Project Title	Expanding FAIR solutions across EOSC
Project Acronym	FAIR-IMPACT
Grant Agreement No.	101057344
Start Date of Project	2022-06-01
Duration of Project	36 months
Project Website	fair-impact.eu

M5.5 - Initial repository registry support for discovery of repositories, policies, and interfaces

Work Package	**WP5 - Metrics, certification, and guidelines**
Lead Author (Org)	**Robert Ulrich (KIT)**
Contributing Author(s) (Org)	**Maaike Verburg (DANS), Mike Priddy (DANS), Robert Huber (UBremen), Hervé L'Hours (UKDS), Charlotte Neidiger (KIT), Gabriela Meijas (DataCite), Joy Davidson (UEDIN-DCC)**
Due Date	**2024-03-31**
Date	**2024-03-28**
Version	**V1.0**
DOI	**10.5281/zenodo.10847707**

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)

Funded by the European Union

Versioning and contribution history

Version	Date	Author	Notes
0.1	07.02.2024	Maaike Verburg (DANS), Mike Priddy (DANS), Robert Ulrich (KIT), Charlotte Neidiger (KIT)	TOC and V0.1
0.2	12.03.2024	Maaike Verburg (DANS), Mike Priddy (DANS), Robert Ulrich (KIT), Charlotte Neidiger (KIT), Robert Huber (UBremen)	First draft of content
0.5	27.03.2024	Maaike Verburg (DANS), Mike Priddy (DANS), Robert Ulrich (KIT), Charlotte Neidiger (KIT), Robert Huber (UBremen), Hervé L'Hours (UKDS), Gabi Meijas (DataCite), Joy Davidson (UEDIN-DCC)	Complete version
1.0	28.03.2024	Maaike Verburg (DANS), Robert Ulrich (KIT)	Final edits, Published and submitted

Disclaimer

FAIR-IMPACT has received funding from the European Commission's Horizon Europe funding programme for research and innovation programme under the Grant Agreement no. 101057344. The content of this document does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of such content.

Table of Contents

Funded by the European Union

TERMINOLOGY

Terminology/Acronym	Description
API	Application Programming Interface
CKAN	Comprehensive Knowledge Archive Network
DCAT	Data Catalog Vocabulary
DCAT-AP	Data Catalog Vocabulary Application Profile
DQV	Data Quality Vocabulary
DRAWG	Data Repository Attributes Working Group
FAIR	Findable, Accessible, Interoperable, Reusable
FDP	FAIR Data Point
GeoDCAT	A general geospatial profile of DCAT
IDCC	International Digital Curation Conference
IRI	Internationalized Resource Identifier
JSON-LD	JavaScript Object Notation for Linked Data
LDP	Linked Data Platform
LOD	Linked Open Data
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OGC	Open Geospatial Consortium
PID	Persistent IDentifier
RDA	Research Data Alliance
RDF	Resource Description Framework
REST	Representational State Transfer
SPARQL	SPARQL Protocol And RDF Query Language
TDR	Trustworthy Digital Repository
XML	Extensible Markup Language

Funded by the European Union

1 Introduction

1.1 Role of the Milestone

The purpose of this Milestone is to initiate the prototyping phase, embodying FAIR-IMPACT's approach to implementing the principles outlined in the initial version of the *"Guidelines for repositories and registries on exposing repository trustworthiness status and FAIR data assessments outcomes"*[1] *(Milestone 5.2)*. The proposed standards and technology stack are based on insights gathered through the activities in the earlier project FAIRsFAIR "Fostering FAIR Data Practices In Europe"[2] and cross-work package collaboration and exchange among the FAIR-IMPACT project partners. This application of this stack is aimed at complementing the guidelines' technology and tool-agnostic character by establishing an initial linking and information structure for the prototype to be developed. It is intended to bridge repositories with registries and discovery services, thereby facilitating the better exposure and discovery of repository information. The registry role within the prototype will be covered by re3data - registry of research data repositories[3]. In addition, it will enable the provision of information on the object level as well as quality measures due to the extensibility of the chosen technology stack. This work will serve as the foundation for the development of future guideline versions. Suggestions collected from the community will be used for revising, extending, and refining the transparent exposure of repository information and its support in registries like re3data as well as other services, such as F-UJI and CoreTrustSeal. Thus the approach is similar to the development of the guidelines and ensures that not only the prototype but also guidelines themselves remain responsive to the needs and insights of stakeholders.

1.2 Means of verification

The means of verification for this Milestone is to have a report available detailing the Milestone. The current document establishes this verification, and is made publicly available on Zenodo.

[1] Verburg, M., Ulrich, R., L'Hours, H., Huber, R., Priddy, M., Davidson, J., Gonzalez-Beltran, A., Meijas, G., & Neidiger, C. (2023). M5.2 - Guidelines for repositories and registries on exposing repository trustworthiness status and FAIR data assessments outcomes (1.0). Zenodo. https://doi.org/10.5281/zenodo.10058634
[2] https://fairsfair.eu/
[3] https://www.re3data.org/

Funded by the European Union

2 Description of the Milestone

The Milestone focuses on the technical aspects of making repository information accessible to repository registries and discovery portals. It will detail the selected technology stack and architecture and indicate how information can be exposed at the entity level, respectively the source of its origin. Furthermore, it will discuss integration into the research data landscape, including Persistent Identifiers (PIDs), assessments, and metrics. The objective is to establish connections between entities as a foundational step for developing the prototype, where linking repositories and discovery services are only one important component.

In addition, the Milestone will offer insights into the desired policies, attributes of repositories, and the trust through transparency model that will be incorporated into the prototype. Therefore, this Milestone adopts a bottom-up approach concerning technical implementation, while employing a top-down perspective for the desired semantic artefacts that enable the description of repositories from different perspectives, e.g. organisation, catalogue or web services.

2.1 Exposure and harvesting of repository information

2.1.1 Semantic Web and Linked (Open) Data

The Semantic Web and Linked Open Data (LOD) transfer the concept of the hypertext system[4] by Tim Berners-Lee, we know as WorldWideWeb, to data and forms a substantial foundation of how data is published, shared, and integrated across the internet. These technologies offer a structured and interconnected web of data, enabling machines to understand and process the semantics of information. The semantic web encompasses a comprehensive technology stack including the Resource Description Framework (RDF)[5] and SPARQL Protocol and RDF Query Language (SPARQL)[6] (see Figure 1). RDF offers a remarkable versatility as a standard for metadata description and provides a foundational layer for semantic interoperability among diverse data systems.

By choosing and utilising RDF, resources can be described in a way that is both machine-understandable and tailored to their specific domain requirements, leveraging RDF's ability to model complex relationships between entities. Its flexible approach to data representation allows those entities to be expressed as triples, a structure comprising subject, predicate, and object. RDF is backed by a rich ecosystem of vocabularies[7] that can be used jointly and enable discoverability and interoperability to cover generic as well as specialised metadata descriptions.

[4] https://www.w3.org/Protocols/HTTP/AsImplemented.html
[5] https://www.w3.org/RDF/
[6] https://www.w3.org/TR/sparql11-query/
[7] https://lov.linkeddata.es/dataset/lov/

Funded by the European Union

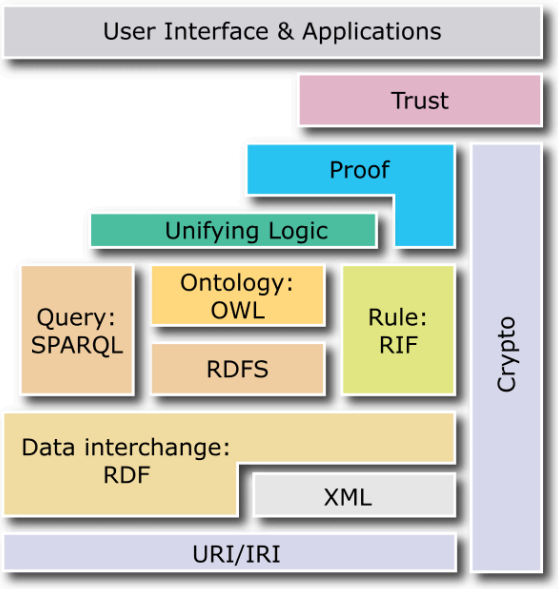

Figure 1 - The Semantic Web Stack[8].

A widely adopted vocabulary is the Data Catalog Vocabulary (DCAT)[9], which standardises the description of digital catalogues, datasets and data services. DCAT will serve as core structure in the prototype to expose basic repository and resource information.

Linked Open Data and the Five Star Open Data initiative promote principles[10] that ensure data is open, linkable, and machine-readable. RDF perfectly aligns with these goals, facilitating the creation of high-quality datasets that are accessible and useful to both humans and machines. This approach breaks down data silos and links information across disparate infrastructures, enhancing transparency and enabling new insights through data integration. The semantic web and linked open data communities are active and growing. This large user base contributes to the ongoing development of standards, tools, and best practices, ensuring the ecosystem remains innovative and responsive to existing and new challenges.

While RDF offers a flexible and powerful framework for data interoperability, integrating it with existing infrastructures can be challenging. Not all systems and technologies are designed with semantic web compatibility in mind. Technologies such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[11] remain popular for e.g. publication repositories, illustrating that semantic web technologies are not universally preferred solutions across all domains and use cases. The steep learning curve and complexity of the technology stack can be barriers to adoption, requiring significant investment in skills and understanding. In addition, ensuring the quality and consistency of RDF data and ontologies

[8] https://commons.wikimedia.org/wiki/File:Semantic_Web_Stack.png
[9] https://www.w3.org/TR/vocab-dcat-3/
[10] https://5stardata.info/en/
[11] https://www.openarchives.org/pmh/

Funded by the European Union

is an ongoing challenge. The open nature of the semantic web can lead to variability in data quality, impacting the reliability of linked data applications. So despite its potential and growing application, these technologies have not yet achieved widespread adoption in the research data landscape. This suggests a need for continued advocacy, tool development, and education to realise the full benefits of these technologies.

The decision to adopt semantic web and linked open data technologies is driven by a balance of considerations. While the challenges of complexity, data quality, and integration cannot be overlooked, the compelling advantages of an extensive technology stack, expressivity, community support, and the promotion of openness and interoperability present a strong case for their adoption. Another major argument for the adoption in the prototype is DCAT as it provides a matured solution to connect discovery services, such as registries, with data providers, research data repositories but also deeper insights for the exposure of data quality, persistent identifiers etc. This is reflected in existing and growing adoption within and outside FAIR IMPACT (see section 2.1.7). As the guidelines continue to evolve, it is expected that the choice of RDF/DCAT shows up to be the right one easing the complexity and implementation of the prototype.

2.1.2 Architecture & Data Model

The guidelines as well as their implementation and testing within the prototype are aiming for the following three main concepts:

Transparency & Standards: Data regarding FAIRness (Findability, Accessibility, Interoperability, and Reusability) will be provided in accordance with open technical standards and community recommendations such as RDF and RDA recommendations. This will apply to digital objects, (meta)data services, and registries, aiming to promote interoperability and harmonisation.

Evidence & Assessment: (Meta)data services will offer information about their functions and activities, supported by evidence and links to authoritative third parties like CoreTrustSeal. This approach enables validation and assessment by both human and automated processes, enhancing evaluation and quality.

Linking & Aggregation: As information varies in terms of level of detail, contexts, and time, it should be linked across different services and data providers to foster discovery and aggregation by registries, assessment tools etc. to provide additional value and insights to users.

The Milestone focuses on linking between the registry and repositories as the initial step towards the prototype. Utilising DCAT and the related open and standardised technology stack, the architecture itself is following the first two recommendations of Milestone 5.2 recommending transparency and standards. As RDF is based on Internationalized Resource Identifiers (IRIs), evidence, resources and context can be linked for validation and

Funded by the European Union

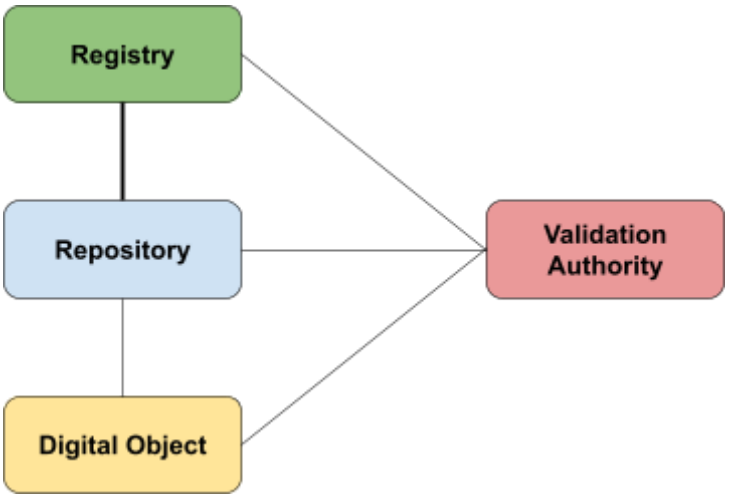

Figure 2 - Prototype schematic.

assessment, e.g. link to a CoreTrustSeal certification (see Figure 2). With the choice of DCAT to be harvested by registries and discovery services, e.g. re3data in the prototype, it enables the discovery of metadata descriptions of the digital objects itself as well as the exposure and linking of related information, e.g. PIDs or quality measures. This information can be backed or utilised by validation authorities, e.g. assessment tools validating that the PID actually resolves to the dataset. With the provision of information directly at the source, namely through the repository itself, the metadata can be reused by multiple services. Updating the information at the source of origin is supposed to be more accurate and complete as opposed to information gathered by third parties. It also enables harvesting and updates across multiple infrastructures. The decentralised approach allows to support use cases that require subject specific information and a level of detail that otherwise would be difficult or impossible to be curated by centralised generic registries, like re3data (see Figure 3).

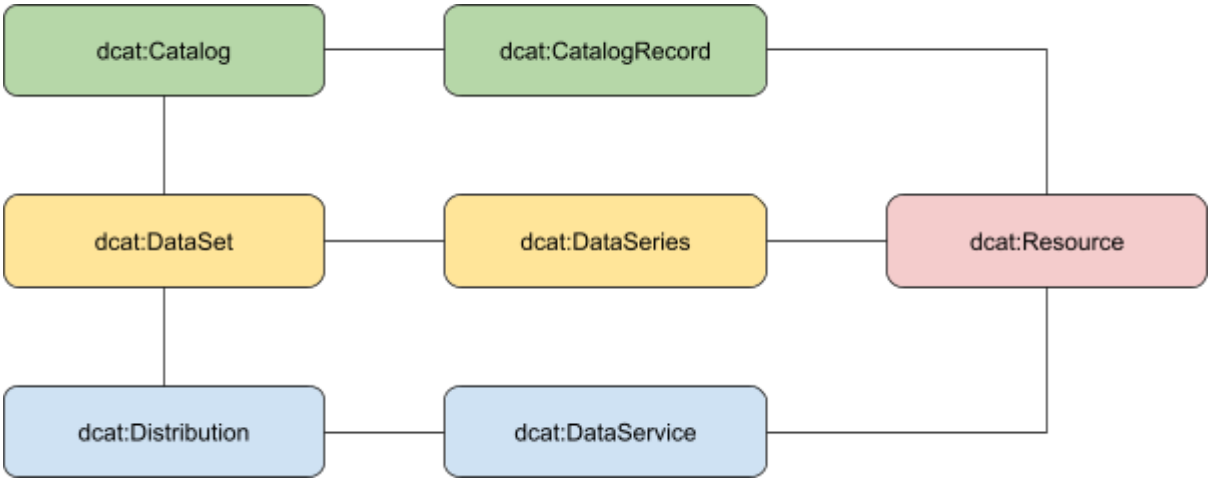

Figure 3 - Overview over the DCAT model.

Funded by the European Union

dcat:Catalog: This class represents a collection of datasets or data services. A catalogue can contain one or more datasets and provides metadata about them, such as title, description, keywords, and access information.

dcat:CatalogRecord: This class is optional and describes a specific record within a catalogue. It typically is used to capture metadata about the catalogue entry itself, such as its creation date, publisher, and other administrative details.

dcat:Resource: This is a generic class representing any entity that can be described in a data catalogue. It serves as a superclass for more specific types such as *Dataset*, *DataSeries*, *Distribution*, and *DataService*.

dcat:Dataset: This class represents a collection of data, often organised and presented in a structured format. A dataset typically includes metadata describing the data, such as its title, description, keywords, temporal and spatial coverage, licensing information, and access methods.

dcat:DataSeries: This class represents a collection of separated datasets that can be grouped or belong together.

dcat:Distribution: This class describes a specific way in which a dataset or data service is available, such as a file format, access method (e.g., download, API), or endpoint.

dcat:DataService: This class represents a service or API that provides access to data. It enables repositories to expose descriptions of their technical interfaces. The properties *dcat:endpointURL*, *dcat:endpointDescription*, *dcterms:conformsTo*, *dcat:servesDataset* provide automatic discovery of the provided interfaces to access the datasets.

2.1.3 Exposing Persistent Identifiers

RDF respectively DCAT relies on IRIs to identify resources. In the scholarly context the scientific and research data communities rely on a variety of different identifiers to reference entities, like publications, authors or data. This can be expressed utilising *dcterms:identifier* or *adms:identifier*. Exposing PIDs in RDF is bridging the semantic web approach and with Persistent Identifiers Providers. Not only does this enable identification in services using PIDs, e.g. DataCite Commons[12], but is expected to contribute to persistent identification. Even so it is possible to keep IRIs[13] resolving for a long time, efforts in long term preservation and curation of identifiers by organisations like DataCite are more likely to be carried on. DataCite Commons shows the connections between researchers, outputs, research organisations, and funders – also known as the PID Graph of scholarly resources identified through persistent identifiers (PIDs) and connected in standard ways.

[12] https://commons.datacite.org/
[13] https://datatracker.ietf.org/doc/html/rfc3987

Funded by the European Union

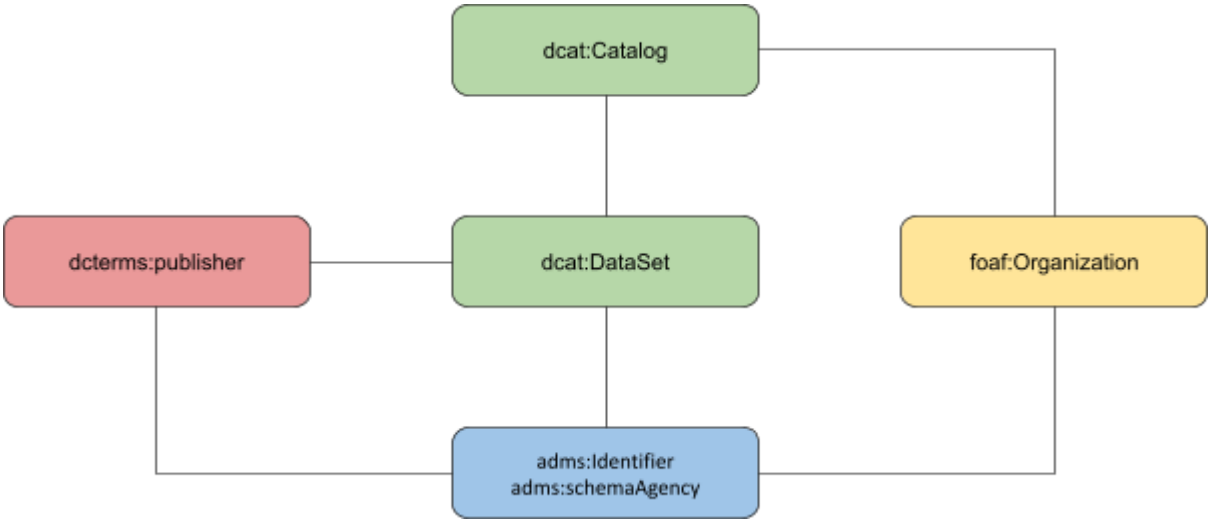

Figure 4 - Using PIDs in DCAT.

```
example:catalog a dcat:Catalog;
 dcterms:title “Example Catalog”;
  adms:identifier example:org ;
  .

 example:id a dcat:Dataset;
  adms:identifier example:iddoi ;
  dcterms:publisher example:EddieExample ;
  .

 example:iddoi a adms:Identifier ;
  skos:notation "https://doi.org/10.1337/data.42"^^xsd:anyURI;
  adms:schemaAgency "DataCite" ;
  .

 example:EddieExample a foaf:Person;
  foaf:name "Eddie Example" ;
  adms:identifier example:EddieExample ;
  .

 example:EddieExample a adms:Identifier;
  skos:notation "https://orcid.org/0000-0000-0000-0000"^^xsd:anyURI ;
  adms:schemaAgency "ORCID" ;
  .

 example:org a foaf:Organization ;
  rdfs:label "Example University" ;
  foaf:homepage <https://www.university.example/> ;
  .
```

Funded by the European Union

```
example:org a adms:Identifier;
 skos:notation "https://ror.org/012abc345"^^xsd:anyURI ;
 adms:schemaAgency "ROR" ;
 .
```

Example linking a dataset with DOI, ORCID and ROR

2.1.4 Exposing quality indicators and certificates

To expose quality indicators and certificates related to repositories or individual data objects, some extensions proposed by the data on the web working group for the Data Catalog Vocabulary[14] (DCAT) can be used: Since *dcat:Catalog* is a subclass of *dcat:Dataset*, we can use the same practice to annotate quality information for data repositories represented as *dcat:Catalog* that is recommended for *dcat:Dataset*. The DCAT documentation proposes using the Data Quality Vocabulary (DQV)[15] to indicate quality information related to resources. The *dqv:QualityCertificate* class is the appropriate DQV class which can be used to indicate an "*annotation [...] that certifies the resource's quality according to a set of quality assessment rules*." We therefore define e.g. a CoreTrustSeal certificate as a *dqv:QualityCertificate* as it expresses the quality of a data repository with respect to its ability to act as a trustworthy long-term data repository.

The DCAT documentation proposes the use of a *dqv:hasQualityAnnotation* property to e.g. link to an individual *dqv:QualityAnnotation*. Since a *dqv:QualityCertificate* is a subclass of *dqv:QualityAnnotation* we utilise this property to point to CoreTrustSeal certificates.

CoreTrustSeal publishes assessment results from successfully evaluated data repositories as PDF documents, which in turn are referenceable as Dataverse datasets via a DOI. *dqv:QualityAnnotation* is a subclass of *oa:Annotation*, a class of the Web Annotation Ontology[16]. Therefore, these DOIs can be used to represent a *dqv:QualityAnnotation* which links an individual *dqv:QualityCertificate* Instance representing a CoreTrustSeal certification via the *oa:hasBody* property. Since the CoreTrustSeal Certificate is stored as a Dataverse Dataset it can additionally be typed as a *dcat:Dataset* and then use the appropriate properties to indicate e.g. date and responsibilities of the certification process. Alternatively, since the use of the PROV Ontology is encouraged by the DQV group, the PROV-O[17] ontology allows to specify a *prov:Activity* to represent the Certification process which links to a *dqv:QualityCertificate* using a *prov:wasGeneratedBy* property.

Example:

```
<https://www.pangaea.de> a dcat:Catalog ;
```

[14] https://www.w3.org/TR/vocab-dcat-3/
[15] https://www.w3.org/TR/vocab-dqv/
[16] https://www.w3.org/TR/annotation-vocab/
[17] https://www.w3.org/TR/prov-o

Funded by the European Union

```
   dqv:hasQualityAnnotation :PANGAEACoreTrustSealCertificate .

:PANGAEACoreTrustSealCertificate
a dqv:QualityCertificate ;
   oa:hasTarget <https://www.pangaea.de> ;
   oa:hasBody <https://doi.org/10.34894/TFRLXN> ;
   oa:motivatedBy dqv:qualityAssessment ;
   .
```

Similarly, FAIR evaluations at the repository level which may have been derived from a representative sample of FAIR assessments of datasets of this repository, can be expressed. As recommended in the DQV documentation a *prov:wasDerivedFrom* relation could link dataset level FAIR assessments, potentially grouped within a *dqv:QualityMeasurementDataset*, with a *dqv:QualityAnnotation* expressing an overall FAIR status or indicator of an individual data repository.

2.1.5 RDF formats and content negotiation

RDF can be expressed in different formats[18] like JSON-LD (application/ld+json), XML (application/rdf+xml), Turtle (application/x-turtle) etc. Each presentation has its own pros and cons. For example Turtle is easy to read and write by humans, JSON-LD more likely be adopted in environments where JSON is used already and the XML representation as one of the oldest formats is well supported even in old rdf libraries. Other formats may have advantages in regards to performance and transmission. It is recommended to provide the formats most likely to be used by the consumers. Different formats can be provided via content negotiation. Not only does content negotiation allow for better machine-actionability, but may serve human readable websites, too. Additionally setting up a SPARQL endpoint enables the execution of queries and constitutes another API to expose information via RDF/DCAT.

2.1.6 Cascaded Harvesting

Registries such as re3data can easily be considered as catalogues. The same is true for repositories exposing their information and dataset description. Yet many of them are service providers harvesting other repositories and data sources themself and even re3data could be considered a repository and datasource for other discovery portals. Not only does this imply linking repository and registry but also the requirement to model and represent this potential catalogue cascades. In the context of CKAN it is named “transitive harvesting” and modelled as sub catalogues[19] and within the realm of DCAT-AP named “super catalog”[20] . Both make use of the property *dcterms:hasPart* (and its reverse relation *dcterms:isPartOf*).

[18] https://ontola.io/blog/rdf-serialization-formats
[19] https://extensions.ckan.org/extension/dcat/#transitive-harvesting
[20] https://doc.piveau.eu/hub/user-guide/#super-catalogue

Funded by the European Union

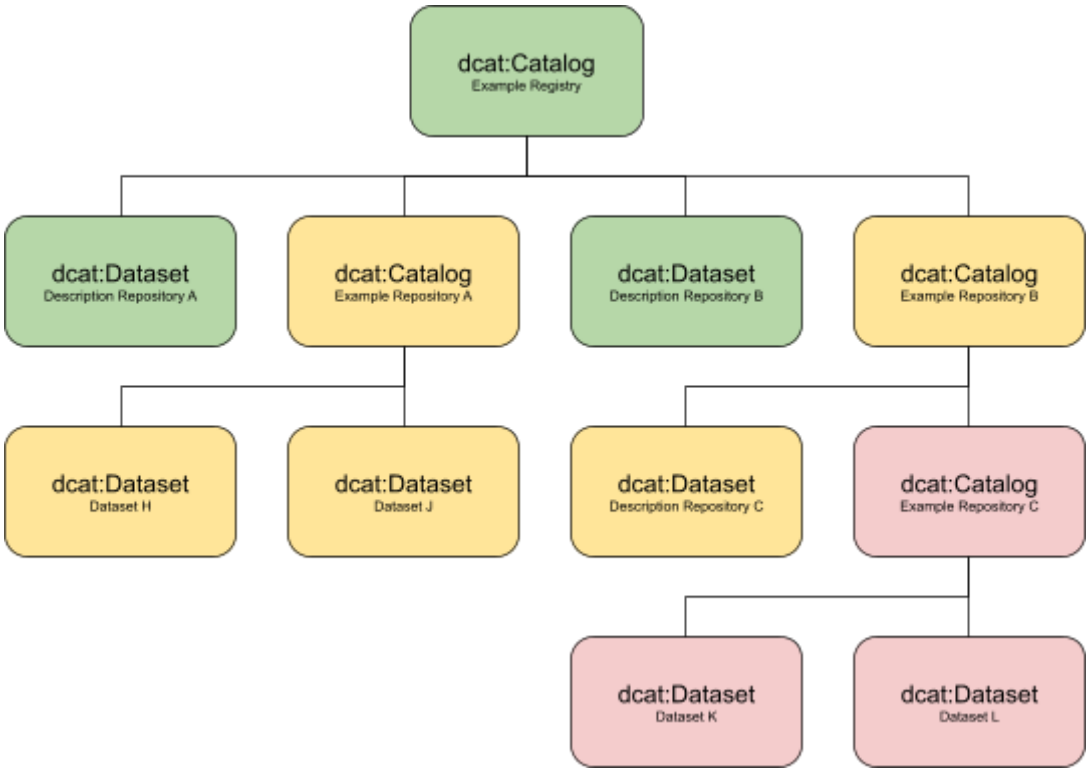

Figure 5 - Connecting repository and registry.

```
example:registry a dcat:Catalog;
 dcterms:title "Example Registry";
 dcat:dataset example:repositoryDescriptionA;
 dcat:dataset example:repositoryDescriptionB;
 dcterms:hasPart example:repositoryA;
 dcterms:hasPart example:repositoryB;
   .

 example:repositoryDescriptionA a dcat:Dataset;
  dcterm:title "Example Description of Repository A";
  .

 example:repositoryDescriptionB a dcat:Dataset;
  dcterm:title "Example Description of Repository B";
  .

example:repositoryA a dcat:Catalog;
 dcterm:title "Example Repository A";
 dcterms:isPartOf example:registry;

example:repositoryB a dcat:Catalog;
 dcterm:title "Example Repository B";
 dcterms:isPartOf example:registry;
 dcat:dataset example:repositoryDescriptionC;
```

Funded by
the European Union

```
dcterms:hasPart example:repositorC;
.
```

Cascaded harvesting

2.1.7 Related activities

DCAT as a standard has a growing number of adopters within the scientific and open (governmental) data communities. For example, the FAIR Data Point (FDP)[21] specification and its reference implementation form a metadata service that enables the exposure and consumption of metadata. As it provides functionality for creating, storing and serving metadata based on standards like REST (representational state transfer) and the Linked Data Platform (LDP)[22] It also relies on DCAT as an important component. Another use case is the DCAT Application profile for data portals in Europe (DCAT-AP)[23] which is a specification to enable the discovery of public sector information that is used in the European Data Portal[24]. A similar initiative is DCAT-US[25]. The Comprehensive Knowledge Archive Network (CKAN) is an open source software by the Open Knowledge Foundation for storage and distribution of open data and it also provides the exposure of metadata via a plugin using the DCAT standard. In 2023 the Open Geospatial Consortium (OGC)[26] formed a Standards Working Group[27] to develop GeoDCAT, which adopts DCAT-AP for geospatial use cases. Given this example DCAT may leverage data exchange and interoperability not only within the research data landscape but also connect beyond and foster the discovery of public sector information and subject specific knowledge for the scientific community.

As the semantic web is not the only technology available, signposting[28] and especially FAIRiCat[29] is another way to enable the discovery and exposure of scholarly metadata via different standards, e.g. OAI-PMH. With that it can also enable the discovery of DCAT based interfaces.

2.2 Trust through Transparency - Certification and Beyond

As outlined in the previous Milestone M5.2 “transparency between parties including repositories, object depositors, object users, and other (meta)data services is a critical precursor to trusted relationships.“[30] Transparency of repository certification status is a candidate for self-declaration by a repository (e.g. through inclusion in re3data repository registry metadata). This and other similar self-declarations would ideally be supported by

[21] https://www.fairdatapoint.org
[22] https://www.w3.org/TR/ldp/
[23] https://semiceu.github.io/DCAT-AP/releases/3.0.
[24] https://data.europa.eu/
[25] https://resources.data.gov/resources/dcat-us/
[26] https://www.ogc.org
[27] https://www.ogc.org/press-release/ogc-forms-new-geodcat-standards-working-group/
[28] https://signposting.org
[29] https://signposting.org/FAIRiCat/
[30] Verburg, M., Ulrich, R., L'Hours, H., Huber, R., Priddy, M., Davidson, J., Gonzalez-Beltran, A., Meijas, G., & Neidiger, C. (2023). M5.2 - Guidelines for repositories and registries on exposing repository trustworthiness status and FAIR data assessments outcomes (1.0). Zenodo. https://doi.org/10.5281/zenodo.10058634 [p.10]

Funded by the European Union

confirmation of status through validation by some authoritative body. In the case of CoreTrustSeal Trustworthy Digital Repository (TDR) certification the CoreTrustSeal Board.

The current scope of CoreTrustSeal certification only covers those repositories undertaking long term active preservation responsibility. However, many of the CoreTrustSeal Requirements are applicable to a wider range of data and metadata services, and not all in-scope repositories are currently certified or prioritising certification as a goal.

With or without certification in place it is desirable for organisations providing services around data and metadata to share characteristics about themselves, including the activities and functions they undertake. These characteristics include existing and proposed metadata about repositories while relevant activities and functions can be identified, both within and beyond CoreTrustSeal. Characteristics range from repository names and contact details (existing) to the levels of curation and preservation they provide (to be proposed). Activities and functions of repositories range from their mission/scope, to their digital object management approaches including preservation where relevant. These functions and activities may imply further specific supporting artefacts e.g. mission statement, policy/policies or preservation plan.

Provision of 'levels of care' metadata at the repository level ("organisation X offers the following different levels of care") is directly informative to users seeking to select a repository, but is not sufficient in a connected research infrastructure ecosystem. Ideally each digital object would include metadata about the level of care it receives alongside information about the repository that provides the care.

As noted in the previous Milestone document "the public assertion of information about an organisation or an object demonstrates transparency and supports mutual trust between human actors (e.g. researcher or funder) and, increasingly, interoperability between machine agents."[31]

Registries such as re3data and FAIRsharing[32] provide a means of sharing some of the supporting artefacts described above. However, there could be benefits associated with developing a standard model for repositories to share more of this kind of information in a consistent manner. This might involve mapping to common repository characteristics (Re3data, FAIRsharing, DRAWG etc) or activities and functions (e.g. CoreTrustSeal, COAR) with standard supporting information (self-assertions statement, links to supporting artefacts) and could support harvesting by a number of existing information consumers. With a clear extension and elaboration model it could support novel use cases, including the automation of portions of the CoreTrustSeal application process, or other assessment mechanisms.

[31] Verburg, M., Ulrich, R., L'Hours, H., Huber, R., Priddy, M., Davidson, J., Gonzalez-Beltran, A., Meijas, G., & Neidiger, C. (2023). M5.2 - Guidelines for repositories and registries on exposing repository trustworthiness status and FAIR data assessments outcomes (1.0). Zenodo. https://doi.org/10.5281/zenodo.10058634 [p.10]

[32] FAIRsharing database schema: https://fairsharing.github.io/JSONschema-documenter/?schema_url=https://api.fairsharing.org/model/database_schema.json

Funded by the European Union

2.3 Data Repository Attributes

The RDA Data Repository Attributes Working Group (DRAWG), that "seeks to produce a list of common attributes that describe a research data repository and to provide examples of the current approaches that different data repositories are taking to express and expose these attributes"[33]. The rationale behind this work was to improve repository discovery and the understandability of repository qualities through common exposure of repository attributes. The working group is currently wrapping up and has presented its final output: the list of Common Descriptive Attributes of Research Data Repositories[34], receiving valuable community input and feedback. The output presents seventeen attributes and related examples, schemata, and notes on how difficult it generally is to find this information about a repository. Through engagement in the DRAWG, this list of attributes was identified as an important community output to interact with in this related work on the exposure of repository information. In their case statement, the DRAWG also mentions that one of the challenges in the current landscape is that information is expressed and exposed in different ways. By aligning our work with the DRAWG attributes, we fill the gap of the mechanism of exposure.

Ongoing work includes the alignment of the proposed DRAWG attributes to repository functions and activities. These align with the CoreTrustSeal Requirements, but also with a much wider crosswalk of trust-related repository and data service criteria. A selected subset of these were presented during the IDCC Transparency Guidelines workshop[35].

2.4 International Digital Curation Conference

The initial version of the guidelines (M5.2) were presented at the International Digital Curation Conference 2024 (IDCC24) in the workshop 'Guidelines on transparent exposure of repository information: informing decisions of trustworthiness'. The aim of the workshop was to introduce a targeted audience to our guidelines and plans for the prototype, and to gather specific and critical feedback on how to improve the guidelines to the realistic experience of the community. The workshop was structured to first allow participants to be introduced to the relevant information and background of the work, followed by hands-on work to consider their own organisations, using a selection of the aforementioned DRAWG attributes. Based on the input of the eighteen participants, a post-workshop report was created to separately showcase the feedback received and lessons learned from the

[33] https://www.rd-alliance.org/group/data-repository-attributes-wg/case-statement/data-repository-attributes-wg-case-statement

[34] Witt, M., Cannon, M., Lister, A., Segundo, W., Shearer, K., Yamaji, K., & Research Data Alliance Data Repository Attributes Working Group. (2024). RDA Common Descriptive Attributes of Research Data Repositories (Version 1.0). Research Data Alliance. https://doi.org/10.15497/RDA00103

[35] Verburg, M., Priddy, M., Ulrich, R., Huber, R., L'Hours, H., Neidiger, C., & Dillo, I. (2024). Guidelines on transparent exposure of repository information: Informing decisions of trustworthiness. 18th International Digital Curation Conference (IDCC24), Edinburgh, Schotland. Zenodo. https://doi.org/10.5281/zenodo.10794116

Funded by the European Union

workshop. The workshop materials[36] and the post-workshop report[37] can both be found on Zenodo.

Participant feedback gave valuable insight into the inclusive presentation of the work to the relevant stakeholders, to ensure everyone feels adequately informed to interact with the work and the guidelines and prototype can thus get the desired level of engagement by the stakeholders we envisioned. The interactive workshop exercise also gave insight into the current state of information exposure, which showed to be often lacking in machine-actionable qualities and sometimes even hard to evaluate by humans as well. Participants did see many clear added values of improving the transparent exposure of such information, and estimated their organisation would also be willing to improve their information exposure according to the guidelines. However, there is an important distinction to make between willingness and capability. Accurately estimating whether the organisation has the capabilities to improve their information exposure depends for a large part on how clear the instructions will be on how to improve the current status. With adequate guidance and support, more repositories could be able to significantly improve their information exposure. The feedback taken from this workshop has helped to inform this Milestone, and also helped shape the upcoming support action related to this topic[38].

[36] Verburg, M., Priddy, M., Ulrich, R., Huber, R., L'Hours, H., Neidiger, C., & Dillo, I. (2024). Guidelines on transparent exposure of repository information: Informing decisions of trustworthiness. 18th International Digital Curation Conference (IDCC24), Edinburgh, Schotland. Zenodo. https://doi.org/10.5281/zenodo.10794116

[37] Verburg, M., Neidiger, C., Ulrich, R., L'Hours, H., Huber, R., Priddy, M., & Dillo, I. (2024). Post-workshop report - "Guidelines on transparent exposure of repository information: informing decisions of trustworthiness". Zenodo. https://doi.org/10.5281/zenodo.10848994

[38] https://fair-impact.eu/support-offer-3-recommendations-trustworthy-and-fair-enabling-data-repositories

Funded by the European Union

3 Conclusions and next steps

This Milestone described FAIR-IMPACT's approach to implementing the previously published guidelines on repository and registry transparency, initiating the implementation phase of these principles through proposing standards and a technology stack. The technical exposure and harvesting of repository information was detailed and approaches for exposing persistent identifiers, quality indicators, and certificates were proposed. Other related activities and plans for further developments in this implementation phase were also described, covering aspects of trustworthiness and FAIR assessment. These developments will continue to be crystallised and tested as the project progresses.

Aside from technical development, the focus of this work will also remain strongly on community outreach and interaction. In this Milestone, the workshop at the International Digital Curation Conference was described as a useful way to inform participants about the work and associated ideas and principles, as well as a valuable mechanism for learning more about the current practices in the research landscape and receiving feedback on how the guidelines and future prototype will likely fit the landscape best. We will continue to gather input like this to gauge community interest and test the flexibility of the guidelines and implementation by considering a wide variety of scenarios. Up next is the FAIR-IMPACT financial support offer *"Recommendations for trustworthy and FAIR-enabling data repositories"*, in which participants will have the opportunity to engage more extensively with the guidelines and consider the current status of their organisation. The deadline for applications is on 31 March 2024, after which we will be able to select participants for the support action that will challenge the flexibility of the guidelines by presenting scenarios that may be less common or more complex. Participants will receive financial compensation for their effort, and learn more about the guidelines and how to evaluate the added values of improving their organisation's transparency. Another outreach activity will take place at Open Repositories 2024, where the work will be presented in a poster session, and we will have the opportunity to engage directly with an important part of our audience.

Funded by the European Union