How FAIR are CMC Corpora?

Jennifer-Carmen Frey, Alexander König, Egon W. Stemle
Institute for Applied Linguistics, Eurac Research
E-mail: {JenniferCarmen.Frey, Alexander.Koenig, Egon.Stemle}@eurac.edu

Abstract

In recent years, research data management has also become an important topic in the less data-intensive areas of the Social Sciences and Humanities (SSH). Funding agencies as well as research communities demand that empirical data collected and used for scientific research is managed and preserved in a way that research results are reproducible. In order to account for this the FAIR guiding principles for data stewardship have been established as a framework for good data management, aiming at the findability, accessibility, interoperability, and reusability of research data. This article investigates 24 European CMC corpora with regard to their compliance with the FAIR principles and discusses to what extent the deposit of research data in repositories of data preservation initiatives such as CLARIN, Zenodo or Metashare can assist in the provision of FAIR corpora.

Keywords: research data management, computer-mediated communication corpora, reusability, FAIR principles

1. Introduction

Over the last few years, both the scientific community and the public demonstrated a growing awareness of the necessity to make research reproducible and research data reusable (see, for example, Cohen et al., 2018; Wieling, Rawee, & van Noord, 2018; or the proceedings of the second dedicated 4REAL workshop, Branco, Calzolari, & Choukri, 2018). As part of general research ethics, the scientific community commits to making research transparent, to sharing and reproducing results, and to enabling the repeated use of costly created research data. However, this has various implications for research data management that regard the way research data is collected and preserved. In order to address these issues, Wilkinson et al. (2016) published the FAIR Guiding Principles (FAIR)[1] for data management and stewardship as a result of a joint workshop on the matter. The principles provide a universal framework for data management based on findability, accessibility, interoperability and reusability that can be utilized to establish community-standards for research data management (Mons et al., 2017). Over the last few years, FAIR have received international support, for example, at the G20 International Summit in Hangzhou[2], and have been adopted within individual domains (e.g. Boeckhout, Zielhuis, & Bredenoord, 2018) as well as within important funding schemes like Horizon 2020 (European Commission, 2016). However, FAIR as such have barely been discussed in the field of language resources, although also costly created language corpora need clear and well-planned research data management.
In this work we take a look at FAIR in the context of language corpora of computer-mediated communication (CMC). We identify the FAIR principles' implications for the CMC community and describe the current state of affairs by reviewing a list of European CMC corpora and assessing their compliance with FAIR.

2. FAIR & CMC corpora

FAIR aim at describing the characteristics of research data that are beneficial for their re-use in the scientific community. They provide added value to the scientific community by facilitating knowledge discovery and ensuring the transparency and reproducibility of research results as well as the long-term preservation of funded research.
FAIR are divided into the four main groups F, A, I, R (Findability, Accessibility, Interoperability and Reusability), each of which is subdivided into sub-items, for example, F1 or A1.1. We will address them in turn and interpret the principles for CMC corpora.

2.1. Findability - F

The most important precondition for having reusable and FAIR research data is to inform others of their existence. This aspect is addressed by the Findable principle of FAIR. It requires that data is described with rich metadata (F2) and both data and metadata are assigned globally unique and persistent identifiers (F1) that link to each other (F3). Additionally, the data should be registered or indexed in a (usually field-specific) search engine (F4).
For CMC corpora, metadata can be provided on dedicated corpus web-pages or in research articles. However, in order to comply with FAIR, metadata should be "machine-actionable", this means they must be represented in a structured and machine readable format and have a persistent identifier. Research data repositories for language corpora such as CLARIN centres (Hinrichs & Krauwer, 2014) or other data repositories such as META-SHARE[3], zenodo[4], and figshare[5] provide the infrastructure to store metadata in one or multiple specific metadata formats and automatically assign persistent identifiers. To find CMC corpora, general purpose search engines like Google and Bing or specialized search engines for language resources like the CLARIN Virtual Language Observatory (VLO)[6] and the Open Language Archives

[1] https://www.go-fair.org/fair-principles/
[2] https://www.consilium.europa.eu/media/23621/leaders_communiquehangzhousummit-final.pdf
[3] http://www.meta-share.org/
[4] https://zenodo.org/
[5] https://figshare.com/
[6] https://vlo.clarin.eu/

Community (OLAC)[7] can be used.

2.2. Accessibility - A

According to FAIR, research data are accessible if they can be automatically retrieved (A1) by their unique identifier (e.g. PID, URL) using a free and open protocol (e.g. HTTP) (A1.1). However, the retrieval method should also handle authentication and authorisation for non-public data (A1.2). Furthermore, even when access rights are restricted, metadata should still be accessible (A2).
For CMC corpora, this means that access to the data does not depend on individual, personal communication (e.g. mail requests), but that the data can be retrieved autonomously by standardised methods – usually via the internet. Furthermore, conscious steps should be taken to secure the long-term preservation of the metadata. Note that all these points can usually be addressed by depositing data in a research data repository.

2.3 Interoperability - I

In order to be Interoperable, both data and metadata have to use widely accepted standards for knowledge representation that are properly and openly documented. Proprietary or undocumented formats should be avoided (I1). If vocabularies are used to populate certain fields, they should comply with FAIR (I2) and cross-references should be provided whenever possible (I3).
For CMC corpora, there is no explicit knowledge representation format for data, ultimately also because it is still unclear what is to be represented at all. But as long as the format is open, broadly used and well documented, we see this as a step in the right direction. In this respect, the TEI standard (Burnard & Bauman, 2007) and other typical formats for corpora such as XML, JSON or CSV and CMDI (Broeder, Van Uytvanck, Gavrilidou, Trippel, & Windhouwer, 2012) for metadata are good examples. Cross-references between different data are not always necessary, but become relevant in the presence of similarly named corpora, related projects, different versions of a corpus, or the publication of different sub-corpora.

2.4. Reusability - R

To comply with the final principle of reusability, data should be properly described, with the information provided being both accurate and comprehensive (R1).
Relevant and therefore necessary metadata is dependent on the specific domain and existing community standards (R1.3). However, detailed provenance is an important part of this point (R1.2). It has to be clear where the data came from and who should be acknowledged for having played a part in its creation. For CMC corpora, for example, we assume that information on the type of communication (e.g. microblog, blog, forum), the origin of the data (platform e.g. Twitter, Facebook), the year of provenance as well as the corpus creator, possible updates and version numbers are crucial for corpus reusability.
Finally, the data should have a clear and accessible usage license, so potential users know what they can and cannot do with the data (R1.1).

[7] http://search.language-archives.org

3. Assessment of FAIR data management in existing CMC corpora

3.1 Methodology

For the empirical part of this study, we investigate a list of European CMC corpora and evaluate where and to what extent they comply with FAIR.
Our selection of corpora is based on the CLARIN CMC Resource Family[8], a publicly accessible and easily findable list of corpora dedicated to computer-mediated communication. Although the list is published via the CLARIN infrastructure, it contains language resources within and outside the CLARIN community, and corpora of various sizes (from 600,000 up to 670 million tokens), sources (Twitter, Facebook, Blogs, etc.) and languages (e.g. Slovenian, Dutch, German, English, Lithuanian). Of the 24 corpora listed in the CLARIN Ressource family at the time of this study, around 50% (13) were deposited within research data repositories of the CLARIN infrastructure (12) or similar providers (these corpora are marked with an asterisk in the table). This shows a relatively high awareness of the benefits of using established infrastructures for data management. However, as depositing data in a repository does not necessarily fulfill all the requirements for FAIR, we analysed the detailed compliance with FAIR (see Section 2) for each corpus of the list. Whenever applicable, we evaluated the compliance for both metadata (abbreviated as m/M) and the data itself (abbreviated as d/D). For the evaluation of metadata characteristics, we only considered machine-actionable, structured metadata, as prescribed by FAIR and further elaborated in Mons et al. (2017), as fully compliant. Corpus websites or scientific papers dedicated to the description of the corpus, which can be considered as additional metadata (availability listed separately in the table in column *Docu*) were investigated if no other metadata was available, but would only resolve to *partially compliant*. Furthermore, we added columns to indicate the size of the corpus in tokens (Size), the openness of the data and its license (Open+Lic). We interpret the general Reusability principle (R1) as whether the – in our opinion – important information on the data provenance, author, version and year of production of texts are provided. On the other hand, we have omitted the column for the use of FAIR vocabularies in the Interoperability principle (I2) because we believe that it is not (yet) applicable to the domain of CMC corpora. We also omitted A2 (preservation of metadata after data is not available anymore) because we cannot evaluate this point.
In order to check the rather abstract principle of Findability, we queried the search engines and data repositories mentioned in section 2.1.

3.2 Results

Below we summarize the results of our investigation. The detailed evaluation for each corpus can be seen in Table 1.

3.2.1 Findability of CMC corpora

Regarding the findability of the analysed CMC corpora, we observed the expected differences between corpora that were deposited in a research data repository and those that were not. The FAIR principle F1 requires metadata and data to have a persistent identifier (PID). Although the existence

[8] https://www.clarin.eu/resource-families/cmc-corpora

of such is not always obvious, the deposited corpora all provided a PID. Similarly, machine-actionable metadata (F2) was only available for deposited corpora, while other corpora where described mainly via corpus websites or research papers dedicated to the description of the corpus. For a few corpora, neither machine-actionable nor other types of data descriptions were available. The link between metadata and data (F3) was ensured for deposited data through PIDs in the metadata. Links provided on websites or in scientific publications were in some cases outdated. Concerning the findability of corpora via search interfaces (F4) we noticed the use of a data repository greatly increased findability because most add the information to special search engines like the VLO or OLAC. To our surprise, some of the corpora did not yield any results (apart from the CMC resource family website itself) with any of these search engines.

3.2.2 Accessibility of CMC corpora

We found a similar situation for compliance with the Accessibility principle in the investigated corpora. Deposited corpora were usually more accessible in terms of the retrievability of data and metadata via standardized protocols that are open, free and universally implementable (A.1.1), and that allow for authentication and authorisation when needed (A1.2). While accessibility does not necessarily mean open or free, most deposited corpora use Creative Commons or academic licenses. For the latter, an institutional user account valid for the CLARIN infrastructure[9] (e.g. a university login) suffices to retrieve data from CLARIN repositories.

For non-deposited corpora, metadata can often only be retrieved online via the HTTP protocol, while the data is not accessible or its accessibility is not clear and standardized (e.g. mail requests). Only sometimes there is specific information on how and under which conditions the corpus can be accessed and reused.

3.2.3 Interoperability of CMC corpora

With regard to the interoperability of corpora, that is, whether they use a formal, accessible, shared, and broadly applicable language for knowledge representation and vocabulary that complies with FAIR for metadata and data, and whether meaningful cross-references are provided, the division between deposited and non-deposited corpora is not so clear.

Non-deposited corpora often do not provide metadata in a standardised format (I1), but only describe the corpus on webpages or within a research paper, having deposited the corpus in a research data repository usually includes the availability of structured metadata files. However, while CLARIN enforces the repositories to use the CMDI standard, its inherent flexibility does not ensure comprehensive and appropriate documentation. CMDI only enforces a certain way of encoding information, but there are no mandatory metadata fields, meaning that even fully compliant CMDI metadata can contain very little information. With regard to the data itself, there are no clear instructions as to the data format in which a corpus should be uploaded to CLARIN[10] or any other data repository. Hence, some of the encountered formats do not comply with the FAIR requirements of being "formal, accessible, shared, and broadly applicable".

We have also found that the vocabularies used for data and metadata (I2) are rarely standardised or even documented and therefore do not comply with FAIR.

Although the need for appropriate cross-references (I3) is a rather subjective matter, we have found some corpora that would benefit from clear cross-references to other projects, different versions, or related corpora.

3.2.4 Reusability of CMC corpora

The availability of extensive metadata is essential for the reuse of CMC corpora, this includes metadata that goes beyond the needs of the original corpus project. But since there is no clear community standard about which information has to be provided and which metadata fields have to be filled in, there is still a lot of room for improvement.

FAIR also require licensing information and information on data provenance. The deposited corpora analysed in this work were all explicitly licensed. In most of the cases, a common licensing framework like the Creative Commons licenses was used to provide clear and comprehensive licensing information. For non-deposited corpora, the licensing is less coherent. Sometimes the article describing a corpus also covers the usage license (e.g. it states that the corpus is openly available but then does not state whether it can be reused and under which conditions).

Regarding the data provenance most corpora indicated an author, however, the concrete source of the data, its year of provenance, and especially the versioning information was not always clear.

Finally, FAIR recommends using domain-relevant community standards, but there are no clear standards for CMC data that are adhered to by the majority of corpora. This regards standardised vocabularies, minimum sets of metadata as well as data formats for CMC corpora. Note that there is a TEI SIG[11], but only few corpora were actually using TEI. Moreover, although CLARIN provides a list of recommended formats[12], there are no strict rules on using them and in case a non-standard format is chosen, there is no obligation to document choices, tags or structure. This leads to relatively free data formats, that might not be well documented (e.g. custom XML formats).

4. Discussion

In general, it can be said that depositing a corpus in a data repository helps to enforce Findability and Accessibility of corpora, while non-deposited corpora, in contrast, were often less findable (e.g. they were listed in the CMC resource family but not findable via any link or paper outside of this registry) and accessible. Given the lack of PIDs and structured metadata, these corpora were generally less compliant with FAIR.

In terms of Interoperability and Reusability, however, deposited and non-deposited corpora require further steps in order to comply with FAIR. This regards especially comprehensive documentation and the use of interoperable

[9] https://www.clarin.eu/content/federated-identity

[10] CLARIN provides some guidelines on data formats (see 3.2.4) but these are very generic.

[11] https://wiki.tei-c.org/index.php?title=SIG:Computer-Mediated_Communication

[12] https://www.clarin.eu/content/standards-and-formats

and reusable vocabularies and formats for knowledge representation, which apparently have not yet been established in the community. This lack of standardised formats might be self-induced by the many different corpus tools used by the community (e.g. different formats needed for different software packages that are used in parallel, or the software might be flexible enough to use semi-standardised data structures like custom XML, JSON, or CoNLL). One could argue that the CMC community does not need such common standards because the field is very close to computational linguistics, and people are sufficiently proficient in data conversion and data handling to work with their own standards. However, this usually leads to diverging definitions of identical terms, different terms for identical concepts, or even to different underlying schemata altogether. But to achieve true *conceptual interoperability* (Chiarcos, 2012), common terms and schemata linked with a common vocabulary and embedded into an encompassing ontology are paramount.

Also the data's provenance is a critical point for reusability. Documentation and the corpus description should comprise all steps from data collection, (pre-)processing and eventual transformations and modifications. Versioning should be explicit, that is, the scope and origin of different sub-parts of a corpus and their versions must be clear and the date of any update should be indicated, especially for corpora which are being constantly refined. Furthermore, in order to be reusable CMC data also needs to provide information on the time of data collection[13], as well as on the people involved in the collection, processing, and publication of the corpus, including an up-to-date contact address.

5. Conclusion and Future Outlook

Our study analysed the data management policies for CMC corpora in Europe according to the FAIR principles introduced by Wilkinson et al. (2016). Through a detailed investigation of 24 CMC corpora listed in the CLARIN resource family, we have shown that the currently prevalent data management policies are often only partly and almost never fully compliant with FAIR principles. While depositing a corpus in repositories for data preservation (e.g. via the CLARIN infrastructure or other data repositories) helps to ensure the findability and accessibility of research data, interoperability and reusability are exclusively driven by implicit (community) standards. However, such implicit community standards are not necessarily known to everyone when creating a CMC corpus for the first time, which may lead to non-interoperable or non-reusable data. In order to promote FAIR data management for CMC corpora, we see two necessities for the future: first, (continued) interest and efforts for depositing CMC corpora at (institutional) repositories for long-term research data preservation; and second, community-driven efforts to raise awareness for all stages of FAIR research data management.

In this respect, the already ongoing efforts within the community to introduce a TEI-CMC are particularly welcome and should be supported and the creation of a CLARIN K(nowledge)-Centre[14] for CMC could formalise and centrally register already existing expertise even further. All in all, this could make research on CMC corpora truly FAIR.

[13] Note that not all data repositories provide appropriate fields for such information.

[14] https://www.clarin.eu/content/knowledge-centres

6. References

Beißwenger, M., Wigham, C. R., Etienne, C., Fišer, D., Suárez, H. G., Herzberg, L., … Zesch, T. (2017). Connecting Resources: Which Issues have to be Solved to Integrate CMC Corpora from Heterogeneous Sources and for Different Languages? In E. W. Stemle & C. Wigham (Eds.), Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (pp. 52–55). https://doi.org/10.5281/zenodo.1041877

Boeckhout, M., Zielhuis, G. A., & Bredenoord, A. L. (2018). The FAIR guiding principles for data stewardship: Fair enough? European Journal of Human Genetics, 26(7), 931–936. https://doi.org/10.1038/s41431-018-0160-0

Branco, A., Calzolari, N., & Choukri, K. (2018). LREC 2018 workshop proceedings: 4REAL 2018 Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language.

Broeder, D., Van Uytvanck, D., Gavrilidou, M., Trippel, T., & Windhouwer, M. (2012). Standardizing a component metadata infrastructure. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), 1387–1390.

Burnard, L., & Bauman, S. (Eds.). (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange.

Chiarcos, C. (2012). Interoperability of Corpora and Annotations. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata (pp. 161–179). https://doi.org/10.1007/978-3-642-28249-2_16

Cohen, K. B., Xia, J., Zweigenbaum, P., Callahan, T. J., Hargraves, O., Goss, F., … Hunter, L. E. (2018). Three Dimensions of Reproducibility in Natural Language Processing. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 156–165.

European Commission: Directorate-General for Research & Innovation. (2016). H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020 (No. 3).

Hinrichs, E., & Krauwer, S. (2014). The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), 1525–1531.

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017).

Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use, 37(1), 49–56. https://doi.org/10.3233/ISU-170824
Wieling, M., Rawee, J., & van Noord, G. (2018). Reproducibility in Computational Linguistics: Are We Willing to Share? Computational Linguistics, 44(4), 641–649. https://doi.org/10.1162/coli_a_00330
Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … 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

APPENDIX

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	CC-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)*	25m	n	n	n	md	--	--	--	MD	NA	AS-Y	md	MD	MD	on request (partly download)	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[2]	-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[1]	mD	mD	mD	D	--	m	A---	--	m-	--	download	W
INT_KOMETARAI_INDV2 corpus (lt)	4m	n	n	y[1]	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)	160m	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)
[1] There is no structured/machine readable metadata, but the corpus website provides a link to the data
[2] Only query, web page claim CC-BY-SA