Practical webby FDOs with RO-Crate and FAIR Signposting

Document Version
Submitted manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):

Published in:
Open Conference Proceedings

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester’s Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.
Practical webby FDOs with RO-Crate and FAIR Signposting
Experiences and lessons learned

Stian Soiland-Reyes, Peter Sefton, Simone Leo, Leyla Jael Castro, Claus Weiland, and Herbert Van de Sompel

1 The University of Manchester, UK
2 University of Amsterdam, The Netherlands
3 University of Queensland, Australia
4 Centro di Ricerca Sviluppo e Studi Superiori in Sardegna (CRS4), Italy
5 ZB MED Information Centre for Life Sciences, Germany
6 Senckenberg – Leibniz Institution for Biodiversity and Earth System Research, Germany
7 DANS, The Netherlands

*Correspondence: Stian Soiland-Reyes, soiland-reyes@manchester.ac.uk

Abstract: Research Object Crate (RO-Crate) is a lightweight method to package research outputs along with their metadata. Signposting provides a simple yet powerful approach to navigate scholarly objects on the Web. Combining these technologies form a "webby" implementation of the FAIR Digital Object principles which is suitable for retrofitting to existing data infrastructures or even for ad-hoc research objects using regular Web hosting platforms. Here we give an update of recent community development and adoption of RO-Crate and Signposting. It is notable that programmatic access and more detailed profiles have received high attention, as well as several FDO implementations that use RO-Crate.

Keywords: FDO, Web, RO-Crate, Signposting, metadata, linkset, community

1 Signposting

The Signposting [1] effort started around 2015 in an attempt to address a long-standing problem regarding machine interaction with scholarly objects on the web. Omnipresent landing pages support human interaction with scholarly objects by providing descriptive metadata and links to content. But those pages are not optimised for use by machine agents that navigate the scholarly web. For example, how can a robot determine which links on a landing page lead to content and which to metadata? And, how can a bot distinguish those links from the myriad of other links on the page? Signposting addresses this by suggesting some purposely simple, yet standard-based patterns for using typed web links [2] to unambiguously point from a landing page to descriptive metadata (describedby link), content resources (item link), persistent identifier (cite-as link), etc.
1.1 FAIR Signposting

It is fair to say that the Signposting effort initially did not gather a lot of attention nor momentum. But that changed when, in 2020, the FAIR Signposting Implementation Guideline was published. FAIR Signposting goes beyond merely suggesting patterns as it specifies concrete, well-documented and illustrated recipes that developers can follow to add Signposting support to repository platforms or to implement compliance checks in FAIR assessment tools. It provides details regarding the precise semantics of the select set of link relation types, their cardinality, best practice recommendations for license and vocabulary URLs, levels of support, etc.

1.2 Web technologies in Signposting

Web links [2] used by Signposting primarily take the form of additional Link: headers in the HTTP response to a given resource, with rel= indicating the link relations as standardised in the IANA registry, e.g.:

```
Link: <https://schema.org/LearningResource>;rel="type"
Link: <https://w3id.org/a2a-fair-metrics/103-fdo-gr2-pid-record/>; rel="cite-as"
Link: <https://example.com/103-fdo-gr2-pid-record/metadata>; rel="describedby";type="text/plain"
```

FAIR Signposting is a profile for systematic use of a handful of link relations in level 1, adding bi-directional completeness in level 2. This non-intrusive addition of links means Signposting can be added to any kind of Web resource (including large data), and then inspected by using the regular HTTP method GET, or HEAD to avoid potentially costly content retrieval. However, adding HTTP-level Signposting requires control of the Web server, and injecting headers can be cumbersome in some Web server frameworks.

Signposting can also be embedded in HTML-based landing pages by adding the corresponding `<link href="..." rel="..." />` element within the page `<head>`, just like the familiar rel="stylesheet" link relation. This option is useful when developer control is limited to HTML templates, e.g. in a content delivery network (CDN) or GitHub Pages [3], however the HTML approach do not as easily provide inverse links back to the landing page (e.g. rel="describes").

In addition to the two previously mentioned Signposting approaches, typed links can also be provided in stand-alone documents named link sets. Two serialisations for linksets are defined in [4]: the text-based application/linkset format that is equivalent to the content of the HTTP Link header, and the JSON-based application/linkset+json format. Linksets are an attractive option when large collections of links need to be shared. They are made discoverable by means of links with the rel="linkset" relation (or rel="api-catalogue" for service descriptions). In theory, linkset representations can also be used in non-Web contexts such as the Handle system or local filesystems.

1.3 Signposting as web-centric implementation of FDO

An extra boost for Signposting came when it became clear [6] that FAIR Signposting essentially provides a web-centric model (Figure 1, right) for the notion of the FAIR Digital Object (Figure 1, left) [7]. FAIR Signposting exposes the topology of a FDO on the web and does so by using the simplest possible ingredients. The choice for
Figure 1. An identified FAIR Digital Object (left) captures a digital object in a PID Record through attributes and associated types, profiles and metadata, each of which may be other FDOs. FAIR Signposting (right) implements FDO using link relations to the persistent identifier (cite-as), the type PID (type), data/code (item), and metadata (describedby). FDO attribute shown as signposting include PIDs of author and license (not shown). Figure adapted from [5], [6].

Simplicity is motivated by a desire to minimise implementation and maintenance costs and is inspired by experience with numerous standardisation efforts over the years. The choice for web-centricity is about the availability of off-the-shelf tools and expertise as well as a realistic prospect of long-term sustainability.

Table 1 shows the correspondence between FAIR Signposting link relations and FDO concepts [8]. Note that while kernel attributes author and license are part of FAIR Signposting, any additional type-specific attributes should be provided by the metadata resource (e.g. a RO-Crate Metadata File), although extension relations are also possible.

<table>
<thead>
<tr>
<th>Link relation/attribute</th>
<th>FDO concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any rel= after PID redirect</td>
<td>FDO/PID Record (PID Profile: Signposting)</td>
</tr>
<tr>
<td>rel=cite-as</td>
<td>PID (persistent identifier)</td>
</tr>
<tr>
<td>rel=linkset</td>
<td>FDO Record (PID Profile: Linkset)</td>
</tr>
<tr>
<td>rel=describedby</td>
<td>Metadata, or Metadata FDO if PID</td>
</tr>
<tr>
<td>rel=describes</td>
<td>FDO Record that has this metadata</td>
</tr>
<tr>
<td>rel=item</td>
<td>Bit-sequence</td>
</tr>
<tr>
<td>rel=collection</td>
<td>FDO Collection, or FDO Record for a bitstream</td>
</tr>
<tr>
<td>rel=type</td>
<td>FDO Type</td>
</tr>
<tr>
<td>type= on rel=item or rel=describes</td>
<td>MIME type of bitstream</td>
</tr>
<tr>
<td>rel=profile</td>
<td>PID Profile</td>
</tr>
</tbody>
</table>
| profile= on rel=describes       | FDO Type of Metadata (e.g. PID Profile: RO-Crate)
|                                | https://w3id.org/ro/crate                         |
| rel=author or rel=license       | Kernel attributes                                 |
| rel=api-catalogue to linkset    | FDO Collection of FDO Operations                  |
| rel=service-desc or rel=service-doc | FDO Operation                               |
1.4 Signposting improves FAIR consumption

The EOSC task force for FAIR Metrics and Data Quality organised a series of “Apples to apples” hackathons and workshops, and have reported on FAIR Signposting and its uptake [9]. It highlights Signposting as a mechanism for guiding FAIR consuming machine-agents to locate the globally unique identifiers, data records and metadata records, and the challenge for FAIR assessment tools to make consistent evaluations without these elements.

A series of detailed benchmarks for Signposting have been created by the hackathons [10], that include identified test cases for each FAIR principle. For instance, 55-rda-r1-01m-t5-type-unresolve is a negative test case of a Signposting FDO where each of the rel=type has resolution errors. Such benchmarks are used by evaluation tools like F-UJI, as well as Signposting clients like the signposting Python library and command line tool.

Recently, additional tests for “Webby FDO” using Signposting have been added, e.g. 108-fdo-gr3-bit-sequence-metadata-pid verifies FDO principle FDO-GR3 to reference a bit-sequence, alongside a PID, type PID and metadata PID. Work to further mature FDO benchmarks continues as part of a FDO Forum task force.

1.5 Growing adoption through Support Actions and Hackathons

The FAIR-IMPACT project’s first support action funded 18 participants contributing to FAIR assessments as well as 14 participants implementing Signposting and RO-Crate in their repositories and code bases. Among the latter, Signposting support was developed for University of Novi Sad’s Research Information System DOSIRD UNS, Simula created a Signposting extension for CKAN to be used by the Norwegian Research Data Archive [11], ZB MED developed Signposting for GitHub Pages [3], [12].

Signposting linkset support was developed for InvenioRDM (used by Zenodo) and using HTML headers in Tunisia’s Scientific and Technical Information Portal by Tunisia National University Center of Scientific and Technical Documentation. EURAC Research’s Environmental Data Portal also added detailed signposting and new metadata serialisations, here the use of profiles in signposting was highlighted. The Digital Repository of Ireland (DRI) has implemented Signposting [13], adding detailed linksets at Level 2.

Several FAIR-IMPACT participants had planned to work on Signposting support in Dataverse, but that was concurrently implemented by the Dataverse community and relatively complete. For these participants, focus moved to upgrading and testing Signposting in their instances, as well as contributing to RO-Crate support in Dataverse to support richer metadata (see section 2.6).

When reaching out to the US NIH’s Generalist Repository Ecosystem Initiative (GREI), among its members (including Dataverse and InvenioRDM/Zenodo, Figshare, OSF, Vivi, Mendeley Data and Dryad), many indicated they were developing support for Signposting as well as user-provided metadata.

A growing list of adopters of Signposting, including repositories and data platforms, shows the interest in and feasibility of implementing Signposting. In addition to the above, earlier adopters include repositories WorkflowHub (Signposting to RO-Crate using profile), DSpace, DSPACE-CRIS, and the DANS Data Stations. Journal system and preprint server support include Episciences Overlay Journals, HAL Science and Open Journal System.
As part of the ELIXIR Biohackathon [12], the Signposting Sniffing plugin for Chromium-based browsers was developed and tested with signposting in the HoloFood Data Portal. The plugin supports both HTML- and HTTP-based Signposting, and can be used to debug and discover signposting.

These adopters highlight that it is easy to retrofit Signposting, but that the learning curve of other FAIR standards means that providing detailed metadata is more demanding.

## 2 RO-Crate community updates

The RO-Crate community supports and promotes FAIRification of research outcomes and provides tools for researchers to improve FAIRness. RO-Crate [14] makes it easy to package research outcomes in the form of digital objects together with their corresponding metadata, using and extending schema.org to describe an RO-crate package as a dataset and all its elements (i.e., research outcomes). Thanks to the use of profiles, researchers can choose a description that best suits their needs. A Profile Crate provides a specification including types and properties tailored to a specific domain or community (effectively a PID Profile that lists FDO Types). In this section we provide an update on recent new community adoption of RO-Crate, see also [14] for existing adoptions.

### 2.1 Workflows and provenance

Workflow Run RO-Crate [15] is a set of three RO-Crate profiles to capture the provenance of computational workflow execution: Process Run Crate, which deals with possibly composite computations not necessarily described by a formal workflow; Workflow Run Crate, which extends Process Run Crate with specifications to represent computations orchestrated by a workflow; Provenance Run Crate, which extends Workflow Run Crate with guidelines on how to represent the execution of the various tools used by the workflow. The profiles are developed by a working group that includes workflow developers and users, workflow engine developers and users and other members interested in the provenance of workflow execution. Workflow Run RO-Crate is already implemented by six workflow engines: Galaxy, COMPSs, StreamFlow, WfExS-backend, Sapporo and Autosubmit, enabling interoperable comparisons between heterogeneous workflow executions.

#### 2.1.1 Executable papers

LivePublication [16] is a proof of concept of an executable paper, whose interactive visualisation and statistical calculations can be regenerated on the fly taking into consideration data sources updated after the paper’s publication date. Here RO-Crate enables execution on the Globus infrastructure through an innovative use of individual RO-Crates and containers for each computable element of the paper, nested within a top-level Crate for the paper. This novel approach shows how it is possible to use RO-Crate as a machine-actionable object, which does not rely on bundling an underlying workflow representation in an existing workflow language.

#### 2.1.2 TRE-FX and TREs

Trusted Research Environments (TREs) are used to provide computational analysis access to sensitive data. An architecture for federated analytics running workflows
across such TREs [17] developed the Five Safes Crate profile [18], where the crate goes through manual and automatic review processes which are tracked within the crate and abstracts the workflow run requests so the federated analytics APIs are interoperable across different workflow engines [19].

2.2 RO-Crate and FDO

RO-Crate has previously been proposed as a mechanism for implementing FDO [20], in particular for metadata FDOs. The FDO approach fosters self-contained and machine-actionable knowledge units which persistently bind necessary contextual metadata and typed operation semantics to enable processing of their actual content across different data spaces [21].

One way to implementing FAIR Digital Object is using dedicated network protocols including the Handle system and the Digital Object Interface Protocol (DOIP) [22], [23]. In addition, an implementation path for web-based or “webby” FDOs [24] has been proposed to build on common web technologies [25]. This approach uses RO-Crate to provision structured metadata based on schema.org and its extension Bioschemas [26], deployed on the Web with persistent identifiers and FAIR Signposting [12].

2.2.1 Distributed FDO architectures

The DeSci Nodes system has developed a Distributed Persistent Identifier (dPID) concept which acts as an overlay of the Interplanetary File System (IPFS) [27]. APIs for the DeSci use detached RO-Crates with dPID references. This is a novel FDO implementation that challenges both the traditional centralised FDO approach using the Handle system, as well as the mostly web-based RO-Crate ecosystem.

Further considerations of FDO and Linked Data as distributed systems [28] evaluates FDO approaches in terms of capabilities based on existing frameworks (including the FAIR and FDO principles themselves), and highlights that the FDO community should take advantage of lessons learnt by the Linked Data community (avoiding similar mistakes, e.g. with relation to rigidity), and utilise fully the mature and modern capabilities of the Web technology stack, e.g. multiplexed streams.

2.3 Webby FDOs in biodiversity

The Biodiversity Digital Twin (BioDT) project is developing a specimen profile [29] that combines the Handle system and RO-Crate. A similar hybrid approach is taken for “retrofitting” Signposting to a FDO approach based on the digital object server Cordra in Senckenberg’s WildLIVE Portal [24]. This deployment injects Signposting headers as part of the public nginx server without modifying the proxied Cordra instance, except adding a single API method to generate the Signposting links [12] (Figure 2). Semantic mappings across biodiversity vocabularies [30] can also benefit from the use of RO-Crate, as a way to capture multiple mappings and to version and attribute the mapped semantic artefacts.

2.4 Using the Crate-O editor and building ad-hoc vocabularies

The Crate-O tool has been developed by Language Data Commons of Australia (LDaCA) as a general-purpose RO-Crate editor as a successor or alternative to Describo. This browser-based tool can use the file system access API, currently implemented in the
Figure 2. Signposting detected in the WildLIVE portal by the Signposting Sniffing browser detailing a data capture event. The link relations provide the set of the images contained in a capture event, metadata like license and authorship, and an RO-Crate comprising the actual content such as annotations (e.g. AI-based taxon identification).

Chrome and Edge browsers to describe any local folder with resources from the Web as an RO-Crate, or be embedded in an online service to work with any browser with server-side crates. It supports any schema.org type and property, pluggable with other rdfs vocabularies. A mode file selection indicates recommended and required properties; mode files combine schema information on classes, properties and defined terms, with RO-Crate Profile rules on how they may be defined. Notably this tool is also intended for creation of such “ad-hoc” vocabularies without need of Semantic Web, and is effectively a lightweight user interface for building Profile Crates using “Schema.org style Schemas” (SOSSs). There are a number of tools associated with Crate-O which can create HTML documentation and Crate-O mode files from SOSSs, provide simple validation against a profile and infer schemas from example documents.

2.5 Cultural heritage preservation in LDaCA

The Language Data Commons of Australia Program (LDaCA) contributes to language research and cultural heritage preservation by archiving Australian Indigenous languages, regional languages of the Pacific, and Australian English, and making such data and annotations publicly available. LDaCA and the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) use RO-Crate with an RO-Crate profile, as an interchange and archive format for language data, and is providing data discovery portals and API access to data using RO-Crate-centric APIs. For instance, the LDaCA data portal uses detached RO-crates for FDO-style navigation of centralised API resources.

2.6 Dataverse integration

In the research data repository software Dataverse [31], RO-Crate support is implemented by multiple plugins. The AROMA (ARP RO-Crate Manager) [32], used by Hungarian repository CONCORDA, extends Dataverse for dynamic metadata editing using the Describo Crate Builder Web component. The FAIR-IMPACT support action (section 1.5) saw community development of import, export and preview plugins for RO-Crate
in Dataverse, and also for the Electronic Lab Notebook format (ELN), which is based on RO-Crate and also supported by the research platform RSpace [33].

3 Discussion

“Webby FDOs” are being implemented by a range of adaptations that use Signposting and/or RO-Crate, in many cases accelerated by relatively modest support action funding (5 days FTE) or implemented at hackathons. This hints that these technologies are easy to deploy and developed, however there are some divergence between the approaches, which could be seen as a hindrance for full machine actionability.

Further documentation of the “Webby FDO” approach should specify concrete and testable recommendations (e.g. specifying a Signposting PID Profile), and the community should further develop the use of Profile Crate as a mechanism to formalise domain-specific PID Profiles, which should be made discoverable in a profile registry (in development).

The hybrid FDO approaches from biodiversity that combine the Handle system, Signposting and RO-Crate shows possible alignment across “FDO flavours”. Further development of the Signposting benchmarks and a common FDO testing framework should guide the FDO community towards a more unified ecosystem of FDO implementations.

Underlying and related material

Training material [34]: https://doi.org/10.5281/zenodo.10892090
Presentation slides [35]: https://doi.org/10.5281/zenodo.10847062
Signposting Benchmarks [10]: https://w3id.org/a2a-fair-metrics

Author contributions

Contributions according to CRediT taxonomy: SSR: Conceptualization, Project administration, Software, Visualization, Writing – original draft, Writing — review & editing. PS: Conceptualization, Project administration, Software, Supervision, Writing — review & editing. SL: Software, Supervision, Validation, Writing — review & editing. LJC: Investigation, Supervision, Validation, Writing — review & editing. CW: Investigation, Software, Supervision, Validation, Writing — review & editing. HVS: Conceptualization, Project administration, Visualization, Writing – original draft, Writing — review & editing.

Competing interests

SSR & HVS organised the FAIR-IMPACT support action, where LJC was one of the awardees. The authors declare that they have no other competing interests.

Funding

This work was partially funded by the European Union programme Horizon Europe under grant agreements HORIZON-INFRA-2021-EOSC-01-05 101057344 (FAIR-IMPACT), HORIZON-INFRA-2021-EMERGENCY-01 101046203 (BY-COVID), HORIZON-INFRA-2021-EOSC-01 101057388 (EuroScienceGateway), HORIZON-INFRA-2021-TECH-01 101057437 (BioDT), and by UK Research and Innovation (UKRI) under the UK government’s Horizon Europe funding guarantee grants 10038992 (FAIR-IMPACT), 10038963
(EuroScienceGateway), 10038930 (BioDT). LJC’s work is partially supported by the NFDI4DataScience project, funded by the German Research Foundation (DFG), project number 460234259.

Acknowledgements

We would like to acknowledge the RO-Crate Community, the Signposting community, the Apples-to-Apples hackathon participants and the FAIR-IMPACT Support Action #2 participants.

References


