International FAIR Digital Objects Implementation Summit 2024

Extended Abstracts

https://doi.org/10.52825/ocp.v5i.1118

Published: 18 Mar. 2025

FDOs and Service Brokering, the CLARIN Switchboard Use-Case

Daan Broeder[1,*], Willem Elbers[1], Dieter van Uytvanck[1], and Michal Gawor[1]

[1]CLARIN ERIC, NL

*Correspondence: Daan Broeder, d.g.broeder@uu.nl

Abstract. Even though considerable progress with making research data and also research data processing services more FAIR, there is still lacking sufficient detail in the description of workflows, data and services for their immediate application and they are mostly not machine-actionable. CLARIN has developed the Language Resource Switchboard helping users find and invoke specific services for specific types of data. We consider and explore mapping this approach onto an FDO architecture and see what additional gains can be made.

Keywords: Research Data Management, Service Brokering, Metadata, SSH, CLARIN, FAIR, FDO

1. Background

The current landscape of general and thematic data and services has become more FAIR with more and improved registration and description by both thematic Research Infrastructures (RI) and more generic infrastructures, such as EOSC at the European level. Researchers and RI managers have, in principle, access to many more tools and services. However, this is also not without complexities in the sense that research workflows are not described in sufficient detail for immediate application, either alone or in combination (composition or orchestration). Especially regarding the composition of services, the problems are often underestimated. In the CLARIN domain, there is ample experience with what it takes to orchestrate Natural Language Processing (NLP) services, and usually such services must be especially designed for orchestration. E.g., special data exchange formats are required and designed, multiple input and output files are involved, etc. Nevertheless, advances w.r.t. brokering between available services and data have been achieved and can, to our expectation, also be provided beyond the CLARIN domain. It is especially interesting to see how this approach can also be mapped onto the FDO architecture if we expect a transition of the different RIs into using FDO infrastructure.

CLARIN has been exploring and developing a brokering infrastructure, the CLARIN Language Resource Switchboard (LRS), which can help users choose relevant data processing tools for specific types of language data. The infrastructure consists of (1) a registry of ‘known’ processing tools together with metadata that specifies which type of data a specific service can process, e.g., language; (2) a vocabulary of data formats (currently equal to IANA media types, aka. mime-types); and (3) a portfolio of web applications that integrate the LRS and utilise it to offer the user a choice of processing services matching the user's selected resource. e.g., a repository system UI allows the user to select a resource and invoke a matching processing service. Alternatively, a user can also directly upload or provide a link to data files or objects without a web application integrating LRS.

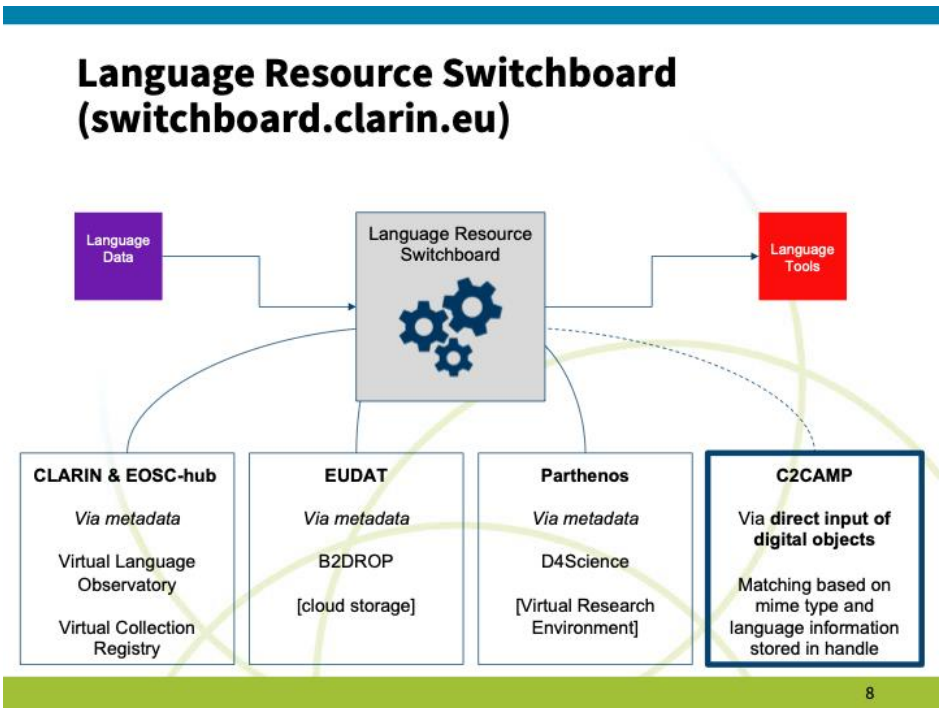

Figure 1. *The Language Resource Switchboard was integrated in several web-services, some from the CLARIN infrastructure itself, e.g., Virtual Language Observatory and Virtual Collection Registry or provided by collaborating initiatives, e.g., B2DROP and D4Science (Parthenos project)*

Note that, while this setup profits web applications to process the data by offering users a choice of processing services, it is not difficult to envisage limited changes that would support automatic orchestration scenarios.

While in the SSHOC project [1] the LRS was discussed, introduced, and tested in the wider SSH domain [2], there is no reason why its application should be limited to any particular thematic domain as described in Buddenbohm et al. [3]. In the FAIRCORE4EOSC project [4] the vocabulary of classic media types will be augmented by extended media types using a Data Type Registry (DTR) [5] for improved specificity and to allow taxonomical search of tools suitable for given data types.

Clearly, the effectiveness of this brokering infrastructure relies on, and can be increased by providing high-quality metadata and useful, understandable service descriptions along with detailed data types. Domain-specific aspects would be present in the detailed service metadata, e.g., CLARIN's Service Description (CSD) profile as described by Odijk [6], which is excellently suited for describing linguistic tools and services, especially NLP tools. CSD is much more specific than the general approach offered by, e.g., CODEMETA [7]. The domain specific aspects of data descriptions are partly in the community-specific descriptive metadata schema but also found in vocabularies, e.g., for media types, file formats, value-schemes for languages, text encoding, etc., which are already relatively well standardised. Crosswalks between such metadata schemas are described in the SEMAF report [8] and are a goal of the FAIRCORE4EOSC project.

2. Translating into FDO Architecture

We are now moving to FAIR Digital Objects (FDOs) described by Schwardmann [9], which are a model for identifying with PIDs, bundling and encapsulating a DO's content and metadata, and increasing interoperability by prescribing the essential information structure and access methods of a FDO. FDOs have machine-understandable typing as an essential characteristic.

Not only in the sense of typing the content by media types, but also by typing intrinsic properties of the FDO, e.g., if the object represents a collection or just a metadata record without content. Registries for such type definitions are called Data Type Registries (DTRs) and have been suggested already since long for other data management solutions by Lannom et al. [10]. Current DTR implementations are based on Cordra [11] and used in the FAIRCORE4EOSC project.

How would the FDO model further support and improve the brokering of data processing services such as the LRS? Most importantly, if FDOs are adopted by multiple RIs, it will make cross-infrastructure brokering of services possible, provided they share the FDO typing registries and service registries. Note that a suggested way of binding processing services to FDOs by linking them at birth, identical to the model of Object-Oriented Programming languages as mentioned by DeSmedt et al. [12], is acceptable for some basic services but cannot be exclusive since new services are created continuously and many useful services exist outside the direct knowledge of the FDO creator.

The machine actionability of FDOs, as specified in Weiland et al. [13], fits perfectly with the approach taken by the LRS. “Machine readability” is facilitated by the resolution of the PID into the bitstream of the object. This is already very well supported by making use of Handles and DOIs. “Machine interpretability” is facilitated via well-defined data types from Data Type Registries. This can be improved through closer integration with such DTRs, which provide more detailed types. "Machine actionability” is facilitated via well-defined services that can operate on specific data types and are managed in service registries. The LRS is already offering such a service registry, and the adoption of such an approach by other RIs will pave the way towards cross-infrastructure brokering and eventually orchestration. In the current landscape, the APIs for these data types and service registries are not yet standardised. The current CLARIN strategy is to implement adaptors and connectors for the various endpoints across registries and RIs and offer a single API to the LRS.

In general, the FDO model fits well with infrastructures such as CLARIN that rely primarily on several strong federated centres to provide discovery and access to the data and services in an agreed-upon, coherent way. CLARIN B-centres are required to provide data and metadata identification and accessibility using PIDs and standardised metadata using a repository system. Already in FDO Forum discussions, the use of repository FDO adaptors and connectors was suggested to create an FDO view on such repositories. Current work in CLARIN follows the adaptor strategy of developing the Digital Object Gateway (DOG) [14], which functions as a community proxy service and offers coherent access to the metadata and data of (part of) the whole set of CLARIN centres. DOG can be one basis for a community wide FDO adaptor earlier described by Broeder et al. [15].

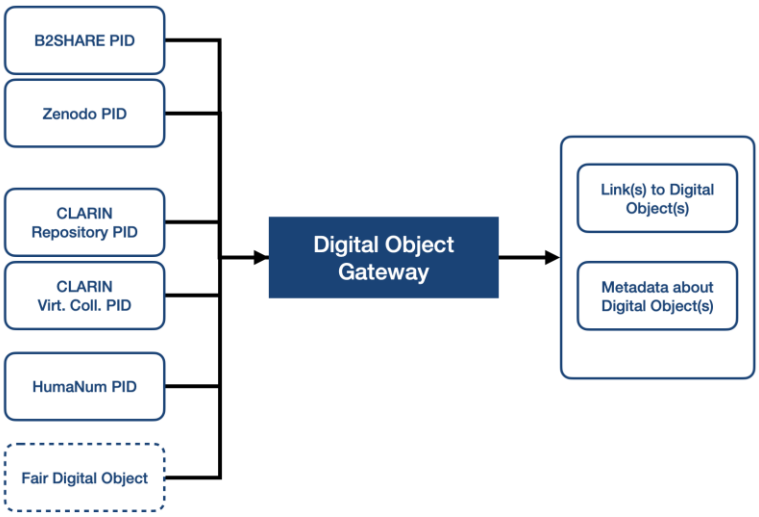

Figure 2. *The Digital Object Gateway functions as a community proxy service and offers coherent access to the metadata and data of (part of) the whole set of CLARIN centres*

The current landscape is one where data provider organisations often must meet the requirements of multiple RIs or dataspaces. A strategy to develop specific RI adaptors at the repository or community level seems like a reasonable solution while waiting for repository systems that can meet FDO requirements from the start. Adaptors for different infrastructures, developed for FDO interoperability should be able to enforce consistent typing of data content and structure, which would considerably expand the applicability and scope of tools such as the LRS over different RIs. However, it does also require the existence of service registries that offer detailed metadata, sufficient for mapping services and tools onto specific data types. For covering a broad scope of content data types (formats) and data processing services, a broadly agreed-upon extended media type registry would be needed, along with vocabularies to describe the FDO structure and processing service APIs in sufficient detail.

3. Conclusions and Future Work

Experiences within CLARIN and SSHOC show that there are challenges in obtaining such high-quality metadata for both data and services, as well as finding the resources required for curation. And while we have some experiences gained over the years on how and what the costs of curating high-quality metadata for data are, doing that for data processing services, including required brokering information, is maybe not yet that far. Beyond that consideration, we have a basis with the current DTR and plans for it to host extended media types, as well as the many already accepted value schemes for describing research data formats. However, a basis for a shared (or federated) set of registries of data processing services should be subject to further work.

Author contributions

Daan Broeder: Conceptualization, Writing original draft; Willem Elbers: Software, Writing – review & editing; Dieter van Uytvanck: Conceptualization, Writing – review & editing; Michal Gabor: Software, Writing – review & editing

Competing interests

The authors declare that they have no competing interests.

Funding

This study is supported by the FAIRCORE4EOSC (Core Components Supporting a FAIR EOSC) project, funded by the EU's Horizon Europe Research and Innovation Programme under Grant Agreement No. 101057264.

References

[1] Social Sciences & Humanities Open Cloud (SSHOC) project website. Available at https://sshopencloud.eu/project (Accessed 24 February 2024).
[2] Broeder, D., Elbers, W., Buddenbohm, S., Smilde, W., Dima, E., Durco, M., Concordia, C., Sanesi, ., & Degli'Innocenti, E. (2021). D3.8 Implementation report and available SSHOC Switchboard and VCR services (v1.0). *Zenodo*. https://doi.org/10.5281/zenodo.5608542
[3] Buddenbohm, S., Broeder, D., Eisner, M. I., Illmayer, K., & Durco, M. (2020). Collaborative Use Cases between SSH Open Marketplace and the Language Resource Switchboard and Virtual Collection Registry. *Zenodo*. https://doi.org/10.5281/zenodo.4442320
[4] Developing EOSC-core components to enable a FAIR EOSC ecosystem (FAIRCORE4EOSC) project website. Available at https://faircore4eosc.eu/ (Accessed 24 February 2024)
[5] EOSC Data Type Registry in the DTR in the FAIRCORE4EOSC project. Available at https://faircore4eosc.eu/eosc-core-components/eosc-data-type-registry-dtr (Accessed 24 February 2024)
[6] Odijk, J. Discovering software resources in CLARIN. In Inguna Skadina, Maria Eskevich, (Eds.), *Selected papers from the CLARIN Annual Conference 2018*, *Pisa, Italy, October 8-10, 2018. Volume 159 of Linköping Electronic Conference Proceedings* (pp. 121-132). Linköping University Electronic Press, 2018. Available from https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=159&Article_No=13 (Accessed 24 February 2024)
[7] The CODEMETA project website. Available at https://codemeta.github.io (Accessed 24 February 2024)
[8] Broeder, D., Budroni, P., Degl'Innocenti, E., Le Franc, Y., Hugo, W., Jeffery, K., Weiland, C., Wittenburg, P., & Zwolf, C. M. (2021). SEMAF: A Proposal for a Flexible Semantic Mapping Framework (1.0). *Zenodo*. https://doi.org/10.5281/zenodo.4651421
[9] Schwardmann, U. (2020). Digital objects–fair digital objects: Which services are required?. *Data Science Journal*, 19(1), 15. https://doi.org/10.5334/dsj-2020-015
[10] Lannom, L., Broeder, D., & Manepalli, G. (2015). RDA Data Type Registries Working Group Output. *Zenodo*. https://doi.org/10.15497/A5BCD108-ECC4-41BE-91A7-20112FF77458
[11] Tupelo-Schneck, R. (2022). An Introduction to Cordra. *Research Ideas and Outcomes.* https://doi.org/10.3897/rio.8.e95966
[12] De Smedt, K., Koureas, D., & Wittenburg, P. (2020). FAIR digital objects for science: From data pieces to actionable knowledge units. *Publications*, *8*(2), 21. https://doi.org/10.3390/publications8020021
[13] Weiland, C., Islam, S., Broeder, D., Anders, I., & Wittenburg, P. (2022). FDO Machine Actionability. *Zenodo*. https://doi.org/10.5281/zenodo.7825650
[14] Digital Object Gateway on CLARIN Website, Available from https://www.clarin.eu/dog (Accessed 24 February 2024)
[15] Broeder, D., Elbers, W., Gawor, M., Concordia, C., Larrousse, N., & Van Uytvanck, D. (2022). Towards FAIR Data Access. *Research Ideas and Outcomes*. https://doi.org/10.3897/rio.8.e94386