Towards FAIR Data Access

Daan Broeder[‡], Willem Elbers[‡], Michal Gawor[‡], Cesare Concordia[§], Nicolas Larrousse[|], Dieter Van Uytvanck[‡]

‡ CLARIN ERIC, Utrecht, Netherlands
§ CNR-ISTI - Institute of Information Science and Technologies "Alessandro Faedo" , National Research Council of Italy, Pisa, Italy
| Huma-Num, Paris, France

Corresponding author: Daan Broeder (daan.broeder@gmail.com)

Abstract

Background

In the past decade many different national, EU and global projects have been successful in raising awareness about Open Science and the importance of making data findable and accessible such as stated in the FAIR principles (Wilkinson et al. 2016).

In this respect, there have been many advances with respect to options for discovering data. A multitude of either thematic or general catalogues are providing faceted browsing interfaces for humans and Application Programming Interfaces (APIs) for use by machines and similarly, data-citations in publications offer references to resources hosted by repositories. However, using such catalogues and data-citations, researchers are not guaranteed to obtain access to the data itself. Mostly the resource link in the catalogue (and also in the metadata) or citation is a “landing-page”, a description of the resource meant for human consumption. The landing-page may contain instructions how to access or download the resource itself but usually it is difficult to parse by machines.

FAIR data access

Thus the approach sketched above does not meet the requirements in scenarios where applications need assured and quick access to data. Also the FAIR principles interpretation from GO FAIR states*[1] that these “emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.” The requirement for providing a Persistent Identifier (PID) for a resource*[2], is mostly interpreted as meaning a PID for the resource’s metadata or landing-page only.

Note that we ignore the need for user authentication and authorization prior to accessing data, here we will only consider data that is ‘freely’ accessible.

To improve the situation with respect to machine data accessibility a number of technologies and approaches that have been discussed in the CLARIN and Social Sciences and Humanities (SSH) infrastructure domain can be useful. We present some and comment on their suitability.

Signposting

Signposting*[3] is a technology proposed by van de Sompel (Sompel and Nelson 2015) to release relevant technical and bibliographical attributes from a resource URI. It's well described, and uses the HTTP protocol to provide additional information via HTTP Link Headers*[4]. Alternatively, for HTML type resources, the information may also be provided in HTML Link elements.

In the CLARIN community the signposting concept was accepted, but its proposed implementation deviated from van de Sompel and made it less dependent on the HTTP protocol (Arnold et al. 2021). However on the downside, the signposting information is embedded in the CLARIN specific Component Metadata (CMDI) (Broeder et al. 2012), and so makes it CLARIN specific, or at least requires clients to have specific knowledge about CMDI.

CLARIN Digital Object Gateway (DOG)

One approach that is currently worked on for the CLARIN research infrastructure is the creation of a DOG library*[5] and (later) a service that provides a proxy gateway from the resource PID to the actual data. DOG uses implicit knowledge about the different repository solutions that are used by the CLARIN B-type centres*[6] and some repositories outside the CLARIN infrastructure.

DOG works in two steps: first obtaining metadata from the resource PID and secondly extracting resource links from the metadata. Each of the repositories registered within DOG has a minimal configuration specifying how to parse fields of interest from the resource's metadata. For B-type CLARIN centres DOG uses content negotiation as the primary way of obtaining the metadata in CMDI format. For repositories outside the CLARIN infrastructure, DOG primarily relies on the API provided by the repository in order to access metadata and data resources.

The DOG solution does have scalability problems, but within the limited domain of CLARIN centres, it can offer a solution until a better one becomes available.

Limited PID kernel information

The (limited) PID kernel information approach assumes that for every Digital Object (DO) (Berg-Cross et al. 2015) and its metadata a Handle type PID (CNRI 2020) is issued and that the Handle information record can be used to store and associate additional important information with the (meta) data PID using handle value types such as for example a checksum and references to the data or metadata. This is a simplification of the architecture proposed in the work done in RDA context: PID Kernel Info recommendations

(Weigel et al. 2018). Consistent use of Handle information records could solve the data access problem, but just as for the signposting strategy, it requires strong discipline to maintain the additional information source. Examples from smaller projects and repositories exist that do manage this information in the Handle record eg. the DARIAH-DE repository*[7].

FAIR Digital Objects (FDO)

FDOs*[8] attempt to overcome the data management challenges posed by the heterogeneity and complexity of data using a combination of abstraction, virtualization and encapsulation (Schwardmann 2020). In practice, in the context of our access to data problem, the FDO solution can be seen as both a generalization and upgrade of the PID kernel information approach. The key characteristics here are the (conceptual) encapsulation of data objects with data structure and services that allow aware applications to recognize the data objects metadata and bitstream format, and process as intended by the programmer. Eligible data processing services, either general ones from communities, can be found through the FDO typing mechanism, or can be directly linked from the FDO.

A rich set of FDO attributes permit signaling machines processing FDOs where and how to access bitstream data including for instance additional information about supported protocols and APIs.

What to do?

For our community and in our collaboration with others, we need solutions now but would prefer not to invest and get closed in unscalable technologies.

We would propose to combine the DOG approach with signposting. First testing URIs (obtained by resolving the Handle PID) for the presence of HTTP Link Headers. If these are missing, (extended) DOG could use its idiosyncratic workflow. Long term we see advantages of the general, scalable and protocol independent approach that FDOs offer. Hybrid solutions are conceivable where FDO proxies can sit between the FDO machinery and data hosted by signposting compliant repositories.

Keywords

data management, metadata, CLARIN, PIDs, repositories, Signposting, Fair Digital Objects

Presenting author

Daan Broeder

Presented at

1st International Conference on FAIR Digital Objects, presentation

Conflicts of interest

References

- Arnold D, Fisseni B, Trippel T (2021) Signposts for CLARIN. Linköping Electronic Conference Proceedings https://doi.org/10.3384/ecp1803
- Berg-Cross G, Ritz R, Wittenburg P (2015) Core Term Definitions Data Foundation and Terminology, Work Group Products. RDA https://doi.org/10.15497/06825049-8CA4-40BD-BCAF-DE9F0EA2FADF
- Broeder D, Uytvanck Dv, Gavrilidou M, Trippel T (2012) Standardizing a component metadata infrastructure. In *Proceedings of the 8th Conference on International Language Resources and Evaluation (LREC 2012)*, Istanbul, 2012. *Conference on International Language Resources and Evaluation (LREC 2012)*, Istanbul, 2012.
- CNRI (2020) Handle.net Registry. http://www.handle.net. Accessed on: 2022-7-06.
- 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
- Sompel H, Nelson ML (2015) Reminiscing About 15 Years of Interoperability Efforts. D-Lib Magazine 11 (12). https://doi.org/10.1045/november2015-vandesompel.
- Weigel T, Plale B, Parsons M, Zhou G, Luo Y, Schwardmann U, Quick R, Hellström M, Kurakawa K (2018) RDA Recommendation on PID Kernel Information. Research Data Alliance https://doi.org/10.15497/rda00031
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18

Endnotes

*1 GO-FAIR, FAIR principles, accessed on: 2022-7-6 www.go-fair.org/fair-principles/

*2 GO FAIR, FAIR principles F1: (Meta) data are assigned globally unique and persistent identifiers, accessed on: 2022-7-6 http://www.go-fair.org/fair-principles/f1-meta-data-assigned-globally-unique-persistent-identifiers/

*3 signposting website, accessed on: 2022-7-6 https://signposting.org

*4 RFC5988, accessed on: 2022-7-6 doi:10.17487/RFC5988

*5 CLARIN ERIC GitHub, Digital Object Gate library README, accessed on: 2022-7-6 https://github.com/clarin-eric/DOGlib

*6

CLARIN website, CE-2013-0095: Checklist for CLARIN B Centres, accessed on: 2022-7-6 https://hdl.handle.net/11372/DOC-78

*7 DARIAH-DE repository documentation, Resolving and Persistent Identifiers, accessed on: 2022-7-6 https://repository.de.dariah.eu/doc/services/resolving.html

*8 FAIR Digital Objects Forum website, coordinates the FDO work and keeps track on related publications, accessed on: 2022-7-06 https://fairdo.org/library/