D6.6: 2nd Report on Data Interoperability

Author(s)	Ari Asmi (ICOS-ERIC), Bas Cordewener (JISC), Carole Goble (ELIXIR - UNIMAN), Donatella Castelli (CNR), Eileen Kühn (KIT), Fabio Pasian (INAF), Franco Niccolucci (UFlorence), Helen Glaves (UKRI-NERC), Keith Jeffery (UKRI-NERC), Massimiliano Assante (CNR), Matthew Dovey (JISC), Natalia Manola (Athena RC), Nick Juty (ELIXIR - UNIMAN), Niklas Blomberg (ELIXIR - EMBL), Ricardo Arcila (ELIXIR - EMBL), Rafael Jimenez (ELIXIR - EMBL), Volker Beckmann (CNRS), Giorgos Papanikos (Athena RC)
Status	Final
Version	v1.2
Date	29/06/2017

Dissemination Level

X	PU: Public
	PP: Restricted to other programme participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

Abstract:
The objective of the EOSCpilot data interoperability task (6.2) is to demonstrate how to ensure availability of scientific data to users and services through an open cloud infrastructure. To do so, this task has produced a first draft of the strategy and recommendations to help users and services to find and access datasets across several scientific disciplines. Four data interoperability demonstrators have been proposed to test components of the strategy. This report provides a status update and highlights recommendations from the demonstrators and feedback from EOSCpilot partners.

The European Open Science Cloud for Research pilot project (EOSCpilot) is funded by the European Commission, DG Research & Innovation under contract no. 739563

Document identifier: EOSCpilot -WP6-D6.3	
Deliverable lead	**ELIXIR**
Related work package	**WP6**
Author(s)	Ari Asmi (ICOS-ERIC), Bas Cordewener (JISC), Carole Goble (ELIXIR - UNIMAN), Donatella Castelli (CNR), Eileen Kühn (KIT), Fabio Pasian (INAF), Franco Niccolucci (UFlorence), Helen Glaves (UKRI-NERC), Keith Jeffery (UKRI-NERC), Massimiliano Assante (CNR), Matthew Dovey (JISC), Natalia Manola (Athena RC), Nick Juty (ELIXIR - UNIMAN), Niklas Blomberg (ELIXIR - EMBL), Ricardo Arcila (ELIXIR - EMBL), Rafael C Jimenez (ELIXIR - EMBL), Volker Beckmann (CNRS), Giorgos Papanikos (Athena RC)
Contributor(s)	Yin Chen (EGI), Brian Matthews (UKRI-STFC) Nuno Ferreira (SURFsara), Juan A Vizcaino (EMBL-EBI), Henning Hermjakob (EMBL-EBI), Heinrich Widmann (DKRZ), Vicky Schneider (Amazon), Susanna A Sansone (University of Oxford), Peter McQuilton (University of Oxford), Sarala Wimalaratne (EMBL-EBI), Cristina Duma (IFN), Valentino Cavalli (LIBER)
Due date	**30/06/2018**
Actual submission date	**29/06/2018**
Reviewed by	Yin Chen (EGI), Brian Matthews (UKRI-STFC) and Nuno Ferreira (SURFsara)
Approved by	Mark Thorley (UKRI-STFC)
Start date of Project	**01/01/2017**
Duration	**24 months**

Versioning and contribution history

Version	Date	Authors	Notes
1.0	15/06/2017	Rafael C Jimenez (ELIXIR - EMBL) and Nick Juty (ELIXIR - UNIMAN)	First draft taking into account the work done by the data interoperability demonstrators and

			feedback provided by the partners
1.1	22/06/2017	Ari Asmi (ICOS-ERIC), Bas Cordewener (JISC), Brian Matthews (UKRI-STFC), Carole Goble (ELIXIR - UNIMAN), Donatella Castelli (CNR), Eileen Kühn (KIT), Fabio Pasian (INAF), Franco Niccolucci (UFlorence), Helen Glaves (UKRI-NERC), Keith Jeffery (UKRI-NERC), Massimiliano Assante (CNR), Matthew Dovey (JISC), Natalia Manola (Athena RC), Nick Juty (ELIXIR - UNIMAN), Niklas Blomberg (ELIXIR - EMBL), Ricardo Arcila (ELIXIR - EMBL), Rafael C Jimenez (ELIXIR - EMBL), Volker Beckmann (CNRS), Giorgos Papanikos (Athena RC)	Second draft including feedback from the partners.
1.2	29/06/2017	Yin Chen (EGI), Brian Matthews (UKRI-STFC) and Nuno Ferreira (SURFsara)	Reviewed version and final draft including feedback from reviewers.

TABLE OF CONTENT

1 EXECUTIVE SUMMARY

The objective of the EOSCpilot data interoperability task (6.2) is to demonstrate how to ensure availability of scientific data to users and services through an open cloud infrastructure. To do so, this task has produced a first draft of the strategy and recommendations to help users and services to find and access datasets across several scientific disciplines. Four data interoperability demonstrators have been proposed to test components of the strategy:

- Evaluation of the EDMI[1] metadata guidelines to find and access datasets
- Discovery of compliant data resources and metadata catalogues
- Research schemas for exposing dataset metadata
- Description and guidelines per metadata property

This report provides an update and highlights suggestions from the demonstrators, and feedback from EOSCpilot partners to shape the direction of the EOSCpilot data interoperability strategy. The first report on data interoperability was specially dedicated to work on Findability and Accessibility. In the second report we work on demonstrators aimed to shed more light on aspects of Interoperability and Reusability. This report especially focuses on feedback about how to expose EDMI properties, discover EDMI compliant resources, provide guidelines for describing metadata properties and establish an ecosystem of metadata catalogues.

[1] EOSC Datasets Minimum Information

2 INTRODUCTION

The objective of the EOSCpilot 'data interoperability' (task 6.2) is to demonstrate how to ensure the availability of scientific data to users and services through an open cloud infrastructure. EOSCpilot, and specifically the EOSCpilot task 6.2, produced a first draft of the strategy and recommendations to help users and services to find and access datasets across several scientific disciplines. This strategy is described in more detail in the EOSCpilot "1st report on Data Interoperability"[2].

The strategy relies on three main ideas:

- Agreeing on a common and minimum dataset metadata properties to be exposed by data resources.
- Supporting a coordinated ecosystem of dataset metadata catalogues[3] which work together to efficiently manage and exchange their metadata.
- Demonstrate the applicability of these recommendations by implementing them in data resources to allow user facing services to find and access data.

These recommendations are driven by a set of guiding principles which can be grouped into three categories:

- Reuse - Leverage upon the rich legacy of Research Infrastructures.
- Least - Converge upon the minimum set of metadata from which we can derive maximum benefit.
- Practical - Recommend solutions that are sustainable, pragmatic and easy to deliver.

As a result, three main components have been proposed to implement this strategy:

- EDMI (EOSC Datasets Minimum Information) metadata guidelines - A set of metadata properties[4] and a metadata crosswalk (equivalence) across existing metadata models.
- Metadata catalogues strategy - Recommendations about how to support an ecosystem of metadata catalogues.
- Demonstrators - A set of demonstrators to validate and iteratively improve the proposed recommendations.

The work planned in the EOSCpilot 6.2 task is divided in 3 phases aligned with the 3 deliverables proposed for this task:

- Draft strategy - During the first phase we worked on a draft strategy based on the feedback collected from partners, from a series of open community meetings. The resulting defined set of principles guided the scope of our work. The first phase was specially dedicated to work on Findability and Accessibility. This was reported in the 1st Report on Data Interoperability[5].
- Review strategy - During the current second phase, we have been evaluating the draft strategy. To do so we have proposed 4 internal demonstrators to test components of the strategy. This 2nd Report on data Interoperability provides a status update, and reviews aspects of the strategy based

[2] https://eoscpilot.eu/themes/wp6-interoperability/1st-report-on-data-report-findability-interoperability

[3] The ecosystem of dataset metadata catalogues is introduced in the EOSCpilot "1st report on Data Interoperability", in section "5.2 Better coordination among existing dataset metadata catalogues".

[4] See "Annex H" for more information.

[5] https://eoscpilot.eu/themes/wp6-interoperability/1st-report-on-data-report-findability-interoperability

on the work done in these demonstrators. In the second phase the demonstrators aimed to shed more light on aspects of Interoperability and Reusability.

- Propose final strategy - During the next phase, we aim to make a final EOSCpilot data interoperability strategy proposal based on the results of demonstrators, together with how this strategy aligns with the outcomes of other EOSCpilot tasks, and related community initiatives.

This document starts with this introductory chapter and continues defining the goal, scope and tasks proposed in the data interoperability demonstrators. The following two chapters highlight feedback and recommendations from the demonstrators, and feedback collected from partners. We finish with a conclusion chapter to guide the final work of this task.

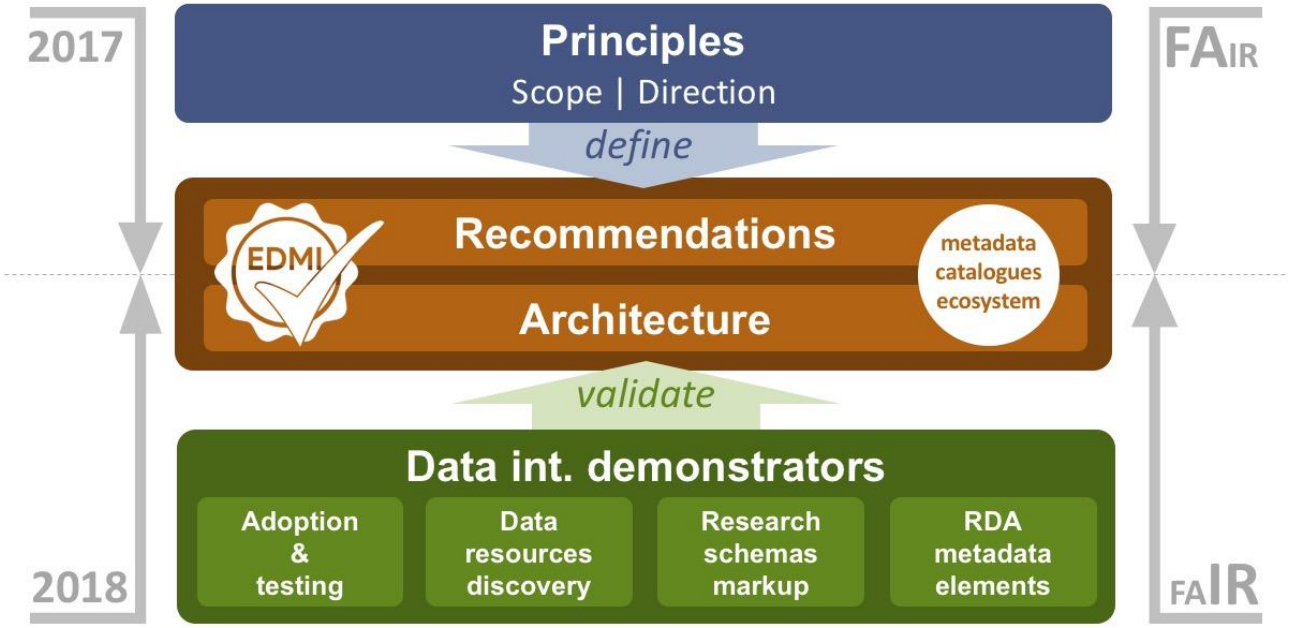

Figure A. "EDMI" and the "ecosystem of metadata catalogues" are the core components of the EOSCpilot data interoperability strategy. The EOSCpilot data interoperability principles helped to define the scope and direction of the recommendations and architecture proposal. The four data interoperability demonstrators aim to validate and refine the proposal. This initial work was focused on improving the findability and accessibility of datasets, while subsequent work done through the demonstrators aims to extend on the aspects relating to interoperability and reusability.

3 The EOSCpilot Data Interoperability Demonstrators

The data interoperability demonstrators aim to test and evaluate the feasibility of the recommendations proposed in the EOSCpilot data interoperability task. Four demonstrators were proposed as a result of the discussions and feedback collected during the three EOSCpilot data interoperability workshops organised in 2017. The feedback from these demonstrators will inform the necessary changes and improvements required in order to have a practical data interoperability strategy by the end of the EOSCpilot project. The demonstrators, initiated in February 2018, are also supported by the EOSCpilot task 6.3 (Interoperability Testbeds). The data interoperability demonstrators aim to be very inclusive, counting on partners as well as

other stakeholders interested on testing and participating in the strategy proposed by task. These four data interoperability demonstrators are:

- Evaluation of EDMI metadata to find and access datasets
- Research schemas for exposing dataset metadata
- Discovery of compliant data resources and metadata catalogues
- Description and guidelines per metadata property

3.1 Evaluation of EDMI metadata to find and access datasets

For services it is not easy to find, access, transfer and keep updated copies of data hosted by third party data resources. It is challenging since there are many data resources, often highly distributed, and employing different data models and a diversity of access interfaces. Operational metadata, provided at the level of a dataset, would help services to efficiently access data. The EOSCpilot data interoperability task provided metadata recommendations and a strategy to make datasets from third party resources more findable and accessible for services. This demonstrator aims to test and get feedback on the functional and operational metadata proposed in the EDMI guidelines, particularly on how it will help services to find, access, transfer and replicate data available in third party data resources. This demonstrator involves participation from data resources, dataset metadata catalogues and services, which will be used to test EDMI. This task explores how compliance of EDMI properties can be measured and could be used to evaluate 'FAIRness'. The following activities were proposed for this demonstrator:

3.1.1 Expose EDMI metadata

This activity is about engaging data resources and dataset metadata catalogues to expose EDMI metadata such that it can be accessed programmatically. This will be done reusing existent programmatic interfaces provided by the resources or adopting an existing standard or interface (eg. schema.org or ResourceSynch[6]). The metadata should be exposed in such a way it can be used programmatically by services.

3.1.2 Index EDMI metadata

This activity engages dataset metadata catalogue providers to test the indexing of EDMI metadata that has been exposed by data resources. This activity will specifically look at collecting/indexing the EDMI metadata exposed by data resources involved in the activity described in 3.1.1.

3.1.3 Use EDMI metadata

This activity is about testing how useful EDMI dataset metadata is for consumption by programmatic services, to find and access the datasets that are available from data resources. This activity will specifically look at the metadata exposed by dataset metadata catalogues from the activity described in 3.1.2.

3.1.4 Explore how to monitor compliance of EDMI

EOSCpilot activity 3.2 is working on metrics and ways to monitor data resources. Since EDMI compliance can be measured, it could be used to evaluate 'FAIRness' of datasets (and of data resources), especially focusing on Findability and Accessibility. This activity is about exploring with EOSCpilot task 3.2 how to measure and monitor EDMI compliance.

[6] http://www.openarchives.org/rs/toc

3.2 Research schemas for exposing dataset metadata

As highlighted in our survey findings[7], around 30% of the metadata catalogues do not provide a programmatic interface that could help services to find and access data. Furthermore, they do not provide all the properties which are considered minimum in our EDMI recommendations. This demonstrator focuses on the need to provide a simple and quick way to implement a solution which allows metadata catalogues to expose this structured metadata. Schema.org (Mika 2015) provides a simple mechanism to expose structured metadata on datasets, using the existing web interfaces of metadata catalogues and data resources. We would like to explore how to use Schemas.org, in a manner akin to that used by Bioschemas[8], to facilitate exposing EDMI metadata properties through data resources and through dataset metadata catalogues. Part of the goal is also to drive adoption, which could be facilitated by providing several examples of use, and by demonstrating how a dataset metadata catalogue can index schema.org metadata provided by a data resource. The following activities were proposed for this demonstrator:

3.2.1 Expose EDMI properties using schema.org

This activity is about engaging with data resources to expose EDMI metadata properties using schema.org. Though many metadata catalogues and data resources are already exposing structure metadata via programmatic interfaces, we suggest they additionally implement Schema.org to expose dataset structured metadata, in compliance with EDMI guidelines.

3.2.2 Harvest EDMI metadata exposed in Schema.org

We would like to demonstrate how dataset metadata catalogues can harvest EDMI metadata exposed with Schema.org by data resources and dataset metadata catalogues.

3.3 Discovery of compliant data resources and metadata catalogues

Catalogues of datasets index and integrate metadata from data resources making it easier for users and services to find datasets. However, it is more difficult to ascertain which data resources have been indexed by a particular metadata catalogue and which resources are compliant with EDMI. This demonstrator aims to help users and services to better find existing catalogues and the data resources indexed by these catalogues. It also aims to recognise which catalogues and data resources comply with EDMI. To do so it aims to involve an existing metadata catalogue of data resources to:

1. index and highlight dataset metadata catalogues compliant with EDMI,
2. index and highlight data resources compliant with EDMI and
3. make the link between dataset metadata catalogues and data resources.

3.3.1 Create a collection of data resources per dataset catalogue

This activity is about facilitating the discovery of dataset catalogues and the data resources they index by creating collections per dataset catalogue in FAIRsharing. This activity will specifically look at the dataset metadata catalogues identified in the first demonstrator.

[7] The survey summary is introduced in the EOSCpilot "1st report on Data Interoperability", in section "2.8 Survey". https://eoscpilot.eu/themes/wp6-interoperability/1st-report-on-data-report-findability-interoperability

[8] http://bioschemas.org/

3.3.2 Highlight resource compliance with EDMI

This activity is about creating an entry for the EDMI recommendations as a 'MI' (minimum information) guideline in FAIRsharing, and to highlight data resources and metadata catalogues compliant with the EDMI guideline, by 'tagging' them in the FAIRsharing data catalogue[9] (Sansone et al. 2018).

3.3.3 Strategy to expose indexed data resources

Dataset metadata catalogues index several data resources. Is there a programmatic way to know which resources are being indexed by a particular data catalogue? This activity is about proposing a simple strategy for dataset metadata catalogues to expose the list of resources they are indexing.

3.4 Description and guidelines per metadata property

The objective of the RDA Metadata Interest Group (MIG)[10] is to provide detailed descriptions on individual metadata properties, including those of use for describing datasets, and to provide recommendations for their use. This EOSCpilot data interoperability task seeks to contribute to this global initiative, and reuse those descriptions and recommendations which focus on the metadata properties identified by EOSCpilot as being part of EDMI. The goal of this demonstrator is therefore to collaborate with the RDA MIG group to describe and provide recommendations for dataset metadata properties.

3.4.1 Contribute to the RDA MIG guidelines

This activity aims to engage with the RDA MIG working group to define guidelines for dataset properties. Participants of this activity will focus on a set of dataset properties with special emphasis on EDMI properties that need clearer and comprehensive definitions, such as 'identifier'. These guidelines should aim to provide generic as well as domain specific recommendations and examples of use.

3.4.2 Gap analysis and proposal of new properties

This activity focuses on a gap analysis, comparing EDMI and RDA MIG dataset metadata properties, with a view to introduce important operational metadata properties to RDA MIG that are potentially missing in the latter.

4 Feedback from The EOSCpilot Data Interoperability Demonstrators

Some of the activities planned for the data interoperability demonstrators are ongoing activities and some of their results will not be reported until the end of the project. However, we have already some preliminary results which are helping us to take some actions. The list of recommendations presented in this section are meant for the participants of the EOSCpilot data interoperability task to help shaping the final EOSCpilot data interoperability strategy.

4.1 Exposing EDMI properties

During a workshop organized in Pisa (EOSCpilot All Hands meeting, March 9th 2018), the topic of exposing EDMI metadata was discussed. Mappings were identified between various catalogue specific schemas and EDMI properties, as well as schema.org types that can be used to present the respective information.

[9] https://fairsharing.org/bsg-s001135/

[10] https://www.rd-alliance.org/groups/metadata-ig.html

Based on workshop material gathered for the purpose, some metadata catalogues volunteered to provide examples showcasing the use of Schema.org and EDMI properties. The Examples were provided for the following data resources:

- OpenAIRE – Linked research publications, project, dataset and author information
- BlueBRIDGE – Earth Observation datasets
- EBI Metagenomics – Life Science domain
- DataCite - Research datasets
- PRIDE - Life Science domain
- OMICsDI - Life Science domain

Examples of usage and output can be found at GitHub[11]. We are working on a We are working on a GitHub repository to collect guidelines and examples to facilitate community feedback and engagement. The GitHub repository will include documentation about EDMI as well as examples and descriptions of tools and resources that support EDMI. A preliminary site is available at https://eosc-edmi.github.io/.

4.1.1 Remarks

- Though this exercise focused primarily on the minimum EDMI metadata properties, there were many cases where the metadata catalogues could not provide all the minimum metadata.
  Recommendation 1: The minimum properties should not be considered a mandatory set but as an ideal state to facilitate findability and accessibility. EDMI should include a core set of fewer properties easy to comply with. This way we could encourage providers to move from "Core"[12] to "Minimum" and enrich the metadata with "Recommended" properties. "Core" properties could be aligned to the DataCite mandatory properties.

- The examples proved to be very useful. Currently, most of the examples use Schema.org. We need to focus on providing a variety of examples of how to expose EDMI using other standards.
  Recommendation 2: Look for more examples using other competing standards like DATS, DataCite, DCAT and other domain specific standards such as CERIF and W3C HCLS, and provide mappings of equivalence to enable users to move between different solutions.

- After evaluating the mappings between EDMI metadata guidelines and the schema.org dataset type we found some EDMI properties did not have schema.org equivalent properties[13]. This can be addressed through extensions to schema.org or through the use of other schemas or vocabularies where a property is missing.
  Recommendation 3: Make a proposal and show examples of how to expose an EDMI property when a property is missing in an existing standard.

- We identified inconsistencies in the way dataset properties such as identifier, access rights and licenses are represented. Controlled vocabularies and specific recommendations for these

[11] https://github.com/madgik/schema2jsonld

[12] We use the term "Core" but it could be any other term we chose to describe a more restrictive set of properties.

[13] Some of the schema.org/Dataset properties selected are still not part of the core schema and are in pending state (e.g. http://pending.schema.org/measurementTechnique & http://pending.schema.org/variableMeasured). Schema.org also has a way to include list of terms relevant to specific domains using DefinedTerms https://dataliberate.com/2018/06/18/schema-org-introduces-defined-terms/

properties could help to increase interoperability.
Recommendation 4: Keep working on the RDA metadata guidelines specially focusing on identifiers, access rights and licenses.

4.2 A conversion tool to help exposing EDMI with Schema.org

In the process of practically examining the feasibility of exposing metadata (EDMI and schema.org) from metadata catalogues, a simple tool was created to handle the transformation between metadata catalogue specific schemas to structured schema.org structured metadata. This tool handles the crosswalk to EDMI properties through catalogue specific transformations, maps the extracted information to EDMI profile schema.org/Dataset properties and generates a JSON-LD document representing the extracted information. The tool uses the notion of profiles to distinguish between the supported metadata catalogues, enabling a single service instance to act as gateway for various metadata catalogues. Each registered profile brings along the catalogue and schema specific logic that is required to access the metadata and perform the required transformations from domain specific schema to EDMI properties. From then on, the process of generating the JSON-LD document is shared and reused across various implementations. The tool, as well as examples of usage can be found at GitHub[14].

4.2.1 Remarks

- The conversion tool offers a quick way to have several catalogues exposing EDMI properties with schema.org. This can be a practical intermediary solution to get all the dataset catalogues exposing metadata the same way while the catalogues evaluate the active uptake of schema.org to present EDMI properties. This could facilitate not just the consumption but the validation of EDMI properties.
  Recommendation 5: Consider the Schema.org conversion tool as a way to quickly get adoption and showcase the benefits of Schema.org and EDMI.

- As a value added side effect, given the nature of the schema.org usage, the metadata catalogues that choose to enhance the data they expose with the JSON-LD schema.org structured documents will gain better findability not only in the context of EOSC enabled services, but also through widely used search engines.
  Recommendation 6: Encourage the adoption of Schema.org and compliance to EDMI in EOSC data resources.

- Other tools like the “Mapping Memory Manager (3M)” are used by RI communities for managing mapping definition files. These tools could also be used for mapping existing schemas to EDMI. https://github.com/isl/Mapping-Memory-Manager

4.3 Discovery of EDMI resources

We have evaluated how to display data resources and their compliance with EDMI in FAIRsharing. We have started looking at the data resources used in the EOSCpilot scientific demonstrators and the data catalogues identified by the EOSCpilot data interoperability task. So far we are working in collaboration with WP5 to identify and register the data catalogues and data resources participating in EOSCpilot. We

[14] https://github.com/madgik/schema2jsonld

have also created an EDMI record in FAIRsharing (https://fairsharing.org/bsg-s001135) which we plan to link to EDMI compliant resources.

4.3.1 Remarks

- The data model of a data resource might be compliant with EDMI minimum properties however the datasets within might not expose all the metadata properties defined in the model. Should the EDMI compliance for a data resource reflect the compliance of the model, the compliance of the content, or both?
  Recommendation 7: Make a proposal of how to display compliance based on the data resource model and based on the metadata content.

- Evaluation of the EDMI compliance can be achieved at the level of the dataset, but not the level of the data resource which could have several datasets with different levels of compliance.
  Recommendation 8: Think about how to show compliance for data resources. For instance for each data resource we could select a few datasets that we could evaluate to identify the compliance profile of data resources.

- At the moment compliance can be evaluated manually, but an automated method would be desirable.
  Recommendation 9: Work with EOSCpilot WP3 on how to monitor compliance with EDMI. EOSCpilot WP3 is working on defining and developing the EOSC Open Science Monitor Framework that could help to evaluate the compliance with guidelines such EDMI.

- In the examples mentioned in 4.1 we realised the metadata exposed by data resources do not comply with all the minimum EDMI properties. Many resources are compliant with the functional minimum set but not the operational.
  Recommendation 10: The minimum set should be seen as a goal to achieve. To be more inclusive and promote EDMI we recommend EDMI to have a subset of properties which could be core (or mandatory). Thus, we could display several levels of compliance (Core, Minimum and Recommended and Optional) as suggested in 4.1.1.

4.4 Description and guidelines per metadata property

Metadata properties and metadata models are of interest to numerous international projects and interest groups (RDA, Force11, BD2K, BioCADDIE, etc), which may result in duplication of effort. There are a number of metadata properties that are common to different metadata models (for example, DataCite). These properties, while conceptually identical, seem often to be defined incongruously, leading to potential ambiguity in their assignment. Additionally, since these properties are being earmarked for use across efforts internationally, the precise details around what each property entails (with respect to attributes for the property, and what constitutes valid values for these attributes) becomes more important to ensure downstream interoperability between these efforts.

During the EOSCpilot All Hands meeting (March 9th 2018, Pisa) a workshop was organized to prioritise the RDA property metadata guidelines on which to focus (and improve). ‘Name’, ‘description’, ‘identifier’ and ‘license’ were the four EDMI properties identified by the community as being important. From these four properties, the ‘identifier’ and ‘license’ properties were voted to be the most crucial for improvement.

Consequently, we are producing a draft proposal for both of these properties. These drafts are including the feedback from EOSCpilot WP6 and other participants of the EOSCpilot All Hands workshop. In addition, as part of this proposal drafting exercise, we are defining a common structure to organise the content of these and future property guidelines. The current draft of the guidelines are open for comments[15] (Annex K include a copy of the current draft of the license as an example).

4.4.1 Remarks

- The current drafts for License and Identifiers are a good start. However, for consensus, further feedback is required to reach an appropriate level of maturity. While some guidelines are quite simple like 'Name' or 'Description', others like 'identifier' require more work, and need input from experts from different domains.
  Recommendation 11: Engage metadata experts from different communities. Identify the most challenging properties and look for existing projects and communities willing to contribute to define the guidelines. eg. For identifiers: FREYA, ELIXIR identifiers, RDA identifiers, etc.

- The current content structure proposed for the guidelines might evolve based on the feedback for other properties.
  Recommendation 12: Make sure we have a template with the structure and we update the rest of the properties with changes.

5 REVIEW OF DATA INTEROPERABILITY RECOMMENDATIONS AND STRATEGY

The aim of this section is to provide an update about the strategy draft presented in the 1st Report on Data Interoperability[16]. This update is based on the feedback collected from partners and the data interoperability demonstrators, and focuses on the recommendations about the strategy of metadata catalogues and datasets for EOSC.

5.1 Metadata catalogues, data repositories and datasets

The 1st Report on Data Interoperability included a description of the terminology used when referring to "Metadata catalogues, data repositories and datasets" and provided an overview of the main stakeholders involved in the process of data sharing.

5.1.1 Remarks

- The terminology used in 1st report has been integrated into the EOSC glossary[17]. Sixteen terms described by the EOSCpilot data interoperability task have been incorporated so far. Terminology defined in the glossary by other work packages has not yet been used in the EOSCpilot data interoperability documents.
  Recommendations: The final EOSCpilot data interoperability strategy needs to be consistent and reuse terminology defined in the EOSC glossary.

[15] https://drive.google.com/drive/folders/1F6uRaofJRYTLKg233hZULrlYXUON3CLU

[16] https://eoscpilot.eu/themes/wp6-interoperability/1st-report-on-data-report-findability-interoperability

[17] https://docs.google.com/spreadsheets/d/1NeKhxiAOAESkZFAmFbVKWhT_JQ3ENGTKhO26csmx8L8/edit?usp=sharing

5.2 Metadata catalogues and datasets in EOSC

The "Metadata catalogues and datasets in EOSC" section within the 1st Report on Data Interoperability[18] provided a review of existing metadata catalogues and highlighted the role of data resources and datasets in EOSC. It emphasised the importance of relying on a common minimum information standard for dataset metadata (EDMI guideline), which respects existing community practices, formats and user interfaces.

5.2.1 Remarks

- EDMI is presented as a Minimum information guideline, however users become confused and see EDMI as a new metadata standard.
  Recommendation 13: In the final recommendations, it must be made clear that EDMI aims to be a crosswalk guideline, encouraging the use of existing standards to describe datasets like DataCite or DCAT for generic datasets, and CERIF or HCLS for domain specific datasets.

- The EDMI guidelines are minimum information guidelines designed to agree on a minimum set a properties to make datasets findable and accessible by humans and machines. Other minimum information guidelines might have other purposes and could be easily combined with EDMI. For instance, the minimum information guidelines of DataCite consider citation an important aspect of the metadata. More domain specific information guidelines like the ones used by CERIF or PARTHENOS might be more detailed and restrictive in the way they describe dataset and important aspects of the metadata, like provenance.
  Recommendation 14: Highlight and make clear the focus and scope of EDMI and show how EDMI complement other minimum information guidelines.

5.3 Strategy

The main strategy of the EOSCpilot relies on the collaboration of metadata catalogues and the adoption of EDMI to make datasets more findable and accessible.

5.3.1 Remarks

- The recommendations describe what needs to be done to create an ecosystem of metadata catalogues. However, it does not provide any guidelines of how it could be achieved. EOSCpilot was not meant to define how to implement such a strategy, nor does it have the resources to do so. The ecosystem of metadata catalogues has been identified as a priority service by EOSCpilot WP5 and EOSCpilot WP6. An agreement on how to implement such an ecosystem, especially for dataset metadata catalogues, requires proactive engagement with metadata catalogues. This engagement and implementation would require some minimum funding, which is currently unavailable.
  Recommendation 15: Bring together metadata catalogues participating in EOSC (catalogues from e-infrastructures and Research Infrastructures) to agree and shape the strategy proposed by EOSCpilot data interoperability, build the case to show the ecosystem of dataset metadata catalogues is one of the key blockers to making EOSC work, and persuade funders that its implementation requires active engagement with, and funding for, metadata catalogues.

[18] https://eoscpilot.eu/themes/wp6-interoperability/1st-report-on-data-report-findability-interoperability

6 CONCLUSION AND NEXT STEPS

The data interoperability demonstrators have been useful to evaluate the EOSCpilot data interoperability strategy for finding and accessing datasets. The demonstrators are also contributing to the identification of issues, the refinement of recommendations, and to the shaping of strategy. The demonstrators are an ongoing activity. They already provided some insights and we hope they will continue to provide feedback. In the coming months, before the end of the project, we should finalise the strategy in such a way that it also considers other recommendations and strategies proposed in EOSCpilot and through other EOSC projects. For instance we want to put this strategy into context with the recommendations from the FAIR data expert group, the EOSCpilot service architecture proposal, the EC interoperability framework, as well as the strategy and recommendations from FREYA and EOSC-hub.

7 ANNEXES

- Annex H - List of minimum, recommended and optional metadata properties.
- Annex K - Copy of the draft proposal for the licence RDA metadata guideline.

8 REFERENCES

Mika, P. 2015. “On Schema.org and Why It Matters for the Web.” *IEEE Internet Computing* 19 (4): 52–55.

Sansone, Susanna-Assunta, Peter McQuilton, Philippe Rocca-Serra, Alejandra Gonzalez-Beltran, Massimiliano Izzo, Allyson Lister, Milo Thurston, and Fairsharing Community. 2018. “FAIRsharing: Working with and for the Community to Describe and Link Data Standards, Repositories and Policies.” https://doi.org/10.1101/245183.

Annex H - List of minimum, recommended and optional metadata properties

List of EDMI metadata properties[19]. On the left column the name of the property, in the middle column the description of the properties and in the right columns the identification of functional and operational metadata and its classification into minimum, recommended and optional properties. M/F: Minimum functional metadata. M/O: Minimum operational metadata. R/F: Recommended functional metadata. R/O: Recommended operational metadata. O/F: Optional functional metadata. O/O: Optional operational metadata.

Properties	Description	M/F	M/O	R/F	R/O	O/F	O/O
MINIMUM							
name	A descriptive name of the dataset	yes					
description	A short summary describing a dataset	yes					
identifier	The identifier property represents any kind of identifier for any kind of dataset	yes					
url	The location of a page describing the dataset	yes			yes		
creator	The creator/author of this dataset	yes			yes		
dateCreated	The date on which the dataset was created	yes					yes
license	A license under which the dataset is distributed		yes	yes			
dataStandard	The standard in which the content of the dataset is represented		yes	yes			
dateModified	The date on which the dataset was most recently modified		yes				
structure	The description of the structure of the dataset		yes				
accessUrl	The link to download the dataset		yes				
accessInterface	The type of interface to present the dataset		yes				
RECOMMENDED							
includedeIn	A dataset or data catalog which contains the dataset			yes	yes		
measurementTechnique	A technique or technology used in a dataset corresponding to the method used for measuring the corresponding variables			yes			
keywords	Keywords or tags used to describe the dataset			yes			
variablesMeasured	The variables that are measured in the dataset			yes			
format	The format in which the content of the dataset is encoded to present the information, typically a MIME format				yes		
scientificType	Scientific domain or type of the information provided in the datataset				yes		
includes	A dataset or data catalog contained in the dataset				yes		

[19] https://tinyurl.com/dats-cats-edmi

contentType	Type of content provided in the dataset based on its origin and type of processes (raw, processed, summarised)				yes		
size	Size of the dataset using a digital information multiple unit byte siymbol (MB, GB, PT, ...)				yes		
authentications	Type of authentication required to access the dataset				yes		
OPTIONAL							
version	The version of the dataset					yes	yes
metric	Metric to provide some quantitative or qualitative information about the dataset					yes	yes
sameAs	Other URLs that can be used to access the dataset page					yes	
spatialCoverage	The location depicted or described in the content					yes	
temporalCoverage	The property indicates the period that the content applies to					yes	
citation	A citation or reference to another work that describes the dataset					yes	
referenceCitation	A citation or reference to that describes the dataset						yes
compression	Type of compression used in the dadataset						yes
authorisations	Type of authorisation required to access the dataset						yes

Annex K - Copy of the draft proposal for the licence RDA metadata guideline

This is an example and work in progress.

'Licence'

DESCRIPTION

A licence describes the conditions which must be met, and under which a dataset may be distributed. This property should be mapped precisely to one of a list of publicly available and accessible licence terms and conditions, with an additional option to add a free text string if the licence is not in the defined list. If the latter, it is recommended that the free text description be submitted for approval as a new licence to an existing authority. The licence property is a recommended property; Access Rights associated with the data, are mandatory.

Tooling to map licences into Access field - so community adds licence and then access is automatically added. If a vague licence, needs to fill in Access manually.

Equivalent or closely related properties:

Property name	relationship	Data model link (URL)
licence	sameAs	http://dublincore.org/documents/dcmi-terms/#terms-license
licence	sameAs	http://schema.org/license

Note: do we want to capture 'relatedTo'? Do we need a list of valid 'relationship'? Or, just want equivalence?

EXPECTED VALUE

List all valid types that can be provided as a value, with an example for each.

Property	Allowed Type Value	Example
Access Rights	Restricted list	open, restricted, embargoed, closed

Licence	From restricted list, with option to add other licences.	CC-BY-SA 3.0

Note: Do we want to restrict to schema.org types? (subclasses of schema.org datatypes)

CARDINALITY

One

GENERAL RECOMMENDATIONS

Examples

DOMAIN SPECIFIC RECOMMENDATIONS

Examples

AUTHORS/CONTRIBUTORS

Authors and contributors are noted in the revision history, where the first row is the initial property definition.

Authors include:

Peter McQuilton

Henning Hermjakob

Juan A. Vizcaíno

Carole Goble

Sirarat Sarntivijai

Shaun de Witt

REVISION HISTORY

Describe what constitutes a revision/version, maintain a log of changes (describe which in scope), and describe how to access different versions. Revisions should be tracked with author, preferably in a table.

Eg. For this property (P), significant structural content changes to the current version (P.1.0) require a change in version (P.2.0) while, for example, descriptive changes of text to the current version (P.1.0) are deemed revisions and reflected in the identifier (P.1.1).

'Property'	Description of change (free text)	Author (txt or ORCID)	Date (mo/da/year)	New Version/ revision
P	Initial document	Bloggs, J., Joggs, B.	10/25/2005	P
P	Clarification with no structural change	Bloggs, J.	12/12/2005	P.1.1
P.1.1				

ACCESS AND IDENTIFIER DEFINITION

Describe identifier pattern and web access.

Identifier: Property.version.revision.

Access https://domain/metadataDefinitions/'Property/version/revision

https://domain/metadataDefinitions/'Property will always resolve to the latest version of the definition...

REFERENCES

References (publication and web links) used in this document listed here, numerically referenced from the text.

1.

2.

3.

...