D4.1 An assessment of FAIR-uptake among regional digital repositories

Author(s)	Andreas O Jaunsen, Mari Kleemola, Tuomas J. Alaterä, Heikki Lehväslaiho, Josefine Nordling, Adil Hasan, Pauli Assinen
Status	Submitted
Version	1.0
Date	2020-08-20

Document identifier:	
Deliverable lead	Andreas O Jaunsen
Related work package	
Author(s)	Andreas O Jaunsen, Mari Kleemola, Tuomas J. Alaterä, Heikki Lehväslaiho, Adil Hasan, Josefine Nordling, Pauli Assinen
Contributor(s)	WP4 team
Due date	2020-08-31
Actual submission date	2020-08-31
Reviewed by	Truels Rasmussen, Damien Lecarpentier
Approved by	
Dissemination level	Public
Website	www.eosc-nordic.eu
Call	H2020-INFRAEOSC-2018-3
Project Number	857652
Start date of Project	2019-09-01
Duration	36 Months
License	Creative Commons CC-BY 4.0
Keywords	FAIR, maturity evaluations, open data, machine-actionable metadata, project outputs

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

DRAFT NOT YET APPROVED BY THE EUROPEAN COMMISSION

Abstract:

The EOSC-Nordic project aims to facilitate the coordination of EOSC relevant initiatives within the Nordic and Baltic countries and exploit synergies to achieve greater harmonisation at policy and service provisioning across these countries, in compliance with EOSC agreed standards and practices. The project brings together a strong consortium of 24 complementary partners including e-Infrastructure providers, research performing organisations and expert networks, with national mandates and experience with regards to the provision of research data services, and a unique capacity to realise the outcomes of the EOSC design as outlined by the EOSC Implementation Roadmap.

The project has pledged to implement FAIR in the Nordic and Baltic regions and aims to encourage, support and assist the research community to FAIRify their data. This will be achieved by communicating the benefits of going FAIR to a broad scientific community. Uniquely, the project has selected a hundred repositories and evaluated them consistently according to their FAIR maturity.

This report describes the first measurement of FAIR-ness for a reasonably sized sample of Nordic and Baltic scientific repositories. The document starts with a description of how the datasets were selected and then describes how the FAIR score was measured followed by observations of the results of the exercise.

The major sections of this Report focus on outlining the FAIR Maturity evaluation process, an initial analysis of the state of FAIR open data in the region and the intention of monitoring the development of the FAIRness among these repositories during the project period.

EOSC-Nordic project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857652

Disclaimer: The content of the document herein is the sole responsibility of the publishers and it does not necessarily represent the views expressed by the European Commission or its services.

While the information contained in the document is believed to be accurate, the author(s) or any other participant in the EOSC-Nordic Consortium make no warranty of any kind with regard to this material including, but not limited to the implied warranties of merchantability and fitness for a particular purpose.

Neither the EOSC-Nordic Consortium nor any of its members, their officers, employees or agents shall be responsible or liable in negligence or otherwise howsoever in respect of any inaccuracy or omission herein.

Without derogating from the generality of the foregoing neither the EOSC-Nordic Consortium nor any of its members, their officers, employees or agents shall be liable for any direct or indirect or consequential loss or damage caused by or arising from any information advice or inaccuracy or omission herein.

Table of contents

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

1. Introduction

Science is one of the greatest collective endeavours and the most important channel we have for gaining knowledge. Scientific research depends on the collecting of data to investigate and explain a phenomenon or theory. By collecting empirical data, scientists can learn about a phenomenon reliably and accurately. Research is critical to societal development. It generates new knowledge, improves education, increases the quality of our lives, and helps in decision-making to name a few examples.

A crucial premise for research, therefore, is to ensure transparency of the data, enable the reproducibility of research results, and to build trust in the results and the scientific method. However, a major caveat is that research data is largely inaccessible to others. Speaking at the World Economic Forum in Davos, EC President Ursula von der Leyen, mentioned that 85% of the data that could be used for research are never used. Some studies indicate that only about 20% of data generated by peer-reviewed research is being deposited in suitable repositories[1]. Another warning sign is that "60% of data scientists spend most of their time cleaning and labelling data."[2]. Similar findings by R&D divisions in the private sector support this claim. If correct, this means that research is tremendously ineffective. To address this and the former problem of building transparency, reproducibility, and trust in science it is necessary to share data in a way that makes the data findable, accessible, interoperable, and reusable. In a word – FAIR[3].

The EOSC-Nordic project has pledged to implement FAIR in the Nordic and Baltic region. This will be achieved by disseminating the benefits of going FAIR to a broad scientific community, providing an evaluation-based recommendation on how to FAIRify a specific data repository, and supporting communities to become more FAIR by hosting dedicated community events that address specific FAIRification challenges.

2. Landscaping

2.1 Sample selection criteria

The project's landscaping initiative is intended to measure the state of FAIR uptake among scientific data repositories in the Nordic and Baltic region and to monitor the development of FAIRness during the project period. We consider a minimum of 18 months of monitoring from the first epoch of FAIRness measurements to be a realistic goal within the project total runtime of 3 years, providing time for the development of a suitable evaluation procedure and for the execution of multiple epochs of FAIR evaluations.

The first step of the task was to define the criteria for repositories to be included in the study. From the outset, these included geographical criteria to **Nordic and Baltic countries** and topical criteria of **digital scientific repositories**.

[1] Federer et al. 2018, Data sharing in PLOS ONE: An analysis of Data Availability Statements https://doi.org/10.1371/journal.pone.0194768

[2] Data Science Report 2016, http://visit.crowdflower.com/data-science-report.html

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

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

There are several studies and reports that have outlined the specifications of a data repository[4]. The most relevant ones are from CoreTrustSeal[5], re3data.org[6], FAIRSharing and publishers[7], Wu et al. 2019[8], and Science Europe[9]. Although we have attempted to follow similar criteria as those used in the listed studies, we arrived at the sample selection criteria listed in Table 1 for this study.

Table 1. Candidate data repository sample selection criteria

1.	The candidate data repository must contain research relevant data
2.	The candidate data repository must not primarily or exclusively be hosting publications or documents (such repositories were excluded from the sample)
3.	The candidate data repository must contain data relevant to Nordic or Baltic countries (either covering the geographic region or be part of a Nordic research project)

2.1.1 Criteria 1

This is a rather obvious criterion that specifies that our survey is tailored for *research data*. Data repositories that do not exclusively contain research data have in some cases been included in the sample, but the majority of the repositories are research data archives.

2.1.2 Criteria 2

We are interested in evaluating the FAIRness of data repositories and therefore exclude a sizable sub-sample of repositories that primarily (or exclusively) hold documents, articles, reprints etc. This is to avoid the meddling of different types of repositories, which likely have very different incentives and ambitions on becoming FAIR.

[4] Group of European Data Experts in RDA (GEDE-RDA) Available at: https://github.com/GEDE-RDA-Europe/GEDE/tree/master/Repositories. See esp "About Digital Repositories-v2-2.docx", FAIR Converge Matrix Working Group. Available at https://osf.io/xqfb9/. See esp. "Formal Repository Descriptions.docx"

[5] CoreTrustSeal requirements. Available at https://doi.org/10.5281/zenodo.3638211
Extended guidance https://doi.org/10.5281/zenodo.3632533

[6] Rücknagel, J., Vierkant, P., Ulrich, R., Kloska, G., Schnepf, E., Fichtmüller, D., Reuter, E., Semrau, A., Kindling, M., Pampel, H., Witt, M., Fritze, F., van de Sandt, S., Klump, J., Goebelbecker, H.-J., Skarupianski, M., Bertelmann, R., Schirmbacher, P., Scholze, F., Kramer, C., Fuchs, C., Spier, S., Kirchhoff, A. (2015): Metadata Schema for the Description of Research Data Repositories: version 3.0, 29 p. DOI: http://doi.org/10.2312/re3.008

[7] The pre-print article: https://osf.io/m2bce/. Commenting available: https://tinyurl.com/RepoCriteriaFeedback

[8] Wu, M., Psomopoulos, F., Khalsa, S.J. and de Waard, A., 2019. Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories. Data Science Journal, 18(1), p.3. DOI: http://doi.org/10.5334/dsj-2019-003

[9] Science Europe: Practical guide to the international alignment of research data management D/2018/13.324/4 (2018). Available at: https://www.scienceeurope.org/wp-content/uploads/2018/12/SE_RDM_Practical_ Guide_Final.pdf

2.1.3 Criteria 3

This restriction is to limit the survey to the geographical region of the Nordic and Baltic countries as the project has a regional focus to promote the FAIR principles.

2.2 Evaluation requirements

The FAIR principles focus on aspects of data and metadata that make discovery, access, interoperability and reuse of data practical and possible. The principles were created with the intention not to force technologies or other specifics onto the data providers and/or users. However, in order to implement the FAIR principles, it is necessary to make some explicit choices. In our context, this implies certain requirements in order to perform the FAIR maturity evaluation.

Table 2. The requirement for repository evaluation

	The repository's datasets must have a Globally Unique Identifier (GUID)[10]

The first requirement (see Table 2) is related to **identification**. In order to perform a machine-actionable analysis of a dataset content it must be possible to uniquely identify it. This means the datasets must have a globally unique identifier (GUID) and is a requirement that stems from the choice of the FAIR Data Systems tool we have chosen, in order to execute the FAIR maturity evaluations (see the relevant section on the evaluation tool).

Another consideration is related to **access to metadata**. Although the FAIR principles themselves do not require *any* metadata to be openly available, a prospective repository for evaluation *should* contain descriptive machine-actionable metadata to enable the evaluation and potential re-use to be considered.

In fact, we can not see why *certain* descriptive metadata would not be available in every conceivable scenario for a dataset. The existence of minimal metadata that provides some (albeit limited) context to the content of the dataset is needed in order to provide prospective users with the tools to evaluate the suitability of the dataset in relation to their requirements. However, we can find no critical argument to set a formal requirement in order to perform the evaluation. In the absence of machine-actionable descriptive metadata, the evaluation will simply not pass the various FAIRness indicator tests, resulting in a low FAIR-score.

The GUID is therefore required in order to execute a machine-actionable FAIR maturity evaluation.

[10] Globally Unique Identifier (GUID): http://bioimages.vanderbilt.edu/pages/guid.htm

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

2.3 Survey sample selection

The survey to collect relevant and representative scientific digital repositories was conducted in two parts; the *first part* was performed using re3data.org as a source by selecting repositories from the Nordic/Baltic region, while the *second part* consisted of an internal survey amongst the team members of the WP4 project group (utilising the extended network and knowledge of the group).

The re3data.org site collects repositories from DataCite as well as via individual repository submissions. However, not all relevant data repositories are registered in re3data. Each member of our survey team approached contacts in their countries (such as research funding agencies or national research councils) to identify additional local or national repositories subject to the criteria described above.

In total, the surveying team collected 74 candidates from the re3data source and 63 from the internal survey. Among these 53 and 44, respectively, passed the selection and evaluation criteria and were consequently included in the sample. In other words, a total of 40 repositories were discarded from the sample due to them not fulfilling the sample selection criteria, some of which were excluded on account of them only hosting publications (typically pre-prints or reprints) and some of which have no obvious potential use in research.

The sample selection yielded 98 repositories associated with one or more of the Nordic/Baltic countries. Among these repositories, in accordance with the evaluation criteria, 24 were discarded on account of not providing a Globally Unique Identifier (GUID) to individual datasets. Thus our final sample consisted of 74 repositories. It is worth mentioning that many repositories that do not currently provide a GUID, are becoming more aware of the value of such a feature and hopefully we will see a few of these repositories adding this during the project period.

Each repository was evaluated based on typically ten randomly selected datasets, each with a unique valid identifier. See Table A1 for a complete list of repositories that were included in the sample. We have marked those that were fit for evaluations in green and those that could not be evaluated in red (details provided in Table A1).

2.4 Quality control

Although the evaluations of the digital objects (DOs) are fully automated once an identifier is provided, the process of selecting the datasets from the list of data repositories is not. The selection of datasets for FAIR maturity evaluations is a tedious and heavily manual effort, which requires quality control by independent members of the team in order to ensure that the selection criteria have been applied correctly.

Below are the major verifications / quality control checks for repositories and related datasets, that were provided to evaluators in the form of instructions.

2.4.1 Verify selection criteria (is repository a valid research data deposit)

- The repository must contain research relevant data and allow for the deposit of such data. Repositories that are not primarily intended for research data (such as government data archives of public records, statistics, etc) shall be excluded. The same applies to repositories that exclusively

www.eosc-nordic.eu

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

host publications, scientific articles, conference proceedings etc (e.g. archive of institution's publications).

2.4.2 Verify evaluation requirement

- It must be possible to uniquely identify a dataset / data record using a so called globally unique identifier (GUID)

2.4.3 Verify the repository attributes

1. Verify name, email, domains covered and check that datasets can be identified using a GUID. Also, try to determine whether the repository supports APIs and note down this information (e.g. REST, OAI-PMH, SOAP etc)

2. For repositories that have previously been marked RED, check if they are now providing GUIDs. If GUIDs are found, proceed to item 3.

3. Having found a way to identify individual datasets, please select N=10 random datasets from the site. These datasets should sample different domains (if possible) or different types/categories (if present). For domain-specific repositories select different dataset formats or types if possible. Add the URL to each dataset in the list of datasets.

4. If PIDs are available for DOs (e.g. DOI, Handle or others), please add the corresponding PIDs for each of the datasets

The selection process was executed by one assigned individual, while the quality control was performed by a different member of the task force. The majority of the repositories had no discrepancy between the two individuals (selection agent and quality controller). A few datasets showed discrepancies, mainly those that originated from repositories that were interfaces to databases. In these cases, the dataset was the result of a database query. Although the identifier existed, it was valid only for a short period of time (deprecated). Trying to access the identifier at a later time point resulted in an error. These repositories were removed as we require permanent identifiers that would at least last long enough for the assessment to be made.

Since one of the major goals of our project is to trace the change in the adoption of the FAIR principles over time (by evaluating regularly the FAIR maturity), it makes sense to cast the net as wide as possible and select repositories that represent diversity in geography, domain of science as well as technical approach and proficiency.

3. FAIR evaluation methodology

Evaluation and assessment of FAIRness of digital objects or data repositories is a new and evolving topic. In the last few years, several assessment tools with different approaches have emerged. The Research Data Alliance (RDA) Working Group 'FAIR data maturity model'[11] was established in January 2019 with the aim to

[11] https://www.rd-alliance.org/groups/fair-data-maturity-model-wg [Accessed 2020-08-11]

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

build on existing initiatives and to identify core elements for the evaluation of FAIRness that will hopefully increase the coherence and interoperability of FAIR assessment frameworks and ensure the compatibility of their results. Their recent report defines a set of indicators, their priorities and evaluation methods that can be used as a common approach across assessment methodologies (FAIR Data Maturity Model Working Group 2020)[12] that will guide the work in the future.

At the moment, most FAIR evaluation frameworks are based on manual questionnaires or checklists[13,14] that are not scalable. Also, studies exploring the FAIRness of repositories and their data have been based on manual collecting of information. Dunning et. al. (2017)[15] collected a sample of 37 repositories that was either affiliated to the Netherlands or was popular with the Dutch research community, and scored their FAIRness based on the information available on the repository's website and in the published data records. They found out that less than half of the repositories complied with many FAIR facets and that I and R were the most difficult ones. Ivanović et. a. (2019)[16] created two questionnaires and collected responses from managers and/or librarians of 29 repositories and from technical staff of 14 repositories. Their results show that the FAIR principles were only partially implemented, that several principles are complicated and that there are some misunderstandings and even misleading implementations. They also concluded that training, guides and best-practice examples as well as checklists to help ascertain FAIR compliance are needed. Stockholm University has evaluated four research data platforms (Dataverse, Figshare, Zenodo and SND) and assigned them FAIR evaluation scores of 17, 10, 14.5 and 18 (out of 22) respectively[17]. The evaluation by the Stockholm University is partly based on Wilkinson et. al. (2018)[18] that summarises the self-evaluations of nine repositories that tested the usability of the core set of FAIR metrics. The ongoing FAIRsFAIR project will implement both manual and automated FAIR data object assessment tools although they note that "automated assessment is essential but might be difficult to be fully achieved during the active phase of the project due to heterogeneity of standards and requirements of the various science communities, and lack of machine-readable resources (e.g., registries and standards) to support the assessment" (Devaraju et. al. 2020).[19]

In order to improve the sharing of research data, it is important to know how to measure this. Since most datasets today are shared through generic or domain-specialised digital repositories, we made an early and obvious choice to direct our attention towards these research data repositories. Given the high number of data repositories in our sample, manual assessment of the FAIRness of their datasets was not feasible. Nor

[12] FAIR Data Maturity Model Working Group. (2020, June 25). FAIR Data Maturity Model. Specification and Guidelines (Version 1.0). http://doi.org/10.15497/rda00050

[13] RDA FAIR Data Maturity Model Working Group (2019). Survey of existing FAIR assessment tools and approaches. https://docs.google.com/spreadsheets/d/14ojMSXVOITg3RoJn-PuDaPj8zuIGQz2Li-kl97HOBH4/edit#gid=0 [Accessed 2020-08-11]

[14] Bahim, Christophe, Dekkers, Makx, & Wyns, Brecht. (2019, May 23). Results of an Analysis of Existing FAIR Assessment Tools. http://doi.org/10.15497/rda00035

[15] Dunning, Alastair; de Smaele, Madeleine; Böhmer, Jasmin (2017). Are the FAIR Data Principles Fair? International Journal of Digital Curation 12 (2) , p. 177 –195. https://doi.org/10.2218/ijdc.v12i2.567

[16] Ivanović, Dragan, Schmidt, Birgit, Grim, Rob, & Dunning, Alastair. (2019). FAIRness of Repositories & Their Data: A Report from LIBER's Research Data Management Working Group. Zenodo. http://doi.org/10.5281/zenodo.3251593

[17] https://www.su.se/english/staff/services/research/research-data/data-repositories [Accessed 2020-08-11]

[18] Mark Wilkinson, Erik Schultes, Luiz Olavo Bonino, Susanna-Assunta Sansone, Peter Doorn, & Michel Dumontier. (2018, July 4). FAIRMetrics/Metrics: FAIR Metrics, Evaluation results, and initial release of automated evaluator code (Version v1.0.3). Scientific Data. Zenodo. http://doi.org/10.5281/zenodo.1305060, document Evaluation_Of_Metrics/Supplementary Information_ FM Evaluation Results.pdf

[19] Devaraju, Anusuriya, & Herterich, Patricia. (2020). D4.1 Draft Recommendations on Requirements for Fair Datasets in Certified Repositories (Version v1.0_draft). Zenodo. https://doi.org/10.5281/zenodo.3678715

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

was it desirable, due to the inferior reliability and demanding resources required to execute manual evaluations over automated ones.

Assessing the FAIRness of digital objects (i.e. datasets) in a repository has a number of challenges that we have tried to take into account. First of all, there is no single resource that lists all such data repositories. Further, many of the repositories have been operating for a number of years and the metadata has evolved over time. Thus, there is a large discrepancy in the maturity of what metadata is provided with datasets and how this metadata is presented. In addition, early datasets may have poorer quality metadata than more recent datasets due to the maturing of the general data management aspects of research projects. In our selection of datasets, we try to select datasets covering the full lifetime of the repository to avoid clustering effects, which could bias the results.

Some repositories span multiple disciplines. To account for any potential non-uniformity arising from disciplines having more or less mature metadata management frameworks we have tried to randomly select datasets from different disciplines.

From early on in the project, we realised that the study would require having an automated (machine-actionable) method for evaluating the FAIRness of the datasets in these data repositories, leading to a FAIR score for each repository. In practice, this implies that each dataset must be evaluated according to the FAIR principles. The methodology must be accurately defined in order for it to be executed in a consistent way, which allows the evaluations to be repeated multiple times allowing us to track their FAIR-ness development over time. Additionally, we want to avoid subjective interpretation leading to inconsistent evaluations and results. We would also need to be able to archive test results for future reference and comparison, so an administrative tool for managing the evaluations would be beneficial.

At the time the EOSC-Nordic project started (Sep 2019) there were no alternatives to automated FAIR evaluations than that provided by the tool of Mark Wilkinson (Wilkinson et al. 2019). This tool is based on the concept of FAIR metrics (Wilkinson et al 2018). In fact, during the EOSC Symposium 2019 other initiatives that were evaluating repositories and datasets were doing this using manual evaluation forms executed by a group of evaluators or by requesting repositories to self-evaluate according to a schema. The EOSC-Nordic choice of an automatic evaluation tool and indeed the execution of a systematic study of repositories and datasets appears to be unique and a first. Many are averse to automated evaluations and solely relying on the machine's evaluation, claiming that what is tested is far too limited and that machine-actionability cannot be a requirement. However, FAIR will have little or no impact without the requirement of machine-actionability as a core driver in order to enable sustainable and scalable data management. The aspect of machine-actionability is indeed embedded into FAIR principles.

The results and lessons learned from this first semi-automated systematic study may therefore gain substantial interest among data providers all over the world.

3.1 FAIR Maturity evaluation tool

The FAIR Maturity evaluation of digital objects (datasets) consists of scoring certain tests based on harvesting specific metadata and in some tests also data from the digital object. To be able to harvest this information the tool must succeed in retaining information from machine-actionable metadata, as detailed in Wilkinson et al. 2019. Machine-actionable means that a 'bot' or a *harvester* ("the machine") can recognise what kind of service (potentially a data repository) it is interacting with, including what kind of data is hosted (e.g. research domain, formats), what the usage license is, provenance details, external links to other FAIR digital resources and in certain tests also retrieving information about the data content. In

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

short, the use of machine-actionable metadata and data is ultimately intended to obtain the goal “the machine knows what I mean”. This all relies on a set of common descriptive metadata that can be understood by ‘the machine’.

The goal of the FAIR Maturity evaluation of digital objects is to demonstrate the benefits of machine-actionable metadata, their crucial role in adding value to a digital repository and raise general awareness of FAIR data. An additional ambition is to monitor the evolution of FAIR metrics for individual digital objects (and data repositories) to trace the (positive) development of FAIRer research data and, if possible, to gauge its impact on data reuse and scientific quality.

3.2 FAIR metrics

The FAIR maturity indicators used in the EOSC-Nordic project have been adopted from the FAIR Maturity evaluator tool developed by FAIR Data Systems and described in Wilkinson et al (2019) as implemented in version 1.0.20 of the fairdatasystems/fair-tests.

The list in Table 3 provides an overview of the indicators and their relevance to specific FAIR principles. Apart from missing indicators for R2 and R3 principles, it is quite evident that some principles (i.e. F1) are well represented by tests, while others are poorly covered. It is not necessarily possible to quantify or formulate an indicator (test) that reliably indicates whether the relevant aspect is satisfied or not. There may be technical and/or conceptual challenges associated with this. To determine whether certain features are present, it may be necessary to test whether a certain combination of indicators has been met and provide additional credits for that in the scoring process (if that is an objective).

Table 3. FAIR Maturity Indicators

Id	Metric name	Principle association	Principle description
1	UNIQUE IDENTIFIER	F1	(Meta)data are assigned a globally unique and persistent identifier
2	IDENTIFIER PERSISTENCE	F1	(Meta)data are assigned a globally unique and persistent identifier
3	DATA IDENTIFIER PERSISTENCE	F1	(Meta)data are assigned a globally unique and persistent identifier
4	STRUCTURED METADATA	F2	Data are described with rich metadata (defined by R1 below)
5	GROUNDED METADATA	F2	Data are described with rich metadata (defined by R1 below)
6	DATA IDENTIFIER EXPLICITLY IN METADATA	F3	Metadata clearly and explicitly include the identifier of the data they describe
7	METADATA IDENTIFIER EXPLICITLY IN METADATA	F3	Metadata clearly and explicitly include the identifier of the data they describe
8	SEARCHABLE IN MAJOR SEARCH ENGINE	F4	(Meta)data are registered or indexed in a searchable resource
9	USES OPEN FREE PROTOCOL FOR DATA RETRIEVAL	A1.1	The protocol is open, free, and universally implementable
10	USES OPEN FREE PROTOCOL FOR METADATA RETRIEVAL	A1.1	The protocol is open, free, and universally implementable

EOSC-Nordic project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857652

11	DATA AUTHENTICATION AND AUTHORIZATION	A1.2	The protocol allows for an authentication and authorisation procedure, where necessary
12	METADATA AUTHENTICATION AND AUTHORIZATION	A1.2	The protocol allows for an authentication and authorisation procedure, where necessary
13	METADATA PERSISTENCE	A2	Metadata are accessible, even when the data are no longer available
14	METADATA KNOWLEDGE REPRESENTATION LANGUAGE (WEAK)	I1	(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
15	METADATA KNOWLEDGE REPRESENTATION LANGUAGE (STRONG)	I1	(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
16	DATA KNOWLEDGE REPRESENTATION LANGUAGE (WEAK)	I1	(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
17	DATA KNOWLEDGE REPRESENTATION LANGUAGE (STRONG)	I1	(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
18	METADATA USES FAIR VOCABULARIES (WEAK)	I2	(Meta)data use vocabularies that follow FAIR principles
19	METADATA USES FAIR VOCABULARIES (STRONG)	I2	(Meta)data use vocabularies that follow FAIR principles
20	METADATA CONTAINS QUALIFIED OUTWARD REFERENCES	I3	(Meta)data include qualified references to other (meta)data
21	METADATA INCLUDES LICENSE (STRONG)	R1.1	(Meta)data are released with a clear and accessible data usage license
22	METADATA INCLUDES LICENSE (WEAK)	R1.1	(Meta)data are released with a clear and accessible data usage license
		R1.2	(Meta)data are associated with detailed provenance
		R1.3	(Meta)data meet domain-relevant community standards

3.3 Methodology and strategy choices

3.3.1 Repositories

We have gathered repositories, containing datasets and metadata records, from various sources that are geographically linked to our region and intend to evaluate the digital objects (research datasets) in these repositories using a systematic method that is reproducible. The framework of our study addresses digital repositories, although we do not assume that such repositories must be actively curating or preserving their content.

On rare occasions, data repositories provide restricted access to data and metadata, requiring authentication before metadata can be searched or queried. This is in conflict with FAIR principles F4 and A1 and pretty much prevents any evaluation (at least using the FDS-tool) to be carried out. Such repositories are nevertheless evaluated provided datasets are uniquely identified using a GUID, but are unlikely to pass most of the test indicators due to the lack of access to the (machine-actionable) metadata.

3.3.2 Metadata

What we mean by metadata being openly available is that *descriptive metadata* should be open and available, intended to enable (or even maximise) data discovery. The most important element of the descriptive type of metadata is the resource identifier that uniquely identifies the digital object. Other elements include title, author, date of publication, subject, publisher and a description. There are numerous types of metadata; descriptive, structural, administrative, reference and statistical metadata (cf. Wikipedia) to name a few. Requiring that all metadata be available in the context of search and discovery is not a constructive or meaningful requirement, both because far from all metadata are relevant without the data in hand and certain metadata may be sensitive or even restricted.

3.3.3 DOIs in evaluations

The Digital Object Identifier[20] is one of the persistent identifier (PID) services available. It is maintained as a commercial service and is in a position to provide guarantees on the sustainability of the identifiers. It is one of the most frequently used PID services by data repositories which is also supported by the FAIR maturity evaluator tool. However, DOIs behave somewhat differently than other services in that they provide centralised metadata for the DOI. The FAIR evaluation tool author, Mark Wilkinson, describes how content-negotiation is intercepted in the case of DOIs:

“DOI is a special case because it *hijacks* content-negotiation. When you request e.g. text/turtle on a DOI, the response is served by CrossRef or DataCite, NOT by the source data provider. When you allow default content negotiation, you end up at the landing page of the data provider... that landing page MIGHT ALSO RESPOND to content negotiation! But that requires a specific workflow that knows how to capture the final URI in the DOI's redirection chain, and then re-start the content negotiation process on that URI.”

For the reason described above, a dataset with a DOI is treated as a ‘special case’ by the evaluator, as the metadata is retrieved from multiple sources. Evaluation results when providing a DOI reflect a combination of features provided by the data repository itself and the hosted (DOI) metadata provided by Crossref/Datacite. Since DOI-provided metadata is not harvested directly from the source, there is a risk that the metadata may be outdated. It is therefore suggested that we rely on a repository intrinsic evaluation (using direct URI for testing datasets with a DOI). We have evaluated both DOI and URI identifiers for digital objects, hence duplicating these objects in the evaluation. However, it seemed reasonable to include the additional indicator test scores the provision of a global unique identifier (GUID) offers.

3.3.4 The generalisation of evaluation results from digital object level to repository level

The aim is to evaluate the FAIR maturity of digital data repositories. These repositories consist of various digital objects (e.g. datasets and related metadata records). Intuitively, it may seem reasonable to evaluate all digital objects contained within a repository, in order to determine its FAIR maturity. Technically this is

[20] Digital Object Identifier: http://www.doi.org/

EOSC-Nordic project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857652

also possible, but extracting the identifiers for each dataset would require a large effort (there is no generic way in which we can list/extract all datasets in a repository).

Our initial assumption was that the evaluation of one dataset within a repository using the FAIR Maturity evaluation tool would give identical results to any other evaluation of datasets within that same repository, since we did not expect there to be features that should vary on an individual dataset level. For example, if the licensing key is provided for one dataset it seems safe to assume that it should be provided for all other datasets within the repository. This was based on the assumption that the underlying software platform would be the deciding factor on whether a specific key is provided or not. However, this was found *not* to be the case. In fact, we found that in a number of repositories, only a small sample of the datasets were equipped with the mentioned licensing key. This variation in scores between datasets within the same repository leads to our scores containing a certain degree of stochastic effects, which is reflected in the errors (measured standard deviations) listed in Table A2.

The lesson is therefore, that it is a good practice to randomly select subsets from a repository and to evaluate those datasets (the more the better) to obtain an *indication* of the FAIR maturity of the data repository itself. We can not really get a quantifiable estimate of how accurate this approximation is without evaluating a substantially large proportion of datasets in a repository. In addition, it is worth noting that a digital object cannot really be made FAIR or evaluated for FAIRness in isolation from its context - in our case, the data repository (L'Hours et. al. 2020)[21]. For example, the persistence of an identifier is determined by the commitment of the organisation that assigns and manages it. This is an area where work is ongoing, for example, the FAIRsFAIR project aims to develop requirements and tools to pilot FAIR assessment of research data objects in trustworthy digital repositories (Devaraju et. al. 2020)[22].

3.3.5 Execution of evaluations

The results presented in this report were executed using the FAIR maturity evaluator (v.1.0.20) developed by FAIR data systems. The public version of this evaluator was experimentally used from the onset of the project in September 2019 until it was determined that its stability and service level was not adequate for dependable regular use by the project. It was therefore agreed with FAIR data systems that a license would be purchased by the project for the duration of the project lifetime (until 2022), and development / installation assistance by Mark Wilkinson would be provided. The version iterations of testing and providing feedback have been numerous during the first half of 2020, but the code now seems to have converged to an acceptable level for our evaluation purposes. This project has contributed back continuously with bug reports and a number of performance and content related suggestions, leading to many of the version increments listed in Appendix B.

In addition to improving the evaluation tool itself, we have also developed (in-house, by partner ETAIS) scripts to assist with the automation of the execution of hundreds (and now over a thousand) dataset evaluations. These evaluations will be repeated every 3 months during the project lifetime. The scripts enable, amongst other features, parallelisation of the evaluation executions.

[21] Hervé L'Hours, Ilona von Stein, Frans Huigen, Anusuriya Devaraju, Mustapha Mokrane, Joy Davidson, … Patricia Herterich. (2020). D4.2 Repository Certification Mechanism: a Recommendation on the Extended Requirements and Procedures (Version 01.00). Zenodo. https://doi.org/10.5281/zenodo.3835698

[22] Devaraju, Anusuriya, & Herterich, Patricia. (2020). D4.1 Draft Recommendations on Requirements for Fair Datasets in Certified Repositories (Version v1.0_draft). Zenodo. https://doi.org/10.5281/zenodo.3678715

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

The first evaluations were executed over 4 days in January 2020. The evaluations presented here were executed on July 20, 2020 and took 12,5 hours to complete. From February 2020 to present the project has executed 7 recorded epochs of full sample evaluations, but continuously with incremental versions with improvements and various procedural changes.

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

4. Results

In Appendix A, Table A2 we present the resulting average FAIR scores for the survey and corresponding averaged evaluations per data repository. The methodology is described in the preceding sections. Average FAIR Maturity scores are calculated based on typically N=10 dataset evaluations from one and the same repository that are then averaged per the F, A, I and R categories. In the following sections, we present the distribution of F, A, I and R scores for the sample and in the final section the averaged total FAIR scores.

In Appendix A, Figure A1 we display all the individual FAIR score results from dataset evaluations (more than a thousand datasets). Note that some of the datasets are evaluated twice because they were evaluated both using the direct URI and if present using the GUID (DOI, Handle etc). The lowest score is 13.64% and corresponds to 3 out of 22 passed tests, which is the minimum number of indicators passed whenever a GUID or URI is given to the evaluator (the indicator numbers for these tests are 1, 10 and 12 in Table 3).

4.1 Findability

The average F-score results for evaluated repositories are shown in the histogram in Figure 1. It shows that the F-scores has a minimum of 12.5% (corresponding to 1 out of 8 tests). The results vary from this minimum value of 12.5% to 62%, but are otherwise scattered between these values.

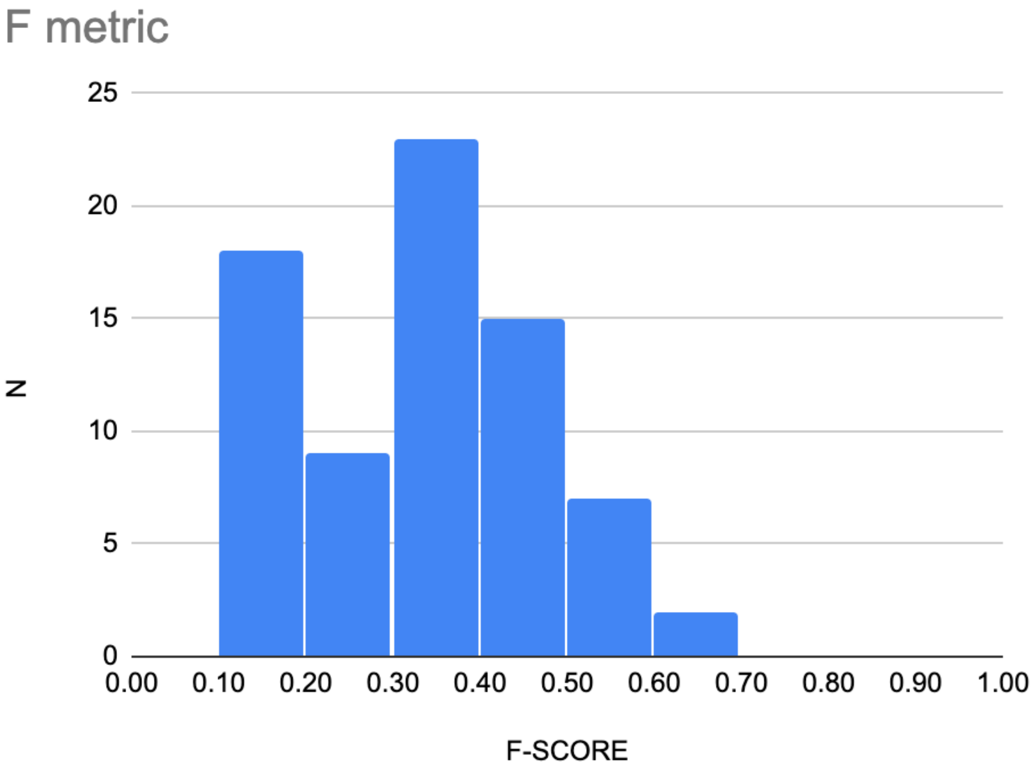


Figure 1. Histogram of F metric scores

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

4.2 Accessibility

The average A-score results for evaluated repositories are shown in the histogram in Figure 2. It shows that the A-scores has a minimum of 40% (corresponding to 2 out of 5 tests). The results vary from this minimum value of 40% to 80% and seem to be clustered mostly around 40% with a secondary peak from 60-80%.

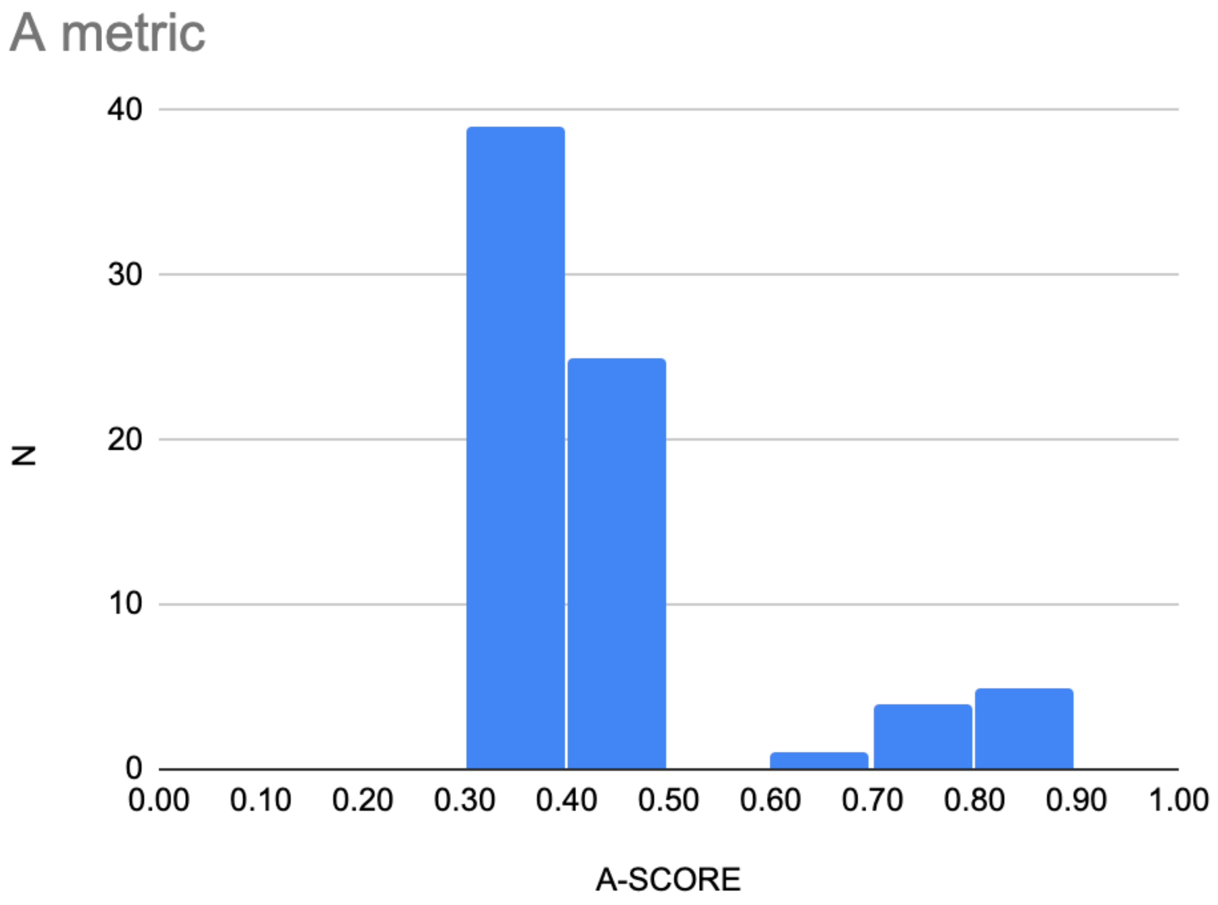


Figure 2. Histogram of A metric scores

4.3 Interoperability

The average I-score results for evaluated repositories are shown in the histogram in Figure 3. It shows that the I-scores are scattered quite well between the minimum of 0% (corresponding to 0 out of 7 tests) to approximately 70%. The results show peaks around 0% and 70% in addition to a broader peak in the range 20-50%.

EOSC-Nordic project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857652

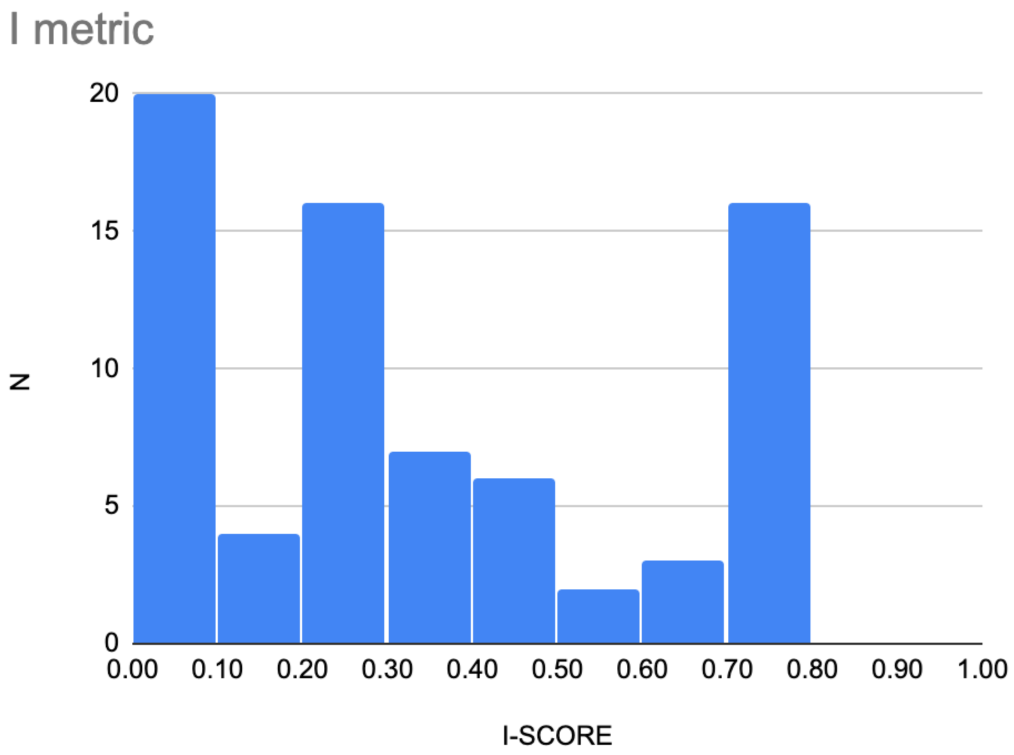


Figure 3. Histogram of I metric scores

4.4 Reusability

The average R-score results for evaluated repositories are shown in the histogram in Figure 4. It shows that the R-scores are mostly clustered around the 0% (no support for any of the two reusability tests). A small number of repositories score from 0-100%, but the low number of counts indicate that the tests for licensing keys are only supported by a small number of repositories and in some cases only for a small number of datasets within those repositories.

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

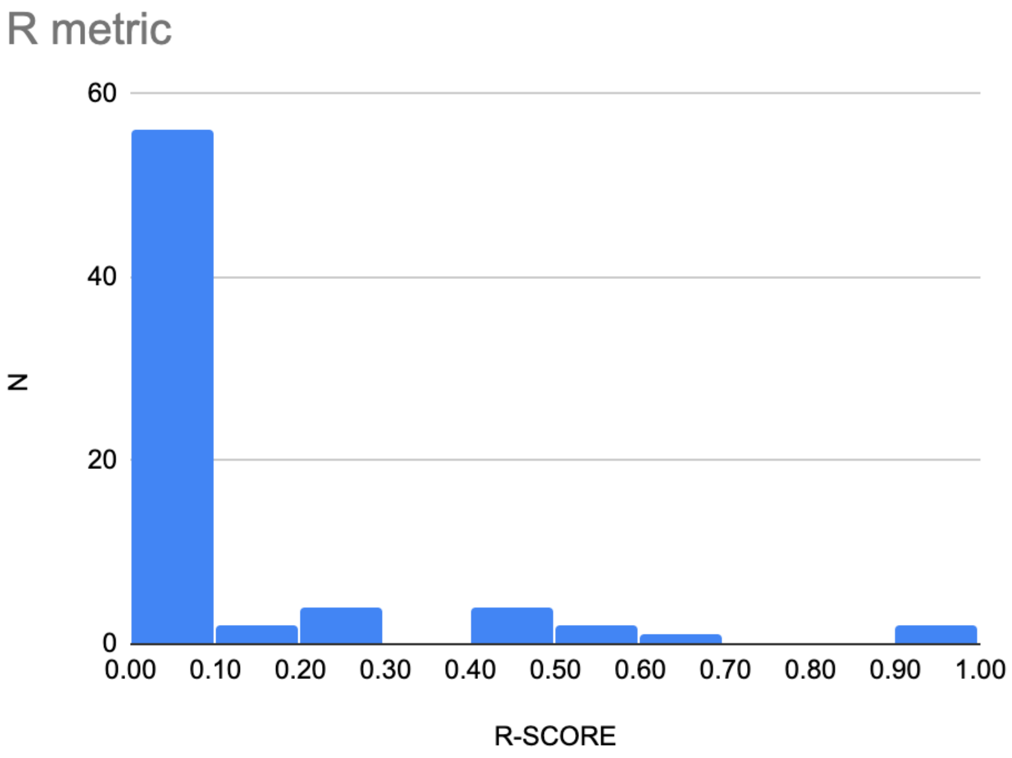


Figure 4. Histogram of R metric scores

4.5 How to interpret the results

The evaluation results and accompanying FAIRification recommendations can be used to guide any efforts to make a repository FAIRer.

4.5.1 Maturity level 0

A null score means that the evaluator could not be run on the repository due to the lack of a GUID. Repositories that do not provide a unique identifier for each dataset can not be tested using the evaluator.

4.5.2 Maturity level 1 (low)

Three indicators will pass by simply providing a URI as an identifier and are; *a unique identifier*, *open free protocol for metadata retrieval* and *metadata authentication & authorisation*. A substantial number of repositories fall in the category that only pass 3 out of 22 tests.

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

A majority of repositories score in the low category with less than 33% of the tests passed (7/22 or less). Although the datasets are equipped with GUIDs, there is not much trace of machine-actionable metadata compatible with the FAIR principles.

4.5.3 Maturity level 2 (medium)

For this medium maturity level, scores fall between 8-11 passed tests, obtaining between 33-50% on the FAIR score. Repositories at this level are employing some machine-actionable metadata and data that score in this category has an advantage over less FAIR data. Often their FAIRness can be improved by adopting more extensive use of FAIR vocabularies, a license predicate, identifying the digital object explicitly in the metadata to name a few.

4.5.4 Maturity level 3 (high)

Very few repositories reached the high maturity level, with 12-17 passed tests. The few repositories that did obtain 50-72% on the FAIR score, depending on how consistent those results are over multiple datasets. In this category the use of machine-actionable metadata, and in some cases, the data itself is a result of strategic decisions and priorities to enable open and accessible data.

Beyond the 72% score we could characterise the data as being of very high maturity level, but we have set this limit somewhat arbitrarily at our maximum score to encourage repositories to reach further towards full FAIR compliance. At the moment there are not many additional tests in the current generation of the FAIR evaluator that can explore the FAIRness aspects much further, but this will change as the field evolves. The most natural advancement would likely be related to principles R1.2 and R1.3 – provenance metadata and domain specific standards, where no tests are currently implemented. However, it is not a coincidence that it is these two that lack tests, as the challenge of defining provenance related tests currently lack established standards and the domain standards are highly variable in maturity due to the differing advancement of metadata standardisation for data within the various science domains.

4.6 Combined FAIR score

The sample consists of 98 Nordic/Baltic repositories of which 24 repositories could not be evaluated due to lack of GUIDs for the hosted datasets. For whatever reason, the service provider has chosen not to identify each individual dataset using PIDs, URI or other valid identification schemes (see the section on Evaluation requirements). That means we will not be able to evaluate any datasets from approximately 25% of the repositories in the sample.

The remaining 74 repositories were evaluated using 732 digital objects with URI identifier and an additional 286 digital objects for which PIDs were available. For the datasets with PIDs the evaluations are replicated as the evaluator is run first on the URI identifier, then again using the corresponding PID (DOI, Handle etc). Although some of the test results are duplicated with this procedure, we consider that acceptable as we include the positive effects of PID based evaluations.

To obtain the averaged FAIR score by averaging the individual category scores by the number of indicators in each category. This gives a single number score for FAIR maturity for each data repository in the sample.

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

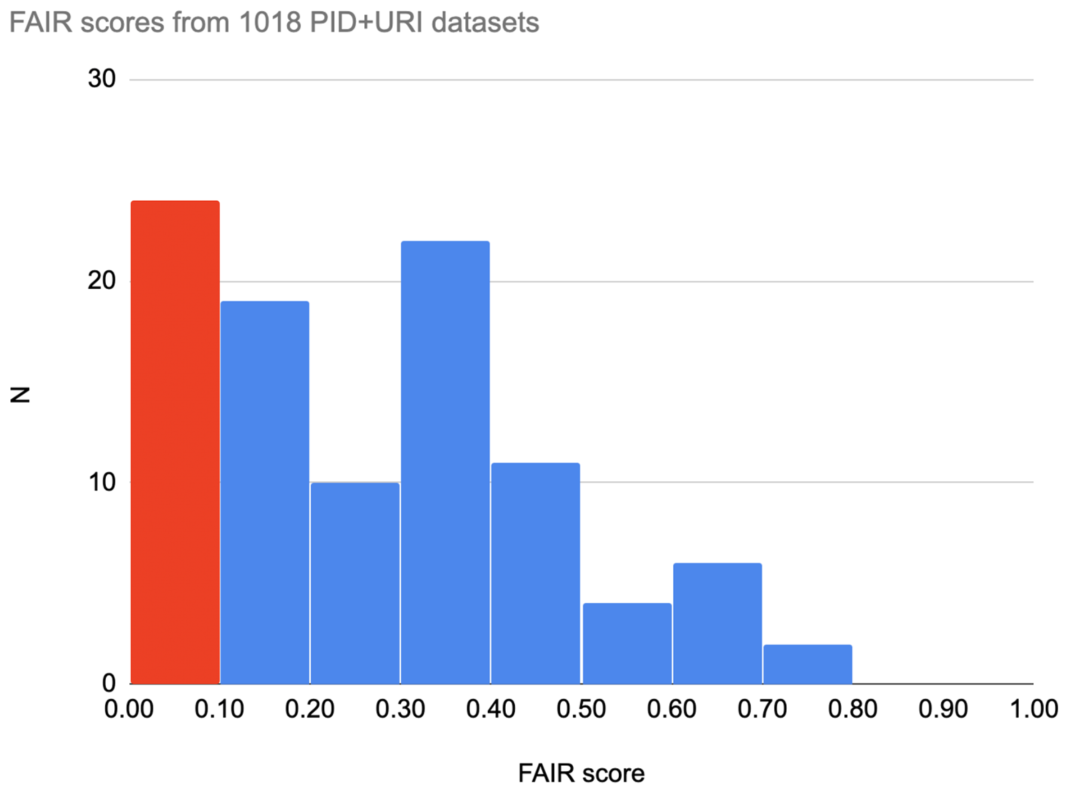


Figure 5. Averaged FAIR score histogram for data repositories that were evaluated in the sample (blue bars) and repositories

While some information is lost by collapsing the individual category scores into a single score, it does provide an indication of their overall FAIR maturity. Condensing measurements into a single number is preferable, if we at some point wish to perform correlation analysis relative to other variables.

As seen in Figure 5, a majority of the 74 FAIR maturity evaluated repositories score in the lower category, meaning they have not much support for machine-actionable metadata. This is not surprising given the recent onset of FAIR adoption globally. A previous meta-study[23] based on re3data.org repositories found that adoption of FAIR in the Nordic region was still in its infancy in 2018. However, we are pleased to note that a small selection of repositories has quite advanced support for machine-actionable metadata and data.

The collection contains DOs using both direct URIs and PIDs. If we exclude the (duplicate) DOs provided with a PID references and keep only those DO references that are given by a direct URI (732 DOs), we see that the average FAIR score drops to 31.2%. Table 4 gives the averaged results for various sample selections, including using the duplicate *URIs+PIDs* and using *URIs* only. The data suggest that the overall FAIR score is reduced when applying only the direct-URI evaluated datasets. Overall the FAIR score drops from 34.5% to 31.2%. As seen in Table 4, the drop is contributed by reductions in the F, I and R segment of the FAIR score (there is no measured variation from the A segment between the two samples). It is also visually evident, comparing the two graphs in Figure 6, that the FAIR score drops appreciatively when one

[23] A O Jaunsen, Nordforsk/NeIC report 2018 (https://doi.org/10.5281/zenodo.2563733)

excludes the evaluations based on PIDs. It is particularly DOIs that contribute to the boosting of the FAIR scores, due to the fact that doi.org is providing some of the (machine-actionable) metadata.

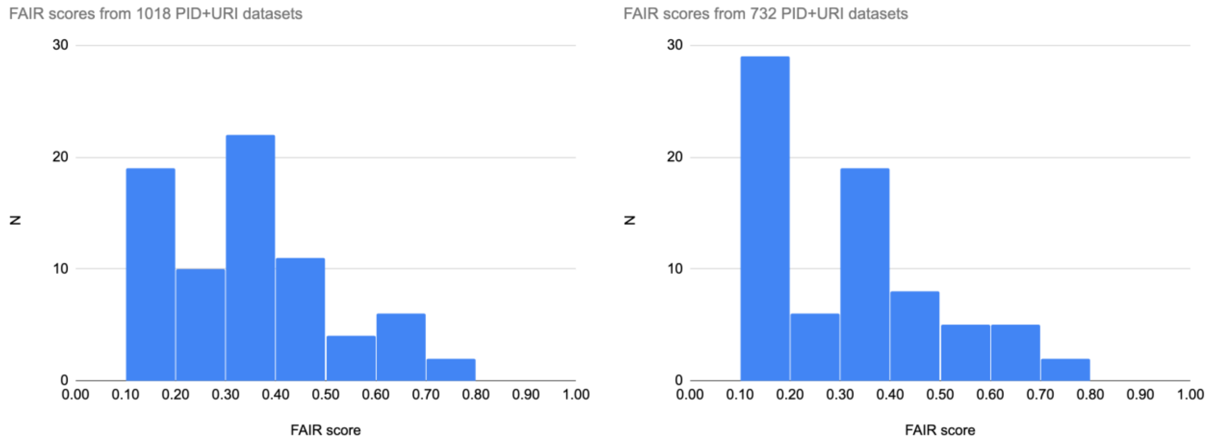


Figure 6. Histogram of FAIR scores for 74 repositories, based on evaluations of 1018 URI+PID (left) and 732 URI (right) digital objects.

Among the 74 evaluated repositories, 17 of them have been developed on or are run on established software platforms. As mentioned earlier (and shown in Table 4), the full sample has an average FAIR score of 34.5%. The sub-sample of 17 repositories with a known software platform has an average FAIR score of 46.2%. This is substantially higher than for the full sample, indicating that repositories running on known or established software platforms generally score higher when using this FAIR scoring method. This is an interesting finding and was already indicated by early evaluations made more than 6 months ago.

Furthermore, we have also evaluated a sub-sample of repositories (9), for which certifications are known to exist. The result shown in Table 4 indicates that there is also a small increase in the average FAIR score for this sub-sample, compared to the full 74 repository sample. Due to the small number of certified repositories, we are reluctant to conclude that a certified data repository is more likely to provide a higher FAIR score, but note here that there seems to be some support in our data to that effect.

It is also noted that, although some of the services or portals evaluated here may not be certified, some of the incorporated hosting services providing them may in fact be certified. This may lead to situations in which the hosting organisation is certified, although the repository service is not and therefore listed as not certified in our Table A2. An example would be Data Service Portal Aila that is not certified but is provided by a CTS certified organization (FSD).

Table 4. Results for various samples resulting from using different selection criteria. N represents the number of datasets (DOs) and # represents the number of repositories relevant to the sample. The averaged scores for F,A,I,R and combined for the relevant samples are given with standard error.

Description	N	#	<F>	<A>	<I>	<R>	err(F)	err(A)	err(I)	err(R)	<FAIR>	err(FAIR)
URIs+PIDs	1018	74	0.334	0.454	0.338	0.132	0.011	0.004	0.018	0.024	0.344	0.008
URIs	732	74	0.301	0.452	0.288	0.091	0.006	0.003	0.007	0.014	0.312	0.004
Known platf.	1018	17	0.438	0.580	0.462	0.269	0.021	0.014	0.033	0.058	0.462	0.018

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

Certified	1018	9	0.382	0.447	0.391	0.117	0.030	0.010	0.051	0.025	0.376	0.016

5. Community engagement

5.1 FAIR Maturity workshop

A one-day EOSC-Nordic workshop entitled “FAIRification of Nordic and Baltic data repositories“ was held online on April 22, 2020 with 122 registered participants and approximately 90 attendees.

The overall goal was to raise awareness of the EOSC-Nordic project to the research data providers in the Nordic and Baltic region, and to disseminate preliminary results to the repositories that had been evaluated. Further to this, we offered FAIRification guidelines and recommendations to those entities hosting research data in order to maximise the reusability of this data. The FAIR maturity evaluations of approximately 100 repositories in the region provide a unique insight into the FAIR uptake. The results were offered to the communities that had signed up to the event and further details offered on an individual basis. Specific recommendations for each of the FAIR maturity indicators was prepared and presented during the workshop in a dedicated session. The guide was provided to participants (and also linked in invitations to all communities that were part of the survey) along with the individual evaluation results in order to enable the community to select relevant areas that they wish to make FAIRer for their data.

The importance and added value of FAIRifying data repositories to adhere to the FAIR principles was presented, as well as information on how the evaluations were carried out using FAIR metric indicators and machine-actionable metadata. Based on the evaluation results, individually tailored recommendations were provided to the communities. Finally, a presentation was given on current project plans and how the EOSC-Nordic project can facilitate limited support through community hackathons, networking, and guidelines to improve the FAIR scores.

The purpose of the community workshop “FAIRification of Nordic and Baltic data repositories” was to provide guidelines and specific recommendations towards data repositories in order to maximise the reusability of their research data. We targeted not only repository managers, but also developers, community stakeholders and other interested parties who define curation activity needs and/or represent the repository host institutions. The initial FAIR maturity evaluation results were shared with the invitation to the webinar with an idea that they would help to further spark the interest in support modes available through the EOSC-Nordic project at later stages (T4.1 and 4.2).

The event was initially planned to be a full day workshop in Sweden. Due to COVID-19 the workshop was turned into a shorter webinar instead. This reduced the options for community building in the form of discussion and one-to-one guidance by the certification experts. However, the event was a success given the high number of attendees and interaction during the webinar. There were 122 registrations, more than twice the number of people that we expected to join an on site event. This shows that there is a large demand for information and guidance on FAIRification of data and services amongst key stakeholders.

Although well visited, what suffered during the community event (due to it being an online event) was the lack of direct interaction with the participants that do not raise questions or ask for help. Although there was a good dialogue with some of the participants during the Q&A / open discussion session on the event day, we were left with the distinct notion that we had no clue about the needs and FAIR proficiency of the majority of the participants. In the next phases of the project we need to find ways to tackle the risk in failing to engage with the majority of communities, taking into account the presence of continued COVID-19 restrictions. This should be done at the project level and at the national/local level in cooperation with other stakeholders.

6. Uptake of FAIR in the Nordic+Baltic region

6.1 Qualitative analysis of FAIR scores

For the full URI+PID sample (1018 digital objects) the results show that 53% of the sample score below 0.33 (one third of the 22 maturity indicators). About one third (31%) of the sample scores between 0.33 and 0.5. This means the large majority of the datasets in the sample (84%) pass less than 50% of FAIRness indicators (passing up to 11 out of 22 maturity indicators). Only 16% of the repositories in the sample can be considered to score in the 'mature' segment (above 0.5). The currently best scoring repositories pass 73% of the indicators (16 out of 22).

If we reduce the sample to the 732 digital objects referenced only using the URI identifiers, these numbers show a similar reduction seen in Table 4 with 69% of the sample scoring below 0.33, 15% between 0.33 and 0.5 and (interestingly) the same percentage (16%) scores above 0.5 as for the full URI+PID sample.

Since the FAIR score is a single number representing the combined effects of the F, A, I and R scores we lose some of the information contained in those measurements when using the FAIR score to categorise different repositories and how they score relative to others. For instance, there is no way to discern a repository with good findability, accessibility and reusability that scores 0.6, 0.8, 0.0, 1.0 (for F,A,I,R respectively) and one that is highly interoperable that score 0.2, 0.4, 1.0, 0.0 – as both repositories score 0.5. Denoting a repository for 'mature' for scores above 0.5 here is entirely arbitrary.

As the project continues to re-evaluate the sample of digital repositories we expect to see these statistics improve.

6.2 Preliminary uptake analysis

The EOSC-Nordic project began evaluating samples of datasets in early 2020, albeit while making adjustments and continuously improving the tool (in collaboration with the author Mark Wilkinson). During this period the sample of data repositories were also in the process of being established and varied somewhat depending on the actual selection criteria. Table 5 provides an overview of the evaluation dates, software version of the FDS-tool, sample size and results for F, A, I, R and the combined score, including standard errors.

Only between epoch 3 and 4 do we see a significant change in the results, e.g. changes larger than the error bars. We know that some repositories made changes to their service in this period, suggesting that at least some of the score difference is due to internal changes in those repositories. In Figure 7 the evaluations for all epochs are shown in the left panel (a total of 4958 evaluations) and in the right panel the corresponding averaged results per epoch. Here it is easy to identify the mentioned increase in the FAIR score from epoch 3 to 4. We know that a few repositories made changes to their services between these two epochs. With this knowledge and the fact that the evaluation results support an increase in scores between these epochs

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

it is tempting to accredit the increased scores to these changes. Furthermore, the change log for the code shows that there were both edits made in the harvester and a number of the test indicators (see Appendix B for changes introduced in v.1.0.14) and the code-induced score variations are evident from the left panel of Figure 7. It is not possible to conclude on the origin of the variability of the results, as code changes could account for the majority of the detected score differences. Therefore, we can not claim that the measured increase is primarily due to these improvements, as opposed to being a side-effect of the upgrade of the FDS-tool.

Table 5. Preliminary evolution of FAIR scores from multiple evaluation runs since Feb 2020, but using different versions of the FDS-tool.

Epoch	Date	Ver.	N	#	<F>	<A>	<I>	<R>	err(F)	err(A)	err(I)	err(R)	<FAIR>	err(FAIR)
1	4-5 Feb 3-4 Mar 26-27 Mar	1.0.5	658	67	0.311	0.432	0.252	0.089	0.007	0.006	0.013	0.021	0.299	0.007
2	4-5 Apr	1.0.8	692	70	0.274	0.435	0.272	0.088	0.006	0.002	0.012	0.014	0.293	0.005
3	17-18 Apr 28-29 Apr 11-12 May	1.0.11 1.0.13	1033	76	0.293	0.451	0.269	0.075	0.008	0.004	0.013	0.016	0.302	0.006
4	15-16 May	1.0.14	995	75	0.337	0.458	0.337	0.122	0.011	0.003	0.018	0.023	0.345	0.008
5	25-28 May	1.0.16	999	72	0.337	0.454	0.337	0.132	0.011	0.004	0.019	0.025	0.345	0.008
6	20 Jul	1.0.20	1018	74	0.335	0.453	0.337	0.130	0.011	0.004	0.018	0.024	0.344	0.008

In summary, we cannot yet conclude on any substantial development in FAIR uptake since the start of the project. To achieve a robust and reliable detection of score variability in the evaluation results, it is necessary to exclude the possibility that code-changes are present, something that can be achieved by fixing the version used for executing the evaluations.

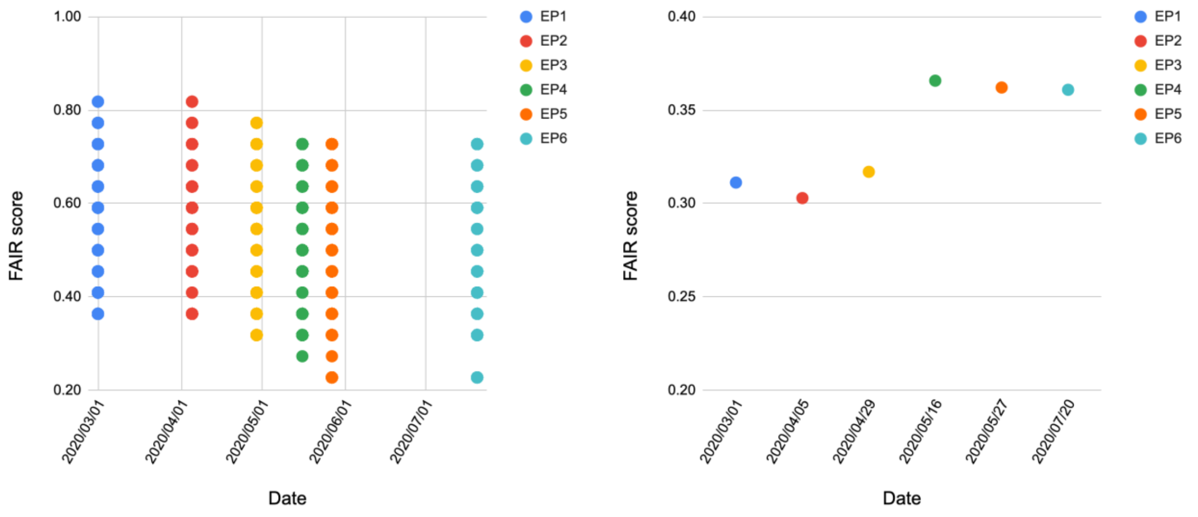


Figure 7. All 4959 data points over 6 epochs (left) and the epoch averages of those data points (right). Note the quantisation of FAIR scores in the left panel due to the limited number of indicators (metrics) and correlation of the scores.

6.3 Regular monitoring of uptake

As repositories gradually mature and FAIRify their services and the associated datasets we expect that FAIR scores will improve. This evolution should be detectable by regularly evaluating the sample of the selected digital objects, provided we perform the evaluation using the same software tool and version, as discussed.

The hope is that with regular evaluations we can uncover changes in the repositories during the project lifetime and quantify in what direction this uptake is moving (for instance, more in the direction of improving discoverability or towards interoperability?). Our limited set of indicators will at the very least allow us to set some estimates on the rate of uptake, development trends and possibly uncover new requirements that can be added to the existing tests in order to further advance the FAIR maturity evaluation of datasets.

6.4 FAIRification of repositories

EOSC-Nordic wishes to encourage and support the FAIRification efforts that repositories may decide to do. Some repositories and communities have already made a strategic choice to strengthen open data by following the FAIR principles. We see several examples where this strategy is efficiently implemented by choosing to use established providers or setting up their own service using software platforms that have already proven to give respectable results when considering steps to obtain good FAIR results.

However, the challenge is not to help those communities that already have chosen to go down the path of achieving FAIR data. The real challenge lies in reaching out to those communities that do not see the benefits for their use-case or those that need a little assistance or support in order to get started. The project will therefore continuously work on outreach related activities and to provide the communities in our sample (or new ones) updated on our activities and eventually also to showcase some of the impacts for repositories that have invested in making their data FAIR.

Although limited support can be provided by the project we plan on hosting a series of hackathons or metadata-4-machines events (in collaboration with the international GO-FAIR office in Leiden). Such events are now being planned for late 2020 and early 2021 and are likely to be executed as online events. In order to maintain a certain degree of interactivity, we may end up with a series of events that can be targeted for narrow and concrete tasks in order to keep communities engaged with the project throughout the project lifetime.

In our next report due towards the end of the project, we will provide detailed analysis and results from the ongoing monitoring of our sample (98) of Nordic+Baltic data repositories.

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

6.5 Exemplary FAIRification

In the spring of 2020, the Finnish Social Science Data Archive (FSD) carried out an internal FAIRification project which was motivated by the FAIR maturity evaluation efforts within the EOSC-Nordic project. The project led to certain changes in the provided metadata to ensure that the services offered to the researchers would become more FAIR.[24]

The evaluation process in the EOSC-Nordic project has been described in detail elsewhere in this report. For FSD's purposes both the public FAIR Evaluator results and EOSC-Nordic FAIR scores were examined.

FSD provides high-quality metadata about its datasets, persistent identifiers and makes machine readable XML available. A previous self-assessment[25] suggested that the services were relatively FAIR. Operating on the assumption that the FAIR scores provide a high-level indication on how well the repository can present its holdings in a machine-readable form, FSD's goal was to score well.

Initial maturity level tests gave results that only barely exceeded the minimum level (4/22). It was evident that several tests failed because the evaluator either did not recognise the way that had been chosen to express what was sought in the test or that the information provided simply was not machine-readable. For example, the license information provided was not identified by the evaluator. When the information was presented in a standardised way using Creative Commons Rights Expression Language, the test passed successfully.

Similarly, each test result was studied in detail. It was apparent that even though the relevant information was present in the metadata, the evaluator had difficulty interpreting it: For instance stating the persistent identifier for the data or describing where the data file can be accessed. In addition, keywords or vocabularies used were not expressed in a machine-readable format. The solution was to embed Linked Data providing essential descriptive information into the metadata descriptions. For this purpose, JSON-LD and schema.org datatypes were used. This resulted in a considerable improvement in the evaluation (17/22).

While competence cannot be condensed to a simple maturity level test, it is still beneficial to openly consider how that competence is made accessible to a machine. Most of the necessary changes were quite simple to implement. They could be carried out by repurposing the existing metadata and by using some of the many existing metadata schemas. The goal is to increase the discoverability and reusability of the data and this can be achieved, in part, by making changes that maximise the FAIR score.

[24] Alaterä, Tuomas J. (2020). Steps Towards Being More FAIR. (https://www.fsd.tuni.fi/en/news/articles/steps-towards-being-more-fair/)

[25] Kleemola, Mari (2017). Tietoarkisto on FAIR. (https://tietoarkistoblogi.blogspot.com/2017/02/tietoarkisto-on-fair.html)

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

Appendix

Appendix A

Table A1 (selected digital repositories)

Table of selected digital repositories with relevance to Nordic/Baltic countries, N=98 as of May 2020. Out of the 98 repositories, 24 are marked in red indicating that they could not be evaluated (due to missing GUIDs).

RepoID	#DS	Short name	Data URL
1	0	DNBC	http://biobanks.dk/?locale=en
2	40	CLARIN-DK	https://repository.clarin.dk/
3	20	DDA	http://dda.dk/simple-search
4	20	Det Kgl. bibliotek	https://soeg.kb.dk/discovery/search?vid=45KBDK_KGL:KGL&lang=da
6	10	Kielipankki	https://www.kielipankki.fi/language-bank/
7	10	Data Service Portal Aila	https://services.fsd.uta.fi/catalogue/search?lang=en
8	10	Fairdata IDA	https://etsin.fairdata.fi/
9	16	NMBU dataverseNO	https://dataverse.no/dataverse/nmbu
10	20	NSD	https://search.nsd.no/
11	10	HUNT Databank	https://hunt-db.medisin.ntnu.no/hunt-db/#/studypart/418
12	0	ESSDA	http://esta.ut.ee/
13	20	CLARINO Bergen Center repository	https://repo.clarino.uib.no/xmlui/
14	0	textlab	https://www.hf.uio.no/iln/english/about/organization/text-laboratory/
16	10	Språkbanken	https://www.nb.no/sprakbanken/repositorium#ticketsfrom?lang=nb&query=alle&tokens=&from=1&size=12&collection=sbr
17	9	ESS Data	http://nesstar.ess.nsd.uib.no/webview/
18	22	TROLLing	https://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/N8KO4O
19	10	EED	https://nsd.no/european_election_database/
20	20	UiT Open Research Data Dataverse	https://dataverse.no/dataverse/uit
24	13	Språkbanken	https://spraakbanken.gu.se/eng/resources

25	20	Lund University Humanities Lab corpus server	https://corpora.humlab.lu.se/
26	20	su.figshare.com	https://su.figshare.com/browse
27	20	SND	https://snd.gu.se/en
28	8	ICES data portals	https://www.ices.dk/marine-data/data-portals/Pages/default.aspx
29	10	JASPAR	http://jaspar.genereg.net/
30	10	STRING	https://string-db.org/
32	22	GBIF	https://www.gbif.org/
39	10	HPA	https://www.proteinatlas.org/
40	0	TRY	https://www.try-db.org/TryWeb/Home.php
41	10	Fishbase	https://www.fishbase.org/search.php
42	0	EMEP	http://ebas.nilu.no/default.aspx
43	0	ECCAD - the GEIA database	https://eccad3.sedoo.fr/
45	10	ISIG	http://isgi.unistra.fr/
46	0	WDC - Geomagnitism	https://www.space.dtu.dk/english/research/scientific_data_and_models
47	10	GERDA	http://www.geus.dk/produkter-ydelser-og-faciliteter/data-og-kort/national-geofysisk-database-gerda/
49	8	ACTRIS	https://actris.nilu.no/
51	0	eKlima	https://seklima.met.no/observations/
52	20	NPDC	https://data.npolar.no/home/
53	0	Norwegian Meteorological Institute	cryo.met.no
54	12	Bolin Centre Database	https://bolin.su.se/data/
55	10	SMHI open data	https://www.smhi.se/data/utforskaren-oppna-data/
57	20	NIRD Archive	https://archive.sigma2.no/pages/public/search.jsf
58	0	DeIC data	https://www.deic.dk/en/data_deic_dk
60	10	GTN-P Database	https://gtnp.arcticportal.org/
62	20	UNITE	https://unite.ut.ee/index.php
63	10	Estonian Biocentre Public Data	https://mpds.io/#start
64	18	DataDOI	http://datadoi.ut.ee/

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

65	20	CELR META-SHARE	https://metashare.ut.ee/
66	10	AHEAD	https://www.emidius.eu/AHEAD/
67	0	tekstlab	https://www.hf.uio.no/iln/english/about/organization/text-laboratory/services/index.html#multi
68	10	USN RDA	https://usn.figshare.com/
71	19	LOAR	https://loar.kb.dk/
72	20	AIDA Data Hub	https://datasets.aida.medtech4health.se/
73	10	QoG Institute's data	https://qog.pol.gu.se/data
75	0	SGO	http://www.sgo.fi/Data/archive.php
76	20	JYX	https://jyx.jyu.fi/handle/123456789/39868
78	12	B2SHARE	https://b2share.eudat.eu/
79	10	DH	https://www.lnb.lv/en/researchers/digital-humanities
80	10	NLL	https://data.gov.lv/eng
82	0	LVM GEO	https://www.lvmgeo.lv/en/data
83	0	LGIA spatial data	https://www.lgia.gov.lv/lv/atvertie-dati
84	10	RTU RIS	https://ortus.rtu.lv/science/en/datamodule/search
85	10	FinBIF	Laji.fi
86	0	OPEN	https://www.sdu.dk/da/om_sdu/institutter_centre/klinisk_institut/forskning/forskningsenheder/open/opens_faciliteter/open+storage
87	10	SARV	http://geokogud.info/
92	15	SSRI	http://fel.hi.is/datice
93	0	ICSRA	http://www.rannsoknir.is/en/request-form/
94	10	IINH	https://en.ni.is/research/scientific-collections
95	0	EERC	https://jardskjalftamidstod.hi.is/services/databank/
96	0	IES	http://earthice.hi.is/
97	0	MFRI	https://sjora.hafro.is/
100	20	QsarDB	http://qsardb.org/
104	20	Bird	https://bird.unit.no/
105	0	UCPH ERDA	http://www.erda.dk/
106	10	Migration Institute of Finland	http://www.migrationinstitute.fi/en/

108	4	Musiikkiarkisto	https://www.musiikkiarkisto.fi/
109	14	SLS	https://www.sls.fi
111	0	Kansanperinteen arkisto (Folklive archives)	https://sites.tuni.fi/kansanperinne/
113	10	SweFreq	https://swefreq.nbis.se
114	10	Metabolic Atlas	https://metabolicatlas.org
115	10	SEAD	https://www.sead.se
116	10	NOW	http://pantodon.science.helsinki.fi/now/
117	10	SNM Digital Assets	http://www.daim.snm.ku.dk/index.php?lang=2
119	0	iPSYCH	https://ipsych.dk/en/research/downloads/
120	10	GEUS	https://eng.geus.dk/products-services-facilities/data-and-maps/
123	10	LARM	https://www.larm.fm/
126	0	Open Data Service	http://data.nationallibrary.fi/
127	10	Garamantas	http://garamantas.lv/repository?lang=en
128	0	LGDB	http://www.latvianbiobank.lv/en/conditions-for-the-issue-of-biological-materials-and-data
129	10	MMB	https://mmp.sfb.uit.no/databases/mardb/#/
130	20	PlutoF	https://plutof.ut.ee/
131	10	MIDAS	https://midas.lt/public-app.html#/midas?lang=en
132	20	NMDC	https://dataverse.no/dataverse/nmdc
133	10	IINH BIOTA	https://en.ni.is/search/biota
134	20	ICOS	https://www.icos-cp.eu/node/1
135	14	CESSDA DC	http://datacatalogue.cessda.eu
136	10	DTU data	http://data.dtu.dk
137	20	CLARIN IS	http://repository.clarin.is
138	11	LIDA	http://www.lidata.eu/en/index_search.php

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

Table A2 (averaged results for FAIR maturity evaluations)

Table of 74 evaluated data repositories, showing averaged results over typically N=10 datasets for repositories with only URI identifiers and N=20 datasets for ones with URI and PID (DOI/Handle). Each of the measured scores have associated errors (standard deviation). Certifications that already exist are provided in the last 4 columns for CTS, DSA, WDS and CLARIN.

rep ID	Name	#DS	Platform	F score	A score	I score	R score	FAIR	Sigma	Sigma (F)	Sigma (A)	Sigma (I)	Sigma (R)	CTS	DSA	WDS	CLARIN
2	CLARIN-DK	40	Dspace	43.75%	40.00%	28.57%	0.00%	34.09%	0.016	0.063	0.000	0.000	0.000	X			X
3	DDA	20		43.75%	40.00%	71.43%	0.00%	47.73%	0.016	0.064	0.000	0.000	0.000				
4	Det Kgl. bibliotek	20		23.75%	44.00%	17.86%	12.50%	25.45%	0.127	0.202	0.123	0.317	0.319				
6	Kielipankki	10	META-SHARE	12.50%	40.00%	0.00%	0.00%	13.64%	0.000	0.000	0.000	0.000	0.000	X			X
7	Data Service Portal Aila	10		57.50%	80.00%	71.43%	100.00%	70.91%	0.016	0.065	0.000	0.000	0.000				
8	Fairdata IDA	10		12.50%	40.00%	0.00%	0.00%	13.64%	0.000	0.000	0.000	0.000	0.000				
9	NMBU dataverseNO	16	Dataverse	61.72%	80.00%	71.43%	0.00%	63.35%	0.008	0.031	0.000	0.000	0.000				
10	NSD	20	NESSTAR	28.75%	42.00%	28.57%	5.00%	29.55%	0.119	0.195	0.089	0.359	0.224	X			
11	HUNT Databank	10		12.50%	40.00%	0.00%	0.00%	13.64%	0.000	0.000	0.000	0.000	0.000				
13	CLARINO Bergen Center repository	20	Dspace	26.25%	40.00%	8.57%	0.00%	21.36%	0.047	0.134	0.000	0.134	0.000	X			X
16	Språkbanken	10		37.50%	40.00%	42.86%	0.00%	36.36%	0.000	0.000	0.000	0.000	0.000				
17	ESS Data	9		37.50%	40.00%	28.57%	0.00%	31.82%	0.000	0.000	0.000	0.000	0.000				
18	TROLLing	22	Dataverse	60.00%	80.00%	71.43%	0.00%	62.73%	0.013	0.051	0.000	0.000	0.000	X			
19	EED	10	Nesstar	12.50%	40.00%	0.00%	0.00%	13.64%	0.000	0.000	0.000	0.000	0.000				
20	UiT Open Research Data Dataverse	20	Dataverse	58.13%	80.00%	71.43%	0.00%	62.05%	0.015	0.061	0.000	0.000	0.000				
24	Språkbanken	13		34.62%	40.00%	21.98%	0.00%	28.67%	0.034	0.055	0.000	0.125	0.000		X		X
25	Lund University Humanities Lab corpus server	20		18.75%	40.00%	0.00%	0.00%	15.91%	0.016	0.064	0.000	0.000	0.000				
26	su.figshare.com	20	Figshare	57.50%	76.00%	70.00%	90.00%	68.64%	0.086	0.063	0.123	0.064	0.308				

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

27	SND	20		46.25%	40.00%	71.43%	0.00%	48.64%	0.015	0.059	0.000	0.000	0.000	X			
28	ICES data portals	8		37.50%	40.00%	28.57%	0.00%	31.82%	0.000	0.000	0.000	0.000	0.000				
29	JASPAR	10		12.50%	40.00%	0.00%	0.00%	13.64%	0.000	0.000	0.000	0.000	0.000				
30	STRING	10		12.50%	40.00%	0.00%	0.00%	13.64%	0.000	0.000	0.000	0.000	0.000				
32	GBIF	22	IPT	48.30%	40.00%	71.43%	100.00%	58.47%	0.011	0.044	0.000	0.000	0.000			X	
39	HPA	10		27.50%	40.00%	21.43%	30.00%	28.64%	0.160	0.242	0.000	0.345	0.483				
41	Fishbase	10		37.50%	40.00%	28.57%	0.00%	31.82%	0.000	0.000	0.000	0.000	0.000				
45	ISIG	10		12.50%	40.00%	0.00%	0.00%	13.64%	0.000	0.000	0.000	0.000	0.000				
47	GERDA	10		12.50%	40.00%	0.00%	0.00%	13.64%	0.000	0.000	0.000	0.000	0.000				
49	ACTRIS	8		34.38%	40.00%	39.29%	25.00%	36.36%	0.153	0.186	0.000	0.356	0.463				
52	NPDC	20		31.88%	42.00%	35.71%	50.00%	37.05%	0.167	0.201	0.089	0.366	0.513				
54	Bolin Centre Database	12		43.75%	40.00%	71.43%	0.00%	47.73%	0.016	0.065	0.000	0.000	0.000				
55	SMHI open data	10		37.50%	40.00%	71.43%	0.00%	45.45%	0.000	0.000	0.000	0.000	0.000				
57	NIRD Archive	20		33.75%	40.00%	35.00%	0.00%	32.50%	0.099	0.186	0.000	0.348	0.000				
60	GTN-P Database	10		12.50%	40.00%	0.00%	0.00%	13.64%	0.000	0.000	0.000	0.000	0.000				
62	UNITE	20		33.13%	40.00%	35.71%	15.00%	33.86%	0.140	0.216	0.000	0.366	0.366				
63	Estonian Biocentre Public Data	10		35.00%	40.00%	25.71%	0.00%	30.00%	0.030	0.079	0.000	0.090	0.000				
64	DataDOI	18		43.06%	40.00%	47.62%	0.00%	39.90%	0.057	0.064	0.000	0.219	0.000				
65	CELR META-SHARE	20		43.75%	40.00%	50.00%	0.00%	40.91%	0.057	0.064	0.000	0.220	0.000	X			X
66	AHEAD	10		12.50%	40.00%	0.00%	0.00%	13.64%	0.000	0.000	0.000	0.000	0.000				
68	USN RDA	10	Figshare	61.25%	80.00%	71.43%	100.00%	72.27%	0.010	0.040	0.000	0.000	0.000				
71	LOAR	19		42.11%	40.00%	47.37%	47.37%	43.78%	0.144	0.095	0.000	0.243	0.513				
72	AIDA Data Hub	20		47.50%	40.00%	64.29%	90.00%	55.00%	0.096	0.077	0.000	0.220	0.308				
73	QoG Institute's data	10		41.25%	40.00%	41.43%	0.00%	37.27%	0.054	0.060	0.000	0.207	0.000				
76	JYX	20		31.25%	40.00%	14.29%	0.00%	25.00%	0.040	0.064	0.000	0.147	0.000				
78	B2SHARE	12	Invenio	29.17%	40.00%	29.76%	41.67%	32.95%	0.165	0.187	0.000	0.368	0.515				

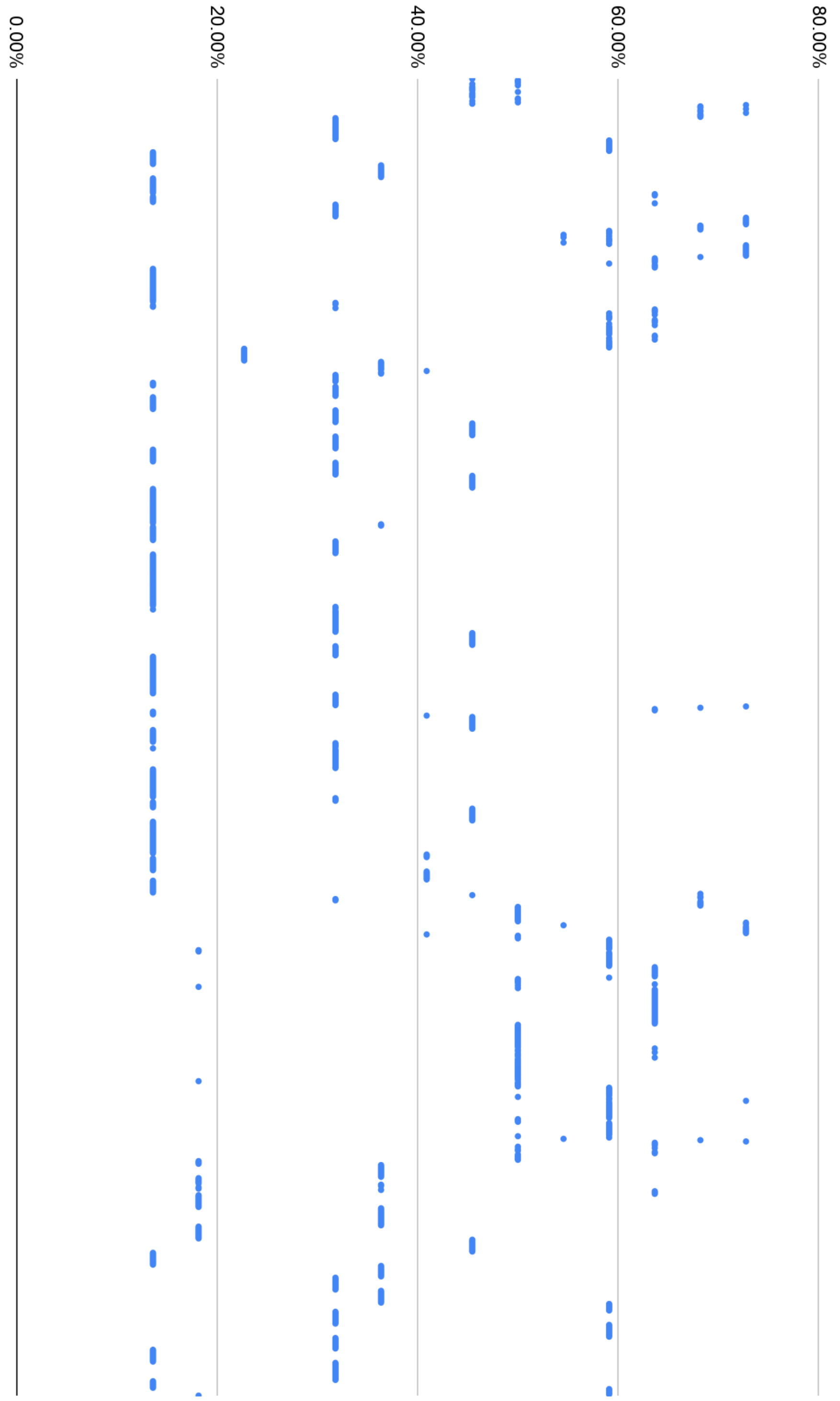
EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

79	DH	10		25.00%	40.00%	14.29%	0.00%	22.73%	0.000	0.000	0.000	0.000	0.000				
80	NLL	10		50.00%	80.00%	71.43%	0.00%	59.09%	0.000	0.000	0.000	0.000	0.000				
84	RTU RIS	10		37.50%	40.00%	71.43%	0.00%	45.45%	0.000	0.000	0.000	0.000	0.000				
85	FinBIF	10		37.50%	40.00%	71.43%	0.00%	45.45%	0.000	0.000	0.000	0.000	0.000				
87	SARV	10		37.50%	40.00%	28.57%	0.00%	31.82%	0.000	0.000	0.000	0.000	0.000				
92	SSRI	15		42.50%	40.00%	54.29%	40.00%	45.45%	0.133	0.063	0.000	0.145	0.507				
94	IINH	10		29.17%	40.00%	19.05%	0.00%	25.76%	0.047	0.125	0.000	0.143	0.000				
100	QsarDB	20		31.25%	40.00%	35.71%	45.00%	35.91%	0.164	0.192	0.000	0.366	0.510				
104	Bird	20		43.75%	40.00%	37.14%	20.00%	38.64%	0.113	0.064	0.000	0.176	0.410				
106	Migration Institute of Finland	10		37.50%	40.00%	28.57%	0.00%	31.82%	0.000	0.000	0.000	0.000	0.000				
108	Musiikkiarkisto	4	CKAN	59.38%	80.00%	71.43%	50.00%	67.05%	0.145	0.063	0.000	0.000	0.577				
109	SLS	14		37.50%	40.00%	28.57%	0.00%	31.82%	0.000	0.000	0.000	0.000	0.000				
113	SweFreq	10		12.50%	40.00%	0.00%	0.00%	13.64%	0.000	0.000	0.000	0.000	0.000				
114	Metabolic Atlas	10		37.50%	40.00%	71.43%	0.00%	45.45%	0.000	0.000	0.000	0.000	0.000				
115	SEAD	10		12.50%	40.00%	0.00%	0.00%	13.64%	0.000	0.000	0.000	0.000	0.000				
116	NOW	10		12.50%	40.00%	0.00%	0.00%	13.64%	0.000	0.000	0.000	0.000	0.000				
117	SNM Digital Assets	10		12.50%	40.00%	0.00%	0.00%	13.64%	0.000	0.000	0.000	0.000	0.000				
120	GEUS	10		20.00%	40.00%	8.57%	0.00%	19.09%	0.046	0.121	0.000	0.138	0.000				
123	LARM	10		12.50%	40.00%	0.00%	0.00%	13.64%	0.000	0.000	0.000	0.000	0.000				
127	Garamantas	10		37.50%	40.00%	28.57%	0.00%	31.82%	0.000	0.000	0.000	0.000	0.000				
129	MMB	10		12.50%	40.00%	0.00%	0.00%	13.64%	0.000	0.000	0.000	0.000	0.000				
130	PlutoF	20		35.00%	40.00%	35.71%	30.00%	35.91%	0.160	0.235	0.000	0.366	0.470				
131	MIDAS	10		12.50%	40.00%	0.00%	0.00%	13.64%	0.000	0.000	0.000	0.000	0.000				
132	NMDC	20	Dataverse	57.50%	80.00%	71.43%	0.00%	61.82%	0.016	0.063	0.000	0.000	0.000				
133	IINH BIOTA	10		38.75%	40.00%	42.86%	0.00%	36.82%	0.010	0.040	0.000	0.000	0.000				
134	ICOS	20		43.75%	40.00%	57.14%	0.00%	43.18%	0.016	0.064	0.000	0.000	0.000				
135	CESSDA DC	14		23.21%	40.00%	20.41%	0.00%	24.03%	0.095	0.176	0.000	0.335	0.000				

136	DTU data	10	figshare	46.25%	68.00%	62.86%	70.00%	58.64%	0.139	0.060	0.193	0.181	0.483				
137	CLARIN IS	20	CLARIN	43.75%	40.00%	28.57%	0.00%	34.09%	0.016	0.064	0.000	0.000	0.000				
138	LIDA	11	Nesstar	37.50%	40.00%	28.57%	0.00%	31.82%	0.000	0.000	0.000	0.000	0.000				

Figure A1 (Individual dataset FAIR scores)

The distribution of the FAIR scores for 1018 evaluated datasets on July 20, 2020



EOSC-Nordic project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857652

Appendix B

FDS-tool (version history fairdatasystems/fair_tests)

v1.0.20

Hvst-1.3.1

Handle new format of DataCite Link headers - iterate over all key/value pairs searching for 'alternate', and allow for absence of 'rel' tag.

v1.0.19

gen2_searchable Tst-0.2.7

if there are multiple possibilities for the title, then test all of them

v1.0.18public - Public version - do not use

v1.0.18 - (DO NOT USE - image deleted)

Hvst-1.3.0

uses a cache for the harvested metadata object

fixed the dublin core elements identifier predicates for metadata self-identification

use the 'convert to url' routine in all cases

add an "INFO: END OF HARVESTING" tag to the metadata comments (can be used in front-end to find the beginning of the test-specific analysis)

gen2_metadata_identifier_persistence Tst-0.2.2

- add the "INFO: END OF HARVESTING" tag to the metadata comments (because this test doesn't use the harvester)

gen2_metadata_uses_fair_vocabularies_strong Tst-0.3.1

- enhanced output reporting

v1.0.17

gen2_searchable Tst-0.2.6

trim overly long Bing requests (max 1500 characters)

catch other Bing failures

v1.0.16

Hvst-1.2.1

catches more kinds of HTTP errors and reports them, rather than crashing

accept RDF::Literal::Long datatypes when reading from Cache

Grounded Metadata Tst-0.2.2

failure message

Searchable in major search engine

can't reproduce reported error - following-up

Structured Metadata Tst-0.2.2

failure message

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

Data Knowledge Representation Language (strong) Tst-0.3.1

failure message

Data Knowledge Representation Language (weak) Tst-0.3.1

failure message

Metadata Includes License (strong) Tst-0.2.2

failure message

gen2_metadata_identifier_in_metadata Tst-0.4.4

report more precisely why an identifier cannot be found

gen2_metadata_contains_outward_links Tst-0.2.3

= 1 is a pass (rather than >1) (note that this is a very weak test!)

v1.0.15

10-cgi.conf updated to pass BING_API key into the Webserver environment

v1.0.14

Hvst-1.2.0

fixed call to distiller

increase the verbosity of output to improve interpretation of results

fix problem with unknown identifier types

added content-type used by DataCite to indicate JSON-LD

don't block on calls to extruct that return large amounts of metadata

gen2_metadata_uses_fair_vocabularies_strong Tst-0.3.0

- rewrote algorithm to test one predicate in each domain, non-stochastically

gen2_metadata_uses_fair_vocabularies_weak Tst-0.3.0

- rewrote algorithm to test one predicate in each domain, non-stochastically

gen2_searchable Tst-0.2.5

- incorrect license key would kill this, and all other tests due to invalid output message

gen2_searchable Tst-0.2.5

- incorrect license key would kill this, and all other tests due to invalid output message

gen2_data_kr_language_strong Tst-0.3.0

- changed algorithm, no longer tests directly but rather checks only the reported content-type header for a Linked Data content type. This test will miss cases where the structured data is embedded as JSON inside of HTML, but it is balanced against the danger of retrieving gigabytes worth of data.

gen2_data_kr_language_weak Tst-0.3.0

- changed algorithm, no longer tests directly but rather checks only the reported content-type header for a structured data (including XML and JSON and xhtml+xml) content type

v1.0.13

Hvst-1.1.4

added the http[s] versions of the dc-element identifier predicates (super property of dc-terms identifier)

EOSC-Nordic project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857652

gen2_metadata_identifier_in_metadata Tst-0.4.2

- improve output by listing all predicates tested

v1.0.12

Hvst-1.1.3

rewrite of caching system

catch e.g. 404 errors and never call again

compare RDF against cache to prevent re-parsing RDF (in some cases)

never call HEAD; it seems to be a waste of time most of the time, and it slows everything down

fix problem with HTTP Link headers when the relation-type is quoted

fix bug where detected GUID type was being overwritten by the URI type at the last step in resolution.

v1.0.11

Hvst-1.1.1

bug fix in RDF caching system

observation that message bodies are not identical, even if called the same way (likely timestamps or something like that in the body) Not a bug, just a limitation on optimization of the RDF caching system, which is content-based.

gen2_metadata_identifier_in_metadata Tst-0.4.1

- bugfix for when the schema.org "identifier" is not a Property-Value, but rather is a string or a URI

v1.0.10 (BROKEN - DO NOT USE)

Hvst-1.1.0

recognition that the semantic versioning should have triggered the second-digit update in Hvst-1.0.8. Apologies

the results that come from this harvester cannot be guaranteed to be the same as the results that came from earlier versions, due to changes in the way identifiers are searched, and the number of different metadata paths that are now accepted (see docker image 1.0.9 notes)

extensive revisions of the caching system

change in the order of cache lookups, versus calling HEAD on a URI

RDF::Graph objects are now Marshal'ed to disk directly in the cache, and then read back as an array of arrays and merged with the current metadata graph. This operation is about 6-10X faster than re-parsing the serialized RDF.

gen2_metadata_uses_fair_vocabularies_weak Tst-0.2.5

- accidentally introduced a bug in the final build... grrrr...

v1.0.9 (BROKEN - DO NOT USE)

Hvst-1.0.8

additional security fixes to cleanse calls

extensive changes to the "self identifier" queries to make search for metadata identifiers more rigorous; allow multiple ids to be found

extensive changes to the harvesting workflow to follow HTTP header 'Link rel=alternate' and to follow all HTML meta links rel='alternate' so long as their defined content-type is one of the structured data types recognized by the Harvester

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

gen2_metadata_identifier_in_metadata Tst-0.4.0

- changed the logic entirely
- uses the shared self-identifier in the Harvester, rather than its own
- requires an exact match on the input GUID for success

gen2_metadata_uses_fair_vocabularies_weak Tst-0.2.4

- hard-code acceptance of dcat, foaf, dublin core and vcard

gen2_metadata_uses_fair_vocabularies_strong Tst-0.2.3

- hard-code acceptance of dcat, foaf, dublin core and vcard

v1.0.8

Hvst-1.0.6

Security fix for harvester's calls to extruct

changed the way the harvester parses extruct output to hopefully kill all future crashes

gen2_metadata_contains_outward_links Tst-0.2.2

- tightened the outward links test so that it now filters-out XHTML structural triples like buttons (they aren't really FAIR metadata).
- fixed the problem where zero/zero outward links found would report as success

v1.0.7

Hvst-1.0.5

catch errors caused when invalid JSON-LD is returned from the RDFa parsers. report a warning in the output.

gen2_metadata_identifier_in_metadata v. Tst-0.3.0

- require that the metadata identifier be the value of a dc:identifier, or a schema:identifier property.

v1.0.6

fixed the version reporting for gen2_metadata_identifier_in_metadata

v1.0.5

Hvst-1.0.4

Changed SPARQL query for schema:mainEntity (used by the CommonQueries::GetDataIdentifier method. Incorrectly attempted to get the value of a bnode. Now looks for schema:identifier inside of the mainEntity object.

activated cron

EOSC-Nordic project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857652