • Created with [Datawrapper](#)

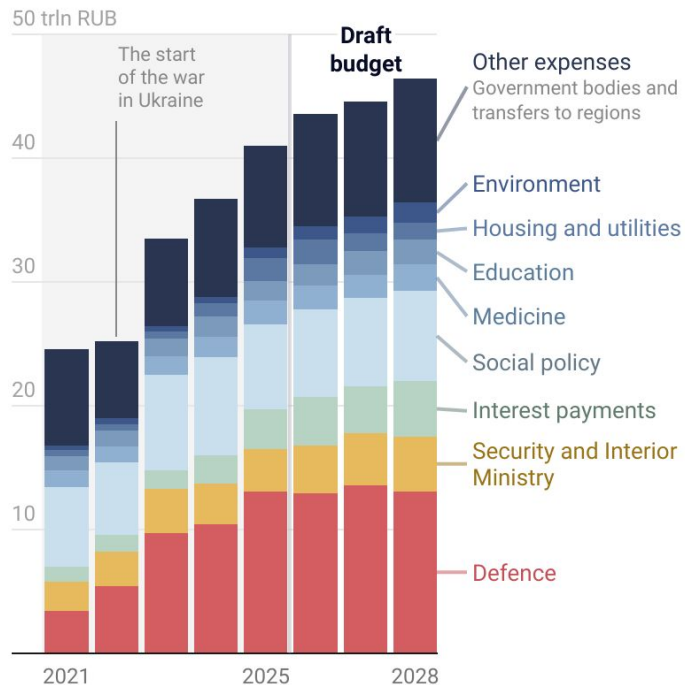
Unter Strom

Anteil verschiedener Energieträger an der Stromerzeugung zwischen 2000 und 2024 in ausgewählten Ländern, Prozent

Andere Sonstige Fossile Kohle Gas Kernkraft Biomasse Wasser
Wind Solar Sonstige Erneuerbare



38% of the Russian budget will go on defence and security from 2026 to 2028



2026–2028 budget: draft as of September 2025.

2025: adjustments to the May 2025 budget.

2022 defence spending: Gaidar Institute estimates; 2023: Reuters data.

Source: Russian Finance Ministry

Legend

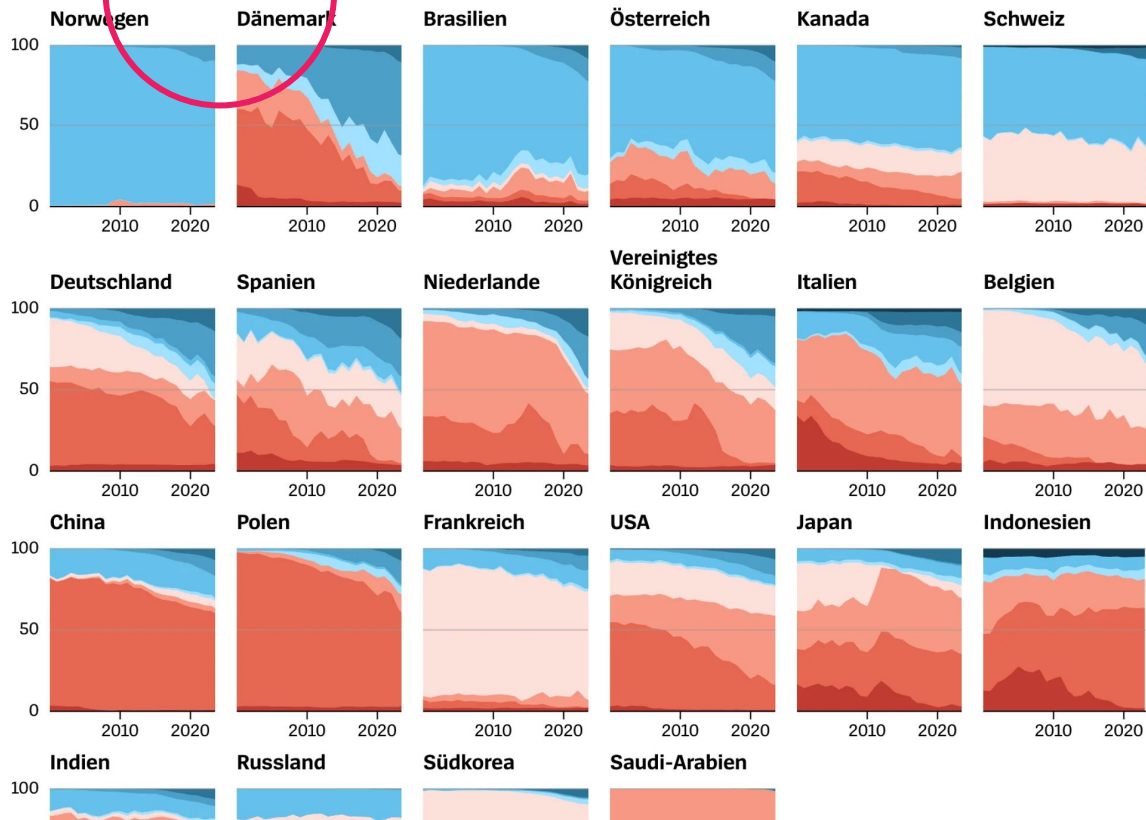
- Use for colors, scale, shapes
- Direct labels or legend in title improve readability
- Encode attributes only once

Anteil verschiedene
ausgewählten Länd

■ Andere ■ Sonst
■ Wind ■ Solar

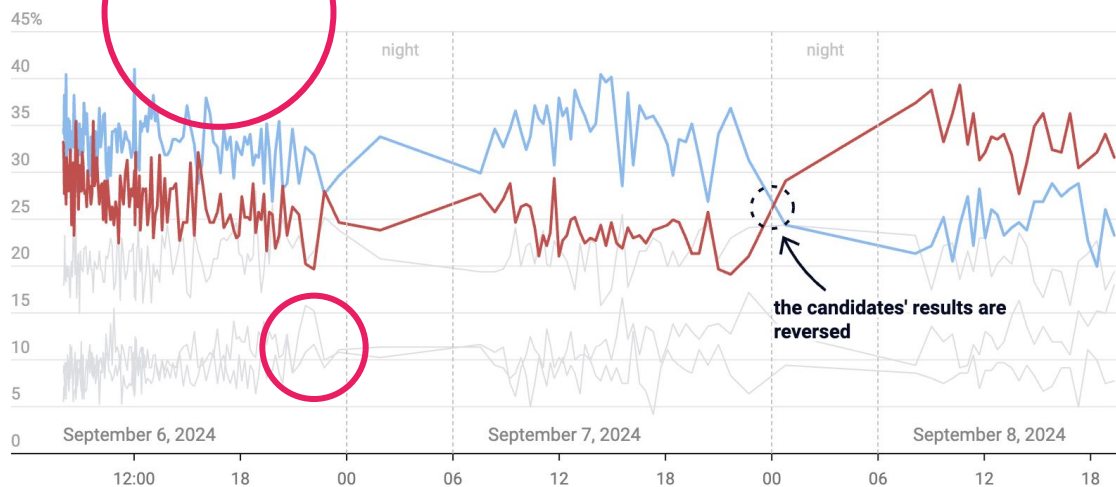
wegen Dänemark

Anteil verschiedener Energieträger an der Stromerzeugung zwischen 2000 und 2024 in ausgewählten Ländern, Prozent



In three electoral districts on Sunday night, **United Russia** candidates began to take the lead, although they had previously been losing to their opponents from the systemic parties.

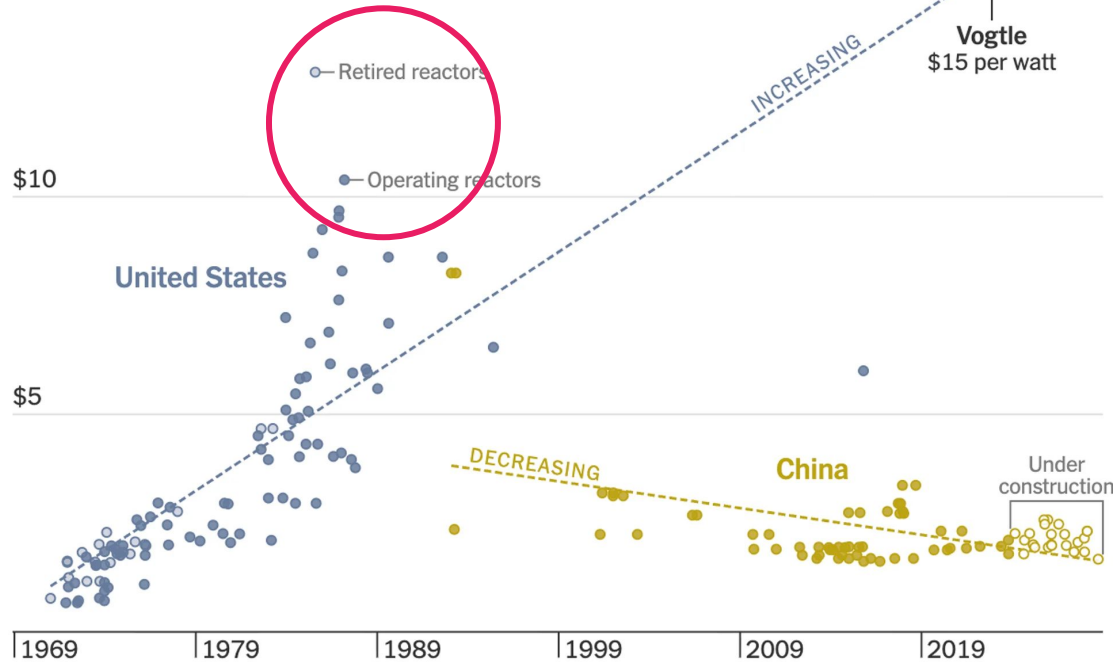
10th District Avdoshkina / Zubrilin (CPRF) / Other candidates



Each point represents a candidate's result across 300 ballots, so the points are not uniformly distributed over time but have the same noise level.
Source: observer.mos.ru

Construction costs of nuclear reactors

\$15 per watt

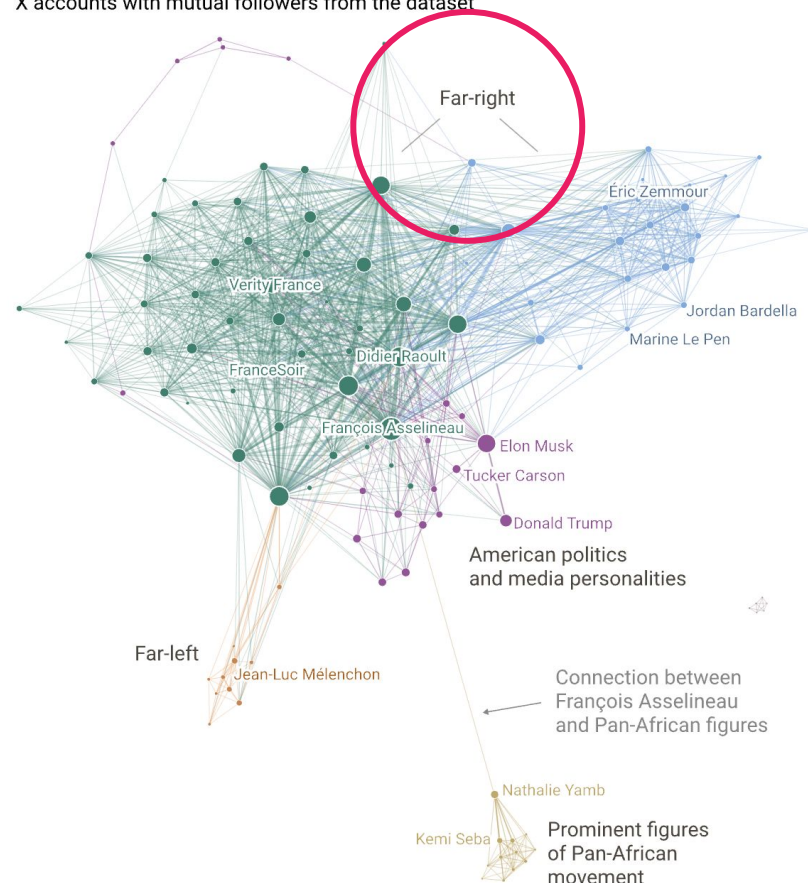


🖱️ Hover to explore the data

Note: Reactors are placed according to the date they entered or are expected to enter commercial operation. Chart shows inflation-adjusted overnight costs, which exclude interest payments. Trend lines show linear regression slope. • Source: [Liu et al. 2025](#)

X comments dataset

X accounts with mutual followers from the dataset

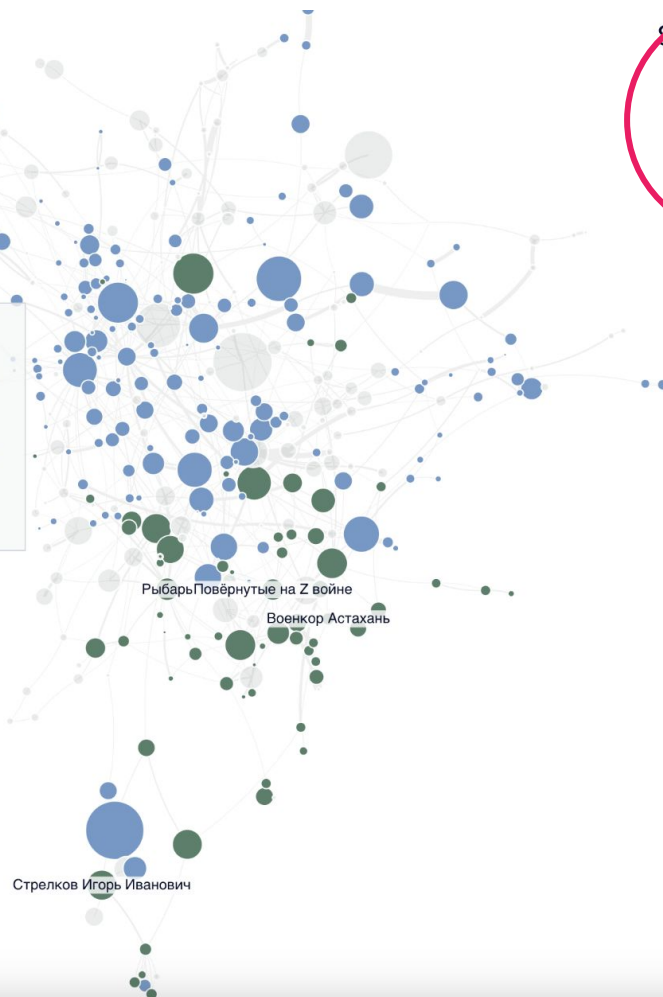


Spring 2022

Average number of
views per post
2 million

500
thousand

In terms of numbers and influence, the "turbopatriots" have begun to rival the "systemic" channels. These include war correspondents, military analysts (such as Igor Strelkov), and anonymous channels whose ties to the state are not always obvious (such as "Rybar" and "Obsessed with the Z War"). This is the most radical faction: they are often dissatisfied with the progress of the war, demanding decisive action on the front lines and a total mobilization of society.



Checklist for good charts

- ☑ No more than **5 categories** / colors
- ☑ Rounded values (100 000 → 100k; 10.0% → 10%)
- ☑ Units shown (€, %, km, hrs, etc.)
- ☑ Year or time period stated
- ☑ Clear title + data source
- ☑ 🌟 **Key insight** (what the viewer should remember)

“Social media” check

Sometimes tables are enough

Ranking of “pro-Putin” MEPs based on their voting records in 2019-2023

	MEP Party	Total number of rejected resolutions*	For	Abstained	Against	Didn't vote
1	Tatjana Ždanoka Latvian Russian Union	20	2	2	14	4
2	Ivan David Freedom and Direct Democracy	19	0	8	9	2
3	Konstantinos Papadakis Communist Party of Greece	19	3	1	17	1
4	Lefteris Nikolaou- Alavanos Communist Party of Greece	18	3	1	17	1
5	Milan Uhrík Republic	19	1	7	12	0
6	Sandra Pereira Portuguese Communist Party	19	2	4	15	0
7	Maximilian Krah Alternative for Germany	18	1	2	10	6
8	Gunnar Beck Alternative for Germany	18	3	6	10	2
9	Christine Anderson Alternative for Germany	18	4	6	11	1
10	Guido Reil Alternative for Germany	18	4	6	11	1
11	Bernhard Zimniok Alternative for Germany	18	4	6	11	1

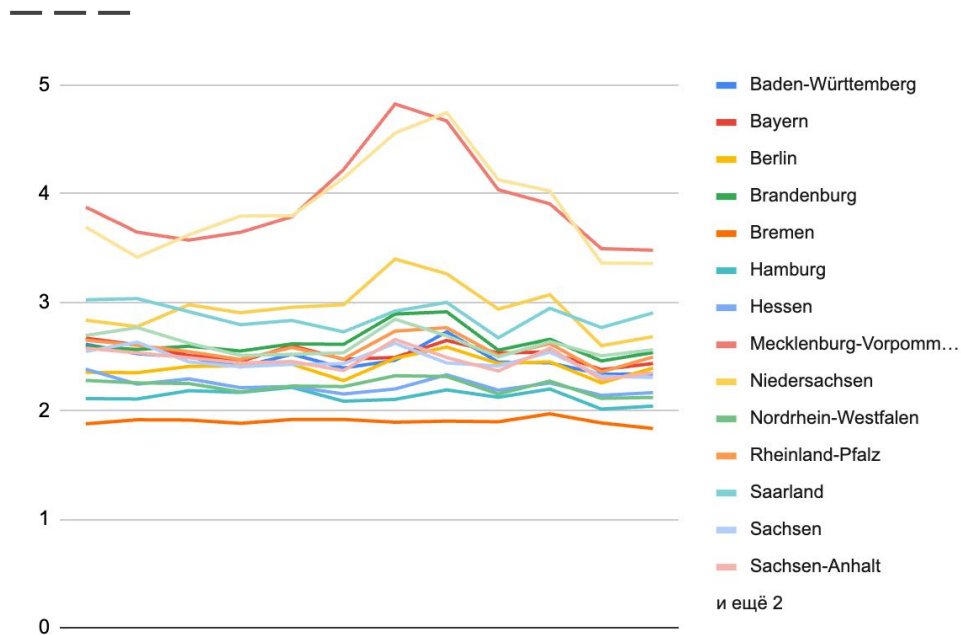
10 most followed personalities from the X comments dataset

Number of followers among users in the dataset

François Asselineau @f_asselineau	155 foll.	Leader of the Union Populaire Républicaine, political party that has far right ideas and is advocating for France leaving the European Union
Idriss Aberkane @idrissaberkane	135 foll.	French scientist and editorialist that is accused of endorsing conspiracy ideas
Didier Raoult @raoult_didier	135 foll.	Microbiologist who became an idol for the French antivax movement
Alexis Poulin @Poulin2012	133 foll.	French journalist and influencer, often appears on SudRadio associated with far right and conspiracy ideas
Mike Borowski @BOROWSKIMIKE	125 foll.	Far right journalist and a politician accused of spreading fake news and endorsing conspiracy theories
Florian Philippot @f_philippot	125 foll.	Leader of Les Patriotes political party clearly identified as far right and souverainiste
Elon Musk @elonmusk	125 foll.	Entrepreneur and the owner of X (ex-Twitter), associated with far right ideas and having close ties to Donald Trump until recently
Verity France @verity_france	106 foll.	“An association of victims of side effects and difficulties linked to the vaccination”
Yves Pozzo di Borgo @YvesPDB	106 foll.	Right wing former senator linked to conspiracy theories
Fabrice Di Vizio @DIVIZIO1	103 foll.	French lawyer who became a popular figure of the antivax movement

Table: Cedar • Source: X (2022–2023) • Created with [Datawrapper](#)

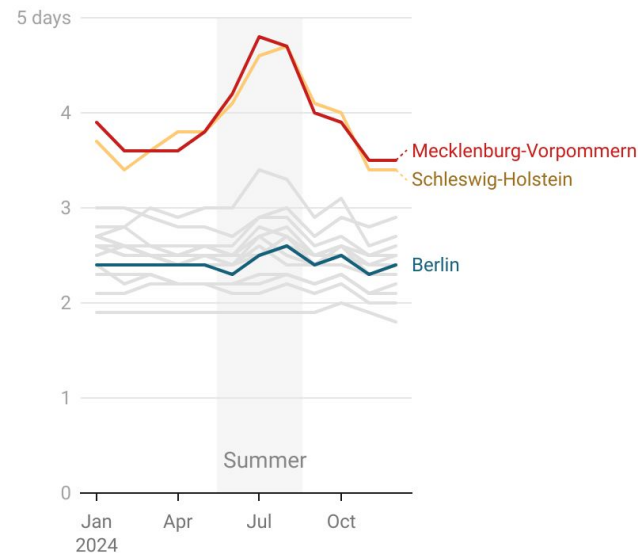
Practice: Datawrapper



Basic Google Sheets

How Long Tourists Stay in Different German Regions

Monthly average number of overnight stays per person



Average overnight stays per person is calculated as total overnight stays divided by total arrivals, where arrivals count all guests checking in and overnight stays count all nights spent by those guests.

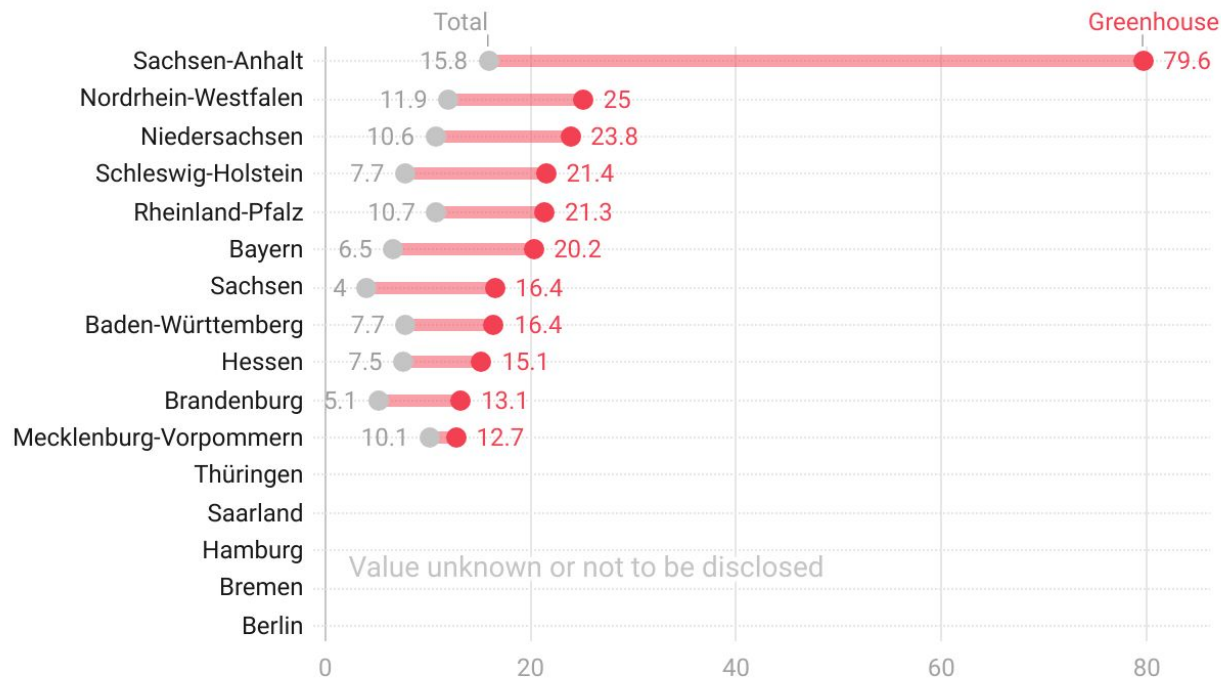
Source: Federal Statistical Office of Germany (Destatis)



Datawrapper

Strawberry Yield in 2024

Tonnes per hectare



Greenhouses = Under high accessible protective cover including greenhouses.

Source: Statistisches Bundesamt (Destatis)

Links

datawrapper.de

Data:

shorturl.at/WjGIP

08.01.2026

#5: Working Efficiently with LLMs

What do you already do with
LLMs in study/research
context?

When you have an
hammer,



ChatGPT



Claude

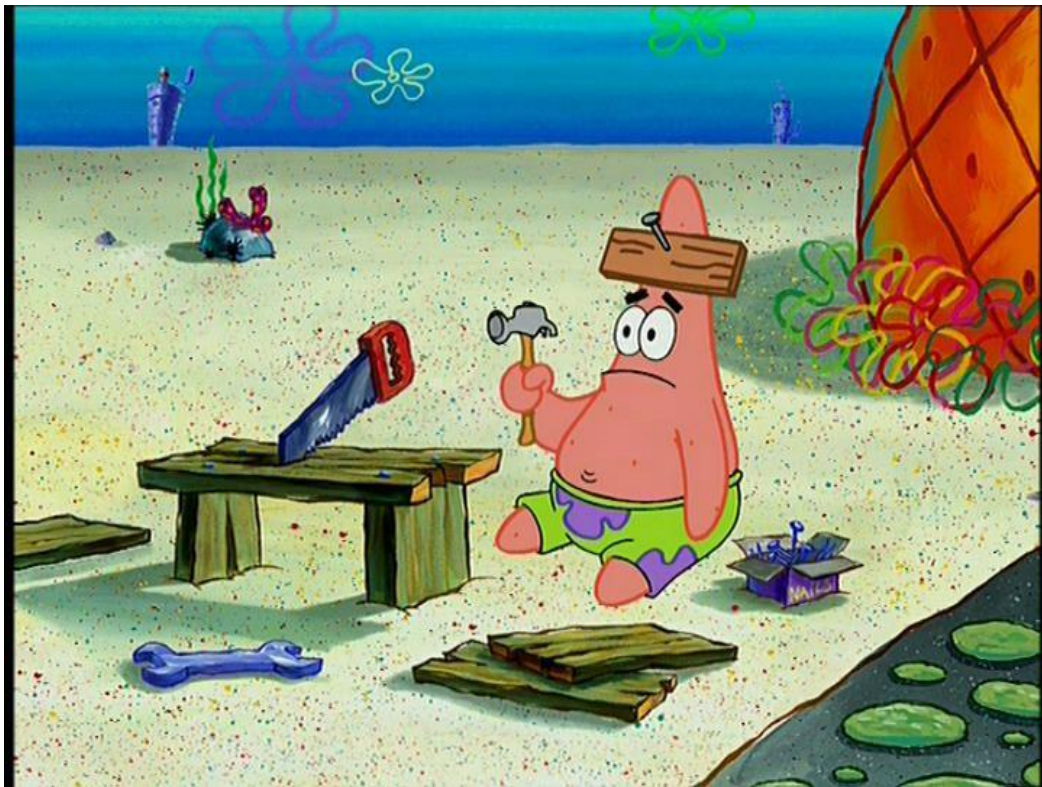


MISTRAL
AI_

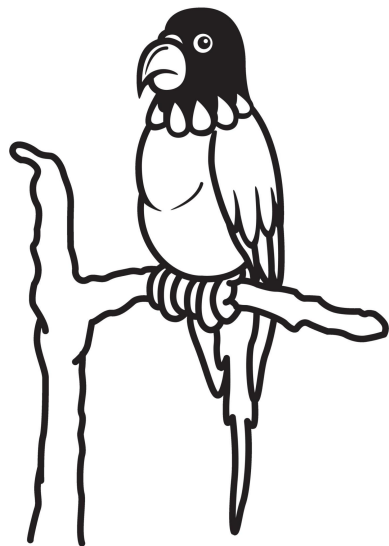


deepseek

everything looks like
a nail...



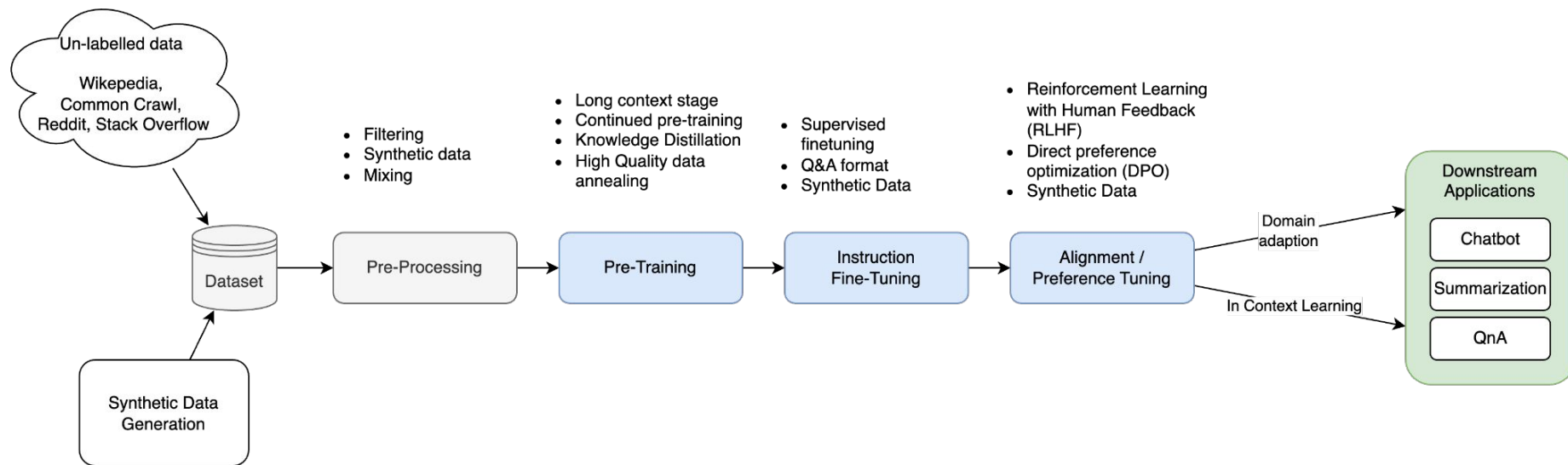
Ok, so what's a large language model?



- Statistical **text prediction and generation system** trained on massive text datasets
- Predicts the most probable next token based on **patterns** learned during training
- Built using **neural networks** (typically **Transformer** architecture)
- Has **no understanding** and **no reasoning** of its own (yet?) → once singled out as “stochastic parrots” ([Bender et al. 2021](#))

Is it really a black box?

— — —



What are current LLMs good at?

— — —

- **Assisting with programming tasks:** Python scripts, regexes, data wrangling, API usage
- **Explaining unfamiliar concepts/code/anything:** quick orientation in new theories, methods, or libraries; infinite step-by-step explanations ‘without social cost’, examples, exercises
- **Explaining papers section by section** into clearer language
- **Summarising and paraphrasing texts:** articles, chapters, transcripts, archival descriptions
- **Interactive exploration and ideation:** suggesting research questions, feature ideas, annotation schemes
- **Generating analysis plans** before committing to full implementation
- **Suggesting operationalisations:** how to turn abstract concepts into measurable features

In the Digital Humanities context LLMs offer ‘on-demand’ classification/annotation of diverse research objects

— — —

Recent advances in instruction-tuned LLM technology make a difference: it has become feasible to use them for **on-demand data classification** and **annotation**, making multi-variable machine-assisted research designs much more feasible. In a **zero-shot scenario**, a generative model is "instructed" to generate output given an input (prompt) including in some form both the analysis instructions and a unit of data. The generated **outputs** can be then **parsed and quantified** as necessary.

Karjus, Andres. ‘Machine-Assisted Quantitizing Designs: Augmenting Humanities and Social Sciences with Artificial Intelligence’. arXiv, 20 October 2024.
<https://doi.org/10.48550/arXiv.2309.14379>.

What are current LLMs bad at? ❌

— — —

- **Reliable factual accuracy:** hallucinations, invented references, plausible-sounding errors
- **Knowing when they are wrong or uncertain:** confidence $\neq$ correctness
- **Source criticism and provenance awareness:** distinguishing evidence, editions, authority, archival contexts
- **Deep interpretive reading:** irony, ambiguity, narrative perspective, historical nuance
- **Consistency over long or complex workflows:** multi-step pipelines, evolving assumptions, long documents
- **Replacing domain expertise**
- **Knowing what needs to be done and why**

Best way to learn this is to try everything:

— — —

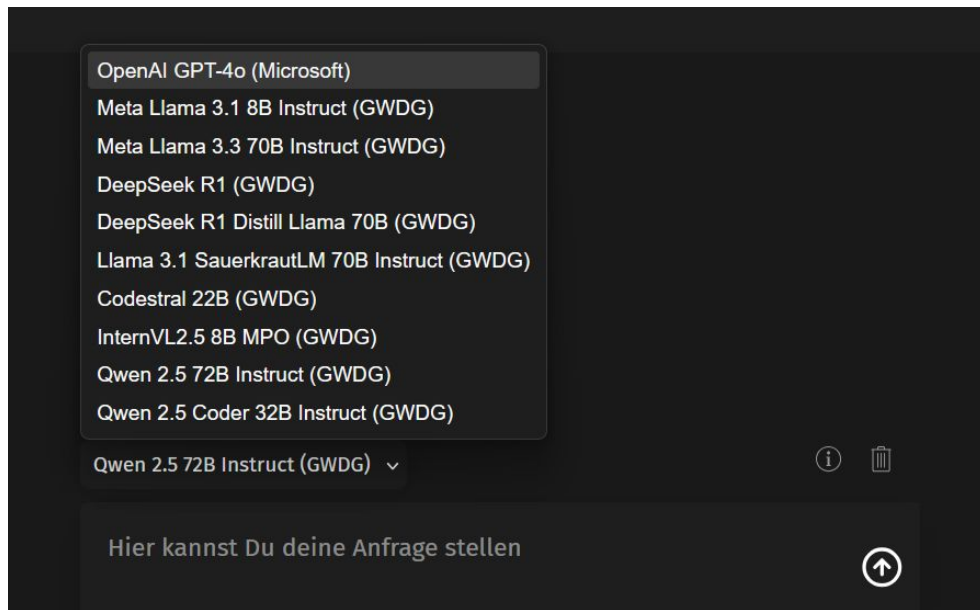
“You should **try inviting AI to help you in everything you do**, barring legal or ethical barriers. As you experiment, you may find that AI help can be satisfying, or frustrating, or useless, or unnerving. But you aren’t just doing this for help alone; familiarizing yourself with AI’s capabilities allows you to better understand how it can assist you—or threaten you and your job. Given that AI is a General Purpose Technology, **there is no single manual or instruction book** that you can refer to in order to understand its value and its limits.”

Ethan Mollick (2024). Co-Intelligence: Living and Working With AI

Let us try things

Setup

- Go to <https://gptup.uni-potsdam.de/login>
- Login with your UP credentials
- Select the **model** you want to try



Why using GPT.UP and not the commercial one?

- better privacy (user prompts and answers models are not saved within [GPT.UP](https://gptup.uni-potsdam.de) and stay in local cache)

Exercise 1: 'LLM vs. Human Annotator' Speedrun (Live Battle)

— — —

- Object: Take this one page of the Barbados Gazette: [AA00047511_01264_p1.txt](#) (link also in the Moodle)
- Task: **find all inanimate objects of trade** (e.g. “codfish” or “ginger” or “dish covers”). List **types of goods** that are being traded on Barbados, according to this one page of the Gazette.
- Conditions: 5-7 minutes work in groups
 1. Group A: manual work only
 2. Group B: LLM-use only (only allowed to prompt, copy, and paste from the LLM output)
 3. Group C: freely combine LLM and manual work
- Output: a list / a table (whatever seems appropriate to the group) – each group should upload via Moodle

*SIXTEEN DOLLARS (Cash) will be given for
any quantity of good GINGER, if delivered between*

Reflection on the exercise

Let's compare performance between teams

- Coverage (recall): how many different items did each team find altogether?
- Reliability (precision): how many obvious mistakes / weird cases did each team have in the results?

Regarding 'Holland' and 'Russia'

— — —

https://en.wikisource.org/wiki/1911_Encyclop%C3%A6dia_Britanica/Holland_%28cloth%29

<https://www.oldbaileyonline.org/record/18140216>

Exercise 2: comparing prompt quality

1. Take the [same page](#) of the Barbados Gazette
2. Compare results for the following prompts:
 - a. “Extract entities from this text: [text]”.
 - b. “Extract persons, places, dates, government entities, ships from the following text, return them as a JSON list: [text]” .
 - c. “Extract persons, places, dates, government entities, ships from the following text, return them as a JSON list. Only include entities explicitly mentioned in the text. If you are unsure about an entity, omit it. Do not infer or guess: [text]”.

(in each case paste the text into the prompt)

Reflection on the exercise 2

— — —

- What were the differences between prompts 1 and 2? Why?
- What were the differences between prompts 2 and 3? (if any). Why (not)?

Exercise 3: Comparing prompt quality

— — —

1. Include the text of [this article](#) in the prompt.
2. Add each time one of these prompts:

Prompt 1 (zero-shot): “Summarize this text in 3 sentences.”

Prompt 2 (minimal prompt engineering): “This is a opinion piece from *Politico*. Summarize the main themes and rhetorical strategies of this text in 3 sentence.”

Prompt 3 (persona modelling): “You are an expert in EU policy with a background in translation studies. What does this text reveal about the author’s intentions and historical context? Summarise in 3 sentences. ”

Exercise 4: LLM as co-author

1. Write 3 sentences of an job cover letter you could imagine writing
2. Let the LLM continue
3. Look at the results. What would you like to change/fix? Why? How do you think HR would assess it?

Exercise 5: Annotation assistance

1. Include this text in the prompt.
2. Use this prompt:

“You are assisting a human annotator. Identify all passages that might express religious conflict. Quote the passage and explain why it may fit the category. Do not decide definitively.”

3. Manually accept/reject

Wrap-up

— — —

- Did this session change/enhance your perspective on LLMs?
- Do you plan to use LLMs *differently* in the future?

29.01.2026

#6: Project Consulting