FAIR Digital Objects research and adoption: Let's pretotype

Fidan Limani[1,3] [0000-0002-5835-2784], Adamantios Koumpis[2,3] [0000-0003-2661-7749], Atif Latif[1,3] [0000-0003-3085-3031]

[1] ZBW – Leibniz Information Centre for Economics
[2] Institute for Biomedical Informatics, Faculty of Medicine and University Hospital of Cologne
[3] NFDI4DataScience consortium

Introduction and document scope

While demonstrators across communities and institutions about FDO adoption exist and are being developed, many communities still need to experiment and explore the value of adopting FDOs. In this document we do not focus on specific demonstrators - technology considerations and implementations, but reflect on how best to explore the value that FDOs can bring to an adopting community. And while most of the demonstrators in the 'FDO demonstrators 2021' document focus rather on the technicalities of adopting FDOs via different technology alternatives, such as the DOIP protocol, we prefer to focus on the perceived benefits from adopting the FDO model - initially focusing on the research artifacts in the context of the NFDI4DS project.

On FDOs and pretotypes

FAIR Digital Objects have been defined in (de Smedt et al., 2020) and have attracted attention and became part of the research agendas of several NFDI projects. Apart from the NFDI4DS (https://www.nfdi4datascience.de/) that the authors of the present article are involved, FDOs have been included in the research agenda of NFDI4Phys (https://nfdi4phys.de/), dating also to older communications and activities of working groups within the Research Data Alliance (https://www.rd-alliance.org/).

Whatever the perceived or associated value that FDOs may have, one may see the difficulty that FDOs face in winning a momentum. Namely, it is difficult to avoid the temptation to compare the way that the World Wide Web has been winning momentum. The latter followed a completely different approach where specifications followed the developments and the commitment for uptake. In contrast, the approach adopted for FDOs is that of building specifications that may see lack of acceptance by the industry or any other adoption and uptake communities.

While all the efforts for coming up with some shared understanding are ongoing, an important - and relatively effective and low cost - approach to gain momentum and further promote the adoption of FDOs is through testing a few ideas based on the FDO that a community would be interested in.

In the NFDI4DS project, as part of the "Research Knowledge Graphs" task area, there is already a recommendation to share tasks across measures via working groups that could start as soon as possible with the project realization. In this context, we can target an artifact collection and model it via FDOs. This should give

us the means to explore the adoption of this model, as well as the assumptions or issues we will face in the process, something that we can, in turn, give feedback to the FDO forum.

This exploration should provide us insights about adopting FDOs as our model of choice for representing research artifacts. In the process, we aim to assess whether this idea is the right one, whether there is a community that will engage in their realisation and, ultimately, if there will be commitment from industry to invest in them, if these will be proposed.

To cope with the above, it is worth considering the concept of *pretotyping* as a tool. The term has been coined by Alberto Savoia who had been Google's first Engineering Director and had introduced the term as a cheap-but-all-worth alternative to products and prototypes. In a talk that had been emblematically called "Skin in the Game"[1], Savoia recommends the use of what he calls *pre*totypes as an alternative to the more time-and-resources-taking prototypes and the even more expensive products. So the idea is to engage other researchers and scientists in the exploration of pretotypes that might demonstrate the usefulness and the value of FDOs.

Below we briefly mention aspects of this process which we expect to attract the interests of other researchers from the research community. For consistency reasons, we present for each of us arguments to the following questions:

- What can be the *Research interest in adopting the FDO model*
- Which are some *Use cases of interest* and, last but not least,
- Which can be some worth-to-consider *Industry application of FDOs*

FDO what?

Coming up with provocative or 'unpleasant' questions is a normal part of the scientific process and the researcher's routine. Two such questions could be:

- Consider that FDOs could mean different things or be manifested differently for different artifacts.Their meaning in this case could be a resultant of any of the FDO instance layers, such as its type, metadata or operations layers, etc.
- In the context of adopting FDOs, how can we distinguish between a paper vs data vs software vs scientific notebook, or any other research artifact?

One may come up with many similar questions that share a common bottomline: why should we use FDOs as a new vehicle, as long as there are no tangible proofs of their value?

FDOs in NFDI4DS

Research interest

FDOs can be used to represent more complex resources that could be considered also for the pretotyping process, consisting of more parts, aspects, etc., and serve as a way to model such resources that make the

[1] Skin in the game, Alberto Savoia; https://ecorner.stanford.edu/clips/skin-in-the-game-2/

resulting instance FAIR. The first ‘feature’ is in a way similar to initiatives to model aggregate resources, such as in the case of ResearchObject[2] or RMap[3].
This may have an impact on at least two types of use cases:

- Showcasing FDOs in practice: How do we represent e.g. an artifact collection (from NFDI4DS, for example) as a set of FDO instances?
- Showcasing the added value and the unique benefits that FDOs bring in terms of e.g. demonstrating how adopting the FDO model gives us more benefits than Digital Objects (Kahn & Wilensky, 2006) as well as their derivations.

Use cases

Within the working routines of data scientists and practitioners, there are several people involved with multiple roles, many contributors, all of them interacting with different systems. There is a plethora of data sources, code, results and, unsurprisingly the quest for reproducibility engages other aspects like e.g. computational resources close to the need for grounding human- and machine-implemented practices to achieve some type of ‘tangibility’ that FDOs – in all their constituent qualities of being findable, accessible, interoperable and reusable may offer to us. This has been the subject of an NFDI InfraTalk (Beyan et al., 2022) studied a specific use case of applying ML to a distributed data set, including a sequence of data science steps that include the development of the code with statistical and ML models to be trained and pushing to the relevant Git repos, creation of the necessary containers and their subsequent registration, sending the container to the first data source for execution, and, after training of the ML model with data from the first source, send it to second data source and execute, and so on to further other data sources and nodes. The scenario presented aimed to demonstrate the potential for applying the concept of FDOs for a case of distributed analytics following the concept of ‘Bringing analysis to data’. The case is based on previous work related to the application of the Personal Health Train paradigm to perform statistical analyses and Machine Learning on skin lesion data distributed among three Germany-wide data providers (Mou et al., 2021).

Industry applications

Instead of counting on or foreseeing the willingness of the industry to invest on FDOs, the effort should be to win commitment from start-ups that may in some cases be spin-outs of academia and research. There will be need to connect their offerings and co-create and build value. Pretotyping may again help here. We consider the case of a French start-up Mecabotix that has built a new generation of innovative reconfigurable robots. As we learn from their presentation, they offer them in the form of mono- (basic) and poly-robots (composite) called m-bots and p-bots respectively, and which can be used for offering services to logistic companies. M- and p-bots are able to aggregate around a load in order to move it, regardless of its size, shape or mass. If one now sees the FAIR qualities as very relevant and applicable to the context they are not wrong at all. The difficulty is that in industry one needs to leave the world of ideas to make a step towards the real world. Again everything mentioned above still hold: there are, again, several entities - not people but robots - involved with multiple roles, many contributors, all of them interacting with different systems. And there is, as expected, a plethora of data sources, code, results that affect the final outcome. One way to increase the value of FDOs for industry is to co-

[2] https://www.researchobject.org/
[3] https://github.com/rmap-project/rmap

build pretotypes with them, where gains and pains by the introduced concept can be examined, as for sure apart the gains there will be also pains.

Funding

All the authors are partially funded by a German Research Foundation DFG grant corresponding to the NFDI4DataScience project.

References

NFDI. (2022, June 20). FAIR Digital Objects within NFDI4DataScience and Beyond, NFDI InfraTalk, [Video]. YouTube. https://www.youtube.com/watch?v=lMeVUA-H3gc&list=PL08nwOdK76QlnmEB659gokiWN3AC-kqFS&index=15&t=1522s

De Smedt, K., Koureas, D., & Wittenburg, P. (2020). FAIR digital objects for science: From data pieces to actionable knowledge units. *Publications*, *8*(2), 21.

Mou, Y., Welten, S., Jaberansary, M., Ucer Yediel, Y., Kirsten, T., Decker, S., & Beyan, O. (2021). Distributed skin lesion analysis across decentralised data sources. In *Public Health and Informatics* (pp. 352-356). IOS Press.

Kahn, R., & Wilensky, R. (2006). A framework for distributed digital object services. *International Journal on Digital Libraries*, *6*(2), 115-123.