The Cross-Domain Interoperability Framework (CDIF): Current Status

Arofan Gregory and Simon Hodson, 14 December 2022

Contents

Overview

The Cross-Domain Interoperability Framework (CDIF) is an effort which has grown out of several years of discussions around how FAIR can be implemented – and the promise of data sharing and integration practically realized – across domain and infrastructure boundaries. This document describes the current state of play, and highlights some of the topics which have emerged as a focus for this discussion.

The main driver for this idea has come from a series of workshops organized by CODATA, GESIS – The Leibniz Institute for the Social Sciences, and the DDI Alliance working with a handful of other institutions and experts from a wide variety of backgrounds at Schloss Dagstuhl. Week-long intensive workshops were held each year starting in 2018 (see https://codata.org/initiatives/decadal-programme2/dagstuhl-workshops/).

Over the course of these workshops, it became clear that in many different domains, some approaches and standards were very common. The workshops were organized around specific use cases, and these gave rise to the idea that a core set of non-domain-specific metadata models and standards existed which could be combined to provide reasonably complete support for data reuse in a cross-domain scenario. Some of the major challenges to this approach have also come into view.

These ideas have been presented and discussed in many other for a during that time: at RDA plenaries, in SciDataCon sessions, in EOSC and GOSC efforts, etc. During the last of the Dagstuhl workshops, held in August of 2022, a first attempt to begin the formalization of the CDIF was launched, with the support of some of the work packages from the WorldFAIR project and others.

CDIF is intended to be a set of practice recommendations based on existing standards and approaches which supports technical implementation across the entire set of FAIR principles. Much attention has been paid to *Findability* as a starting point for implementing FAIR, but less to the other principles. Given that much of the needed metadata is reused across the breadth of FAIR, it was seen as advisable to look at all of the problem space in a single glance. CDIF can help us understand the end goal of full support for FAIR, and to establish a path toward achieving it.

CDIF has proposed the creation of a *lingua franca* to allow for FAIR exchanges across domain boundaries: each domain can map its domain-specific metadata holdings against CDIF-recommended standards, allowing for a reasonable degree of interoperability across domain boundaries without requiring systems to support every possible domain standard involved.

In order to support this idea, a working methodology was established during the 2022 Dagstuhl workshop, based on the determination of design principles for the recommendations, and looking at the functional basis of FAIR exchanges across the set of FAIR principles. Consideration was also given to how a set of information models and standards could usefully be employed as the basis for services development.

The three use cases from the 2022 effort included Findability, based on the application of the approach used in the Oceans Infohub (ODIS) to related work in the UN on Disaster Risk Reduction; an examination of how Access can be programmatically negotiated based on current prototypes in the social sciences archives in the UK, Australia, and elsewhere; and work on data integration and harmonization using environment, social science, and public health data.

The work here has only been started. This document will briefly describe these topics, and suggest some of the standards and models which seem to be candidates as this work moves forward.

Goals and Scope

CDIF is not intended to be a final solution to all the challenges of FAIR implementation. It is focused on creating a practical (that is, a "reasonable" 80%-style) approach to enabling systems to exchange data and metadata across domain boundaries at the application level. It builds on other work around FAIR interoperability, notably the FAIR Digital Object Framework (FDOF) or a similar solution at the communication protocol level.

FAIR is seen as involving granular descriptions of data (that is, at the level of variables or observations, and not just at the data set or service level) and it assumes that data reuse should be automated to the greatest extent possible. This implies a metadata-driven approach in some respects, and the resulting focus is on the use of metadata standards to support exchange.

Ideally, CDIF would present a set of reasonable solutions to the problems of implementing FAIR which could be used as a guide by systems developers. Specific functions should have a small number of recommended solutions, including profiles of metadata standards and specific technical approaches. CDIF likely stops short of providing service-specific APIs, but should form a clear basis on which these can be developed. The exact line to be drawn in regards to the services implementation layer has yet to be determined, but there is clearly a strong relationship between an identified information set for supporting a function, expressed in a standard fashion, and the API used to communicate that information between systems.

The intention of CDIF is to address the technical framework within which FAIR-enabling services can operate at the level of information exchange. It does not look to solve problems of legal or organisational interoperability, or to focus on the challenges of capacity building or "metadata uplift" within communities implementing FAIR. That said, having an agreed set of guidelines will make those topics easier to address, by bounding the discussions: if the goal is a known set of technical implementations, then the capacity required and the demands for legal and organizational readiness can be more easily specified and addressed. This is also true of FAIR assessment. Those aspects of the work lie beyond the scope of CDIF.

The initial work will be carried out under the auspices of the WorldFAIR project. This is an EU-funded project but is international in scope, lead by CODATA, with RDA as a major supporting partner. It is a time-limited project, with the intention of producing an initial draft of CDIF. Once that draft exists, it is hoped that fairly frequent iterations can be made to extend the scope of the recommendations and to keep them current vis-à-vis developments in the technology and standards space, but in line with the established design principles. The institutional basis for on-going development and maintenance beyond the WorldFAIR project has yet to be determined.

Working Method/Process

At the 2022 Dagstuhl Workshop, an initial outline for CDIF was developed. This reflected a working methodology agreed for the purposes of the use cases considered there, and provides a basis for further progress.

Each area of FAIR was broken out into a set of Capabilities – the business-level interactions required between systems. These were then broken down into lower-level component Functions, with requirements in terms of the Information Objects needing to be exchanged. These Information Objects could them be expressed using specific standards and models.

Establishing design principles to inform this work was begun in parallel. Although not an ideal way to establish design principles, it is felt that making these explicit is important in providing direction, clarity, and sustainability to the work.

Not all use cases were at the same level of maturity: discovery was the most mature, and benefitted from existing large-scale implementations (ODIS) which could be used as a model. Data integration and harmonization is also an area where there is some fairly mature prototyping, notably within EOSC Futures WP 9 combining environmental and social sciences data in Europe, and around efforts to implement an ESS-style survey in Australia. Automated support for data access is being prototyped at the UK Data Archive, but is the least well-understood of the areas to be analysed.

In every case, the outputs from the use cases were in an early form – none of this work is yet completed, although initial drafts were started, and can be seen at https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/2832367617/2022+Interoperability+for+Cross-Domain+Research+Machine-Actionability+Scalability. (Follow the links to the Google docs from the sub-pages to see working drafts. Note that these are still very rough.)

Candidate Standards and Models

A number of standards have emerged from the 2022 work and earlier which seem to recommend themselves as candidates for CDIF. This section highlights some of these, with an eye toward further exploration and discussion. There is no claim that what is presented here is comprehensive – this list has grown organically from the cast studies we have looked at up to this point, and there are likely to be others which have not yet been identified.

The FAIR principles are not a functional breakdown of what is required for data reuse – they are aspirational, rather than prescriptive. To implement them, a more functionally oriented view of the interactions needed between systems is used, to better identify the roles played by different standards in these exchanges.

1. Discovery and Cataloguing
2. Assessment of fitness-for-purpose
3. Negotiating and performing data access
4. Obtaining structural metadata
5. Understanding semantics
6. Understanding the context of data reuse
7. Identifying and managing the data reuse process

We will look at each of these areas in turn, and briefly suggest some of the standards which have come up in discussions up to this point in the work.

Communications Protocols

At this time, the FAIR Digital Object Framework (FDOF) seems to be the clear choice here. Without final specifications in hand, it is difficult to know exactly what this will entail, but existing drafts show that this is a viable solution for digital objects to identify themselves as FAIR, and for assembling the needed packages of information to further process them. The work on signposting offers a practical way forward here.

Discovery and Cataloguing

There has been a lot of work done with both Schema.org and DCAT in this area. In each case, communities tend to produce extensions or profiles which may not work across domain boundaries, but there has also been some interesting work done in looking at what a cross-domain set of properties would look like (notably at the Swedish National Archive, SND). It does appear that both standards may prove useful in many implementations, as they offer different features.

In terms of implementation of Schema.org, the inclusion of JSON-LD as a script element in landing pages seems to be a viable approach, and one which is seeing considerable uptake. ODIS provides a good example of how this can be done.

Assessing Fitness-for-Purpose

Assessment of data is often done when data is being discovered, but may also occur after data has been accessed. It is the examination of data to determine whether it is useful for the intended research question, often through integration with other data. Questions regarding periodicity and geographic granularity are paramount, but other methodological considerations may also be important. This is a very metadata-intensive activity (category and other statistics at the variable level are very important here) which can rely on the same standards which describe structural metadata and provenance/processing, as discussed below. Often, some aspect of these (temporal coverage and periodicity, geographical coverage, etc.) can be usefully provided at the point of discovery through other standards such as DCAT and Schema.org.

Negotiating and Performing Data Access

Granting access to confidential data is today often a manual process conducted at the level of entire datasets or databases. Many functions could be automated, however, and with sufficient granularity in data management practices much data which is currently restricted could be made available (it is combined with disclosive data and "poisoned" by it, even though it may not itself be disclosive – a side-effect of giving access at the level of the data set or database). Standards such as the Data Privacy Vocabulary (DPV) and the Open Digital Rights Language (ODRL) may offer some support for automating these processes, and the DATS model and the DUO vocabulary are also of interest here. Given the paucity of practice, some attention should be given to implementations such as that currently ongoing at the UKDA, to see what practical implementation can look like.

Obtaining Structural Metadata

Many domain-specific standards provide a description of structural metadata, independent of semantics, but these are often limited. The only standard designed to work across domain boundaries with a reasonable descriptive breadth is the DDI Cross-Domain Integration (DDI-CDI) spec. Others of interest include SDMX/DataCube and CSV on the Web, but these may not be broad enough or provide

strong enough ties between the concepts employed and the structural roles they play. (While of greater complexity, DDI-CDI does provide explicit concept tie-ins.).

Understanding Semantics

The challenge of harmonizing domain semantics is a daunting one. While we have good standards for describing concept systems (SKOS and XKOS being the most common, with OWL for Ontologies, RDF-S, etc.) these do not solve the problem of mapping semantics. If we have good structural information showing how any given concept is associated with the data, and we can find a rich description of it within a classification or ontology, we are still left with the problem of understanding its equivalence with other, similar representations.

There is on-going work in this space, and some very interesting projects to look at. The Simple Standard for Sharing Ontological Mappings (SSSOM) gives us an idea of how these equivalencies can be described and reused, and there is other interesting work going on in some other initiatives. Projects like BioPortal have shown that large-scale mappings can be established and tooled. These approaches need to be analysed – crosswalks may not be direct descriptions of data, but they are an incredibly valuable form of metadata nonetheless in terms of their ability to support data integration and harmonization.

Understanding the Context of Data Reuse

There are two distinct challenges when it comes to understanding the context of data coming from "external" domains: (1) understanding the purpose and collection method of the data as it was known by the data producer; (2) knowing the full set of information to ensure effective and responsible reuse. The implicit knowledge of data which exists within domains needs to be more explicit in a cross-domain scenario.

PROV gives us a good basis for describing provenance, but needs to be enhanced through specific profiling (PROV-One, RO Crates) or supplemented with additional standard information about processing (VTL, SDTL. etc.).

Further, standards such as I-ADOPT and the Observations & Measurements (O&M) work from OGC show that in order to fully contextualize a variable, we will need a description of the additional values which inform its use.

Tracking and Managing the Data Reuse Process

People often forget about the need to track data reuse, as it is itself an activity requiring management, and is also a way of producing new data. Frameworks exist for understanding how the business of data reuse can be attached to the business context (within a research institution, etc.) and how outputs and costs can be associated with it. CERIF is popular in the environmental domain in Europe, and may be interesting to consider. In the world of official statistics, the GAMSO standard offers a parallel model. We should consider how a technical framework for data reuse can be usefully connected to the business framework within which it operates, and how impacts can be understood in the context of research and policy.

Units of Measure, Geography, and Other Widely Reused Information

Some types of information are central to scientific data integration and reuse, and will need to be addressed. Transformations to and from a *lingua franca* may be less useful here, as it may be possible to

recommend existing standards specific to these types of information – they are “domain independent” in practical terms. This will need further investigation.

The work of the DRUM group as regards units of measure should be considered. The various offerings in the geospatial standards world may also bear consideration. There may be other, similar types of information which can be described in standard forms which are in practice universal, and these should be flagged. The goal here is to embrace existing practice in those cases where it already supports the needed interoperability.

Timelines and Organization

Under the current WorldFAIR project, an initial draft of CDIF is to be delivered in the spring of 2024. (The project work for the entire effort will be completed by end of May 2024.) The intention is to continue WorldFAIR and to support the development of CDIF after that time, and we are already looking for funding and support for this work.

The work will initially be organized into two groups: a working group to produce proposals and drafts, and an advisory group to review and comment on them. The WorldFAIR project will provide support and perform many of the organizational tasks as well as helping with drafting and editing. Further, the WorldFAIR case studies – eleven in different domains – will provide feedback to the proposals, to help keep the work grounded in reality. This feedback process will be facilitated by the WorldFAIR WP 2 staff. (This organization is flexible, and can be modified as needed, but the starting point is as described.)

Open questions remain as to how the CDIF effort will be continued – ideally, it will enjoy broad-based support throughout the FAIR implementation community, and a good institutional basis can be arranged for it. In the immediate term, an implementable draft of CDIF is the critical piece. We have more than a year to progress the work, and feel that this will require a high level of energy and focus if we are to produce something tangible. If we work with a “early and often” mindset, it is hoped that we can progress from a modest-but-sufficient beginning to something more robust in time.

The first draft of CDIF should serve to illustrate that practical guidance around FAIR is useful and will help us to achieve scalable data sharing across domain and infrastructure boundaries. Once established, we feel that it can serve as a solid basis for helping to guide FAIR implementation, in combination with the work of other groups.