Making Research Data FAIR. Seriously?

Reflections on Research Data Management in the Computational Literary Studies

Introduction

Computational Literary Studies (CLS) are an evolving, interdisciplinary field of research combining research questions from the traditional field of Literary Studies with methods and technologies from Computer Sciences and Computational Linguistics. The German Research Foundation (DFG) is funding a priority program to foster the ongoing evolution of the field and the development and establishment of innovative computational methods in literary studies:[1] The priority program comprises eleven research projects in Germany and Switzerland and one central project (Pielström et al. 2021) for improving the interdisciplinary exchange between the projects and developing a common and domain-specific research data management (RDM) strategy.

Research data produced within the CLS is, similarly to many other disciplines in the humanities, heterogeneous (Pempe 2012). The management of this research data is a key element of scientific progress (Bryant, Lavoie & Malpas 2017) and a substantial aspect of good research practice (DFG 2019); in this respect, a major landmark are the FAIR Principles (Wilkinson et al. 2016). Within the central project of the program, we interviewed all projects (Helling et al. 2020) with regard to their discipline specific methods and approaches as well as the data and software they both use and produce during their research. We analysed the interviews qualitatively and quantitatively. The results of the survey (Helling et al. 2021) are used to develop and establish a common RDM strategy for the whole priority program to meet the FAIR Principles and enhance the sustainable findability, accessibility, interoperability and reusability of the data and outcomes of the projects.

In this paper we present our experience in RDM within the program. We will illustrate both pragmatic RDM solutions and major barriers in making research data FAIR. We will show that these barriers are intrinsic in the discipline itself.

Pragmatic Solutions and Barriers in Making Research Data FAIR in the CLS

We recommended Zenodo, which meets the FAIR principles, as a fallback solution for storing the outputs of the projects within the program, because overarching domain-specific infrastructures within the CLS are very rare. In fact, most of our projects were already using Zenodo for publishing outputs that do not fit into other infrastructures. However, the research

[1] https://dfg-spp-cls.github.io/ [last request: 24th of November 2021].

data of the program is also stored and published in institutional, generic and domain-specific repositories (see Figure 1 and 2).

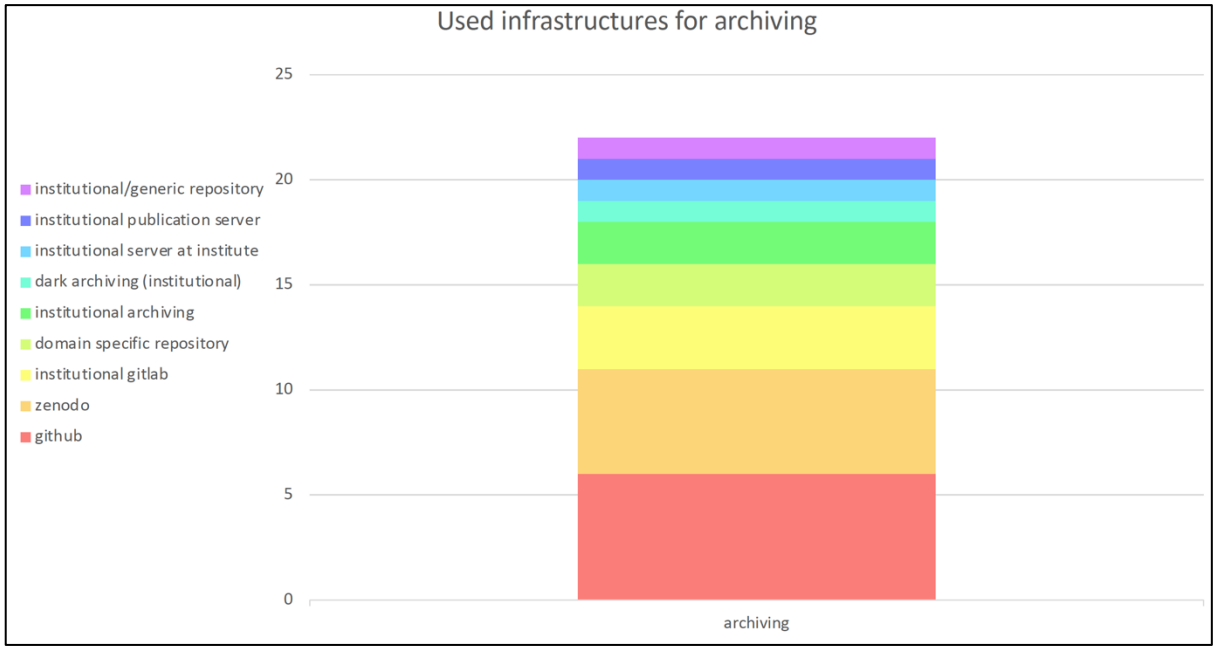

Figure 1: Used infrastructures for archiving within the program.

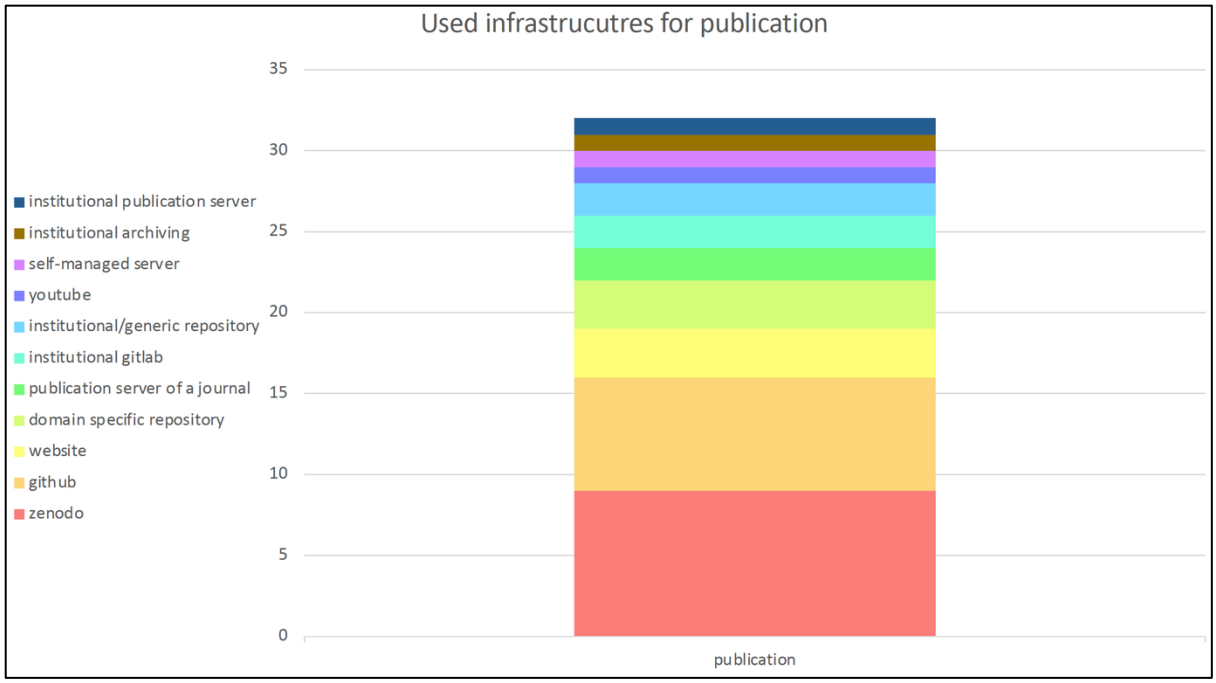

Figure 2: Used infrastructures for publication within the program.

Technical Perspectives on FAIR

From a technical perspective, repositories used within the program address the FAIR Principles fairly well. Regarding the findability and accessibility of research data, outputs of the program are usually registered or indexed in searchable resources (F4), which are accessible via standardized, open, and universally implementable protocols (A1/A1.2).

Besides Zenodo, some of the used infrastructures support the assignment of a persistent identifier (PID) (F1). Moreover, supplied metadata is often based on the generic DataCite scheme (F2).[2] In addition, most of the repositories and infrastructures offer the definition of generic licenses and the possibility of making the research data gradationally accessible (R1.1).

Domain-Specific Perspectives on FAIR

A large set of primary data in the priority program is beyond copyright by age, thus licenses for reuse in research and education are unproblematic. The remaining smaller set of data can be restricted by personal rights (studies) or individual copyright negotiations with authors, publishers, or libraries. In the community, there is a large interest to make data as accessible as possible and provide secure licenses (R1.1). Still for some aspects there are no clear solutions or test cases, such as the context necessary in derived formats (Schöch et al. 2020). The matter is regularly discussed in a working group on copyright within the program.

Schemes to capture provenance metadata (R1.2) are still evolving in the DH domains (cf. Gärtner et al. 2018). In contrast to e.g., the life sciences the objective is not a fully automatically reproducible workflow, but an equal treatment of automatic and manual steps which pose the domain-specific challenge. Individual documentation is available as well as commit histories from GitHub repositories.[3] So overall these aspects are evolving along the lines of FAIR.

Regarding interoperability (I1/2/3) and the relevance of attributes (R1) and standards (R1.3) the CLS requires to distinguish between resource-related metadata, content-related metadata, and data from annotations. Resource-related metadata is and can be based on DataCite (cf. F2 above), whereas more domain-specific metadata ranges from information on time period, genre and author uncertainty over technical settings like encoding (e.g., TEI variants) to the overall application of very different methodologies (see Figure 3) which comprises the existence of specific annotation layers as well as different segmentation schemes. Thereby neither content-related metadata nor annotation categories can come with a fixed, agreed on, common vocabulary since these categories are an integral part of the research (outcome) itself.[4] Still this vocabulary poses the basis for search, exploration and the FAIRness of the built resources.

[2] DataCite Metadata Schema 4.4, https://schema.datacite.org/meta/kernel-4.4/ [last request: 07th of December 2021].
[3] GitHub, https://github.com [last request: 07th of December 2021].
[4] Regarding annotations, Eckart and Heid (2014) argue for a separation of content-related interoperability and representation format-related interoperability. For the latter we found the projects in the priority program to agree on CATMA (Gius et al. 2018-2021) using its own TEI Export Format.

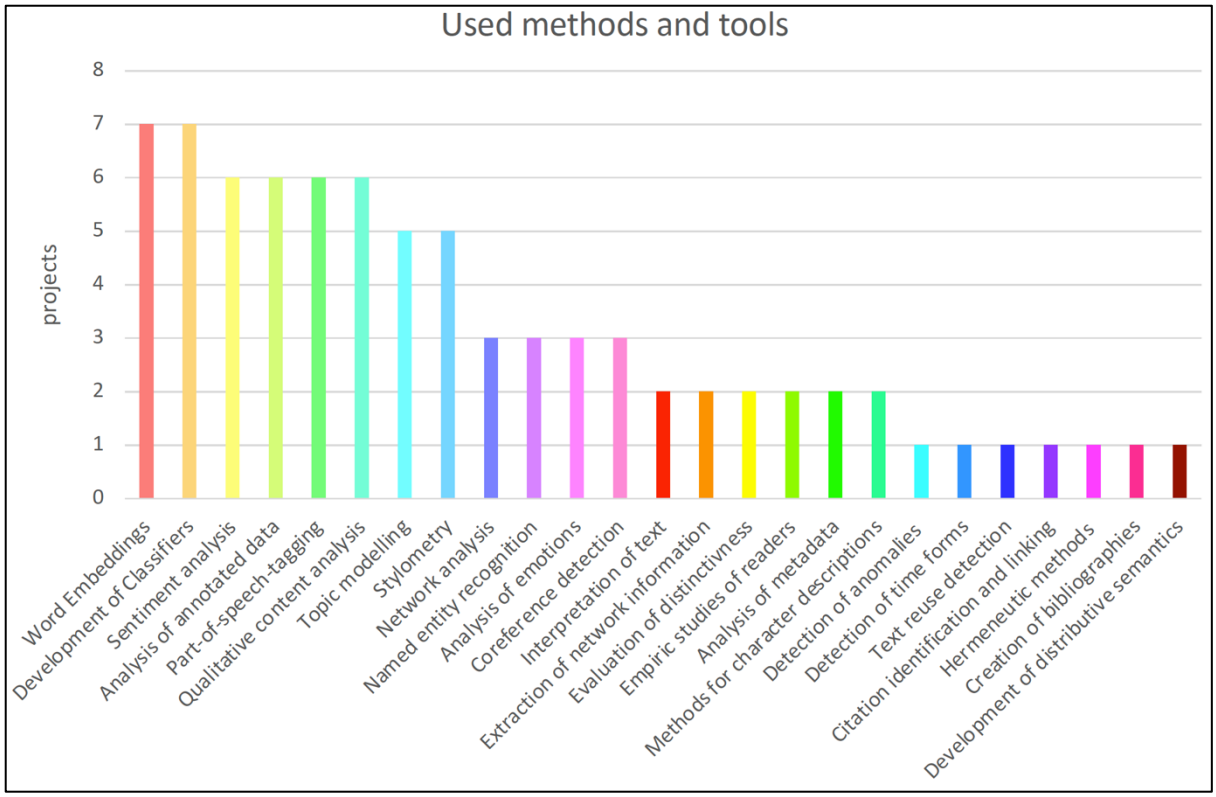

Figure 3: Used methods and tools within the program.

Conclusion

While different domain-specific as well as generic/institutional repositories meet the FAIR principles at least partially, Zenodo (also in combination with GitHub) is the closest infrastructure to the FAIR principles which is used in the context of the program. Nevertheless, it is still difficult to make research data stored in generic/institutional repositories findable for specific research communities, especially since a domain-specific metadata scheme is missing. Moreover, a common vocabulary for such a scheme possibly cannot exist without losing relevant content, differing between research fields within the domain. This problem is of course addressed by some more domain-specific infrastructures but still a comprehensive and domain-specific description model for the CLS is not existing.

In sum, without domain-specific metadata schemes, sustainable infrastructures and guiding legal implementations of copyright handling for the CLS, the FAIR principles can hardly be addressed in their entirety in this research domain.

Currently, pragmatic RDM seems to be the only way to meet the FAIR principles at least partially and to do effective RDM for the research community. In our talk, we will present more of our pragmatic RDM solutions and illustrate our approach to improve FAIRness of CLS research data for the CLS community. In this regard, a pragmatic approach for harvesting the heterogenous achievements of the program will be discussed. Finally, we will address specific RDM requirements for the CLS for fulfilling the FAIR principles and plead for a more domain-specific and measurable interpretation and implementation of the FAIR principles.

Bibliography

Bryant, R., Lavoie, B. and Malpas, C. (2017). A Tour of the Research Data Management (RDM) Service Space. The Realities of Research Data Management, Part 1. Dublin, Ohio: OCLC Research. DOI: https://doi.org/10.25333/C3PG8J.

DFG - Deutsche Forschungsgemeinschaft (2019). Guidelines for Safeguarding Good Research Practice. Code of Conduct. Zenodo: http://doi.org/10.5281/zenodo.3923602.

Eckart, K. and Heid, U. (2014). Resource interoperability revisited. *Proceedings of the 12th edition of the KONVENS conference Hildesheim*, Germany, pp. 116-26. URN: https://nbn-resolving.org/urn:nbn:de:gbv:hil2-opus-2725.

Gärtner, M., Hahn U., and Hermann, S. (2018). Supporting Sustainable Process Documentation. Rehm G., Declerck T. (eds) *Language Technologies for the Challenges of the Digital Age*. GSCL 2017. Lecture Notes in Computer Science, vol 10713. Springer, Cham DOI: 10.1007/978-3-319-73706-5_24.

Gius, E., Meister, J. C., Meister, M., Petris, M., Bruck, C., Jacke, J., Schumacher, M., Gerstorfer, D., Flüh, M., and Horstmann, J. (2018-2021). CATMA. Concept DOI: 10.5281/zenodo.1470118.

Helling, P., Jung, K., Reiter, N. and Pielström, S. (2020). Interviewleitfaden zur FDM-Bestandsaufnahme im Schwerpunktprogramm Computational Literary Studies. DOI: 10.5281/zenodo.4269639.

Helling, P., Jung, K. and Pielström, S. (2021). Disziplinspezifisches Forschungsdatenmanagement - FDM-Bedarfserfassung in den Computational Literary Studies. *FORGE 2021 Konferenz: Forschungsdaten in den Geisteswissenschaften - Mapping the Landscape - Geisteswissenschaftliches Forschungsdatenmanagement zwischen lokalen und globalen, generischen und spezifischen Lösungen (FORGE 2021)*, Cologne. DOI: 10.5281/zenodo.5379629.

Pempe, W. (2012). Geisteswissenschaften. In: Neuroth, H., Strathmann, S., Oßwald, A., Scheffel, R., Klump, J. and Ludwig, J. (eds), *Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme*. Boizenburg: Verlag Werner Hülsbusch, pp. 137-60.

Pielström, S., Helling, P. and Jung, K. (2021). Zentralprojekt des DFG-Schwerpunktprogramms Computational Literary Studies. *Program General Meeting*, virtuell. DOI: 10.5281/zenodo.5041338.

Schöch, C., Döhl, F., Rettinger, A., Gius, E., Trilcke, P., Leinen, P., Jannidis, F., Hinzmann M. and Röpke, J. (2020). Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. *Zeitschrift für digitale Geisteswissenschaften*. Wolfenbüttel. DOI: 10.17175/2020_006

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G.,

Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. In: *Scientific Data 3*, Article number: 160018. DOI: 10.1038/sdata.2016.18.