Open Standards and the FAIR Principles: Three Use Cases

Simon Cox, Arofan Gregory, Simon Hodson, Steve McEachern, Bill Michener, Joachim Wackerow

September 2019

I. Introduction

The "Interoperability of Metadata Standards in Cross-Domain Science, Health, and Social Science Applications" workshop was held at Schloss Dagstuhl – Leibniz Center for Informatics (https://www.dagstuhl.de/en/) in October 2018 to examine three applied interdisciplinary use cases in light of the FAIR principles (the principles that research outputs, and particularly data, should be Findable, Accessible, Interoperable and Reusable).[1] High-level principles such as FAIR are easy to agree with, but implementing them in a practical sense can be challenging. In order to better understand these challenges, the workshop examined three pilot projects, chosen for their need to use data across domain boundaries, and explored in detail the problems faced by those needing to integrate data coming from disparate sources.

In the remainder of this paper, we first describe the FAIR principles in relation to supporting interdisciplinary science. Second, we introduce the three case studies—infectious diseases, resilient cities and disaster risk reduction. Each case study was selected because it requires data from many disparate domains, the relevant data adhere to many different data types and formats and are described by widely varying metadata standards, and the data apply to a broad range of spatial and temporal scales. Third, we describe many of the challenges experienced by researchers and decision-makers associated with the three case studies in relation to each of the FAIR principles. Fourth, and in part based on the assessment of the three case studies, we examine the present status of metadata standards and their potential applicability to resolving many of the interoperability challenges that hinder interdisciplinary science. Finally, we conclude by proposing a research roadmap whereby focused follow-up workshops and research projects can advance state-of-the-art solutions to the challenges identified.

II. The FAIR Data Principles

The FAIR Data Principles were published in 2016 in the *Scientific Data* journal (from Nature Research). The acronym stands for "Findability," "Accessibility," "Interoperability," and "Reuse." They present 15 principles for how to make data agree with each of the terms, accompanied by 14 metrics across these categories as a way of assessing specific data sets/sources. (We will not repeat these here but will refer to them by the designations given in the 2016 article as relevant). The current state of development of the FAIR Data Principles can be found at the GO FAIR site (https://www.go-fair.org/fair-principles/).

The FAIR Data Principles have been well-received across many different communities, and have resulted in the formation of different groups, working both within and across domains to help individuals and organizations understand what it means to make their data more open and shareable. (Perhaps the

[1] Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 (2016), https://doi.org/10.1038/sdata.2016.18

single best starting point for looking into the adoption and popularity of FAIR is the GO FAIR website's description of implementation networks: https://www.go-fair.org/implementation-networks/overview/.)

The FAIR principles are used here as a cross-cutting framework for describing the findings as each of the chosen use cases was examined. Not every FAIR principle was addressed in each Pilot Case Study – while participants were aware of FAIR, it was not used as a means of organizing the workshop, but merely presented as one perspective on the issues facing each group in terms of their particular data challenges.

III. The Pilot Case Studies

Three pilot case studies were chosen because of their relevance to real-world challenging problems and because they all required data from multiple domains including the geophysical sciences, biological sciences, engineering and the social sciences. Furthermore, requisite data adhere to a wide variety of data types and formats, are described by very different metadata schema, and differ markedly with respect to discoverability, accessibility and usability.

A. Infectious Disease Data Observatory (IDDO)

IDDO (https://www.iddo.org/) is a collaboration housed at the Centre for Tropical Medicine and Global Health at the University of Oxford, where researchers are able to share data on infectious disease. Based on the approach used by the Worldwide Antimalarial Resistance Network (WWARN), it also addresses other infectious diseases, notably Ebola, visceral leishmaniasis, and schistosomiasis, as well as considering antimicrobial resistance and the quality of medicine.

IDDO provides a researcher-driven portal. Data can be both accessed and provided by researchers, and research findings can be published. Tools and resources are made available with the goal of promoting a more standardized publication of data for reuse. Many of the tools provided are those developed by WWARN.

IDDO organizes its data into a number of broad categories which correspond to major topics within the research around infectious disease, some of which require data which go beyond what is typically used in public health research and epidemiology. One example of this is quality of medicine: problems such as the sale of illegal counterfeit drugs are potentially significant, but have very different data requirements than more traditional "public health research" topics.

B. Resilient Cities

Resilience.io (https://resilience.io/) is an agent-based systems modeling platform taking into account the human, ecological, and economic aspects of cities, and examining how new and innovative approaches can be made to solving the problems they face in the modern world. Led by Resilience Brokers (an offshoot of the Ecological Sequestration Trust), there are 24 partners including many research funders and institutes as well as organization addressing specific aspects of the problem space

(poverty, ecological issues, etc.) They have a number of case studies ongoing in Africa, the UK, and elsewhere in the world, using the Resilience.io model to support good policy and informed interventions in urban settings.

An important aspect of the innovations promoted by the use of a systems model is the need for data which traditionally have not been combined in examining and formulating solutions to the problems being addressed. This demand for data is fundamentally cross-cutting in terms of domains and disciplines, and it highlights a large number of the challenges faced by all forms of cross-domain research.

Even within the way in which resilience is understood and applied, there are different forms which present different demands in terms of what data are significant. Resilience is often understood as how best to withstand systemic shocks, but it also concerns itself with adaptive and transformational resilience. These are perhaps longer-term considerations, looking at somewhat different factors and ultimately a potentially broader set of actors. With these come new and different data requirements.

C. The Sendai Framework for Disaster Risk Reduction

The Sendai Framework (https://www.unisdr.org/we/coordinate/sendai-framework) is a United Nations accord aimed at promoting good policy for addressing disasters of all types at a national and international level. It spans infectious disease, ecological disaster, terrorism, and a host of other problems. Although it has backing at the highest level, the challenges it faces are many.

Prominent among these is the need for detailed data on a broad range of topics. Although the UN produces the Sustainable Development Goals Indicators (https://unstats.un.org/sdgs/), this is a high-level aggregate data set which lacks sufficient detail to inform policy at all needed levels. While governments collect much of the needed data, there are challenges integrating data within and between national governments, and other data which exist only in the academic or commercial sectors. Challenges are not only restricted to data as such – the classifications for important elements such as risk measurement are non-standard at the national level. Sendai provides an excellent use case for examining the problems of wide-reaching and large-scale data integration across both domain and organizational boundaries.

IV. Adherence to the FAIR Principles

In this section, we identify many of the challenges experienced by researchers, decision-makers and others who were working in the three pilot use cases. Findings are associated with each of the four classes of FAIR Principles.

A. Findability

The ability to discover the existence of data is a fundamental aspect of data sharing and reuse. Within domains, the existing body of data is often navigated by researchers in reference to publications within the domain. This is problematic when cross-domain use of data is considered: researchers will not always be familiar with the literature in other domains, even when they offer useful sources of data.

Data catalogs, repositories, and portals are common mechanisms for finding data, but these are also often deployed within a domain, and are best known to and best suited for those within a specific research community. Data description and identification are also core aspects of the FAIR principles relating to discoverability of data. In each case, issues were encountered by the pilot projects.

The IDDO case study found that in some of the domains which provide useful data, repositories or other searchable data sources did not exist, or were not well-established. (This is seen as a failure in terms of the F4 FAIR principle regarding searchable resources.) Further, in almost all cases the identifiers assigned to data were not persistent ones. (Both the FAIR F1 and F3 principles address identification.) Only those sources linked to indexed publications were found to have persistent identifiers (typically DOIs).[2]

In the Sendai case study, the problems were seen as more fundamental. Although some sources are searchable and well-established, such as the Sustainable Development Goal Indicators, they can also be very sparse, and finding data can be frustrating, especially when coverage across a large number of countries is desired. In many cases data do exist, but not within the governmental sphere. Rather, they would be known to academic researchers or those in the corporate sector. The practical ability to search for data in sources which are not typically used by those involved in government agencies concerned with risk management is extremely limited. Identifiers were seen as a secondary issue, although one that is clearly important.

For the Sendai use case, the ability to discover relevant data became a focus of the workshop, with an emphasis placed on how contemporary technologies facilitating search (the term "data science" was used) could be better employed by those concerned with policy making in support of the Sendai Framework. Major improvements in terms of each of the relevant FAIR principles could be anticipated if the envisioned approaches could be realized. [ADD: Reference to the Sendai paper here.]

A small group looked at existing cross-cutting standards for data discovery – notably the DCAT vocabulary – and how it could intersect with more domain-specific standards to enhance the Findability of data. This work did not produce a final result but provided a promising initial look at how standards could be combined to benefit those searching for data across domain boundaries. [ADD: Reference to the DCAT-DDI profile work here.]

One aspect of the "Findability" of data which surfaced during the workshop, but which does not feature in the FAIR principles as such, was the idea of the "assessibility" of data (https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf). Metadata are always important for determining whether a data source will be truly useful given a particular research question – FAIR recognizes this in terms of identifiers (F3). Some portion of the metadata is also necessary to determine whether data is actually fit for purpose, however – "assessibility" became a topic of some discussion during the workshop as a consequence, even though detailed metadata is typically seen as more of a concern for Interoperability in discussions of FAIR.

[2] Pisani, Elizabeth; Ghataure, Amrita; Merson, Laura (2018): Data sharing in public health emergencies: A study of current policies, practices and infrastructure supporting the sharing of data to prevent and respond to epidemic and pandemic threats. figshare. Journal contribution. https://doi.org/10.6084/m9.figshare.5897608.v1

B. Accessibility

Access to data is clearly a major challenge across domain boundaries. Within specific domains (such as the social sciences and health sciences) standards, mechanisms, and best practices exist for protecting the confidentiality of respondents and patients while still providing access to their data. In other areas of research, data are confidential not because they provide access to information about the subjects of study, but because they are a source of competitive advantage in terms of career or the marketplace. Some industries view data as a saleable resource. In some domains metadata is not seen as confidential, while in others it is. Across domains generally there is little or no coordination around data access.

On the technology side, open standards are needed to facilitate accessibility. There are some existing generic standards for supporting the technical aspects of data access - the technical problem is well understood. Services such as those provided for Digital Object identifiers (DOIs) are an example of how identification of data is a tractable problem (FAIR A1). Standards such as the Extensible Access Control Language (XACL) for describing access policies have been with us for many years. There are many other examples. The use of such standards and technologies is uneven across domains, however. Even in the best case, using a common mechanism cannot solve problems stemming from the nature of restrictions to data driven by domain-specific considerations.

The FAIR Data Principles recognize this reality, by setting criteria for protocols and metadata accessibility, but without crossing into the domain-specific space which presents in some ways a greater challenge.

Some of these issues were highlighted by the IDDO pilot project discussions at the Dagstuhl workshop. Much of the lower-level clinical data are not publicly available because of confidentiality concerns. Often, aggregate data – which are non-disclosive – are buried in research publications, making them difficult or impractical for researchers to access for reuse or replication purposes. While the data exist, there is no standard protocol or practice around data access – often, the principle investigator must be contacted to obtain the data in a useable form.

Only about a third of the publications related to pathogens of epidemic and pandemic concern were accompanied by the data on which they were based.[3]

This pilot also found that other useful data were only accessible for a fee or came with requirements to register with the data owner. Such barriers, while not always insurmountable, discourage easy access to the data. Data access is also difficult even at the level of protocols: specifically, data which exist only in PDF formats are – even if theoretically accessible – practically unusable.

Similar problems were seen in the other pilot projects – for the Sendai use case, especially, lower-level data (such as that from hospitals and clinics) were often distributed by commercial organizations or networks for their own purposes. While the breadth of data in some areas is excellent, the cost for many government agencies would be prohibitive. Other data is collected and held by government agencies with a strong aversion to any form of disclosure risk, and so is practically inaccessible even to other departments within the same government.

[3] Pisani, Elizabeth; Ghataure, Amrita; Merson, Laura (2018): Data sharing in public health emergencies: A study of current policies, practices and infrastructure supporting the sharing of data to prevent and respond to epidemic and pandemic threats. figshare. Journal contribution. https://doi.org/10.6084/m9.figshare.5897608.v1

Although the FAIR principles regarding protocols and metadata for data access may not seem very extreme, they are rendered moot in situations where the data are held by organizations that are not interested in providing them. Simple solutions to some of these problems exist – publishing data in a processible form such as CSV instead of PDF, for example – but even these are not always adopted. It is clear that improvements need to be made in this area if existing FAIR Data Principles are to become a reality.

C. Interoperability

Interoperability is a complex subject. It can be used narrowly to mean that data are sufficiently documented to allow for a potential user to evaluate the data, it can be used to describe a more sophisticated phenomenon in which computer programs are able to act on data because it has been sufficiently described in a machine-actionable fashion. Sometimes the term refers to the ability of data to be accessed at different points across the data lifecycle. There are many different forms of interoperability. What they all share, however, is an emphasis data being useable within their area of coverage, whether this extends beyond human readability to machine-actionability or across the stages of the lifecycle.

In order for data to be useable, it must be both processible and understood. The term "metadata" is used in a general sense to refer to aspects of both of these functions: it describes the structure of data as well as providing a definition of its meaning, without which integration of the data with other sources is impossible. The FAIR principles regarding interoperability apply to both data and their metadata – both are equally necessary.

This raises issues on several levels, aspects of which were encountered by the groups at the Dagstuhl workshop. At the most basic level, a lack of agreement on how best to format and describe data can inhibit interoperability. In the infectious disease pilot, many data sources were compared and it was found that there was no agreement on how data were formatted/structured or described. This is a simple problem, but a very real barrier to interoperability, and one which has been observed in other public health efforts.

In the Resilient Cities project, data collected for efforts in Medellín, Columbia were examined. Although individually of high quality, a more complicated problem is encountered when methodological aspects of data collection are examined. The scope of data in terms of geography, time, completeness and coverage all varied in ways which sometimes made the data non-comparable. One example was that of air quality data: this is measured hourly or daily by some sources but needs to be compared with aggregated hospital records which only provide annual data.

The Resilient Cities pilot project has produced a paper looking at how a group faced with these problems can approach them systematically. [NOTE: Insert reference to Resilient Cities paper here.]

The Sendai Framework case provided an example specifically in the area of metadata standardization. In order for data to be comparable at an international level, a high degree of standardization is wanted: for risk classification, a critical element in the Sendai data context, good standard classifications of risk do not exist.

Thus, it can be seen that while metadata standards do provide a necessary ingredient for interoperability, they are not sufficient to provide a complete solution. Methodological approaches and definitional agreements must also exist.

One interesting idea was presented at the workshop, in which a modern technology architecture for data integration was combined with algorithmic "data science" techniques to provide automated data integration. One resource seen as fueling this mechanism was Schema.org, a site for enumerating the many standards now in existence for describing data, at many different levels (domains being just one). This presentation used the moniker "Plinth," although this term was more a proposal for carrying forward the idea than an acronym associated with a formal effort of some kind. [QUERY: Does PLINTH belong here or in the next section? It applies on some level to both topics.]

More traditional approaches to solving this problem were also identified by the workshop. The existence of provenance metadata would help researchers understand the fitness-for-purpose for the data they may wish to use, and the existence and publication of summary statistics would also be useful. A better description of scope, in terms of geography, and temporal and thematic coverage would also be useful. Existing standards provide a means of doing this (e.g., DDI).

The promotion of metadata standards generally could improve the situation in some areas, as the Resilient Cities group found for regional and city-level indicators. Standard classifications at the highest level need to be agreed in some areas, as for international hazards.

The infectious disease pilot concluded that it would be necessary to initiate a long-term process of consensus-building and adoption of appropriate data and metadata standards and controlled vocabularies to enable epidemiological, demographic, socio-economic, clinical, laboratory, and genomic data to be linked, queried and integrated more effectively.

The problem is not a simple one, but many of the tools needed to solve it do exist. FAIR highlights some initial steps for approaching the issues surrounding interoperability, but it may be that these principles are a foundation on which more work will need to be based.

D. Reusability

The FAIR principles regarding reusability place a strong emphasis on data description. In many ways, reusability and interoperability require many of the same things: reusable data is by definition interoperable data. Data which is interoperable, however – data that can be processed and understood – must be further described to make it truly useful. The level of data description seen when various sources are examined is uneven, although metadata standards and best practices do exist.

At the workshop, the infectious disease pilot found that the level of data description was frequently quite poor: data dictionaries and other useful metadata did not exist, and when they did were often not available in a processible (interoperable) form. Further, agreement on the definitional aspects of data description seemed to be stratified: local, regional, and national-level data often used different definitions of the same administrative and geographic zones, for example, complicating the use of the data.

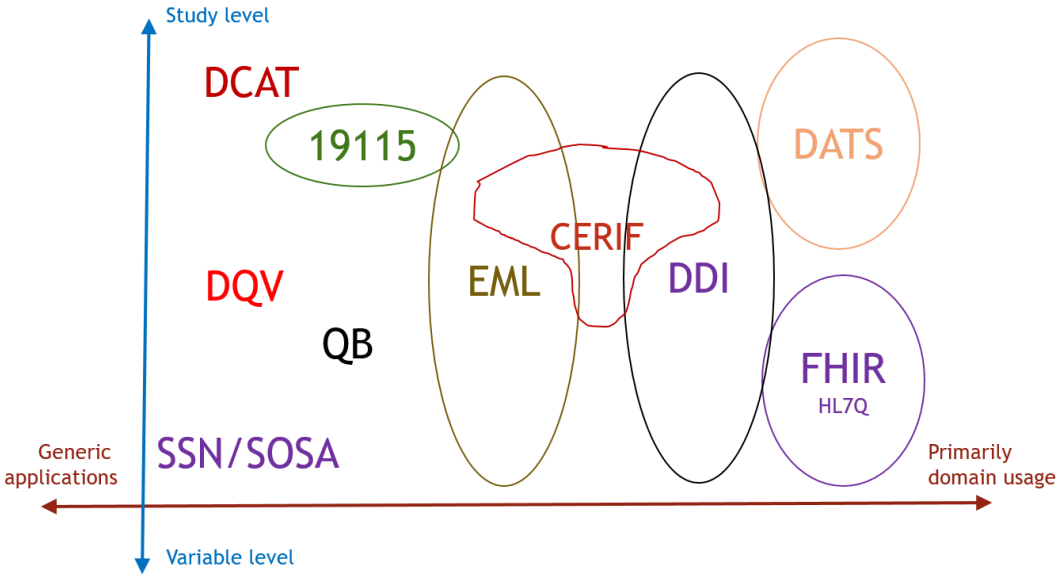
Definitional misalignments can also lead to issues of data quality: without good documentation of how the data are collected and how they perform their measurements, disagreements between what they tell us may not be explicable. The Sendai pilot found this to be the case with data regarding the impacts of different hazards – seeming mismatches between data from different sources could not easily be accounted for.

Ultimately, the need for better documentation, at a detailed level, is highlighted by an examination of reusability. Having a data dictionary is a fundamental need. Further, well-defined semantics, according to recognized standard vocabularies and classifications is also needed. The classifications and definitions of concepts at different levels – international, national, regional, and local – need to be mapped to one another, so that the data at each level can be more easily aggregated and used for purposes which it may not originally have been collected to support.

The metadata standards for achieving these goals exist, at least in exemplary form. The culture of practice for their consistent use seems to be missing.

V. Status of Metadata Standards and Specifications for Supporting Interdisciplinary Studies

Many of the key challenges associated with application of FAIR principles to the interdisciplinary use cases were related to the insufficiency of the metadata. While often promising to allow reuse across domain boundaries, many metadata standards are firmly rooted in the domain from which they originated and may not be relevant to or flexible enough to apply in other domains. Some standards can apply across domain boundaries but may require further modification within specific domains (e.g., additional elements to provide full contextualization for the data and enable determination of fitness-for-use) to be fully useful. The challenges and findings described in Section IV with respect to the FAIR principles help us to understand what further steps might be most useful in making data more broadly useful in cross-domain applications. The relationships among and between existing metadata standards and specifications was discussed at the Dagstuhl workshop. A diagram summarizing that discussion is presented below.



The various standards shown here support different capabilities, most of which are expressly focused on specific domains. It is not the point of this paper to discuss any of these standards in great detail, or even to characterize them all in general terms. However, they do offer a starting point for finding and using data across domains, and as such served as a focus during the workshop.

The diagram was created after presentations by experts at that workshop, attempting to visualize the relationships of these standards to each other in terms of how fine-grained they were in their descriptions of data, and whether they were specific to a particular domain, or were used more generically. This diagram is of necessity an approximation and contains many acronyms/abbreviations. It was not intended to be exhaustive, and, many other specifications were also discussed during the course of the workshop.

The table below spells out the abbreviations of the standards and provides a brief characterization of each specification. URLs to explanatory resources on the Web are provided for those who wish more information.

Acronym	Specification	Description	URL
CERIF	The Common European Research Information Format	A standard model for understanding entities important in the administration of research and their relationships.	https://www.eurocris.org/cerif/main-features-cerif
DATS	The Data Tag Suite to Enable Discoverability of Data Sets	The Data Tag Suite to Enable Discoverability of Data Sets	https://www.nature.com/articles/sdata201759

DCAT	The Data Catalog Vocabulary	Published by the W3C as part of the Linked Data family of vocabularies and technologies.	https://www.w3.org/TR/vocab-dcat/
DDI	The Data Documentation Initiative	A standard used in the Social, Behavioural, and Economic Sciences for documenting and exchanging data.	http://www.ddialliance.org/
DQV	The Data Quality Vocabulary	One of the Best Practices series from W3C, intended for use with the Linked Data family of vocabularies and technologies.	https://www.w3.org/TR/vocab-dqv/
EML	The Ecological Metadata Language	A specification for describing data of interest to the study of biodiversity and ecological sciences.	http://www.dcc.ac.uk/resources/metadata-standards/eml-ecological-metadata-language
FHIR (HL7)	The Fast Healthcare Interoperability Resources specification	Produced by HL7 (Health Level Seven), it is designed for the exchange of healthcare information.	https://www.hl7.org/fhir/overview.html
ISO 19115	The Geographic Information – Metadata standard	A widely adopted standard for describing information about geography, published by the International Standards Organization.	https://www.iso.org/standard/26020.html
QB	The RDF Data Cube Vocabulary	A W3C vocabulary, based on the ISO 17369 Statistical Data and Metadata Exchange Standard (SDMX), for describing statistical data sets. It is part of the Linked Data family of vocabularies and technologies.	https://www.w3.org/TR/vocab-data-cube/

SSN	The Semantic Sensor Network Ontology	A specification for describing sensor network and observations, published by the W3C as part of the Linked Data family of vocabularies and technologies.	https://www.w3.org/TR/vocab-ssn/
SOSA	The Spatial Data on the Web working group.	Not yet a specification, this group is designing a sister vocabulary to SSN for spatial data. This working group operates under the auspices of the W3C.	https://www.w3.org/2015/spatial/wiki/SOSA_Ontology

The Role of Metadata Specifications

Metadata standards and specifications exist in significant numbers, coming from different domains or from the technology space. They are designed to perform specific functions, and often do not focus on issues germane to cross-domain use of data as a fundamental part of their purpose. Regardless, they do offer us elements of a solution to the challenges encountered by the use cases.

Metadata standards and specifications give us a common format and model for needed information at least within the confines of a given domain. This is not universally the case – in some domains there are well-established standards, which are often symptomatic of existing communities of practice. In other domains, there seems to be no regularity to how data are documented or shared, nor any accompanying culture of how such activities are typically performed.

It is also the case that the existence of specifications does not necessarily indicate an accepted best practice within a domain: in some cases, standards are created as a way of championing a specific approach and marketed as a way of establishing a best practice within a community of use. While such standards may be based on excellent and innovative ideas, they are not the same thing as standards which reflect common practice across a broad community.

This is sometimes the case with the recommendations coming from the W3C: they are not designated “standards” in part because they are something different, intended to provide ways of realizing new functionality in a common way. They stand in contrast to domain standards such as EML or DDI, which are essentially agreements about how to describe things which are common practice across the community, expressing common formats and models for doing it.

This distinction may seem unimportant, but it has significant consequences for how the existence of metadata standards and protocols can be leveraged for the purposes of cross-domain data discovery and use.

If we consider the FAIR principles, we can begin to understand the roles played by different metadata specifications.

Findability is in one sense the easiest case: if metadata within a domain are described according to a standard model, then that model can be used as the basis for indexing and searching on that data, even if no other domain uses that model. In this application, the standards serve as a tool for supporting the harvesting of needed metadata. In an ideal world, all domains would describe their data according to a single standard model, and standards such as DCAT offer us tools for doing this. They are not universally adopted, however, and in many cases would need to be mapped against more common domain models before they could be practically used in a widespread fashion in specific domains.

Insofar as Access is a technical challenge, it is one that is primarily solved by specifications which most would not consider metadata standards, but technical protocols or formats. While the use of such protocols and formats needs to be better coordinated across domains, it is a relatively simple problem when contrasted with the challenge of describing domain-specific practice. The different needs across domains in terms of data access cannot be as easily standardized, and the existence of tools and models for approaching these problems are not very mature. They require an understanding of how data relates to other resources within the business models of domain organizations, and while such models as CERIF exist for approaching this space, they are far from universal in terms of their adoption.

Integration and reuse are the biggest challenges from the perspective of metadata standards – they reflect the very real differences in methodology and terminology across domains, and also the techniques used in specific domains for applying data. Some domains manage and use data sets as relatively static resources, which can be stored, versioned, accessed and used in a form that changes infrequently. Research findings are based on specific, known sets of data. In other domains, the data are produced using sensors or other tools which provide data as a constant stream, rather than as a static data set.

The techniques used in the latter for analyzing data are consequently different, focusing on the models used to understand the phenomenon they study, rather than on a specific set of representative or indicative data. Such differences are driven by the practices within the domains – they can be described using metadata standards, but these standards must embrace models which reflect the different types of practices.

Some aspects of research can be standardized across domains relatively easily, however. One example of this is in describing classifications and terms. The standards exist for expressing classifications and similar concept systems, and they are widely adopted. What is needed in some domains is a more disciplined approach to their formal use alongside published data – in other domains such as official statistics classification management is seen as a primary activity of data producers, and this metadata often exists in standard forms today.

Metadata standards offer us the basic tools needed to implement the FAIR principles: they give us a known expression of the metadata used for almost all interactions with data. In each of the different areas, however, the role played by standards may be different, and may be specific to particular domains. For data discovery, mapping existing domain standards to agreed cross-domain standards is sufficient; for interoperation and reuse, a much more nuanced approach to coordinating the expression of useful metadata is needed – the required metadata standards may not even exist. In some sense,

these more difficult cases will force us to start where there is commonality – in the formal description of classifications and concepts – and then to develop cross-domain models which can express the differences in practice.

Fundamentally, metadata standards – however insufficient they may be when judged as a complete solution – offer us the building blocks of one. What is needed is coordination across domains in the use and development of such standards. Metadata specifications offer us the foundations on which a solution can be built. Recognizing where metadata standards provide a useful description of the data practices within a domain - as distinct from areas where common practices need to be encouraged or established – is the challenge which must be faced if we are to realize the promise of the FAIR principles across domain boundaries.

general comment.

VI. Conclusions/Overall Findings/Next Steps (????)

The major, pressing global scientific and human issues of the 21st century (including infectious diseases, sustainable development, disaster risk reduction) can ***only*** be addressed through research that works across disciplines to understand complex systems, and which uses a transdisciplinary approach to turn data into knowledge and then into action. Realizing this depends upon our capacity to extract knowledge from the large and diverse volumes of heterogenous data that are increasingly available that reflect the behavior of complex systems. Yet our ability to combine data from heterogeneous sources and across disciplines remains rudimentary at worst, excessively resource intensive at best. The three use cases highlighted both challenges and potentially promising solutions:

- the increasing recognition in many scientific communities of the vital need for FAIR standards to render data usable and interoperable (Findable, Accessible, Interoperable and Reusable) by both humans and machines;
- improved uptake and sophistication—in certain disciplines—of the metadata and data specifications that allow data to be appraised, linked, combined, integrated, analyzed and reused, and at great degrees of automation and at scale;
- wider application and advances in the use of 'data science', in particular by using machine learning, which offers powerful means of supporting interoperability and data integration, as well as the extraction of knowledge from complex systems.

Building on the current pilot case studies, it is proposed to apply and refine the techniques and processes trialed in the pilot to a far broader range of cases and to explore how, in addition to the application of the FAIR principles and the alignment of data and metadata specifications, machine learning and data science can be brought to bear to assist interoperability, distributed data integration and the extraction of meaning from interdisciplinary data.

The anticipated impacts are more effective and evidence-based solutions for complex global challenges through:

- an increased number of science disciplines and interdisciplinary research areas that adopt rigorous standards and ontologies for their data;

- widespread adoption of replicable, generic approaches to data integration and FAIR data by scientific disciplines and interdisciplinary research areas;
- more effective application of programmatic data linking and integration, facilitated by more effective semantic description of data and the widespread implementation of FAIR data;
- the application of integrated, inter-disciplinary data to the characterization of system complexity in global challenges;
- the take-up by policymakers and other users of solutions for complexity;
- concrete outcomes and exemplars from the interdisciplinary research areas are designed both as exemplars of potential and important contributions to thematic understanding and practice.

As a first step, a follow-up intensive workshop is needed whereby domain and data scientists assess a subset of the data and metadata that are used to address central questions in three or more exemplary use cases. Questions to be addressed relate to the adequacy of the standards and specifications employed, whether or not other more effective standards and specifications exist, and what are the key barriers to adoption and utilization of improved solutions that can better facilitate interdisciplinary research. Furthermore, are there changes in organizational reporting practices that can improve compliance to all four FAIR principles.