ARPHA Preprints

PENSOFT

Project Report

Author-formatted document posted on 06/11/2023

Published in a RIO article collection by decision of the collection editors.

DOI: https://doi.org/10.3897/arphapreprints.e115047

Reuse and Reproducibility: Describing Cross-Domain Research Data in the Science Project *Climate Neutral and Smart Cities*

Arofan Gregory, Joachim Wackerow, Hilde Orten

ARPHA Preprints *Author-formatted document posted on 06/11/2023*. DOI: https://doi.org/10.3897/arphapreprints.e115047

Reuse and Reproducibility: Describing Cross-Domain Research Data in the Science Project *Climate Neutral and Smart Cities*

Arofan Gregory, Hilde Orten, Joachim Wackerow
12 September 2023

Contents

I . Overview 3
II. Characterising Researchers and Infrastructures 4
III. Information Requirements 5
A. General Considerations 5
B. Types of Data and Metadata 6
1. Semantics and Definitional Metadata 6
2. Structural Metadata 7
3. Time and Geography 9
4. Process, Methods, and Provenance 10
C. Researcher Scenarios 11
1. Researchers Using Data from Their Own Domain 11
2. Researchers Using “Raw” Data from Other Domains 12
3. Researchers Using “Integrateable” Data from Other Domains 13
D. The Data Provider Perspective 13
IV. Current Models and Systems for Data and Metadata 15
A. ESS and DDI Lifecycle 15
B. ERA5 Climate Data (Copernicus), GRIB, and NetCDF 16
C. European Environment Agency (EEA) Air Quality Data 17
D. General Considerations 18
V. The Researcher Perspective 19
VII. The Infrastructure Perspective 20

VIII. Trust and Transparency 22
IX. Looking Forward 24
A. Process and Provenance in Cross-Domain Data Use 24
B. FAIR in Cross-Domain Data Integration 25
C. Common Standards for Cross-Domain Metadata Exchange 26
D. The Institutional Framework 26

I . Overview

Science Project 9 (SP 9) *Climate Neutral and Smart Cities* of the EOSC Future (Task 6.3)[1] explores many different aspects of supporting cross-domain research through prototyping the integration and use of data coming from the European Social Survey (ESS[2]), the European Environment Agency (EEA[3]), and Copernicus Climate Change Service ERA5[4], covering social attitudes and behaviours, air quality, and environmental measures for a variety of European cities. The project has produced several outputs, which can be found at https://www.europeansocialsurvey.org/esslabs/. This paper focuses on the reusability of data, and the information needed by researchers and infrastructures to support it in a scenario which cuts across traditional domain boundaries. The findings are based on a practical prototype, which can be seen on the site listed above.

This paper is intended for those concerned with the issues of metadata and documentation within cross-domain research, whether systems designers, metadata specialists, methodologists, or technical implementers. It argues that a coherent set of metadata standards must be employed for information exchange across domains, and that the metadata must include detailed information about the provenance of data and its processing, as well as more typical descriptions of the structure and definitional aspects of data, and the documentation of methodologies. This requirement can only be met if a coherent institutional approach is taken, built on teams of experts from across the involved domains, through a framework such as EOSC and accompanied by necessary resources.

More and more, research projects involve many organizations and researchers working on multi-disciplinary research questions – subjects such as climate change give rise to many “cross-domain” research questions by their very nature. To support such research in an effective way, the research infrastructures which provide the data must account for the range of expertise involved, and the nature and depth of documentation which will be needed by researchers.

The reuse of data is common to such research: while one domain may provide the primary data set, data coming from other domains will be used to contextualize and inform this data. Researchers from these domains will have the expertise necessary to best utilize the data they know well, but must work together to address the full set of data to be analyzed.

Because these multi-disciplinary research projects have not traditionally been the norm in many domains, it is important that the processes and methods applied to the data be as transparent as possible. The choices made by those preparing the data may not be as readily understood by the researchers using it, and so must be thoroughly documented. This is even more true when we consider that research findings may well be subject to questions requiring their reproduction, so that they can be validated and their accuracy understood. This reproducibility is even more demanding of the information made available about how the data was processed and integrated, in order that it can be duplicated.

While there are many projects aimed at establishing collaborative platforms across domains, involving Jupyter Notebooks, data spaces, etc. these do not address the issues with which this paper is concerned: the metadata and documentation required by collaborative, cross-domain research teams, and the

[1] The EOSC Future Science Project Climate Neutral and Smart Cities is funded under the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101017536.
[2] https://www.europeansocialsurvey.org/
[3] https://www.eea.europa.eu
[4] https://cds.climate.copernicus.eu/

standard models needed to express it. These requirements exist within any collaborative environment. Nor is it enough to provide access to a data set through an infrastructure portal, with a modicum of cataloguing metadata – while this is useful, it does not address the fundamental challenges of cross-domain data reuse, or provide a full understanding of the data.

This paper examines the requirements for information in relation to the researchers participating in cross-domain projects and the infrastructures which provide the data and documentation needed. In order to avoid confusion, we examine the problem from a fundamental level – within domains, different assumptions are often made about the relationships of researchers and data providers, and it is felt that a full description is needed to provide clarity. Starting from the information requirements, we look at the different scenarios considered when researchers from different domains interact with data providers. The metadata models used within the different domain systems are considered, and the implications for metadata standardisation and exchange are addressed. The longer-term goals for such cross-domain research are then considered from a metadata perspective, and the relevant issues highlighted.

It is not within the scope of this paper to perform a focused examination of the literature on the roles of researchers within collaborative, cross-domain teams. This paper reflects the needs and actions of researchers which we found in the work on the SP9 project and as we have observed them in other cross-domain areas where metadata and documentation requirements are discussed, notably in the context of FAIR implementation. Discussion of researchers and infrastructures should be understood in this light: the focus of the paper is on the fundamental needs for information, and the way it is represented and used within systems.

In exploring these questions, the EOSC Future Work Package 6.3 Science Project 9 Climate Neutral and Smart Cities (SP 9) has prototyped a system for facilitating data reuse. In this project, the primary data comes from the European Social Survey, but it is integrated with ERA5 temperature data from Copernicus and air-quality data from the European Environment Agency. In order to facilitate the integration, data regarding geographical systems and population density have been drawn from other sources.

Other reports will describe the methods used in the selection and processing of the data. Here, we will focus on the system and metadata requirements placed on the data infrastructure in order to support the reuse and reusability of the data. From the prototype example, we further generalize the needs of such applications in a world where cross-domain data sharing is expected to become a regular aspect of scientific research, as a result of the existence of infrastructures such as EOSC, and in the adoption of the FAIR principles more generally.

II. Characterising Researchers and Infrastructures

In understanding the nature of cross-domain research, it is helpful to establish some basic distinctions between the various players involved. From the researcher perspective, there will be a principal investigator and supporting researchers who typically will have expertise in a single domain, and this domain perspective can be said to align with the primary direction of research. The primary data of interest will be that coming from this domain.

In SP 9, while there is data from the ESS, as well as climate and air-quality data, it is assumed that the researcher will be a social scientist for whom the European Social Survey is a familiar and well-

understood data set. This is the traditional user community for the ESS data, and the target audience for the systems built by the ESS to support the use of its data.

Other types of users are seen as important but secondary audiences, which would be more likely to rely on their own domain infrastructures and repositories as the places where they would find data. Thus, while the ESS might be interested in helping climate scientists reuse the social data from the ESS – especially questions regarding attitudes about climate change, for example – it is to be expected that this is a secondary audience for the ESS. Climate scientists would be more likely to turn to familiar sources of data within their own domain infrastructure for primary data, and use the ESS data as a secondary, contextualizing resource.

Thus, we can understand that research data infrastructures such as data repositories and disseminators can be aligned with the particular domains they have traditionally served. They are the focus for researchers within that domain when it comes to finding data and attendant information about that data. This alignment of both researchers and data providers with domains is typical of the scientific landscape, reflecting the organization of research more broadly.

When we consider the knowledge typical of researchers within a domain, they can be expected to have a familiarity with the literature in their specialty, and a consequent familiarity with the main sources of data which are used in that research. Thus, it can be expected that social scientists in Europe (and across the globe) will be familiar with the European Social Survey, that those who study the environment will be familiar with the data available from the European Environment Agency, and so on.

This simple characterisation of researchers and data providers being clearly aligned with specific domains is, of course, a simplification. In some cases, the alignment is not a clear one, as some research subjects are by their nature cross-disciplinary, and some researchers will have a broader knowledge than just their own narrow discipline. Further, some types of data are used widely across domains, The most important example of such data is geographical data, While geospatial science is a domain in its own right, it also serves as the basis for the integration of data across many other domains, and thus represents a special category which does not neatly fit our simplified model.

This simplified distinction is still useful in understanding the needs of the different actors in this discussion, however and we will make reference to it in looking at the information requirements of different researchers and systems, below, even while recognising that there are significant exceptions. These will be specifically addressed in the places where they apply.

III. Information Requirements

A. General Considerations

When we consider their information requirements, we can understand that researchers within a domain, and those outside of it, will have different needs. This is true not only in terms of what needs to be known about the data to be used, but also about the specifics of the data itself.

We can understand that there are different cases here: a researcher working with data coming from their own domain (either gathered by themselves, or secondary data from familiar sources); a researcher working with un-integrated data coming from outside their domain, in the form presented by the source;

and the researcher working with secondary data from outside their domain, where that data has been processed so as to be integrated with the primary data.

The type of metadata and documentation needed will also be different in these cases, although there are some aspects which remain common. The basic description of data – definition of concepts used as variables and categories, the representation of values, the structures of data sets – is a common requirement regardless of which case is considered. The context regarding that data – the way in which it was produced, and the ways in which it can be used – will differ significantly. Where data has been processed to make it subject to easy integration, as for raw data from outside a domain, there may be questions about the methods used to do so. In various cases, these different types of information will be more or less critical.

Further, while it can be assumed that existing documentation provided by a dissemination platform is sufficient to support use, it may be that improvements could be envisioned. Metadata and documentation and the systems which disseminate them are often expensive to acquire and implement, and researchers are often asked to work with whatever can be provided, even if it is less than ideal.

In this section, we will examine the different cases, using the SP 9 prototype as a reference point, as these considerations emerged during that work.

B. Types of Data and Metadata

1. Semantics and Definitional Metadata

The domain semantics – that is, the definition of the scientific concepts which describe the data – are termed “semantics” for the purposes of this report. (Many different aspects of data description involve semantics, but for this discussion we are using this to refer to scientific meanings important within a domain.)

Formally, we can use the definition of semantics as that information which is "describing statistical units, populations, classifications, data elements, standard questions and question modules, collection instruments, and statistical terminology;" .[5] This should, however, be understood as distinct from structural metadata (see below) which addresses the organization, rather than the definition, of the data.

Formal semantics as used in data management and dissemination systems are typically based on concepts, which provide a term and a meaning. These are organised into collections (concept systems) and are then used in the many places in a data description where a definition is needed.

There are many examples: concepts are used to define variables, the categories within classifications and other coded systems (including the items in a taxonomy or ontology), for the definitions of universes, populations, and unit types, and so on.

[5] "The Role of Metadata in Statistics", Cathryn S. Dippo, and Bo Sundgren, https://www.bls.gov/osmr/research-papers/2000/pdf/st000040.pdf

In order to describe the meaning of data, the concepts are associated with the different structural aspects of it. Thus, a concept such as “air quality” might be used to define a variable, with the concepts “good,” “fair,” “moderate”, “poor” (etc.) as concepts used to define the categories of the codes which represent the values of that variable. Thus, the semantic definitions are assigned roles relative to the structural metadata (see below). Because concepts can be reused, it is typical that a structural description will reference the concepts from a centralised source such as a repository.

In some metadata models, including DDI Lifecycle, concepts are formally modelled and their structural roles described, but semantics are not defined for any particular domain by the standard itself: instead, references are made to external definitions, or these are provided by the user. Thus, DDI Lifecycle is a flexible way to connect domain concepts to data, but does not force users to employ any specific set of semantics.

DDI-CDI functions in much the same way, with the added ability to directly use formalizations of concepts expressed in many other common structural standards such as the Simple Knowledge Organization System (SKOS[6]) published by W3C, which is commonly used across domains.

Many data formats and structures are very weak when it comes to the formalization of concepts. NetCDF (Network Common Data Form[7]), for example, typically provides only the label of the column as a definition of the variable in a data set, beyond some agreed common variables (e.g., time and geography). The challenge of formally defining these labels is left entirely up to the user: definitions may or may not be made available, as with common formats such as CSV, where a simple label provided as a column header may be defined in some other file, or may be assumed to provide sufficient information to the user.

2. Structural Metadata

Structural metadata is the information about how records and variables are organized within a data set, including information about how the variable values are represented. A definition from “The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials” reads:

> “Structural metadata describes the structure of the digital object and the relationships between its components; this information is crucial to assist navigation and ensure that complex objects that belong to a larger collection are linked meaningfully together.” (p53, https://www.ninch.org/guide.pdf)

The description of variables includes links to the formally defined concepts that are used to provide the definitions of the variables and the categories which are represented in their values. Structural metadata also includes a description of the type of unit of study associated with each record, as well as the identification of each unit (one of the variables). Structural metadata may also be provided for the description of survey instruments, by providing the flow logic of the survey and describing its component parts (question text, response domains, and so on.)

Structural metadata is distinct from metadata describing the entire data set, which may describe provenance, contact and access information, methods, population and universe descriptions, and so on. There is an overlap between cataloguing metadata and structural metadata, but the structural metadata

[6] https://www.w3.org/2004/02/skos/
[7] https://www.unidata.ucar.edu/software/netcdf/

typically has more detail at the granular, variable level. Structural metadata is also distinct from controlled vocabularies (including classifications, codelists, etc.) – structural metadata indicates where these are used, and identifies which ones, but does not provide the definitions themselves – it only shows how they are structured, and how they are employed.

There are many examples of standards which explicitly model structural metadata, as distinct from definitional and semantic metadata. Among these are the Statistical Data and Metadata Exchange (SDMX) Technical Specifications[8], and CSV on the Web (CSVW[9]). It is perhaps more typical for structural metadata to be combined with definitional metadata within a single domain standard – practice varies widely.

The European Social Survey uses the DDI Lifecycle[10] standard for describing structural metadata. Descriptions of the variables available from all ESS data files are stored in a data repository (found at https://ess-search.nsd.no/). This is a Colectica Repository[11,12], which provides access to all of the metadata in various forms, but which is closely aligned with the DDI Lifecycle[13] model.

DDI Lifecycle is a standard which is widely employed within the social, behavioural, and economic (SBE) sciences, and is used by many of the CESSDA ERIC archives (Consortium of European Social Science Data Archives European Research Infrastructure Consortium[14]). It is the product of the DDI Alliance[15], a membership consortium of archives and data producers working in this domain. As such, it reflects the needs, conceptual framework, and terminology of the social sciences. Further, it is strongly oriented toward the description of unit-record data files, which are often termed “wide” files because they have a lengthy set of variables (represented as columns in a table) for each of the unit records (represented as rows across the set of columns).

Structural metadata can also be described using the DDI Codebook standard[16], although this is more limited in how it views variables (there is no inherent concept of the reuse of variables or their representations, which makes comparability across data sets more difficult to manage). DDI Codebook was formerly the agreed standard within CESSDA[17] for describing data, and is still widely used in some of the CESSDA archives. It also focuses primarily on the description of wide data sets.

Structural metadata can also be described using a non-domain-specific standard being produced by the DDI Alliance for the purposes of implementing exchanges based on the FAIR Principles (The FAIR Guiding Principles for scientific data management and stewardship[18]) – the DDI Cross-Domain Integration (DDI-CDI) specification[19]. This is a model for the generic description of a wide range of different data types, including the wide data structures typical of the social sciences.

[8] https://sdmx.org/?page_id=5008
[9] https://csvw.org/
[10] DDI Alliance: Overview of Current Products, https://ddialliance.org/products/overview-of-current-products
[11] ESS Processing, https://colectica-ess-processing.nsd.no/
[12] Colectica Repository, https://www.colectica.com/software/repository/
[13] https://ddialliance.org/Specification/DDI-Lifecycle/3.3/
[14] https://www.cessda.eu/
[15] https://ddialliance.org/
[16] https://ddialliance.org/Specification/DDI-Codebook/2.5/
[17] CESSDA ERIC - Consortium of European Social Science Data Archives, https://www.cessda.eu/
[18] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/
[19] https://ddialliance.org/Specification/ddi-cdi

The description of structural metadata in other data sources used by SP9 is described more fully below, but neither of these has the granularity of the ESS data when it comes to structural descriptions. This type of metadata is often combined with formatting information (as in NetCDF[20]) or is addressed in less formal ways (as the documentation of columns in a spreadsheet or similar). DDI Lifecycle provides a rich, standard description of the structural metadata as it has been implemented for the ESS.

A more complete description of the SP9 data sources can be found in other papers produced by the project, available through the ESS Labs site.[21]

3. Time and Geography

Although time and geography are represented in data sets using the same set of structural mechanisms (variables, categories, values) as other important aspects of data, they are considered here in particular as they form an important feature of the integration of data sets. Time and geography are both subjects which are highly standardised: there are many common models for describing them.

Time values in data are generally encoded using a virtually universal ISO standard (ISO 8601 Date and time format[22]), and this does not present a barrier to integration in itself. What is more challenging is the way in which time is associated with data: for the ESS, each respondent is interviewed on a specific day, which is recorded, but their answers are understood to be representative over a longer period. Because the survey is only conducted every two years (since 2002), the data is essentially a snapshot of responses for the specific time. The ESS measures mainly social attitudes and behaviour. The survey questionnaire combines continuity with change through a consistent core module and a series of rotating modules addressing key social themes. The core module is measuring a range of topics of enduring interest to the social sciences as well as the most comprehensive set of socio-structural variables[23]. Because the ESS measures opinions, which are not subject to continuous monitoring, the social sciences use this approach as a way of measuring behaviors and attitudes, and the domain has developed methods for handling time in this way as it relates to the data.

Measurements in climate and environmental science are very different, as a function of the way data is collected: sensors can be used to take routine measurements at short intervals, spread across the geography to be measured. Thus, we see that the temperature data we used from Copernicus is collected and reported hourly, and the air quality data is likewise reported every hour. The relationship of time to the data is very different across domains, and must be accounted for when the data is integrated. Social data is fundamentally different than environmental and climate data as a function of the subject being studied: each domain collects data in a way which makes sense given the phenomenon being studied.

Geography is a more complicated topic, as the systems for encoding geographical data – while highly standardised – are also diverse. There are many different coordinate systems for describing geography, although most rely at their core on the ISO 19115[24] standard which establishes the way geography can be described as points, lines, and polygons in relation to the planet's surface. These coordinate systems

[20] NetCDF, https://en.wikipedia.org/wiki/NetCDF and https://www.unidata.ucar.edu/software/netcdf/

[21] https://www.europeansocialsurvey.org/esslabs/
[22] https://www.iso.org/iso-8601-date-and-time-format.html
[23] Prospectus of ESS ERIC, https://www.europeansocialsurvey.org/docs/about/ESS_prospectus.pdf
[24] ISO 19115-1:2014, Geographic information — Metadata — Part 1: Fundamentals, https://www.iso.org/standard/53798.html

are the usual way that geography is encoded for the climate and environmental sciences. When we consider the focus of these domains, this orientation toward the physical world makes sense.

Social science, however, typically uses a very different approach. While people can be located on the surface of the planet using coordinate systems, it is more important to the social scientist to understand where they are located in terms of their social context. This can be understood as an idea of where people are according to the hierarchy of continents, nations, regions, provinces, states, counties, cities, towns, etc. This hierarchy is a social and political/administrative construct: it is a way of understanding geography according to the identity of the people who occupy the space, and by the way they understand their own locations. In the frame of social science, this social (and political/administrative) context is more important than the coordinates on a matrix which specifically locate a point on the surface of the globe.

Further, the precise location of respondents can also present a risk of identity disclosure, and thus cannot be included in publically available data. This is an issue which stems from the fact that social scientists study people, and is not typically a concern in environmental and climate science.

For social science, the encoding of locations is therefore done using standard classifications representing these less-precise social hierarchies. For the ESS, the NUTS[25] geographical classification is used, which covers all of Europe and the immediately surrounding countries, providing several levels of granularity.

While these two systems of describing location are different, they can be correlated: one can describe the polygon which represents the political boundaries of a country, state, city, etc. in relation to the coordinate system describing the surface of the earth.

While time and geography represent important dimensions over which data can be integrated, it is clear that these issues must be addressed carefully when translating their use in relation to the data coming from specific domains. This is not necessarily difficult to do, but it must be done correctly and carefully, even in cases where there is a high degree of standardisation in their representation.

4. Process, Methods, and Provenance

When looking for data to use, researchers need to understand the precise nature of the possible sources. The term “provenance” covers a wide variety of topics: who collected the data, and how? What were the methods used? What was the purpose of the data collection? How was it processed during the production of the data in its reusable form? Who was the funder? (Etc.)

From our perspective this type of information can be seen as a class of metadata, although often taking a more narrative form. Provenance is often seen as a key component of the metadata needed for discovering data, but it is also important in determining fitness for purpose, and in deciding how specifically it can best be used in an analysis.

An important concept here is that of “data lineage.” This is the answer to the question “How was this data collected, and what specifically has been done to it since?” We can model data lineage as a set of stepwise events, from the point of origination, through a series of processes, and resulting finally in the data we have in front of us. There are a series of actors – whether human or machine – which perform these steps, and they frequently involve the transformation of the data in some aspect. This results in a

[25] NUTS - Nomenclature of territorial units for statistics, https://ec.europa.eu/eurostat/web/nuts/background

series of distinct data sets, or versions of the data, which although non-identical, carry the same overall intellectual content.

Methods determine the steps which will be taken in relation to any given data: we clean and validate data, transform it to be integratable in terms of how it represents the values of variables, and so on. The methods used by the data collectors and processors will determine how it can best be used, and the quality of the data in relation to any specific research question – the “fitness for purpose” of the data.

Data lineage is not the only type of provenance metadata, but it is an important aspect of data. Very often, the data lineage is available to researchers in documentary form – as PDFs – and is not as thorough or easy to navigate as researchers might wish. In some cases, it is very difficult or impossible to determine exactly how a specific variable within a data set has been produced, and researchers are left to assume that it is a “normal” variable of its type, based on the implicit knowledge of the domain as reflected in the literature or the reputation of the principal investigator who was responsible for its production. Within a domain, researchers often have access to the researchers who produced the data, and can ask specific questions directly, rather than relying on documentary sources.

There are many standard models for describing data processing, and indeed much of the code written to perform processing may be familiar to researchers. In the case of SP 9, data processing was performed in Python, which is a widely used programming language for such purposes – it is not unreasonable to expect researchers to be able to understand it, or even take the source and modify it if that is useful to them.

There are also many domain standards for describing data lineage, and especially data collection. These standards often require a detailed knowledge of the possibilities of data collection specific to the domain: for example, survey methodology is a complex subject, and the design of a survey can have a huge impact on the utility of the data collected. Similarly, sensor data can be massively influenced by the sensors used and their configuration and deployment. Within a domain, these factors may be generally understood by researchers using the data, but they are frequently not well documented in relation to a specific data set.

C. Researcher Scenarios

1. Researchers Using Data from Their Own Domain

Generally speaking, researchers are most comfortable working with the data which is common in their domain. For many social scientists, the ESS is a known and trusted source. The ERA 5 data from Copernicus EU is another well-known source, supported by an extensive literature. The European Environment Agency produces high-quality and widely recognized time series on many environmental measures, including air quality.

The details of these data are certainly familiar to their users, and this knowledge is reflected in the domain literature which surrounds each of these sources, covering both their collection and use. The data are clearly documented on the sites which disseminate them. For the ERA 5 data, each of the main variables is documented in a table which can be seen at https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview. For the air quality data from the EEA, users can navigate from the download site (https://www.eea.europa.eu/en/datahub/datahubitem-view/82700fbd-2953-467b-be0a-78a520c3a7ef) to a

"metadata fact sheet" and from there to various references describing the methodology and measures in more detail. The European Social Survey provides access to rich metadata for each variable, as well as providing extensive documentation for each round of data collection. This is available at the ESS Data Portal.[26]

In every case, the researcher wishing to understand the data will – if familiar with the domain and its literature – find a sufficiency of information to support the use of that data in their research. In practical terms, this level of documentation can be understood as the standard against which other cases can be judged.

In many cases there is machine-actionable metadata available in various forms (for example, the EEA metadata pages have a considerable amount of information included in their source as JSON) but this will not necessarily be of great utility to the researcher. In these cases, the definitional metadata – the information about the definitions of variables and categories – is generally not available from these sources. Interestingly, this information does exist in standard form within the disseminating institutions, in some cases, but is not provided to users directly (the ESS is managed using DDI Lifecycle, for example, which provides a standard way of encoding this metadata in a machine-readable form, but it is not directly available for download.) The ERA 5 data is available as a NetCDF file, which combines a basic set of structural metadata with the data itself, but the definitional and semantic metadata is missing. (Other standard formats available share a similar issue – see below).

This is likely not to be a major problem for the researchers using data which is familiar to them, as they will have the background knowledge of how terms are defined and measurements made in their domain, based on the literature and other, similar sources.

2. Researchers Using "Raw" Data from Other Domains

When researchers who are less familiar with a domain encounter these data sources in their "raw" form – that is, as disseminated directly from the sources listed, and not integrated with any other, more-familiar data – then the available documentation may be less useful, depending on the knowledge of the audience, and who the documentation was prepared for.

The primary barrier to understand the data will be a lack of familiarity with the domain itself: terminology and methods may be less familiar, and may not be explained clearly enough for those who lack domain expertise.

For example, the Copernicus temperature data documentation[27] makes the statement: "The framework uses the statistical method known as Kriging to interpolate data from stations to arbitrary locations on the Earth." This method may not be familiar to a researcher who does not have a background in geostatistical methods, and although it can be further explored, a degree of uncertainty will remain as to what the impacts of this method will be on the data.

While it is not reasonable for the data disseminator to go beyond an identification of the methods used, it is clear that domain expertise is of huge benefit in understand what the data actually is, and how it can be processed. (The Copernicus documentation is actually quite good when judged by our domain standard

[26] https://ess-search.nsd.no/
[27] https://datastore.copernicus-climate.eu/documents/insitu-gridded-observations-global-and-regional/C3S_D62.3.5.1.v1_202110_Documentation_observation_data_v1.pdf

described in the preceding section: it gives the link to the paper describing the method, and provides a fairly complete summary of its application to the data being described.)

This example is used here to demonstrate how the barriers presented by domain specialization are difficult or impossible to avoid: the necessity for domain expertise is inherent in the optimal use of virtually any scientific data.

Another challenge in using unfamiliar data regards the collection and subsequent processing. While many standard statistical methods for data cleaning may be familiar, the way in which they are selected and applied may reflect aspects of the data collection which will be unfamiliar. For climate scientists, sensor data and the factors which determine the quality of the data they produce are familiar – to social scientists, this is likely unknown. Social scientists will understand the pros and cons of the particular sampling and survey methods used when collecting data – to the climate scientist, these are unknowns. (As an example, response times in a survey may indicate that some respondents are providing low-quality answers, and the resulting data should be cleaned accordingly. This is unlikely to occur to a climate scientist, as these types of considerations do not exist with sensor data, etc.)

3. Researchers Using “Integrateable” Data from Other Domains

A third scenario we will consider is when a domain data source is providing “integrateable” data coming from another domain – the case that we are exploring in the SP 9 prototype. In this case, it is assumed by the researcher that a familiar source of data is being supplemented with additional data from outside the domain in a form which is pre-processed to allow for integrated use.

In this scenario, the data is being provided in a familiar structural form, but some of the definitional and semantic metadata will need to be provided (domain-external data may be represented with unfamiliar categories/classifications which need explanation, etc.) The integration of time and geography across the different sources of data will have already been performed, but the details of how this is done will be a point of interest, as they will affect the way in which the data is analysed. Other processes such as weighting will also have a potentially major impact on the data, and will be a point of interest.

Data quality is always an issue, and must be understood from the perspective of fitness-for-purpose: any given data may be more or less useful based on the research question.

The researcher will be placing a degree of trust in the familiar data provider, assuming that unfamiliar data has been provided for easy integration in a form that can be effectively used. This trust, however, is not unqualified, and researchers will also want to understand what has been done to prepare domain-external data for integration by the disseminator.

D. The Data Provider Perspective

The organisations which disseminate scientific data for reuse are often focused on specific research areas or types of data, even when these may be reused across many different domains. In SP 9, we have the case of the ESS, which is clearly aligned with the social sciences, while Copernicus ERA5 and the EEA have a somewhat different relationship to domains. Copernicus data is used for air quality research, but is also used intensively in other domains such as energy and public health (to give but two examples). Environmental data is also used by a wide range of academic disciplines, driven by increasing interest in the many aspects of climate change.

The common thread in the use of much environmental and climate data is the geospatial grounding of these data. Typically, measurements are made in relation to specific geographical locations or areas, but the same is not necessarily true in areas such as social science where the primary grounding is the observation and measurement of people and their various social groupings (this depends on the research question being addressed). These different types of data require the use of different data collection techniques, with an impact on the way in which data are structured, described, and processed.

The focus of data disseminators is thus driven in part by the communities the serve – domains or disciplines – and in part by the nature of the data itself, and the purposes for and ways in which it is collected. These organisations – whether they aggregate data from others or collect it directly – develop strong, specialised expertise in the domains or types of data on which they focus. They are often major players in setting domain standards and establishing good practice, as they engage with many different research projects in their area of focus, thus developing a breadth of experience.

There are many common concerns, however. The success of a large data producer is largely determined by the quality of the data produced, and the reputation which develops as a consequence. Long-running data collection which is consistent and which complies with good practice in terms of data management and documentation will be trusted, and perceived as of good quality. It will be more widely used because of these factors.

Concerns about data quality naturally lead to the need for ownership of the data: an organisation cannot be held responsible for data which it does not control. Without control, there can be no guarantee to users that data is consistent and otherwise of high quality.

Reputation also relies on effective citation: the source of the data must be credited when used, so that the data provider can demonstrate impact and utility, which are often key in justifying continued funding.

These motivations are important in understanding why data producers and disseminators will consider collaboration across domains: they do not want their data to be irresponsibly disseminated by others. If they can make it easy to cite and reuse their data, then they can remain the gatekeeper, ensuring the quality of the data and sustaining their reputation and impact.

It is typically the case, however, that organisations in different domains will employ different models for describing their data, reflecting domain practice – if the data can be primarily associated with a domain, as in the case of the ESS – or reflecting the technical capabilities and choices of the data provider's systems. This presents a barrier to data integration and reuse which must be overcome, unless all data providers agree on a single standard set of models and formats for their data and metadata.

Another significant factor is related to the institutional arrangements and relationships within the overall data ecosystem. In SP9, we see clearly that the formation of cross-domain teams of experts is a requirement for providing the best support to researchers, but there was no recognised process or channel for this collaboration. Further, in some cases there were resourcing issues, as the extent of the expertise needed had not been fully anticipated. These issues are very real, and must be addressed in any practical system for supporting cross-domain research.

IV. Current Models and Systems for Data and Metadata

The domain metadata models and the data management and dissemination systems which implement them are the basis for all use and reuse of data: they are what allow researchers to understand the data in all cases except where the data was collected by specifically for the immediate analysis being conducted, by the same group of individuals. While individual publications may provide an indication of what data was used to answer specific research questions, the articles often do not contain sufficient detail to support reuse. While publications often describe the methods of analysis, this alone is not sufficient to serve all the many requirements for metadata of different kinds.

In order to fill this gap, organisations which provide data for reuse will typically have one or more standard models or formats for providing needed metadata to researchers, and for internal use in data management functions. The degree to which “internal” metadata is provided varies from organisation to organisation. What is often lacking entirely is the information which assumed to be known to potential domain users. This background, or context, is often known to researchers and data providers alike within the domain, and so is not necessarily seen as a priority for capture or dissemination within the systems which support secondary use of data. In a scenario where data is coming from sources in different domains, however, this can present a major challenge for reuse and data integration.

Further, different technical approaches within systems may produce barriers of a different kind, stemming from the technical implementation choices made by organisations. Often, these reflect the technology cultures within specific domains or infrastructures, caused by differences in applications used for analysis or for other purposes. These types of barriers may be surmountable by the researchers reusing or integrating data, but often involve a degree of effort which is substantial. (Ideally, any such technical differences could be overcome without specialized knowledge or even in a fully automated fashion, but this is not always the case.)
‘

This section will consider the kinds of metadata described in the preceding section, and the specific models which were encountered during the SP9 prototype work.

A. ESS and DDI Lifecycle

The European Social Survey is managed and disseminated using a system based on the Data Documentation Initiative Lifecycle (DDI-L) standard. The presentation of information is organized around the variables being used, so that a detailed view can be obtained for each variable. This includes the round of data collection to which the variable belongs. In integrated files (those containing more than one wave of data) the changes in categorical variables can be viewed across time. Each round of data collection is documented as a whole, including links to many related documents and providing information about the universe, time, countries involved, surveys used for data collection, methodology, etc. Country specific information is also provided, including sampling information, collection details, and extensive notes to facilitate cross-national comparison of the data. The data themselves are available, as are test data sets and various forms of paradata (data about the data collection).

For each variable, there is a detailed set of information provided, including the frequencies of each response, the countries in which the data was collected, the question text, the data type of the variable,

information about weighting, information about prompts used by the interviewer, and so on. Graphical visualizations of the data are provided, and can be downloaded for use in publications and so on.

What we see is a rich description of the data at the level of each round of data collection, along with detailed structural metadata for each variable. Additionally, there is user functionality for easily making data selections and download (the “Data Wizard”).

In two areas, the metadata is not as granular. We do not see the formal definitions of concepts presented as such on this site. The researcher is required to understand the concepts by looking at the survey instruments used to collect the data. In the social sciences, this is not an unreasonable expectation, since it is in the context of the questionnaire and the specific questions that the categories used as responses can best be understood. In order to access this information, however, the researcher will need to download the copies of the questionnaires and look for the specific questions. (The questionnaires are provided at the country level, which makes this an interesting feature of the ESS for cross-national comparison.)

The second area which is not defined at a granular level is the processing of the data. Here, we have documentation regarding the methods used for data editing and so on, but again these must be downloaded as documents and the specifics for any particular variable determined by the researcher from these sources.

For machine-to-machine access, there is a GraphQL API[28] accompanied by some technical documentation and the information model which can be queried. (GraphQL is a generic way of expressing queries – it can be used for almost any information model, but the queries themselves must be meaningful according to the system being queried.)

It should be noted that ESS is in the process of implementing a new dissemination system for the ESS, so that what is described here will have several features added in the future. Planned improvements include access to the metadata in XML form, according to the DDI Lifecycle standard at several different levels of granularity. Further, enhanced functionality for comparing variables across waves of data collection may be provided to researchers.

B. ERA5 Climate Data (Copernicus), GRIB, and NetCDF

The ERA5 data used in SP9 was downloaded from the page “ERA5 hourly data on single levels from 1940 to present”.[29] This resource provides a general overview of the data, as well as a listing of each of the major variables, with their definitions. In addition, there is some “quality assessment” information associated with each variable, which provides a rich set of information regarding how the re-analysis was performed, the methods used, validation, etc. Links are made to a large amount of technical documentation regarding the provenance of the data. There is a very thorough description of the geo-spatial aspects of the data.

The data itself is available in two standard formats: as GRIB (General Regularly-distributed Information in Binary form[30]), a format defined and maintained by the World Meteorological Organisation (WMO)

[28] https://graphql.org/
[29] https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview
[30] https://confluence.ecmwf.int/display/CKB/What+are+GRIB+files+and+how+can+I+read+them

Commission for Basic Systems (CBS[31]), and as NetCDF (the Network Common Data Form), which is maintained by the Open Geospatial Consortium (OGC[32]). For ERA5, the NetCDF distribution is listed as "experimental".

On the ERA5 site, there is a downloads selection functionality which guides the user in making a selection of the relevant variables according to topic, time, and geography.

The documentation available for the ERA5 data is impressive in terms of both its depth and its granularity. The type of information available is driven by the nature of the data itself: re-analysis data is obtained from a wide range of sources, and these each need to be described in order to understand the data in its entirety. While the overview does provide a link to an article describing the approach to producing the ERA5 re-analysis data as a whole, the researcher will need to dive into the specifics of each variable in order to use the data appropriately. This is, of necessity, an exercise requiring expertise in meteorology.

The standard formats are designed as data interchange formats, with a limited capability to express metadata. While NetCDF provides support for some functionality not found in the GRIB format (notably in terms of multi-dimensional or array based processing), neither format is intended to be documentary. The sheer volume of such data places a strong emphasis on the compactness of data for interchange purposes, and there is a separation between the metadata and documentation, and the data itself toward this end. GRIB data is intended to be "stand-alone" – that is, each value is described independent of others, so that the values can be re-packaged as needed. NetCDF organises the data into arrays. These different design aspects of the formats have implications for how the data can be processed for the purposes of integration.

Further, both formats require software designed specifically to work with them: unlike many XML- or RDF-based standards, there is no "open" expression of the data formats (although it should be noted that there are such formats for the metadata in the case of NetCDF). This is a consequence of the verbosity of the typical "open" formats, which make them unsuited for use in describing very large collections of data.

For the SP9 project, the NetCDF format was used, with freely available software tools used to access the data. This work was performed manually by the developers of the SP9 system, so that the data wanted for integration could be selected and presented in a form which could then be made integrateable through further processing.

For NetCDF, structural metadata is provided by the standard formats. Even though formal definitions of each field are provided, there is no machine-navigable link between the structural metadata and the semantics. Similarly, the information about provenance and methodology is very rich, but there is no automated way for consuming systems to connect this information to the data. In each case, the connections must be established manually by the researchers.

C. European Environment Agency (EEA) Air Quality Data

In SP9, data regarding air quality was obtained from the "Download service for E1a and E2a data".[33] This service provides a form which allows the user to formulate a query expressed as a URL. When resolved,

[31] https://public.wmo.int/en/our-mandate/how-we-do-it/technical-commissions/commission-basic-systems-cbs
[32] https://www.ogc.org/
[33] https://discomap.eea.europa.eu/map/fme/AirQualityExport.htm

the URL will provide the requested data. Parameters include country and city names, pollutants, start and end years, source, output (as HTML or text), as well as features for requesting updates from a specified date and time coverage (year or last seven days).

The service parameters are all described on the downloads page, along with links to relevant metadata expressed in spreadsheet form. Links are provided to many of the coding systems and definitions used in the data. The information regarding the sources of the data – that is, which stations provided the readings – is well documented, often as links to external sources.

Instructions for using the service to download data are clear, and easy to follow. Sample files are provided, to make it easier for users to understand how to work with the data resulting from a query.

There is a listing of all the available variables (“fields”) in the data files, as well as in the metadata spreadsheet. These are accompanied by basic descriptions, although not formal definitions. Links are provided to further documentation at a general level.

There is no standard model being used for this service, even though it is optimised for the repeat collection of air quality data, with features to make on-going updates easy to use in a machine-actionable way. The data is delivered as a CSV file which corresponds to the parameters embedded in the URL. While provenance information is well-documented (especially as regards the collection stations) this requires navigation by the researcher, or specific processes which are designed to read the spreadsheet formats proprietary to the application.

For SP9, the process of collecting data was automated, so that a script could generate a large number of small data files which could then be further processes. This script was developed manually to agree with the information provided on the service page, as described above. Once the many small files were collected, these were then aggregated into a single large file which included necessary metadata.

D. General Considerations

Across these cases, we see a general pattern emerging: standards and expertise are driven in large part by the nature of the data itself. In those cases where standard models exist, these reflect the needs of the domains which produced them. For the EEA data, the infrastructure rather than the domain dictates how the data is organized and made available.

Further, the type of metadata and documentation provided also reflects the concerns within the domain, and according to the kind of data being disseminated. For the ESS, there is a rich set of granular metadata regarding the structures of the data, and the context of its collection with in survey instruments. This reflects the reality of social science data – a single, formal definition would not be as meaningful to the researcher as the information provided regarding the question and survey instrument. For other data, the formal definitions or descriptions are provided, along with links to the wide range of data sources used, whether collection stations or other data sets used as input to the ERA5 reanalysis. Here, the granularity is not in the structural description of the variables within the data, but on the description of methods and provenance. This reflects the importance of the methods and mechanisms of collection, specific to the climate and air quality data.

In each case, there are significant challenges for the researcher in understanding the data, insofar as the needed information must be assembled and organised manually for the data of interest. While each of the

systems exposes some functionality in a machine-actionable fashion, there is no agreement on the models, formats, or technologies employed. Again, these appear to be specific to the domain or infrastructure, presenting an even greater challenge for cross-domain use of the data, where the standards or infrastructures may be unfamiliar.

V. The Researcher Perspective

If we are to make research using data from different domains and infrastructures easier for the researchers themselves, there are several possible approaches to consider. In the scenarios we described above, the current real-world situation is least problematic when researchers are using data coming from familiar sources within their own domain. What we see is that in general terms, there is a rich set of information available, and that for the purposes of a human user, the way in which the data are documented is accessible through the sites we have looked at. While there are no doubt improvements which could be made, they are incremental in nature: the ability of researchers to understand the data within their own domain is not a major barrier for our purposes.

The second scenario we consider is when a researcher is using data from another, unfamiliar domain, and has gone directly to the providers of that data. The barriers here are often based on the researcher's lack of expertise in the domain producing the data. Methods may be unfamiliar, and the background understanding of the data available to those who are regularly involved in its production, processing, and analysis may be missing. These problems are exacerbated by the unfamiliar systems and models used for the data and accompanying metadata. In this case, the researcher must turn to those with greater expertise in the domain – that is, to work collegially with those who do not share the same challenges. The barriers presented by unfamiliar models and systems could be solved through the presentation of the data in a familiar form, through automated transformations, etc., but this may also be beyond the ability of the researcher in question, as it is of a technical nature.

Our third scenario involves the accessing of data from outside the researchers domain through a familiar infrastructure. This is the case being prototyped by SP9, and it provides a way of overcoming the challenges of the second scenario. If the needed expertise and transformations are implemented by the infrastructure, and the data is presented to the researcher in a familiar fashion, then many of the issues may be addressed.

There are still some potential problems, however. A familiar domain infrastructure cannot provide an expert understanding of unfamiliar methods used in other domains to the researcher: the best that can be done is to provide links to the descriptive literature, and to assemble the information to make it easier for the researcher to learn what is needed. The reality here is that the expertise within a research project must include all of the needed domain knowledge, requiring collaboration across domain boundaries at the scientific level.

Ideally, however, the preparation of data *as it is presented to the researcher* can be fully described. Thus, those points where methodological questions are significant can be easily identified, and the expertise within a cross-domain researcher team can be leveraged appropriately.

In SP9, climate and air quality data are presented to the researcher in a form which is familiar to – and usable by – the social science researcher who is the intended audience for the European Social Survey. In order to present "integrateable" air quality and climate variables for use, assumptions are made as to how these data can be processed. The concept in play here is that of "fitness for audience" – unfamiliar

data presented in a fashion suited to its intended user audience, despite the challenges of data coming from outside the domain, employing unfamiliar metadata models, and suited for unfamiliar systems and forms of analysis. (The consideration of methods in SP9 has been described in another paper, also available from the ESS Labs site at https://www.europeansocialsurvey.org/esslabs/.)

These assumptions may not always be correct for a specific research purpose, and so any processing of the domain-external data must be fully described, so that it can be processed differently if that is required. This demands that a clear data lineage be provided, with links to the data at the various stages of processing, a description of the processes, links to the methods, and access to the scripts and program code used to perform it. If this picture connecting the domain-external source data can be presented to the research team, then appropriate consultation with experts from outside the “primary” researcher’s domain can be performed, and any changes needed in the data can be identified before analysis.

This is essentially a layer of metadata/documentation which does not currently exist in any of the data sources considered in the SP9 prototype. While there is documentation regarding the provenance of the data, it is documentary in nature, and requires a significant amount of manual work on the part of the researcher to assemble. What has been explored in SP9 is how such additional documentation can be made available in an easily usable form, so that the additional work required for cross-domain research is reduced to a minimum. Scientific collaboration can be conducted by a team of researchers with varied domain expertise much more easily if the problems of assembling the needed information do not also have to be solved.

Ideally, then, the data infrastructure supporting the primary domain user – for the ESS, the social scientists – would provide a reasonable baseline of “external” (non-domain) data, already in integrateable form, complete with the needed data lineage, so that any changes demanded by the specific research in question could be easily performed. This is thus the target for the data provider when consider support for this type of research: to provide cross-domain data that is fit for the audience being served.

VII. The Infrastructure Perspective

The challenges facing domain infrastructures interested in providing data from other domains to their primary audience are many. The problem itself is not a new one: the ESS serves as an example: “contextual” variables have been provided alongside the ESS data for many years. In the past, however, this data was not fully described: the integrateable variables would be presented with an indication of their source, but the processing to which they had been subjected was not fully documented. Given the nature of the data in question – often, aggregated data from sources such as Eurostat on topics with a demographic basis (i.e., economic data) – this was not a significant problem.

In SP9, however, the climate data and air quality data to be integrated is fundamentally different from the social data, and a simple indication of that data’s source is not sufficient to support effective use by researchers. If optimal support is to be provided to researchers, then a more-robust approach is required.

In order to provide the “data lineage” which is ideal for the researcher, the information available from the different sources of data must be harmonized and presented through a single interface, in combination with the familiar domain data.

It is important that domain-focused infrastructures retain ownership of their own data, and that when it is used, it is appropriately cited. In SP9, the researcher must be able to trace back from an “integrateable”

variable to the source data from which it was generated. The ESS may produce a “new” form of that data, for the purposes of integration, but it is necessary that the foundations of that data and the methods and processes used be evident.

In order to provide this data lineage to researchers, the data provider will need a suitable information model. While several different standard models currently exist, SP9 selected the DDI-CDI Process model as the most suitable one. (This is aligned with the popular W3C PROV ontology[34], which would be an alternative, but one which would require a lot of additional configuration to be useful.) DDI-CDI is designed to work in combination with other DDI metadata, specifically DDI Lifecycle. Since the ESS systems use the DDI Lifecycle model for their documentation, the DDI-CDI Process model was seen as a good candidate for the prototype.

The role of a model for describing data lineage is straightforward: it must describe the chain of steps in the process by which data was taken from different sources (ERA5, EEA air quality data, and the ESS itself) and processed into the form which is presented to the researchers. Links to relevant documentation are required, so that the documentation of the source data can be easily accessed by the researchers. The processes performed on the data must be described, and links provided to the code which was used in this processing. The process model acts as a way of bringing together the scientific methods and the specific implementation artefacts which implement those methods with the data being operated on and produced.

This model must then be presented to the researcher in a useful way. Because there has not historically been a major focus on presenting this type of information to researchers, there is no “typical” form of user interface in the way we see for applications like data catalogues. SP9 has explored how such a user interface can be designed. The process description prototype can be viewed at https://eosc-provenance.sikt.no/, and the source code found in the repository at https://github.com/sikt-no/ddi-cdi_process2web.

The value of having standards which cut across domain boundaries should be considered here. We have seen that different standards are typically used in different domains. When a domain infrastructure is presenting data and metadata from an “external” domain to its users, part of the service it performs is to translate these resources into a familiar, easily usable form. This can, however, present a challenge of scale to the organisation providing the data: for each new type of external data to be provided to its primary users, a new set of models and systems must be accommodated. When we think about the intended scope of initiatives such as the European Open Science Cloud (EOSC), it is clear that lessening this burden is a goal. Where possible, a domain-neutral standard, designed for use across domain boundaries, should be employed. (This topic is explored further in the “Looking Forward” section of this document.)
For SP9, the DDI-CDI Process model offered a domain-neutral way of describing the needed data lineage, and connecting the ESS data with the variables derived from external data sources. This model serves as the basis for the resulting “process browser”.

It should be noted that the process of integrating the air quality and climate data itself becomes part of the overall scientific process. Having a standard way of describing this integration, with links to all of the relevant data, metadata, and documentation will help in establishing good practice in this area, by making the methods employed by different infrastructure players more visible. While researchers see all of the data preparation as a precursor to analysis, regardless of who performs it, for the infrastructure which

[34] https://www.w3.org/TR/prov-o/

provides the service of generating “integrateable” variables for their primary users, there is a clear distinction between the integration work to provide the data, and the subsequent analysis the researcher performs. Having a standard description of this processing provides both necessary documentation to the researchers, and a potential resource for infrastructures looking at cross-domain issues.

When we consider standards, we should also think about not just how a single infrastructure disseminates data from non-primary domains beside its own data, but also how infrastructure players from other domains might use their data for a similar purpose. This is clearly in the interests of the data producer, as it allows for them to retain ownership of the data, even while it is being disseminated through services other than the ones it maintains. In order for this to work, all data and metadata should ideally be available in machine-actionable formats which use standards that are domain-neutral. This will apply to all types of metadata. Such standards will include the Simple Knowledge Organisation System (SKOS) for describing concepts, classifications, and codelists; DDI-CDI for structural and process metadata; DCAT (Data Catalog Vocabulary[35]) or Schema.org[36] for cataloguing metadata, and so on.

In SP9, the focus was on describing the data lineage in combination with the existing domain standard (DDI Lifecycle). While this approach was a practical one, it would be possible – using the kinds of domain-neutral standards mentioned here – to further optimize the ESS data for cross-domain use (see the “Looking Forward” section, below).

VIII. Trust and Transparency

The prototype work within SP9, and analysis of the data integrated as part of the project has highlighted several areas which are of heightened importance in cross-domain science. Notably, the way in which researchers work together, and the kinds of trust and visibility in the data they work with is different than when research follows a more traditional, domain-focused pattern.

These differences result from the scope of the research questions, and the mix of perspectives and expertise needed to answer them. While the requirements around metadata and the systems based on it have been explored, and the scientific methods they support have been addressed in a separate SP9 paper, the dynamics of trust and transparency which naturally result from the use of data across domain boundaries are worth considering.

One aspect of cross-domain research which gains importance is the reliance researchers place on the source of their data. In looking at different scenarios of use, the most practical one is where the infrastructure providing data to the researchers performs the initial steps to make data “integrateable”. Even when an effort is made by a domain researchers to understand the data coming from another domain, there are inherent issues with the knowledge required to effectively use unfamiliar data. A social scientist is not typically a climate or environmental scientist, and will lack the expertise surrounding the data which is needed to take it from its source and use it directly.

In SP9, the social scientist would rely on their familiar data provider – the ESS – to make climate and environmental data available in a form which is trustworthy and reliable. In part, this stems from the provision of sufficient information regarding the integration of these unfamiliar data. By making the process of integration transparent – revealing not just the sources of the data, but describing in detail the

[35] https://www.w3.org/TR/vocab-dcat/
[36] https://schema.org/

methods and processing used – the social scientist can have confidence in the integrateable data presented to them. It is useable in a way which is not true of the source data found by the providers in other domains, but benefits from the engagement the ESS has made with those providers in preparing the data.

This network of trust is not necessary in the same form for the domain data itself: ESS is respected within the social science domain based on merits which are already clear to social science researchers. They are qualified to look at the data and documentation provided by the ESS and to judge its quality and fitness for purpose. With data coming from unfamiliar sources and unfamiliar domains, they are forced to rely on the ESS to have integrated the climate and environmental data in a useable, useful fashion, because they lack the same familiarity with and expertise they have in their own domain's data. Both scenarios demand that the data be documented and of good quality, but the kind of trust asked of researchers is fundamentally different. There is a major emphasis placed here on transparency regarding how the domain-external domain data was integrated with familiar data for use.

From the perspective of the infrastructure players – the data providers – there is a need not only to gain the trust of researchers by providing a heightened degree of transparency, but also an interest in collaborating effectively with data providers, researchers, and research teams in other domains. The infrastructure players are both the best-qualified to help researchers use the data they produce and disseminate, and the ones most interested in making sure that it is maintained and disseminated properly. This logically extends also to working with other domains, to help them with their own data integration challenges.

The idea that data providers would engage with other data infrastructure players, providing data to other domains, is a new one. This has not traditionally been seen as a high-priority activity. In a scenario where EOSC sets the expectation that data sharing across domain boundaries is a practical reality, the infrastructure players will need to consider how best to engage with each other. The EOSC clusters (ESFRI thematic cluster projects[37]), and EOSC itself, provide a framework for defining such interactions, but there is as yet no typical process for doing so.

The need for data providers to retain the confidence of researchers in cases where data comes both from familiar, domain sources, and from "external", non-domain sources provides an incentive for establishing such collaborations. The data providers have a need to manage the potential reputational risk here – they need to establish that they are providing non-domain, "integrateable" data in a useable form, and to be trusted by their researchers - but also stand to provide an additional, valuable service to researchers, by providing access to data which is otherwise more difficult for them to access and use with confidence.

By collaborating with data providers in other domains, to ensure the accurate reuse of their own domain data, it is possible to mitigate the risk of being cited as the source of data which is subsequently mis-used, through the lack of expertise on the part of researchers in other domains. If climate scientists take data from the ESS and mis-use it, because they do not understand how it can be effectively integrated, the ESS could still be seen as providing "bad" data because the climate scientists have used it in an irresponsible or ill-informed fashion. Pro-active collaboration between data providers across the two domains would reduce this type of reputational risk. Researchers would also use data produced by such collaborations with greater confidence.

[37] https://eosc-portal.eu/esfri-thematic-cluster-projects

The key to establishing trust in these collaborations, however, is transparency: the provenance of the “integrateable” data must be clearly communicated. In order to be trustworthy, data must demonstrate that it is informed by a complete understanding of the data, with expertise coming from all relevant domains. This information must be made clear to the end user, with reference to the sites of the domain data providers.

Cross-domain research will produce findings which are likely to be challenged because of the methods used to integrate and analyze data coming from sources spread across domains, as these methods may be new and unfamiliar. Thus, the reproducibility of findings will gain increased importance. This also emphasizes the need for transparency regarding the integration processes and provenance of the data.

It should be noted that the requirements for the reproduction of findings in any research are typically greater than those of documenting even detailed provenance. It must be possible not only to understand what processes and methods were employed, but to actually perform the processes to achieve the same results. An even more complete set of provenance metadata is required. In the SP9 prototype, this distinction became clear. The way in which documentation of the integration process was performed was meant to meet the documentation requirements, but not those of full reproducibility of findings. (A paper describing the full workflow is another deliverable from SP9.)

IX. Looking Forward

A. Process and Provenance in Cross-Domain Data Use

The SP9 prototype explored the changes needed in how data are documented and processed to make them reusable and reliable for researchers working on cross-domain projects. One major discovery was the need for expanded metadata around the provenance of data, to help researchers overcome some of the barriers presented by the lack of familiarity with data coming from outside their own domain. To do this, the ESS Data Portal was used as the basis, and the process of making ERA5 data from Copernicus, and air quality data from the EEA available in a reusable form was implemented. This process description is at a granular, variable level, and it presents several different types of information in an integrated form, so that the researcher can navigate it in a coherent way through a single interface,

The sources of the non-ESS data are identified, and the methods used to create reusable, integrateable variables are thoroughly documented. These are tied to the program code used to execute this processing, as well as being documented in a form which allows the researcher to understand the processes in a step-by-step fashion. The end result is a set of data which is ready to be used by the social science researchers – the traditional audience for the ESS – when they wish to include data from external sources. This external data has been prepared by a cross-domain team of experts, so that it can be trusted, and reliably used by social scientists who may lack some of the needed expertise.

The prototype demonstrates the need for such an approach from a practical perspective: researchers must be able to approach cross-domain research questions within the context of EOSC without having to become domain experts in every area from which they draw their data. But it also highlights some additional issues and further steps which could not be implemented as part of the prototype, but which are worth considering for wide-scale adoption of such an approach to cross-domain research within large-scale data-sharing networks such as EOSC.

B. FAIR in Cross-Domain Data Integration

One such issue regards the nature of FAIR implementation itself. The FAIR principles are often viewed as applying primarily at the level of a domain – each of the sources of data used in the SP9 prototype could be viewed as FAIR to some extent, insofar as they employ domain standards, and are accessible in a machine-actionable fashion. While they may be FAIR within their traditional domains, it is clearly not as true in the case of cross-domain research: a significant amount of effort was required to bring the data together in a meaningful way so that it could be used for research, and this required that experts from each of the domains in question work together to overcome barriers presented by the need for domain-specific knowledge.

Ideally, the concept of FAIR could be applied across domains as well as within them, but this requires a set of standards which are understood and supported not only within individual domains and infrastructures, but across them. Specifically, we see this in the SP9 prototype.

The description of the integration process was implementing using the DDI-CDI standard, which is explicitly designed to express detailed information about workflows and processing in a way which accommodates data from different domains. Notably, it works directly with domain-specific standards to allow identification of referenced resources using their “native” identification schemes. This description of the process/workflow was both implementable (as we see in the process browser) but is also a machine-actionable set of metadata which could be consumed by other applications for different purposes.

The metadata for the ESS was utilized in its existing form: DDI Lifecycle. This is a domain-specific standard which is well-suited for the description of the ESS data, at both the data-set level and at the variable level. While this can be understood as a FAIR standard, it is not one which is likely to be understood outside of the social science domain. What is needed are cross-domain standards for the structural data description and for the definitional metadata. DDI-CDI also provides a description of data in a domain-independent way, at the variable level, and this could be used – because DDI Lifecycle and DDI-CDI model concepts and variables in a similar way, such a transformation would be straightforward. For data-set-level description, cross-domain standards such as DCAT or Schema.org could be employed. For the description of definitional metadata, generic standards such as SKOS and OWL would be needed.

All of these metadata standards would have to be used in combination, to provide a full set of descriptive information which could be used within any domain. Further, the metadata they provide would need to be implementable in systems for researchers to use, and also be available in machine-actionable formats for easy exchange and use within systems. This approach – the use of agreed, cross-domain standards – is being explored in several projects which are looking at FAIR implementation, including the EOSC Interoperability Framework[38], WorldFAIR[39], FAIR Impact[40], and in the FAIR Digital Object Framework.[41] It was not the intent of SP9 to determine which standards are best for this specific purpose, but to explore the metadata requirements from all the different perspectives, and to examine how these are related to the needed scientific expertise to support cross-domain research in a practical fashion.

[38] https://eosc-portal.eu/eosc-interoperability-framework
[39] https://worldfair-project.eu/
[40] https://fair-impact.eu/
[41] https://fairdigitalobjectframework.org/

C. Common Standards for Cross-Domain Metadata Exchange

While the existence of a common set of standards for exchanging metadata in all of these areas across domains does not address the requirement for cross-domain expertise, it does provide a consistent basis for finding the needed information. If all of the data required for a project like SP9 was organized and modeled consistently, the resources needed to answer the specific scientific questions around the integration (and integrateability) would be reduced. In SP9, this required a lot of specific effort at all levels, including the scientific experts, data managers, and technical implementers. The consistent presentation of the needed metadata would have materially reduced the effort needed.

Even better for reducing this effort would be the establishment of standard APIs for the exchange of this information. Such APIs assume an agreement between counterparties on the information model – that is, the standard models described above – but would further establish standards for the interfaces between applications. There are new technologies (such as GraphQL) as well as established ones (such as REST[42]) which could be employed here. The development of standard APIs for exchanging metadata between domains is one which will be a good topic for further investigation in the future.

D. The Institutional Framework

It is important to consider the implications for data providers. There is today no expectation that data infrastructures will engage across domain boundaries: each data provider serves their own users in their own domain. SP9 shows that the data providers must work together across domain boundaries in order to better support researchers who wish to do multi-disciplinary research.

This has a technical aspect, which could be addressed through the use of an agreed set of cross-domain FAIR standards, and some of these have been mentioned. There is a further issue, however, which demands consideration: what is the institutional framework within which cross-domain collaboration can take place? We have described above the reasons why an infrastructure wishes to be seen as a trusted source of integrateable data by the researchers, but how can data providers do this not only for a specific area, but more broadly? SP9 addresses one important topic regarding climate-neutral cities, but that is a relatively narrow focus, and a much wider range of areas would ideally be supported.

EOSC can serve as a framework for facilitating cross-domain collaboration at this level, but the problem has several different aspects. On one hand, data providers must recognize an expansion in their role: they will be required to engage across domain boundaries as an accepted part of their business, and cannot operate only narrowly within their domain. Participation in cross-domain teams, like the one assembled for the purposes of the SP9 prototype, needs to become a regular function among data providers.

This in turn demands that resourcing be dedicated to this expanded function. Data providers cannot be expected to add to the already significant burden of serving their research communities without the expanded capacity to do so.

What EOSC specifically needs to provide is an operational framework within which such collaboration can take place, so that needed resources can be identified, and the roles and responsibilities of the players

[42] Representational State Transfer - https://en.wikipedia.org/wiki/REST

from different data providers can be understood. Both the EOSC clusters and EOSC at the highest level will need to support the identification of high-priority areas for cross-domain research, and engage in establishing the normal process for supporting them. This is a strategic vision, and is well beyond the specific scope of SP9, but the question has been clearly identified as an important one through the work within the project.

The ultimate goal of EOSC is to enable the ubiquitous sharing of data and related resources. SP9 has shown that this can be done in a way that heightens the transparency of the process, a necessary condition for building the needed trust among researchers, and thus confidence in their research findings. SP9 further demonstrates that reproduction of findings is not impossible with a sufficient degree of process documentation, even if the prototype within the project aimed only at the documentation of process for the purposes of provenance.

This demands standardization at a technical level, across domain boundaries, but also coordination at a scientific level: we cannot support multi-disciplinary research without making it possible for teams of experts coming from all of the concerned domains to collaborate effectively. This requires both an organizational framework, and the resourcing to make collaboration practically feasible. The existence of a standardised technical framework could serve to reduce the needed resources. It is clear that EOSC will need to address all of these issues if the long-term vision around data sharing is to be realised.