



**eurac**  
research

## Data Management & FAIR Guiding Principles

Data Management for FAIR CMC Corpora  
Workshop @ CMC Corpora 2021

Egon W. Stemle, Jennifer-Carmen Frey  
<egon.stemle|jennifercarmen.frey  
@eurac.edu>

27.Oct.2021

# FAQ about Linguistics

Linguistics is the scientific study of human language.

## What do you mean by "human language"?

All humans have some language, without exception. It's a distinguishing biological trait of Homo Sapiens. And all human languages have a lot of similarities, and they tell us quite a lot about what it means to be human, which is a big preoccupation of H. Sapiens. Here's what Edward Sapir (considered by many the greatest American linguist of the last century) said:

*Everything that we have so far seen to be true of language points to the fact that it is the most significant and colossal work that the human spirit has evolved – nothing short of a finished form of expression for all communicable experience. This form may be endlessly varied by the individual without thereby losing its distinctive contours; and it is constantly reshaping itself as is all art.*

***Language is the most massive and inclusive art we know, a mountainous and anonymous work of unconscious generations.***

► Sapir, 1921

# FAQ about Linguistics

Linguistics is the scientific study of human language.

## All right, what do you mean by "scientific"?

Roughly, *objective*, *unbiased*, **data-oriented**, and **reproducible**, among other meanings. Simply put, linguists are concerned with how language actually does work, rather than with how (somebody says) it ought to work. This is a fairly new approach to a very old interest, since people have always been interested in language, even though it is a hard subject to talk about.

▶ [FAQ from John M. Lawler's personal web page \(University of Michigan\)](#)

# Background

- A recent trend in Social Sciences and Humanities (SSH) research – and in many other fields – is to become **more**
  - reproducible
  - reusable
  - transparent

*The European Commission unveiled its plans to make all data derived from EU-funded research projects findable, accessible, interoperable and reusable (FAIR). The Commission estimates that €2 billion in Horizon 2020 funding will be allocated to its so-called ‘European Cloud initiative’.*

▶ European Cloud Initiative to give Europe a global lead in the data-driven economy, 2016-04-19

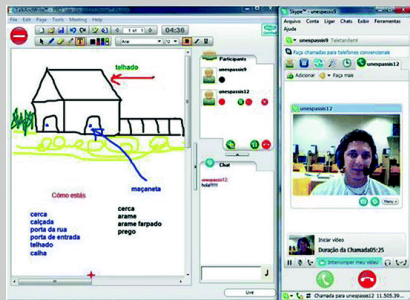
→ *The FAIR Guiding Principles for scientific data management and stewardship – Research Data Management on the basis of Findability, Accessibility, Interoperability, Reusability*

▶ Wilkinson et al., 2016

# Applying the FAIR Principles to CMC Corpora

## CMC Corpora

Computer-mediated communication (CMC) refers to human communication via computers and includes many different forms of interaction that humans have with each other using computers as tools to exchange text, images, audio and video.



► Cardoso and Matos, 2013

# Applying the FAIR Principles to CMC Corpora

## Findability - F

### F - the first step in (re)using data is finding it!

- Globally unique and persistent identifiers (PIDs) for data and metadata
- Rich and accurate metadata
- Metadata include the identifier of the data
- Registered or indexed in a search engine

# Applying the FAIR Principles to CMC Corpora

## Accessibility - A

A - once the data has been found, it needs to be accessed.

- Automatic retrieval of (meta)data on the basis of PID
- Allow for authentication and authorization if necessary
- Metadata should always be public

# Applying the FAIR Principles to CMC Corpora

## Interoperability - I

### I - enable interoperability with applications or workflows.

- Use shared standards for knowledge representation
- Proper documentation
- Cross-references to other data (if necessary)



# Applying the FAIR Principles to CMC Corpora

## Reusability - R

R - describe the data such that it can be used in different settings.

- Appropriate description of data
- Proper attribution of creators
- Clear and accessible usage license
- Extensive documentation of provenance

▶ [The GO FAIR Webpage](#)

## The CLARIN Resource Family for CMC Corpora

- 24 corpora of computer-mediated communication of
  - various sizes (600k to 670m tokens)
  - languages (e.g. DE, LT, SI, NL, ...)
  - and sources (e.g. Twitter, Whatsapp, Blogs).
- Ca. 50% (13) are deposited in a CLARIN Centre or a similar repository (META-SHARE, ..., zenodo, figshare)

▶ [CLARIN website](#)

For each corpus, we checked how well it complied with the four principles and its various subparts

- 1 Findability: via Google/Bing, VLO and OLAC search.
- 2 Accessibility: Was the data accessible?
- 3 Interoperability: Which format was the data in?
- 4 Reusability: Documentation of formats/methodology, licensing.
- 5 Other: Is the data openly available, is there a corpus paper or website.

## Our result table

Corpus	Size	F1	F2	F3	F4	A1	A1.1	A1.2	I1	I3	R1	R1.1	R1.2	R1.3	Open+Lic	Docu
Corpus of contemporary blogs (cs)*	1m	y	y	y	MD	MD	MD	MD	mD	NA	AS-Y	MD	Md	mD	ACA-BY-NC-ND	--
SoNaR New Media (nl)*	35m	y	y	y	MD	Md	MD	ME	MD	m	ASVY	Md	MD	MD	ACA-BY-NC-ND	WP
DIDI - The DiDi Corpus of South Tyrolean CMC 1.0.0 (de, it, en)*	600k	y	y	y	MD	MD	MD	MD	MD	NA	ASVY	MD	MD	MD	ACA-BY-NC-ND	WP
The Mixed Corpus: New Media (et)*	2.5m	n	n	n	md	--	--	--	MD	NA	AS-Y	md	MD	MD	on request (partly)	W-
Suomi 24 Corpus (fi)*	2.6b	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	ACA-BY-NC	WP
CoMeRe repository (fr)*	80m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY	WP
Dortmund Chat Corpus (de)*	1m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY	WP
LITIS v.1 (lt)*	190k	y	y	y	MD	MD	MD	MD	MD	NA	ASVY	MD	MD	MD	ACA-BY-NC-ND	WP
Blog post and comment corpus Janes-Blog 1.0 (sl)*	34m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY-SA	WP
Forum corpus Janes-Forum 1.0 (sl)*	47m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY-SA	WP
News comment corpus Janes-News 1.0 (sl)*	14m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY-SA	WP
Twitter corpus Janes-Tweet 1.0 (sl)*	139m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY-SA	WP
Wikipedia talk corpus Janes-Wiki 1.0 (sl)*	5m	y	y	y	MD	MD	MD	MD	MD	M	ASVY	MD	MD	MD	CC-BY-SA	WP
Flemish Online Teenage Talk (nl)	2.9m	n	n	n	--	--	--	--	--	--	----	--	--	--	no data	--
Dereko - News and Wikipedia subcorpus (de)*	670m	y	y	y	md	Md	Md	NA	MD	m	---Y	MD	MD	MD	CC-BY-SA	WP
DWDS - Blogs (de)	102m	n	n	n	m-	--	--	m-	--	m	A---	--	--	--	only query <sup>2</sup>	-P
Monitor corpus of tweets f. Austrian users (de, en)	40m	n	n	n	m-	m	m	m	--	NA	AS--	--	md	-d	on request	WP
FORUMAS_INDV corpus (lt)	600k	n	n	y <sup>1</sup>	mD	mD	mD	D	--	m	A---	--	m	--	download	W
INT_KOMETARAI_INDV2 corpus (lt)	4m	n	n	y <sup>1</sup>	mD	mD	mD	D	--	m	A---	--	m-	--	download	W
NTAP climate change blog corpus (no, en, fr)	21m	n	n	n	--	--	--	--	--	NA	---Y	--	--	--	no	P
Corpus of Highly Emotive Internet Discussions (pl)	1.60m	n	n	n	m-	m	m	m-	--	NA	AS-Y	--	md	--	on request	P
sms4science (de, it, fr, rm)	0.5m	n	n	n	m-	m	m	m-	--	--	ASVY	--	md	--	only query	W
What's up, Switzerland? (de, it, fr, rm)	5m	n	n	n	m-	m	m	m-	--	NA	AS-Y	--	md	--	no (not yet)	W
The Corpus of Welsh Language Tweets (cy)	7m	n	n	n	m-	m	m	m-	--	--	AS--	--	md	--	on request	W

Table 1: FAIR evaluation of CMC corpora.

(M) fulfilled / (m) partially fulfilled for metadata; (D) completely / (d) partially fulfilled for data; (y) yes; (n) no; (NA) not applicable

R1: (A) author information, (S) data source, (Y) year of data production/collection, (V) version information

Docu: unstructured corpus documentation; (P) scientific publication dedicated to corpus description, (W) corpus webpage

\* Deposited in research data repository (e.g. CLARIN, Metashare, Zenodo)

<sup>1</sup> There is no structured/machine readable metadata, but the corpus website provides a link to the data

<sup>2</sup> Only query, web page claim CC-BY-SA

If you want to study the table in detail: [▶ Frey et al., 2019](#)

## Clear distinction

- Deposited Corpora
  - Provided a PID and machine-actionable metadata
  - Were indexed in domain-specific search engines (CLARIN VLO, OLAC)
- Non-Deposited Corpora
  - Metadata most of the time only available via web pages or corpus papers
  - No use of PIDs
  - Some links to data on web pages/in corpus papers were outdated
  - Some of the corpora (the actual data) could not be found at all

## Again, clear distinction

- Deposited Corpora
  - All provide a clear usage license (mostly Creative Commons)
  - Data is always retrievable (if authorized ) through the repository
- Non-Deposited Corpora
  - Best case: download link on corpus web page, sometimes researchers need to be contacted directly first
  - Often no or unclear usage license

# The Results: Interoperability

## No clear distinction between deposited and non-deposited corpora

- Deposited corpora use structured metadata, but also here often crucial information (data source, retrieval period) is missing
- Data comes in a wide range of formats (TEI, XML, JSON), but the specific format is often not well-documented
- Some corpora would have benefited from clearer version information and cross-references to related corpora

# The Results: Reusability

## No clear distinction between deposited and non-deposited corpora

- Extensive metadata is needed to reuse corpora, but there is no common understanding of which metadata fields need to be present
- Repositories do not demand to fill a certain set of metadata fields
- A lot of non-deposited corpora were lacking a clear license, which is crucial to reuse their data
- A large part of the corpora was missing provenance (collection period, data source) and version information



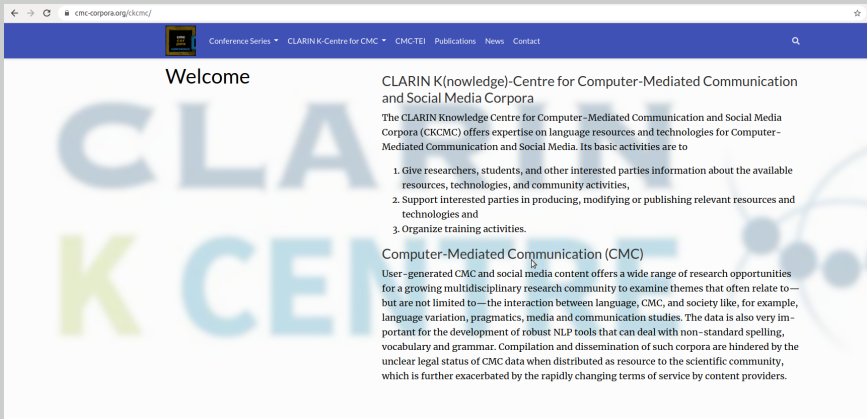
# Conclusions

- High variability of how the corpora comply or do not comply with the FAIR Principles
- In general, deposited corpora provide much better Findability and Accessibility
- Depositing does not *automatically* improve Interoperability and Reusability
- Provenance (source, time, description of the social media site) is crucial for reusing corpus data, but often not described in enough detail
- Notably, the community would benefit from commonly accepted guidelines on how to make corpora FAIR as well as from more agreement on used standards

## Short Checklist for CMC Corpora

- Choose and fix a name with a version placeholder
- Select a license (and start working towards it)
- Write a corpus paper
- Use an established format to represent your data (as early as possible, for example, TEI-CMC)
- Finally, deposit your corpus in a repository

# Have a look at *The CLARIN Knowledge Centre for Computer-Mediated Communication and Social Media Corpora (CKCMC)*.



The screenshot shows the homepage of the CKCMC website. The browser address bar displays 'cmc-corpora.org/ckcmc/'. The navigation menu includes 'Conference Series', 'CLARIN K-Centre for CMC', 'CMCTEI', 'Publications', 'News', and 'Contact'. The main content area features a 'Welcome' heading, a large background watermark of 'CLARIN K CENTRE', and a list of three activities. A 'Computer-Mediated Communication (CMC)' section follows, describing research opportunities and challenges.

← → 🌐 cmc-corpora.org/ckcmc/

Conference Series ▾ CLARIN K-Centre for CMC ▾ CMCTEI Publications News Contact 🔍

## Welcome

### CLARIN K(nowledge)-Centre for Computer-Mediated Communication and Social Media Corpora

The CLARIN Knowledge Centre for Computer-Mediated Communication and Social Media Corpora (CKCMC) offers expertise on language resources and technologies for Computer-Mediated Communication and Social Media. Its basic activities are to

1. Give researchers, students, and other interested parties information about the available resources, technologies, and community activities,
2. Support interested parties in producing, modifying or publishing relevant resources and technologies and
3. Organize training activities.

### Computer-Mediated Communication (CMC)

User-generated CMC and social media content offers a wide range of research opportunities for a growing multidisciplinary research community to examine themes that often relate to—but are not limited to—the interaction between language, CMC, and society like, for example, language variation, pragmatics, media and communication studies. The data is also very important for the development of robust NLP tools that can deal with non-standard spelling, vocabulary and grammar. Compilation and dissemination of such corpora are hindered by the unclear legal status of CMC data when distributed as resource to the scientific community, which is further exacerbated by the rapidly changing terms of service by content providers.

▶ CKCMC Website @ [cmc-corpora.org](http://cmc-corpora.org)



## An example: stop word lists

Open-source software packages for language processing often include stop word lists. Users may apply them without awareness of their surprising omissions (e.g. "hasn't" but not "hadn't") and inclusions ("computer"), or their incompatibility with a particular tokenizer.

► Nothman et al., 2018

## Versioning of linguistic tools

- In linguistic research, handcrafted toolchains (a variety of separate programs) are very common
- Often, it will be difficult to rebuild such a toolchain exactly
  - Some tools might no longer be available (or cannot be found)
  - It might not be completely clear which specific version of a tool was used
  - Some manufacturers do not keep older versions of their software available for download
- One solution is to create a (Docker) container with a "frozen" version of the complete toolchain
- Such a container can also be made available in a public container registry

## FAIR Take-Home Message

- F Deposit corpus/tool in a data repository
- A Deposit corpus/tool in a data repository
  - I Community needs commonly accepted guidelines and standards (TEI-CMC!)
- R As an individual project, describe corpus data & tools to facilitate replication and combination in different settings

# References I

- Cardoso, T., & Matos, F. (2013). Learning Foreign Languages in the Twenty-First Century: An Innovating Teletandem Experiment Through Skype. In A. Moreira, O. Benavides, & A. J. Mendes (Eds.), *Media in Education: Results from the 2011 ICEM and SIIE joint Conference* (pp. 87–95). Springer New York.  
[https://doi.org/10.1007/978-1-4614-3175-6\\_7](https://doi.org/10.1007/978-1-4614-3175-6_7)
- Frey, J.-C., König, A., & Stemle, E. W. (2019). How FAIR are CMC Corpora? *Proceedings of the 7th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora2019)*, 25–30.  
<https://cmccorpora19.sciencesconf.org/resource/page/id/15>
- König, A., & Stemle, E. W. (2019). Technical Solutions for Reproducible Research. In K. Simov & M. Eskevich (Eds.), *Proceedings of CLARIN Annual Conference 2019* (pp. 89–92). CLARIN.  
<https://api.zotero.org/users/332053/publications/items/D2WMT2UL/file/view>
- Nothman, J., Qin, H., & Yurchak, R. (2018). Stop Word Lists in Free Open-source Software Packages. *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, 7–12.  
<https://aclweb.org/anthology/papers/W/W18/W18-2502/>
- Sapir, E. (1921). *Language: An Introduction to the Study of Speech*. Harcourt, Brace.  
<http://www.bartleby.com/186/>



## References II

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018–160018.  
<https://doi.org/10.1038/sdata.2016.18>