The Cross-Domain Interoperability Framework: A Proposed *Lingua Franca* for FAIR Data Reuse (Discussion Draft)

Arofan Gregory, Simon Hodson

15 August 2022 (Draft 0.2)

Contents

Overview

This document provides a summary of the proposed Cross-Domain Interoperability Framework (CDIF). In simple terms, CDIF is a set of recommended best practices for using a coordinated set of domain-agnostic standards – most often as specific subsets or profiles of those standards – to support a core set of functions for cross-domain FAIR reuse.

This document is intended to serve as a basis for work during the "Interoperability for Cross-Domain Research: Machine-Actionability & Scalability" workshop at Schloss Dagstuhl from 28 August to September 2, 2022 (https://www.dagstuhl.de/en/program/calendar/evhp/?semnr=22353).

This idea has emerged over the past several years as a result of several different efforts to identify a practical approach to implementation of the FAIR principles. These have been based on the exploration of a number of different use cases in different domains, but the use cases have not been comprehensive in scope, and much remains to be done to make the work broadly representative. Ultimately, CDIF would become a set of guidelines or best practices for enabling FAIR data sharing, both within and across domains and infrastructures, at a level of specificity which is meaningful to systems developers and implementers. At this point, CDIF is still a proposed framework. This paper intends to document the current thinking, so that further work can be more easily undertaken.

The goal of CDIF is – to the greatest extent possible – to build on standards and models which are already in existence, and which have been widely adopted, or are likely to be widely adopted. CDIF does not represent a new standard itself, but is intended to be a set of guidelines for using existing standards and models in a coordinated way, to ensure a degree of FAIR exchange in as automated a fashion as possible.

Other technical protocols and specifications are needed for the implementation of FAIR, notably the basic protocol stack (currently represented by proposals around the FAIR Digital Object Framework) and the standards and models which are used as standards within specific domains and infrastructures. It is when data and metadata are shared across such boundaries that a *lingua franca* such as CDIF becomes necessary.

This document will describe the benefits of adopting this approach from an implementation perspective, and will describe those standards and models which have been suggested as candidates for the implementation of the full set of FAIR principles at a functional level.

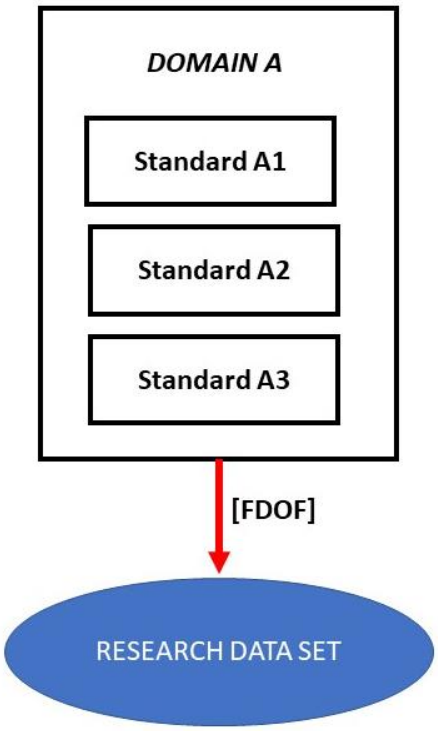
The summary table at the end of the document shows the list of proposed standards, broken out according to functions based on the FAIR principles. URLS to the relevant sites are provided.

The Benefits of a FAIR *Lingua Franca*

At this time, a set of basic protocols for implementing the FAIR principles are being developed as the FAIR Digital Object Framework (FDOF - https://fairdigitalobjectframework.org/). This work describes the way in which FAIR digital objects would be structured and published on the web, such that identification, relevant metadata items and schemas, and other necessary information can be programmatically accessed. CDIF does not operate at this level, but assumes that the FDOF (or a set of protocols for achieving the same functional goals) exists.

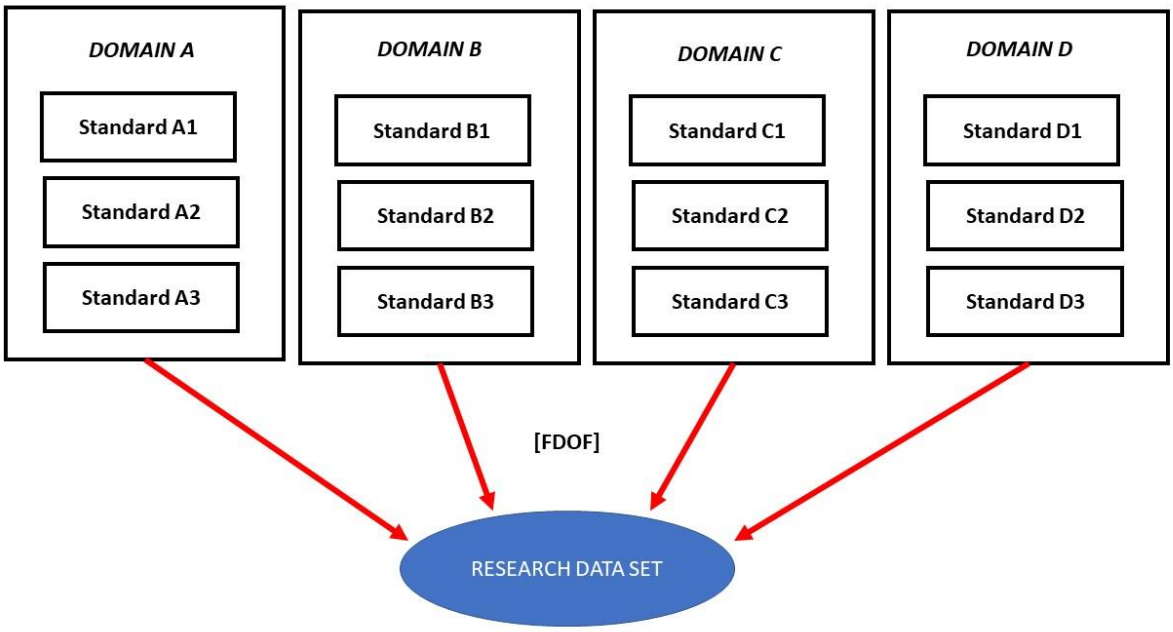
In a scenario involving the exchange of digital objects within a domain, where a recognized set of standards exists, such protocols are sufficient for implementing the FAIR principles. It is often the case, however, that either there is no discrete set of recognized standards within a domain, or the digital objects wanted for reuse are published external to the domain, where unfamiliar standards may be used. In practical terms, this creates a significant challenge for implementers.

In the figure below, we see how the FDOF would enable FAIR reuse within a domain, based on an agreed set of standards:



Here, the research data set can be usefully consumed because the metadata schema referenced by the FDOF are understood by the reusing application: they are agreed domain standards.

In a cross-domain scenario, this becomes problematic:



Here, we see that a much larger number of standards must be understood by the reusing application, because the data making up the research data set is coming not from within a single domain, but from several.

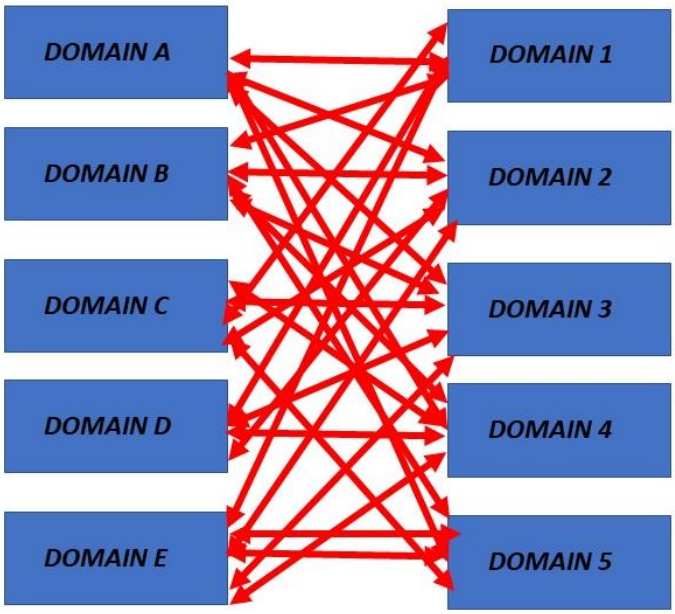
Many widely used integrated data sets today comprise data and metadata produced by a large number of domains, and the demand for integrated, interdisciplinary research data is growing. We see this clearly in the strategic direction of many scientific bodies concerned with research data (*Strategic Research and Innovation Agenda (SRIA) of the European Open Science Cloud (EOSC)*, *Science and Society in Transition: The ISC Action Plan 2022-2024*, etc.).

The wide range of domain standards involved presents a barrier to any organization wishing to assemble an integrated data set, and accounts for why such data sets have traditionally only been produced by a relatively small number of organizations: they are very resource-intensive to implement.

Further, when we consider that each integrated data set will need to support a different selection of domain-specific standards of different types, it becomes clear that it is difficult for technology vendors to provide generic services and applications to meet this need: generic tools, useable across domains, can be built for a set of cross-cutting protocols such as the FDOF, but once you enter the realm of domain standards, each integrated data set will require a different, large selection. The basic tools for data processing exist, but specific tools for simplifying the integration of data – tools which require an inherent understanding of the domain standards in question – do not.

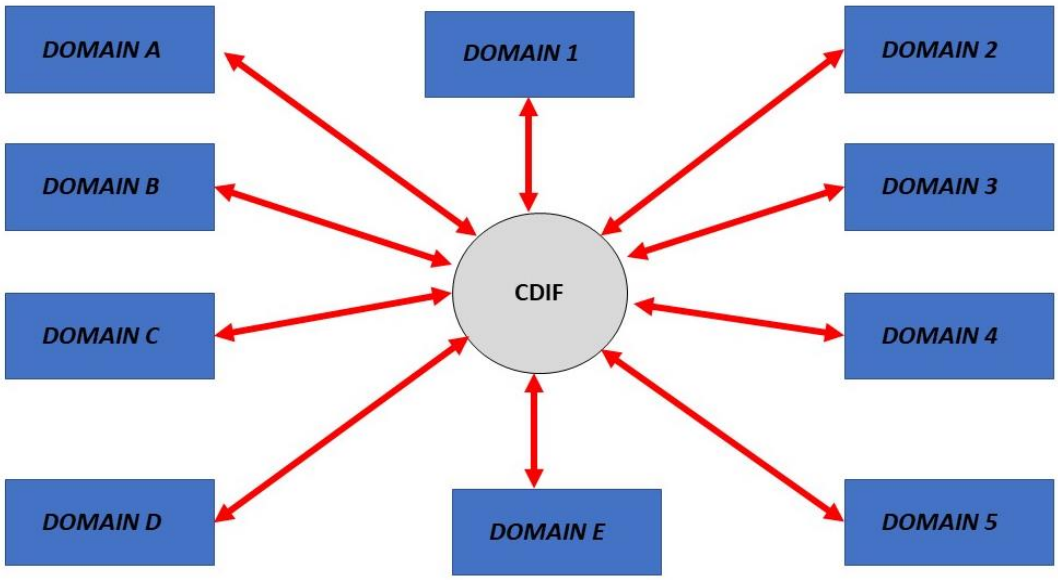
The idea behind a *lingua franca* is to reduce the barriers to integration across infrastructure and between domains by reducing the number of domain-specific standards which require support. The diagram below shows how a multiplicity of direct domain-to-domain integrations operates: each red arrow indicates some set of domain standards, originating at the source of the information, and requiring comprehension by the receiving domain. (Note that this assumes that the FDOF provides

needed information about relevant metadata items and their schemas for each FAIR digital object requiring reuse).



This is a scenario in which each domain must be able to meaningfully process the standards used in every other domain, and it is produces a very large number ("n-to-n") of transformations. (It should be noted that while some of these are semantic in nature, many are not, resulting merely from differences in terminology or in the underlying data and process models built into domain standards).

If a *lingua franca* such as CDIF existed, a single set of standards could achieve the same effect, but require only a transformation from each domain against the common set:



The number of transformations is much smaller ("one-to-n"). This provides – conceptually – a huge benefit: the effort required from any given domain is reduced, and the ability for generic services is enhanced, because vendors can provide services which have an in-built awareness of the *lingua franca* and can be configured by domains to work for them. (This is exactly the scenario which can be seen in other industries involving large-scale exchanges of information between multiple counter-parties, as in supply-chain management, banking, the travel industry, and international trade using UN/EDIFACT, within official statistical reporting at the international level using SDMX, etc.)

In practice, however, there are some considerations which operate to reduce the benefits somewhat. In the first case, not all domains need to use data and/or metadata from every other domain – the patterns of reuse depend very much on the research within each domain, and the kinds of FAIR resources it produces. Another consideration is that transformations between standards are rarely lossless – each step involves a loss of specificity, which can reduce the quality of data.

As a result of these considerations, it may be the case that in scenarios where precision is paramount, or where the volume of reuse is very high, direct domain-to-domain transformations may be desirable. In many of the functions required by FAIR data reuse, however, this scheme is likely to be sufficient to increase the degree of possible automation, and the ease with which data can be reused across domain boundaries. (In the other domains given as examples above, the profusion of hubs and registries operating on this principle is a testament to the fact that this basic notion is sound.)

Ideally, all of the organizations participating in FAIR data reuse would agree to natively support a single set of standards, and avoid any loss whatsoever. This is not a practical proposal, however, due both to existing investments in domain standards, and in the different information needed in different domains to describe FAIR resources. There is no single, universal set of standards at the appropriate level of detail to support the description, use, and management of all possible FAIR resources across all domains. One size does not fit all.

This reasoning may seem obvious, but it is important to bear this basic dynamic in mind when considering what constitutes a *lingua franca* for cross-domain FAIR reuse. The goal is to reduce the burden within domains by lowering the number of "external" (non-domain) standards requiring support, and optimizing the potential efficiency gains through support of automation. It is not the case that a completely automated solution is possible, but it is the case that major gains in efficiency are.

The goal is to identify a minimum of additional required support for external standards to realize a maximum benefit in terms of needed resource expenditure. To do this, CDIF builds on existing investments in data and metadata standards in a minimum set of functional areas – only those required to support FAIR reuse. In those areas where there is already widespread adoption of cross-domain standards (as is the case with Schema.org, PROV, DCAT, etc.) then these standards are preferred. These standards are most often found in those functional areas where there is the lowest degree of domain-specificity, however. The practical identification of the standards which are needed for a *lingua franca* must be based on existing practice, and on the actual patterns of needed FAIR reuse across domain and infrastructural boundaries.

Optimizing CDIF for FAIR Implementers

CDIF consists of a set of recommendations for the coordinated use of existing standards, intended to provide a practical approach to implementation of the FAIR principles. One major question is "which standards will meet the needs of the *lingua franca* described above?" This section describes the criteria by which standards are selected as potential candidates.

(1) *Functionality:* There are several disparate types of information involved in FAIR reuse, supporting a range of different functions. While the FAIR principles themselves suggest a functional breakdown, they are not necessarily complete or sufficiently detailed from an implementer's perspective. One or more standards are needed to support each functional area (a proposed set is given below).

(2) *Applicability across Domains/Infrastructures:* There are many standards which support needed functions but which are specific to the domains for which they were developed. Many standards combine domain-specific and domain-agnostic aspects. The nature of a cross-domain *lingua franca* is that it must, to the greatest possible extent, be domain-independent.

(3) *Widespread Adoption:* If possible, the standards recommended by CDIF should be in current use within the research data community. This not only illustrates their fitness for purpose, but also lowers barriers to adoption.

(4) *Adoptability:* Standards should not place an undue burden on their adopters. Many aspects of FAIR reuse are quite complex, and require complex models to solve (especially Interoperability), but that does not necessarily translate into a standard which systems developers have difficulty working with. The benefits realized by cross-domain FAIR reuse must be sufficient to justify the expenditure on the implementation of new standards needed to support it.

(5) *Alignment with Existing Domain Standards:* The standards selected for CDIF should be as similar as possible to domain standards which are already in use, and should build on existing domain practice. Existing investments in standards for data and metadata reuse may in some cases be easy transformed into less-domain-specific standards, and this can lower the barriers to adoption.

(6) Open/Non-Proprietary: The standards chosen for use in CDIF must be open – that is, free to use – and must be non-proprietary. At the same time, their terms of use must not prohibit proprietary implementation of them, as one of the communities of users will be the providers of commercial services and software applications for enabling FAIR reuse. CDIF should not present a barrier for any potential users as regards its licensing.

It should be noted that some aspects of FAIR – especially vocabularies and other, similar concept systems – are often deeply specific to the domains from which they originate. The standards which

support the use of such information and which meet the above criteria are often those which standardize the structural aspects of such information, rather than the scientific content or semantics.

A good example of this is W3C's Simple Knowledge Organization System (SKOS). While standardizing the structural aspects of concept systems, it does not standardize the concepts themselves – just the way in which they are described. This can be understood as a "meta" approach – that is, a modeling approach based on abstraction of the domain semantics. Many (but not all) of the CDIF standards are similar in this way, as it is often a key aspect of why they can perform a useful function independent of the specific work within particular a particular domain.

One consequence of these criteria is that – while some standards are intended for use with specific technologies (RDF, XML, etc.) - there is no requirement for all FAIR reuse to be driven by a single technology choice. In practical terms, the existing technology culture within a domain is an important part of the foundation on which cross-domain FAIR reuse builds. CDIF should be as forgiving in terms of technology requirements as possible, as often there are good reasons within domains for specific technology use. Such choices must be respected to the greatest extent possible, and can be considered an important part of the "adoptability" of standards for the purposes of CDIF.

Functional Areas for FAIR Implementation and Candidate Standards

General Observations

The functional groupings of candidate standard given here are as much up for discussion as any of the proposed standards. While the FAIR principles suggest a basic functional breakdown, there are additional requirements for the implementation of FAIR reuse which must be recognized. The classification of functions needed to support FAIR in a cross-domain setting will need to be further explored and agreed – what is reflected here is only one of the proposed ways of describing needed functions, but is in no way to be taken as agreed at this time.

Where possible, the functional breakdown implied by the FAIR principles has been followed, but it is recognized that there are other functions which will need support if cross-domain FAIR reuse is to be made practicable.

It is anticipated in almost every case that CDIF would recommend the use of some subset of a particular standard, rather than implementation of the entire thing. Many standards are duplicative in the functions they potentially support, or provide a broad range of uses, not all of which are important for the purposes of CDIF. The goal is to support the range of needed functions with specified subsets of the available, recommended standards.

This section is not comprehensive, but really provides a starting point for further investigation: the standards mentioned are those which have come up in exploration of use cases to date, but will need to be validated and extended through further, more systematic exploration across a more complete and representative set of domain use cases. They are intended to be suggestive, and to provide a useful starting point for further work.

Findability

This functional area addresses identification, search and discovery, indexing and cataloguing, and also the assessment of data (that is, determining fitness for purpose). This last aspect of Findability is not

directly addressed by this category in the FAIR principles, but requires links to the same set of detailed metadata needed for Interoperability – assessment of data is thus in some sense split between Findability standards and those needed to support Interoperability.

Identifiers

There are many widely accepted standards for identification which meet the criteria given above. For data and metadata, Digital Object Identifiers (DOIs) and their use in standards such as DataCite are common.

For identification of researchers, ORCID seems to be the standard which is most widely used, and presents an obvious choice. Other related identifiers such as the Research Organization Registry (ROR – supplanting GRID), and Crossref are used in combination with DataCite and ORCID in many systems, and should be further explored for universal identification of these types of entities.

It should be noted that URIs in and of themselves are not persistent identifiers (PIDs), and thus may not be sufficient for the purposes of FAIR reuse. While URIs can function as PIDs under some circumstances, the simple assignment of a URL to a resource is not enough. (Consider how DOIs can be predictably expressed as URLs as an example of how these systems can be used in a complementary way to provide PIDs.)

Search and Discovery

In this area, the dominant standard is Schema.org. There is some discussion as to how far the use of Schema.org can be taken as the basis for cataloguing and indexing of FAIR resources, and this is still an open question. While the perception in some communities is that the Data Catalog Vocabulary (DCAT) from W3C is duplicative of the support provided by Schema.org, there are differences in how the two standard are implemented and used, and often both are supported within the same system. (To recognize this, we separate out the cataloguing and registration functions of Findability from the search and discovery functions, although this distinction may be an artificial one.) Schema.org is designed to support extensions for particular communities, and such sets of extensions should be taken into account when the use of Schema.org for CDIF is considered.

Cataloguing and Registration

DCAT seems to be the best candidate in this area, with the consideration that it may in some cases duplicate support provided through the use of Schema.org. It is important to consider that DCAT is designed to support specific profiles, and that CDIF will need to address the profile of DCAT to be used, in the same way that extensions to Schema.org need to be further explored.

Packaging and Assessment of Fitness for Purpose

For both DCAT and Schema.org, description of (or links to descriptions of) specific types of observations/measurements ("variables") contained with data sets and provided by data services and sources is a topic requiring some exploration. DCAT can function as a packaging mechanism for some types of resources, giving the potential re-user an entry point into the set of metadata and data for a particular resource. A DCAT entry for a data source might point to a description of the structure of the data found, the information about how the measurements from that source must be described in combination with other values (see "Fully Described Observations" below), detailed descriptions of the variables (including controlled vocabularies used for the representation of values and their definitions),

and so on. The use of standards such as DCAT as a packaging mechanism should be considered – it is an approach seen in some prototypes of FAIR implementations.

The links between description of high-level sources and sets of data, and the detailed description of the observations and measurements they provide is critical for the purposes of assessing the fitness for purpose of data. Often, variables concerning the time (observation period) and geography of data are used to immediately assess whether the data described can be reused. The frequency with which enumerated values appear across the breadth of a data set (or query result set) can also be very useful in assessing fitness for purpose. This functionality depends on metadata described using standards proposed here as candidates for the support of Interoperability, and the connection between search/discovery, cataloguing/registration, and assessment of fitness for purpose should be further explored.

Dublin Core and Related Metadata Schema for Discovery/Cataloguing

It is worth noting that there is a profusion of metadata standards intended for use in cataloguing and discovery based on Dublin Core (indeed, both DCAT and Schema.org use parts of Dublin Core where appropriate). While many of these metadata schema are domain-specific, others are not. Where they fail to meet the CDIF criteria given above is in the extent of their adoption: Dublin Core is pervasive, but it is used in such a wide variety of ways that no single approach can be easily identified as ubiquitous. Still, we must consider Dublin Core and its many implementations when we look at the ways in which Findability can be supported in CDIF.

Accessibility

Accessibility is an area of FAIR which has not received as much attention as some others, apart from the questions around identification (see above). This is in part because it implies (in A1.2) a degree of human involvement which is very much driven by both legal and institutional concerns. Permitted access to sensitive data depends very much on the data in question, and these considerations drive authorization and access processes.

It should be noted, however, that Accessibility is also concerned with the persistence of metadata over time (even when data cannot be persisted) and also with a description of the required processes. We address these functions under the Interoperability and Reusability section, below, insofar as they are currently supported by the standards we have considered.

The extent to which Accessibility can be usefully automated remains an open question, but there are some potential standards which might deserve consideration, whether as a source of documentary information presented in a regular form, or as the basis of automation.

These standards are both produced by the W3C: the Open Digital Rights Language (ODRL) and the Data Privacy Vocabulary (DPV).

The Open Digital Rights Language is a mature standard as used in some other domains (for example, in the distribution of music recordings on the internet). Its use for the control of access to FAIR digital objects has not yet been fully explored, but suggests itself as a useful example in what is really a new frontier for data sharing. Details can be found at the W3C Community Group (https://www.w3.org/community/odrl/).

The DPVCG is a less-mature standard, but one which has more obvious applicability to FAIR reuse of data. It provides a formalization of access conditions to support their automation. Some FAIR implementations are beginning to explore the use of the standard. More information can be found at the W3C Community Group (https://www.w3.org/community/dpvcg/).

The extent to which CDIF can meaningfully provide a best practice in this area is still under consideration – it may be too early to attempt to describe a best practice in the use of such standards, but it something to be examined as thee CDIF develops. The functionality around data access is common to all FAIR reuse scenarios, and standard mechanisms for its automation are needed.

Interoperability and Reusability

Interoperability and Reusability are aspects of FAIR which are distinct in their stated principles, but which in practical terms rely on a common set of standards and a common pool of information which is often found in systems which manage both aspects. The metadata needed for reuse is often the same as the metadata needed for data integration and harmonization. Because of this, both of these functional areas will be addressed in a single section in this document.

From the perspective of the information needed to support these functions, however, the range is broad and the degree of complexity is higher than found in some other areas of FAIR reuse. We have broken the discussion of candidate standards here into several areas, to better reflect this breadth and complexity.

Structural Metadata

Many standards describe the structure of data sets, and the data themselves, but it generally the case that these are tied to a specific domain or set of domains, or to a particular type of data structure. There are several such standards which can be configured with semantics provided by external vocabularies (for example, DDI Codebook and DDI Lifecycle work this way for unit-record data in the Social, Behavioral, and Economic sciences; SDMX and the Data Cube Vocabulary do this for aggregate data, NetCDF does this for large array-oriented data sets, OMOP CDM does this for clinical data, SOSA/SSN for sensor data, etc.).

Other more fully configurable data description standards such as W3C's CSV on the Web and the Metadata Vocabulary for Tabular Data also exist, but they lack the ability to attach semantics to the data in a way that makes them easy to programmatically integrate. (Any table or a CSV file can be described, but these formats are very open as to how they encode semantics and that presents problems for machine-actionability when differently structured data are integrated.)

To meet the criteria described above for CDIF, it is necessary to have a standard which both embraces the ability to use non-domain-specific semantics and describes the very wide range of data structures (data steams, unit-records, event data, multi-dimensional cubes, relational/SQL data, no-SQL data, arrays, geographical coordinate data, etc.).

Given the very large variety of data types, and the tendency within specific domains to prefer a small number of data types suited to the data produced by that domain, this is a challenge.

The one standard which seems designed to address these types of requirements is the DDI Cross-Domain Integration (DDI-CDI) standard. This is a model which can be used either as a stand-alone description of data, or in combination with other domain standards. It can use concepts and semantic

definitions from external controlled vocabularies of different types (SKOS is typical, as are domain ontologies and classifications). Its intended purpose is to show how different datums within any data structure can be assigned roles, making them susceptible to transformation into other data structures which re-assign those roles when the data is re-arranged.

To give a simple example, a field in the first column of a unit-record table (the unit identifier) plays a role as an ID. Taken in combination with the concepts for each column in the table (each variable), these form the basis for identifying each value in the unit record (the "row") and thus the data set (unit identifiers are unique within the data set). When this same data is re-cast as a multi-dimensional cube, several of the variables (the columns) – including the unit identifier – will combine to form identifying keys for each of the values. Since the record structure – the row – is no longer present, the roles of different concepts as identifiers has changed – the values of variables acting as components of a multi-dimensional key are needed to enhance the concepts defining the variables (that is, the column headers). Such patterns are non-obvious, but can be formalized.

DDI-CDI describes these patterns for unit-record ("wide") data, event/streaming data, multi-dimensional data, relational data, and no-SQL/key-value data. The purpose is to allow them to be re-structured programmatically, based on the patterns described.

While initially designed to support data integration of DDI-described unit record data with other domain-external data sources, DDI-CDI can be combined with any other domain standard which describes data structures of the types listed above, regardless of whether DDI Codebook or DDI Lifecycle are used. Further, because it is a model-driven standard which can be implemented in a wide range of syntaxes, it is easy to use with existing domain-specific standards which are based on a particular syntax implementation (e.g., RDF, XML, SQL, etc.), since it is technology agnostic.

The sufficiency of DDI-CDI to serve as part of the CDIF *lingua franca* will need to be explored – it is not designed to cover all of the needed data types (for example, coordinate data for describing geographies) and so may need to be extended for this purpose. It does appear to be the best single candidate for this role within the CDIF, however. Ultimately, it may be necessary to support more than one standard, but the number should be as small as possible.

Semantics

Perhaps the most difficult challenge which faces the establishment of a FAIR *lingua franca* is that of accommodating the huge variety of semantics used to describe data across domains and infrastructures. While progress has been made within (and in some cases, between) many domains in the development of ontologies and formal classifications systems and thesauri, these tend not to be used or understood in all of the domains for which the data they describe is relevant. Data reuse – that is, integration and harmonization – relies on our ability to understand all of the needed semantics, however.

If our description of data structures is able to support the use of whatever semantics are used (as a domain ontology, classification, or other controlled vocabulary), and indicates what role these conceptual definitions play in relation to the data, then we are able to effectively employ whatever mappings or concordances might exist between semantic systems in a programmatic fashion. Failing this we can at least minimize the work required for manual integration.

DDI-CDI and some other structural data descriptions give us the ability to understand the roles semantic concepts play in relation to the data, but we must them be able to process both the concept systems and the mappings/concordances.

For the encoding of concept systems and ontologies, we have many widely adopted standards such as the Web Ontology Language (OWL) and RDF Schema (RDF-S). The Simple Knowledge Organization System (SKOS) and extensions (such as XKOS, for formal classifications) are also widely used. Other models such as the Generic Statistical Information Model (GSIM) and DDI-CDI let us describe classifications according to the Neufchatel model, which is used within many domains which produce and maintain popular classifications. All of these may be useful, especially SKOS and its extensions, as these seem to be very widely adopted.

For describing mappings and concordances, models such as Neufchatel are useful for formal statistical classifications, and standards such as the Simple Schema for Sharing Ontology Mappings (SSSOM - https://mapping-commons.github.io/sssom/spec/) can be employed.

In some specific areas, very important concepts in data integration deserve to be highlighted, and they come with appropriate standards which demand our consideration. One of these areas is the expression and harmonization of units of measure. Several different models and standards are being considered by the Digital Representation of Units of Measure (DRUM) group, hosted by CODATA, and their recommendations should be taken into consideration. Similar models may exist for geographical data, and these should likewise be examined. The importance of these specific areas to data integration demands that they be fully addressed by CDIF.

The question remains as to how fully automated semantic mapping can be, given current approaches. This may be an area where more human intervention is required, but by the same token the many advances in this area which have taken place in the recent past also demonstrate that efficiency gains can be significant when compared with the fully manual approaches commonly used today.

Fully-Described Observations

In cross-domain FAIR reuse, the need for contextual information about data is even greater than it is within domains, because much of the implicit domain knowledge which is often taken for granted may be absent – researchers cannot be expected to be deeply familiar with the literature of disciplines beyond their own, etc.

Here, we address this requirement from two perspectives: that of data provenance and processing, and in respect of the cluster of values which need to be taken together to support accurate data reuse of a measurement or observation.

Process and Provenance

There is one standard which suggests itself for describing provenance, and that is the W3C PROV Ontology. It is widely adopted and used, and has a very generic approach to issues in this area. It is, perhaps, too flexible – it relies on specific configurations (e.g., PROV-One, RO Crates, etc.) for implementation, and this may present a barrier to identifying an appropriate configuration for use in CDIF. Many other standards align with PROV (for example, DDI-CDI uses relevant PROV artefacts to connect the data and metadata it describes with specific processes).

For describing processing – that is, the executed functions which manipulate data – we have different standards such as the Validation and Transformation Language (VTL) – part of the SDMX family of standards – and the Structured Data Transformation Language (SDTL) now maintained by the DDI Alliance and the C2Metadata project. There are a large number of proprietary or system-specific languages as well. Processes described according to these languages can be attached to higher-level processes described using PROV, to provide the needed detail about data provenance at a detailed level. (This approach is used in many domain-specific standards.)

Clusters of Values to Provide Context

In many cases, having an observation by itself is insufficient to support accurate reuse of the data. Several standards are being developed which address the need to understand how a set of related values are connected to an observation or measurement. One of these is the InteroperAble Descriptions of Observable Property Terminology (I-ADOPT) from RDA. Another is the Observations and Measurements (O&M) work from OGC. Whether either or both of these is most appropriate for the needs of CDIF must be considered, but these approaches do suggest that the set of values needed to understand and accurately reuse measurements and observations does have a standard descriptions for CDIF to potentially leverage.

Resource Management

One area which is not directly addressed by the FAIR principles, but which is implicit in them, is the need for resource management at all levels. Support for FAIR reuse – and especially cross-domain FAIR reuse, which may be more demanding in terms of metadata – requires that the costs and benefits be understood, and that the relationship of FAIR resources to research outputs be described.

This may be beyond the remit of CDIF, but it is worth considering that in some domains (the environmental domain in Europe is a good example – ENVRI-FAIR) these issues are addressed through the use of the Common European Research Information Format (CERIF). The need for this type of a standard within CDIF should be further examined.

Open Questions/Known Issues

Many issues remain open in terms of how best to compile a set of recommendations for cross-domain FAIR reuse. Several of these are described below. (Note that other open questions appearing in the sections above are not repeated here.)

How Much Choice Should Be Allowed?

In making recommendations for a limited set of domain-agnostic standards, there is a balance between the additional resources needed to implement new standards and the benefits of their adoption. CDIF intends to keep the barriers to adoption as low as possible, but the demands of FAIR reuse argue for a higher degree of conformance to a smaller set of standards.

CDIF is no more than a set of recommendations, which will in reality act as a point of reference for implementers, who will choose to accept them or not. This raises some questions: What is the most useful form for such a set of recommendations? What criteria do we use to find the right balance? Are there other examples of best practice guidelines which could serve as a model?

Domain Classification

In the development of the FAIR Implementation Profiles (FIPs) at GO FAIR, a classification of domains was used for communities to describe themselves. While not perhaps a direct requirement of the CDIF guidelines, an agreed classification (or classifications) for describing different domains will be very useful in practical terms, as a basis for organizing catalogs and registries, and for automating data integration across domain and infrastructure boundaries.

How can these needs be addressed? Can CDIF usefully recommend anything in terms of its own requirements in this area? There are many examples of domain classifications, used by research publications and libraries, funders, and others. What is available to support automation of cross-domain FAIR reuse?

Generic Services/Services Typology/Services Registry

One the major benefits of having a *lingua franca* for cross-domain FAIR reuse is the possibility of having a class of domain-agnostic services and applications which can be employed to support implementation in the real world. The complexity of across-the-board support for the automation of data reuse, as envisioned here, suggests that many organizations will not be able to develop and deploy the full set of needed services by themselves.

What are the requirements for developing useful and scalable services and applications to support cross-domain FAIR? How could these be organized so that, having determined the existence of FAIR resources and their location (a function presumably provided by FAIR Data Points and the FDOF), the available services can be determined?

Do we need to think about the development of a services typology which can be used to describe the capabilities of different providers of FAIR resources which are available for cross-domain use, and compliant with CDIF? Is a registry needed to associate CDIF-supporting organizations and the services they provide? Could we use DCAT (or a similar standard) to provide this support?

FAIR Evaluation

If CDIF gives us a set of best practice recommendations, how do these fit into the establishment of FAIR metrics? CDIF focuses on cross-domain and cross-infrastructure FAIR reuse, which will only be a part of the overall FAIR evaluation picture. Some functions supported by CDIF may not be needed for the FAIR resources produced by some domains.

The idea that FAIR metrics must be sensitive to the specific needs of domains is already gaining currency in projects such as WorldFAIR – how does CDIF fit into the overall approach to FAIR metrics, and the need to define appropriate conformance on the part of particular domains or classes of FAIR resources?

Ideally, FAIR conformance within a domain should provide an easy path toward compliance with cross-domain FAIR as expressed through the CDIF guidelines.

Conclusions

Practical realization of cross-domain and cross-infrastructure FAIR reuse demands the establishment of a *lingua franca* which can limit the number of different standards requiring support. Even with a framework such as CDIF, a dozen different standards may need to be recommended to cover the full set

of anticipated FAIR functionality. While implementation of FAIR will in many cases be incremental, placing lower demands on adopters at any single point in time, an agreed path forward is still needed.

It is important to remember that CDIF relies on the work going on in other areas around FAIR implementation, notably on the FAIR Digital Object Framework. Without these basic protocols for FAIR reuse, CDIF cannot function. The FDOF development is still ongoing, and there remains a need for coordination with those protocols and standards as they are completed.

This document attempts to summarize the business case for the development of CDIF as a *lingua franca*, and suggests some of the standards which might serve as candidates for inclusion. It is by no means comprehensive, but does include the standards which have come up during many different workshops and discussions on the topic. The details of work to this point are not described here – this document provides a notional checklist of candidate standards, and describes how CDIF has been envisioned, but is intended only to serve as the basis for further work.

The FAIR principles have triggered a keen interest in data reuse which is incredibly timely, given the needs of the research community today. CDIF attempts to formulate a partial answer not to the question of *why* FAIR is a good idea, but the question of *how* it can be achieved. While the practical realization of any such vision is always problematic, the benefits of establishing a shared vision, and an agreed point of reference, are valuable in their own right in helping to drive convergence.

Summary Table of Proposed Standards by Function

Function	Standards	Notes	Links
Findability			
- Identification	Digital Object Identifiers (DOIs), DataCite, ORCID, Research Organization Registry (ROR), Crossref	URLs do not guarantee persistence, but identifiers should be associated with a resolution mechanism	DOI: https://www.doi.org/ DataCite: https://datacite.org/ ORCID: https://orcid.org/ ROR: https://ror.org/ Crossref: https://www.crossref.org/
- Search	Schema.org, Dublin Core	Dublin Core (DC) includes derivatives/ implementations; Schema.org subsets must be considered; well-adopted	Schema.org: https://schema.org/ DC: https://www.dublincore.org/
- Cataloging	Data Catalog Vocabulary (DCAT), Dublin Core	Dublin Core (DC) includes derivatives/ implementations; use of DCAT as a "packaging" standard should be considered; well-adopted	DCAT: https://www.w3.org/TR/vocab-dcat-2/, https://www.w3.org/TR/vocab-dcat-3/ DC: https://www.dublincore.org/
Accessibility	Open Digital Rights Language (ODRL), Data Privacy Vocabulary (DPV)	Still early in development/adoption within the research data arena	ODRL: https://www.w3.org/community/odrl/ DPV: https://www.w3.org/community/dpvcg/
Interoperability/ Reusability			
- Structural	DDI Cross Domain Integration (DDI-CDI), Statistical Data and Metadata Exchange (SDMX), DataCube Vocabulary (QB), CSV	Many of these (all except DDI-CDI) are limited in the coverage of structural types or weak in attaching semantics; all are "meta-standards," which is an	DDI-CDI: https://ddialliance.org/Specification/ddi-cdi SDMX: https://sdmx.org/ QB: https://www.w3.org/TR/vocab-data-cube/ CSV on the Web: https://www.w3.org/TR/tabular-data-primer/ SOSA/SSN: https://www.w3.org/TR/vocab-ssn/

	on the Web, Semantic Sensor network Ontology (SOSA/SSN)	important criteria; varying levels of adoption	
- Semantic	Web Ontology Language (OWL), RDF Schema (RDF-S), Simple Knowledge Organization System (SKOS), Extended Knowledge Organization System (XKOS), Generic Statistical Information Model (GSIM), Simple Schema for Sharing Ontology mappings (SSSOM)	These standards serve many different functions – some are meta-descriptions of controlled vocabularies, some are lower-level ontology languages, others are for expressing mappings and correspondences; the work of the CODATA DRUM group should be considered for units of measure	OWL: https://www.w3.org/OWL/ RDF-S: https://www.w3.org/TR/rdf-schema/ SKOS: https://www.w3.org/TR/2009/REC-skos-reference-20090818/ XKOS: https://ddialliance.org/Specification/RDF/XKOS GSIM: https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model SSSOM: https://mapping-commons.github.io/sssom/
- Provenance/ Process	PROV, Structured Data Transformation Language (SDTL), Validation and Transformation Language (VTL)	PROV uses profiles (PROV-One, RO Crates, etc.) in order to be useful	PROV: https://www.w3.org/TR/prov-overview/ SDTL: http://c2metadata.gitlab.io/sdtl-docs/master/ VTL: https://sdmx.org/?page_id=5096
- Full Description of Observations	Interoperable Descriptions of Observable Property Terminology (I-ADOPT), Observations and Measurements (O&M)	Relatively new standards; O&M is very diverse	I-ADOPT: https://i-adopt.github.io/index.html O&M: https://en.wikipedia.org/wiki/Observations_and_Measurements
Resource Management	Common European Research Information Format (CERIF)	Need to establish exchangeable set of info – look at ENVRI-FAIR use case	CERIF: https://cordis.europa.eu/article/id/8260-cerif-common-european-research-information-format