

How FAIR can you get?

Image Retrieval as a Use Case to calculate FAIR Metrics

Tobias Weber, Dieter Kranzlmüller
Leibniz Supercomputing Centre (LRZ)
Bavarian Academy of Sciences and Humanities
Garching, Germany
{weber, kranzlmueeller}@lrz.de

Abstract—Many providers of research data services officially embrace the FAIR guiding principles for scientific data management and stewardship. To assess the compliance of their services to these principles and to indicate possible improvements, use-case-centric metrics are needed as an addendum to existing approaches. The retrieval of spatially and temporally annotated images can exemplify such a use case. A prototypical benchmark based on that use case indicates that currently no research data repository achieves the full score according to the proposed metric. Suggestions on how to increase the score include automatic annotation based on the metadata inside the image file and support for content negotiation to retrieve the research data. This can lead to an improvement of data integration workflows, resulting in a better and more FAIR approach to manage research data.

Keywords—Research Data Management, FAIR Guiding Principles

I. INTRODUCTION

Let research data be all forms of digitised content used as input for or output of scientific research activities. Metadata are then understood as data about these research data to make them Findable, Accessible, Interoperable, Reuseable (FAIR) amongst other features. Research data products can thus be defined as the research data together with their metadata. Scientific advances will be stimulated and the return of investment for research funding will increase, if research data are reused more often [1]. Enabling others to reuse their data is a major challenge for researchers, since the necessary tasks are diverse, extensive, non-trivial, and often only recently added to their responsibilities [2]. Many institutions, infrastructure projects and service providers support researchers with these data management tasks.

The FAIR guiding principles constitute quality criteria for research data products and thus provide the necessary foundation to assess the services designed to help the researchers [3]. Since they describe the requirements especially for scalable research data services, they mainly focus on machine-to-machine interaction.

Recent proposals to derive specific metrics from these principles focus on single data sets [4][5][6] or on data from single Research Data Repositories (RDR) [7]. While these metrics provide useful insights, they share the shortcoming

that they do not deal with complex scenarios of data integration across RDRs.

Another drawback of current data-centric or repository-centric metrics is the missing disambiguation of some FAIR principles: Two central principles require, that data are richly described with "accurate and relevant attributes" [3, principles F2 and R1]. But the required richness does not have a concise meaning unless the usage context is defined. Since neither data items nor RDRs can be used to derive this context adherence to these subprinciples cannot be measured so far.

This paper proposes a new method to define use-case-centric metrics which do not share the two aforementioned drawbacks: Use cases of data integration across RDRs give the necessary context to decide whether a data set is described richly enough, thus allowing to derive a corresponding metric. This approach complements existing metric frameworks.

A benchmark to exemplify this method has been implemented. The underlying use case is the retrieval of spatially and temporally annotated images across RDRs. It is both relevant to many researchers and necessitates data integration over distributed sources. For five RDRs and 1.408.929 research data products the scores of the metric have been calculated. The first calculation of the metric indicates two major issues: accessibility barriers and the absence of chronoreference. The calculation can be executed automatically; continuous executions will thus give evidence whether the first findings are robust and how the research data landscape changes over time. Since our use case will be relevant to some, but not all disciplines and data-driven research methods, other case studies can be implemented and reuse the methodological approach presented in this paper.

In Section II the relation of this paper to other work is discussed. Section III presents the rationale for the use case selection and states how a use-case-centric metric can generally be calculated. The steps taken in implementing the prototypical benchmark are described in Section IV. In Section V an overview of the FAIR indicators collected is given. Section VI discusses and evaluates the presented approach.

II. RELATED WORK

Two data-centric metrics that can be used to measure FAIRness exist to our knowledge. The metrics of [4], [5] can be applied to a whole RDR by aggregating the measured values.

[4] proposes a five star rating for research data alongside a tool for automatic assessment. The authors of [5] introduce both a framework for measurable FAIRness of meta(data) and tools for semi-automatic assessment. The framework allows to provide additional, possibly community-specific metrics. Currently 14 examples for such metrics are described [6]. Our approach can be integrated into the framework described in [6] and is fully automated.

The authors of [7] focus on measuring the FAIRness of RDRs. 37 RDRs have been manually assessed with a focus on Dutch and international providers. The data to derive the repository-centric metrics is openly available [8]. Our results suggests a more critical perspective on accessibility compared to [7] - at least in the machine-actionable sense.

Relevant shortcomings of data-centric and repository-centric metrics addressed by our approach are their inability to represent retrieval scenarios across RDRs and to disambiguate some of the FAIR principles (see Section I). As far as we know, no other work tried to fill these gaps with use-case-centric metrics.

An architecture for FAIR-compliant research data integration across repositories is described in [9]. This architecture has three main components: The FAIR accessor is the first component (a Linked Data Platform Container), which consists of several MetaRecords (or FAIR profiles, the second component). MetaRecords have themselves links to FAIR projectors, the third component. This allows for machine-actionable access to a data set or even a single data point in it. The publication furthermore lists several community-specific efforts to realise research data integration that could be used to derive other use-case-centric metrics. This approach can be mapped to our generic research data integration workflow (see Section III) and can therefore be extended to calculate another use-case-centric metric. The architecture is based on semantic technologies and focuses primarily on data relevant for the life sciences. The case study in this paper will use another type of technology and focus on a use case that is not specific to the life sciences.

The tools, standards and protocols presented in [10] are of high relevance for our use case since they present state of the art techniques with respect to interoperability of RDRs. Image retrieval is a subset of information retrieval, one of the two major use cases of this primer. Both of the major techniques we use in our implementation (OAI-PMH and DataCite) are listed in [10].

[11] provides an overview over practices in the retrieval of observational data across different disciplines. The concepts

found there provided valuable insights for the development of the generic research data integration workflow presented in Section III. The review focuses more on manual information retrieval and discovery, but it could indicate other possible case studies, which then need to be automated.

The International Image Interoperability Framework¹ provides a convincing alternative to the set of protocols and services chosen by our implementation. Whereas this project concentrates on general tasks of image integration (also outside of the research domain), the rationale behind the presented case study was extendability to other research data types and formats.

In [12] the usage of OAI-PMH for a search index and data discovery service is described. Evaluation of the metadata formats presented in our paper or statistics comparable to ours are not included, though some hints are given how different metadata formats can be used to realise the detection of data retrieval endpoints.

[13] discusses the difficulties in retrieving the data described by a metadata catalogue provided via OAI-PMH. Several metadata formats and their features are discussed, but not DataCite, the metadata format chosen for the presented implementation (DataCite didn't exist at the time [13] has been written).

The discussion of the first calculated metrics in Section V adds to the overall description of the research data landscape given in [14]. Whereas the statistical evaluation presented there discusses the broad range of RDRs listed in re3data.org, we are more specific to data providers supporting our use case.

Research data retrieval can also be categorised as a big data integration task. Further challenges and opportunities in managing big data have been described in many papers (e.g. [15], [16]), The five V's (volume, velocity, variety, value and veracity) of big data are e.g. presented in [17]. Our case study has the management of variety in focus. The proposed measurement of quality criteria can in principle be adapted to be applicable outside of the context of reuse of research data.

III. METHODS

A. Selection of a Use Case

In order to calculate use-case-centric metrics, a use case was needed, which fulfills the following requirements:

- Relevance: the use case has to be relevant for different fields of research and science.
- Complexity: the use case has to include research data integration across different RDRs.
- Effectiveness: the research data products that are integrated and collected due to the use case allow the reproducible calculation of a metric for FAIRness for research data products and RDRs.

¹<http://iiif.io>

These requirements make sure that the use case is "interesting" enough while still being implementable.

As a use case satisfying these criteria, we chose the retrieval of images from RDRs that are annotated with a certain place and time (of their creation). Furthermore they must be licensed in a way that allows to determine whether reusing the images is restricted in any way.

Image processing is a relevant technique for both the sciences and the humanities (cf. for example [18] and [19]). Selecting images on the basis of the date and location of their creation is generic enough to be still interesting for different disciplines and specific enough to be implementable. Images reside in various different RDRs - integration across different data sources is therefore an essential part of the use case.

As will be shown in the following subsection, FAIR metrics can be calculated alongside the implementation of this use case.

B. Calculation of Use-Case-Centric Metrics

In this subsection the calculation of use-case-centric FAIR metrics will be described in a generic fashion applicable to any use case. A Description of our implementation (the benchmark) will be given in the next subsection.

Let D be the set of all research data products of interest which are accessible via a set of repositories. "Of interest" implies the non-consideration of research data products that are in principle not interesting for our use case. The main criterion to separate data products of interest from the rest is their format: An example is the non-consideration of tabulator-separated data for the image retrieval use case of this paper. Calculating the proposed use-case-centric metric for these research data products would be pointless.

A central point for calculating use-case-centric FAIR metrics is the assessment of n quality criteria that need to be met by a research data product $d \in D$ to be fully useable for the use case at hand. Let $f_i : D \rightarrow (0, 1)$ ($1 \leq i \leq n$) be the assessment function which returns 1 when the criterion i is met and 0 otherwise. $d \in Q_i$ if and only if $f_i(d) = 1$. Q_i is hence the set of all data products meeting criterion i .

In the following paragraphs four use-case-centric FAIR scores will be presented, two for a single data product d and two for a Research Data Repository (RDR) R . Together they constitute the proposed metric.

1) *Absolute Score for Research Data Products*: For calculating this score, the weights of the use-case-specific quality criteria are fixed and evenly distributed:

$$score_{absolute}(d) = \sum_{i=1}^n (f_i(d) \cdot \frac{1}{n}) \quad (1)$$

The range $score_{absolute}$ is the interval $[0, 1]$, with $score_{absolute}(d) = 1$ if d meets all quality criteria. Any value between 0 and 1 will indicate to what extent the research data product is useable for the use case. If the benchmark is repeated and all Q_i ($1 \leq i \leq n$) are re-determined,

comparison of this score between two studies easily allows to measure trends.

2) *Relative Score for Research Data Products*: We assume that criteria that are not met by many data products are harder to satisfy and should therefore be weighted accordingly. To achieve this the weight of i is calculated as a function of the size of Q_i (which is the adoption rate at the time the benchmark is executed). We define

$$rareness(Q_i) = 1 - \frac{|Q_i|}{|D|} \quad (2)$$

and

$$weight(Q_i) = \frac{rareness(Q_i)}{\sum_{j=1}^n rareness(Q_j)} \quad (3)$$

The relative score for a research product d is then calculated as follows:

$$score_{relative}(d) = \sum_{i=1}^n f_i(d) \cdot weight(Q_i) \quad (4)$$

The range of $score_{relative}$ is still the interval $[0, 1]$, with $score_{relative}(d) = 1$ if d meets all quality criteria; but if two research data products d_1 and d_2 each meet one and only one criterion, say $d_1 \in Q_1$ and $d_2 \in Q_2$ (with $d_1 \neq d_2$ and $Q_1 \neq Q_2$), d_1 gets a higher score, if $|Q_1| < |Q_2|$. Research data products that satisfy rarely met criteria are hence highlighted positively, whereas criteria met by the major part of the data products have a smaller weight in this score. When repeating the benchmark, it will turn out that $score_{relative}$ - in comparison to $score_{absolute}$ behaves somewhat more difficult to interpret: it may be that $score_{relative}(d)$ decreases even though the number of met criteria for d is constant between the two benchmark executions. This is a consequence of the fact that weights are relative to the adoption rate, which is expected to be non-identical between benchmark re-executions. Comparing two relative scores between two studies nevertheless gives an indication how the research product fares with regard to a common standard of quality.

The values of $weight(Q_i)$ of two or more benchmark executions can furthermore be compared to gain insights how the adoption rate of a certain criteria changed over time, whereas the total sum of all $rareness(Q_i)$, ($1 \leq i \leq n$) gives a hint how well-supported the use case is over all RDRs: the closer total rareness is to n , the less well-supported the use case is.

3) *Average Absolute Score for a Research Data Repository*: This metric is the arithmetic mean of absolute scores for data products d managed in R . Let D_R be the set of all research data products of interest managed in R .

$$score_{avabsolute}(D_R) = \frac{\sum_{d \in D_R} score_{absolute}(d)}{|D_R|} \quad (5)$$

This metric can be interpreted on par with $score_{absolute}$: Any value between 0 and 1 indicates to what percentage the research data products managed in R are useable for the use case at hand with $score_{avabsolute} = 1$ if R only hosts full-quality research data products. This number must be assessed together with the total number of research data products of interest $|D_R|$, since it is possible to get a score of 1 while hosting only one research data product of interest.

4) *Average Relative Score for a Research Data Repository*: This metric complements $score_{avabsolute}$ as $score_{relative}$ complements $score_{absolute}$:

$$score_{avrelative}(D_R) = \frac{\sum_{d \in D_R} score_{relative}(d)}{|D_R|} \quad (6)$$

To actually calculate these use-case-centric FAIR metrics for the chosen use case, the quality criteria Q_i need to be determined and the associated assessment functions f_i need to be defined:

- $d \in Q_{\text{chrono}}$ if and only if d is annotated with the date when the image was taken.
- $d \in Q_{\text{geo}}$ if and only if d is annotated with a reference to the location where the image was taken.
- $d \in Q_{\text{lic}}$ if and only if d is annotated in a way that allows to determine whether d may be used without any restrictions.
- $d \in Q_{\text{ret}}$ if and only if d is automatically downloadable given only d 's metadata.

IV. IMPLEMENTATION

The prototypical benchmark must implement the assessment functions f_{chrono} , f_{geo} , f_{lic} and f_{ret} alongside the use-case-specific data collection system for providing their input. The implemented benchmark must furthermore fulfill the following requirements:

- Non-Creativeness: only tools, standards and techniques that already exist may be used - implementation must be restricted to combining these. This is necessary since the benchmark should measure the status quo, not improve it.

- Automatability: the integration must not include manual effort. This means it is 'machine-actionable' (for an explanation see [3]). This requirement ensures that our implementation scales with the number of research data products and repositories and that it is insightful with regard to automated data integration workflows.
- Repeatability: the benchmark execution has to be repeatable to ensure the ability to measure trends. This requirement furthermore guarantees reproducibility of the measured results.

The implementation is organised along the five steps of a generic research data integration workflow as depicted in Figure 1.² Any use case of research data integration should roughly be mappable to this generic workflow. It is compatible to a generic research data infrastructure as described in [20].

Once the use case of data integration has been implemented, the implementation of the benchmark is taken care of, centering around the steps to calculate the FAIR metrics.

All steps are implemented using the python programming language.³ Shell scripts wrap the python scripts to make sure the scripts are called in the correct order and with consistent parametrisation (the shell scripts also orchestrate the parallelisation). For each step technical options will be discussed and the selected implementation of the use case and the benchmark will be sketched.

1) *Query Data Provider Registries*: The first step of the generic workflow (Figure 1, left-hand side) consists of two tasks:

- 1) Compiling a list of data providers
- 2) Creating a list of APIs supported by them

The compilation of a set of data providers presupposes non-machine interaction with the services: finding and reviewing RDRs, their API endpoints, and relevant policies, assessing certifications, and last but not least, experience in using

²All source code (DOI 10.25927/001) and data (DOI 10.25927/000) necessary to reproduce or replicate our findings are published as accompanying resources to this publication.

³<https://www.python.org>

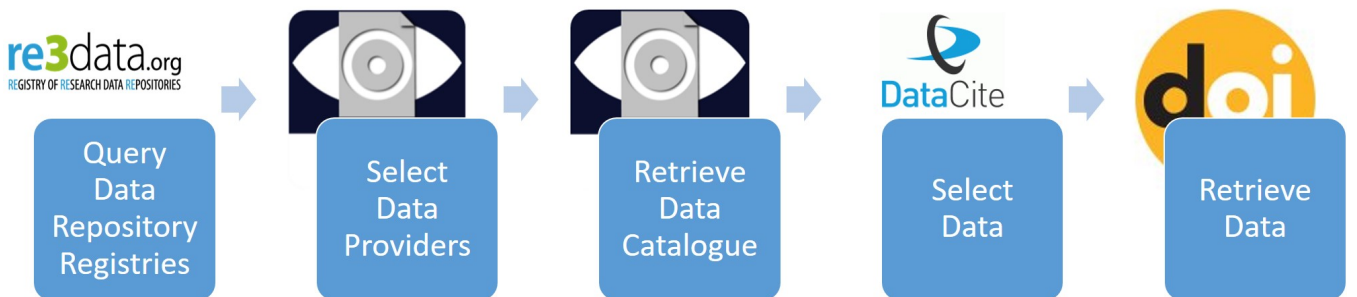


Figure 1. A Generic Workflow for Research Data Integration (images indicate specific solutions to implement the steps)

the provided services. Since the benchmark needs to fulfill the requirements of automatability and repeatability, this information must be provided in machine-readable form via an API.

According to our research, there is only one candidate fulfilling the requirements to a research data provider registry at the moment: The Registry of Research Data Repositories (re3data.org).⁴ The number of repositories registered in re3data.org grows steadily: From 400 repositories listed in July 2013 [21], it held information about 2093 repositories by June 2018. re3data.org also lists data providers that are not research institutions in the classical sense, but nevertheless provide valuable data for researchers such as the Climate Data Centre⁵ or national statistic agencies⁶. Another registry, Databib (databib.org) has been integrated into re3data.org, [22]. The Directory of Open Access Repositories (OpenDOAR)⁷ and the Registry of Open Access Repositories (ROAR)⁸ have also been evaluated. Both provide a means to automatically retrieve a list of RDRs, including API endpoints (exclusively OAI-PMH). Since they are both primarily designed to distribute information about open access publication of scholarly articles we decided to focus on re3data.org

The output of the implemented queries to the re3data.org-API is a list of RDRs including information about their supported APIs and information about their quality management. The raw output is then processed and the relevant information is stored in a Comma Separated Values file (CSV) describing the RDRs. Additional information offered by re3data.org (e.g. about licenses and formats) will not be used, since the calculation of the scores requires that we retrieve them on a data item basis and not aggregated over all data provided by a repository.

2) *Select Data Providers*: Decisive criteria in the selection of suitable RDRs are the feature set supported by their APIs and the metadata format they use to describe their data items. In the following both criteria will be discussed separately.

The **APIs** will be used for three tasks: to retrieve the information which metadata formats are supported, to get the information how many data items are hosted by the RDR and to retrieve the data catalogue (i.e. all the metadata, see step three).

The API which may satisfy our requirements are listed in Table I together with the numbers of RDRs providing the API and the adoption rate as given by re3data.org and retrieved during step one. We left out some listed APIs on purpose since their design rationale is not compatible for our requirements. 1165 RDRs (56% of them) have no API

Table I
SELECTION OF APIS SUPPORTED BY RDRs LISTED IN RE3DATA.ORG

API	Absolute	Relative
REST	304	14.52 %
OAI-PMH	162	7.74 %
SOAP	68	3.25 %
SPARQL	27	1.29 %

Note: n = 2093 (June 2018)

at all registered.

Representational State Transfer APIs (ReST), the Simple Object Access Protocol (SOAP) [23], and the SPARQL Protocol and RDF Query Language (SPARQL) [24] are less viable options for our use case: While each API endpoint of a certain type could support the generic workflow of research data integration, each does it in a different way: Whereas one ReST-API might provide access to "collections", the other one has "datasets" as the basic data type. This makes it very cumbersome to utilise the APIs across RDRs, for example to retrieve the data catalogue (see next step) in a uniform way. Notwithstanding proposals to achieve the necessary level of homogeneity [25], we could only find one example where this has been achieved for SOAP in the context of cancer research [26]. Only extensive implementation work could allow for an uniform access across RDRs. In the context of the present work this effort would violate the non-creativity and the automatability requirement.

The Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH)⁹ has an adoption rate big enough to be relevant and supports data catalogue retrieval. There is a positive trend since 2015 for OAI-PMH: both the absolute number of RDRs supporting it and the adoption rate increased (in 2015 there were 85 RDRs, which entails an adoption rate of 6.2 %, cp. [14]). OAI-PMH provides the necessary semantics to uniformly implement step two and three of both the use case and benchmark. The protocol is embedded into HTTP and supports several operations to retrieve information about a RDR and the research data products managed in it:

- ListMetadataFormats: returns a list of metadata formats.
- ListRecords: returns a list of metadata records describing the data items.

The prototypical implementation presented here is therefore based on OAI-PMH.

Suitable **metadata formats** provide the necessary information to implement the assessment functions f_{chrono} , f_{geo} , f_{lic} and f_{ret} and they need to indicate whether the described data item is an image and therefore of interest for the use case. The 49 metadata formats that are offered by the RDRs providing an OAI-PMH interface have been evaluated

⁴<https://www.re3data.org>

⁵<https://cdc.dwd.de>

⁶e.g.: <https://www.ons.gov.uk>

⁷<http://www.opendoar.org/>

⁸<http://roar.eprints.org/>

⁹<http://www.openarchives.org/OAI/openarchivesprotocol.html>

alongside the following three criteria (If a metadata format is available in multiple versions, the most expressive one has been evaluated):

- **Existence:** There are predefined fields in the metadata format which have the necessary semantics to implement f_{chrono} , f_{geo} and f_{lic} . f_{ret} is not implementable with the metadata alone, but there need to be sufficient information to determine the protocol and endpoint for data retrieval. Furthermore, there has to be a field indicating whether the described data are research data of interest, i.e. a field for the data type.
- **Unambiguity:** The information is placed at a uniquely defined place in the format and can be used as is (i.e. without case distinction, additional retrieval, e.g. of ontologies).
- **Guaranteed Coverage:** The fields are mandatory i.e. we can assume they are always part of a valid metadata record (this is an optional requirement).

DataCite is the most widespread standard fulfilling all requirements when we sort the metadata formats in the decreasing order of number of RDRs supporting it: With 10 RDRs supporting it, DataCite has been adopted by 10.87 % of the RDRs providing an OAI-PMH interface. In fact, DataCite is more relevant than these numbers suggest, since the RDRs supporting it are among the biggest RDRs in terms of the number of research data products (see Section V).

With qualified elements for date, license and georeference, DataCite provides the necessary input to implement the assessment functions. Additionally DataCite's identifier field is mandatory and has to contain a Digital Object Identifier (DOI)¹⁰. Since DOIs can be resolved to URLs to retrieve the data item, an implementation of f_{ret} is possible. This features stands out in comparison with other metadata formats. DataCite supports two elements to detect whether the metadata describe images: On the one hand the "generalResourceType"-field, with "Image" as a possible value (these are determined by a controlled vocabulary), on the other hand the "formats"-element of DataCite. This element can specify the mime-type of complex research data, which can be used here to include those data sets whose formats match the "image/*" pattern. DataCite has therefore been chosen to be the metadata format for the implementation of the benchmark.

With all these choices in mind, step two ("Selection Data Providers") is implemented as an iteration of all RDRs selected in step one, and a ListMetadataFormats-OAI-PMH query to filter out non-functional RDRs and those which do not support DataCite. The output from the previous step is updated accordingly.

3) *Retrieve Data Catalogue:* The retrieval of the data catalogue acts as a necessary precondition to select data of interest and to assess the metadata according to the Quality

criteria. During this step all RDRs providing metadata in a DataCite format are harvested via the OAI-PMH protocol. The harvesting has been distributed to four parallel working harvester processes, but any number of harvest workers can be used. The code allows us to balance the harvest of the RDRs (allocating special resources to bigger RDRs and bundling smaller RDRs together).

Often, a trade-off has to be made: On the one hand investing too much in handling errors caused by missing compliance with the OAI-PMH standard or poor service quality would violate the non-creativity requirement. On the other hand getting more metadata records was necessary to calculate meaningful metrics. To name but one example: During the harvest the timeout for an HTTP request has been set to 20 seconds and the harvester has even been programmed to retry once more after a timeout. Nevertheless some RDRs couldn't be harvested or not to the full extent.

The output of the step three is a distributed data catalogue in the DataCite format of all RDRs selected in step two.

4) *Select Data:* Considering the use case, this step consists of the processing of the retrieved data catalogues, its merging and filtering out data items that are not of interest. For the use case, this would lead to a subset of a merged data catalogue, only comprising records for research data products that match certain search criteria (e.g. images annotated with a specific time of creation or license). For the benchmark, in contrast, all research data products are enlisted with the specific values and evaluated with the assessment functions:

- f_{lic} checks whether the attribute "rightsURI" is filled at least once with a valid URL. Only this field allows for a unambiguous identification of the effective license (the rights field itself allows free text).
- f_{geo} checks whether the geoLocations element has at least one child with valid contents.
- f_{chrono} checks whether the dates element has a child with dateType attribute set to "Created".

Step four has been executed on twelve worker processes, but the code allows scaling up to any number. Output of this step is a list for each data item of interest including the corresponding accessor function result and some additional administrative information.

5) *Data Retrieval:* The fifth step of the generic workflow of research data integration is the retrieval of the data based on the list of data items defined in the previous step. The harmonisation of the data is out of scope of our use case and benchmark.

For five RDRs the data catalogue included data of interest in the sense defined in Section III. These five are not homogeneous in their support for data retrieval. None of the APIs registered at re3data.org offered a uniform and automatable way to determine the protocol and endpoint for data retrieval based on the metadata alone.

¹⁰<https://www.doi.org>

Table II
SCORE OF RDRs

RDR name	Items of Interest	$score_{avabsolute}$	$score_{avrelative}$	$f_{chrono} = 1$	$f_{geo} = 1$	$f_{lic} = 1$	$f_{ret} = 1$
figshare	1.224.071	0.0000004	0.0000004	0	0	0	2
Zenodo	184.796	0.2500000	0.2245688	0	0	184.796	0
PANGAEA	35	0.6642857	0.6558059	0	29	32	32
PUB Data Publications	18	0.2500000	0.2245688	0	0	18	0
GFZ Data Services	9	0.5277778	0.5230702	8	5	6	0

Note: n = 1.408.929 (June 2018)

Since DOIs are mandatory for DataCite metadata, resolving the DOI to a URL results at least in a request to a human-readable page (a so-called splash or landing page). But these pages provide no machine-actionable retrieval mechanism: a human being can identify a download link, if it is present on the page, whereas a crawler only sees a number of unqualified links. Screen-scraping all landing pages would violate all three requirements for the prototype.

The most viable and machine-actionable option therefore is to use HTTP Content negotiation [27] to get the access to the images, which did work for research data products from two RDRs (Pangaea and figshare). Content negotiation allows to systematically retrieve the data, in principle even if the protocol has to change (e.g. from HTTP to FTP). Content negotiation also allows to retrieve different representations of the research data products (e.g. bibtex-formatted citations of the research data products).

f_{ret} can therefore be implemented on the basis of content negotiation: The DOI URL is requested via HTTP with the additional header set as "Accept: image/*". If the request results in 200 HTTP status code eventually (iterating over all redirecting HTTP status codes) and a Content-type header field is set that matches the image/* wildcard, f_{ret} returns 1. If this client-initiated content negotiation fails, the availability of server-sided content negotiation can be checked: If the server sets a HTTP Link header (see [28]) in its last reply, the value of the header can include the necessary data retrieval information: If one of the link-values has a type-field set to a mime-type that is identical to a annotated format in the metadata, the URI-reference is requested. If the request results in a 200 HTTP status code eventually (again after following possible redirects) and a Content-type that matches this specific mime-type, f_{ret} returns 1. In all other cases f_{ret} returns 0.

If we assume that the check for f_{ret} takes about two seconds on average a sequential calculation for 1,4 million records would take more than 32 days. Parallelisation is therefore a necessity. With 34 retrieval workers checking f_{ret} in parallel, step 5 took less than one day. The number of retrieval workers can again be scaled, if the number of records to be checked might increase in the future.

Table III
RARENESS AND WEIGHT FOR QUALITY CRITERIA

Quality Criteria	$ Q_i $	Rareness	Weight
Q_{lic}	184852	0.8687996	0.2245688
Q_{geo}	34	0.9999759	0.2584755
Q_{ret}	34	0.9999759	0.2584755
Q_{chrono}	8	0.9999943	0.2584802

Note: n = 1.408.929 - total rareness = 3.87

V. RESULTS

In total we retrieved DataCite-metadata for 1.408.929 images, i.e. research data products of interest. The numbers aggregated over the RDRs are shown in Table II. Since these data rely on only one run in June 2018, they need to be interpreted with care (cf. the corresponding considerations in subsection VI-A). Nevertheless, a first preliminary interpretation will be given in the following paragraphs. This interpretation is up to revision when the benchmark has been executed more often.

With regard to the calculation of the FAIR metrics the method proposed in Section III leads to reasonable results. Since the images managed in Pangaea and the GFZ Data Services support three out of four quality criteria almost completely, their score is high. The relatively low compliance of images managed in figshare manifests in small scores.

The most surprising result of the first benchmark execution is the low number of temporal annotations. It was to be expected that existing metadata in the image files would lead to a high number of annotations: Many formats can be automatically analysed with open source tools, like the exif tool suite.¹¹ This possibility has been checked with random samples out of the images with the mime type "image/jpeg". In all cases this turned out to be a viable option to programmatically determine the date of creation of the image.

Similar considerations apply to georeferential annotations, but to a lesser extent: Whether geo-tagging of images is part of the image-file's metadata is dependent on the feature set

¹¹see <https://www.sno.phy.queensu.ca/~phil/exiftool/>

supported by the device used for its creation (date-tagging should be more common).

The number of data items that are annotated with a license-URI is large; hence it brings out the difference between the absolute and relative scores. Only license information that is given by a URL are accepted, which is the reason why $|Q_{lic}|$ is lower than could be expected (the free-text license field has been ignored).

The value of $|Q_{ret}|$ has been expected to be low, since automatic retrievability of heterogeneous research data is a challenging task. The first benchmark execution indicates that this assumption is justified.

According to the data collection no single research data product satisfies all quality criteria. Some steps to improve this situation are proposed in the following section.

The calculated rareness and weight numbers can be read off from Table III. They are heavily affected by the size of the two biggest RDRs, figshare and Zenodo. The total rareness is the sum of all rareness values. Its value (3.87) is very close to the maximum (4, or number of quality criteria). This value can be taken as a general indicator how well the set of all tested RDRs supports the use case and its tech stack.

VI. DISCUSSION

The prototypical implementation of the use-case-centric FAIR metrics presented here accounts for 10 out of the 15 FAIR principles [3], as can be seen in Table IV. A checkmark depicts principles that are covered by the overall execution and not by a specific quality criteria. The main claim of the presented approach consists in the hypothesis that the use case chosen provides a specific meaning to the vague criteria F2 and R1 and therefore makes adherence to these principles measurable. The implementation is a proof-of-concept to justify this claim.

F1, F3 and A1 are accounted for, since DataCite requires DOIs as a research data identifier. Since only HTTP and open and freely available protocols based on HTTP are used in the benchmark A1.1 is covered. F4 is covered, since re3data.org is queried as a searchable resource. To check authentication or authorisation is not part of the use case, but A1.2 is nevertheless covered, since HTTP allows for this functionality. Metadata in DataCite format are taken to be a language for knowledge representation, therefore the score also checks for I1.

R.1.3 is not the focus of the generic use case chosen for the implementation. Nevertheless the checked quality criteria of the use case could be extended to cover domain-relevant community standards, hence disambiguating another vague criterion.

A query to the "Bielefeld Academic Search Engine" (BASE)¹² [29] in June 2018 with comparable search parameters (restricting to type "Still image") resulted in about 8,5 million results. This indicates that the first run of the prototype covered an essential part of the available images in the context of research data. BASE included more than 129 million records in June 2018, primarily focusing on publications (which includes but is not limited to research data as understood in this paper).

A. Caveats and limits of validity

Although our use case is quite generic and should therefore capture many aspects of research data integration, it will definitely not capture all. There will be disciplines that find another use case or another technology better-suited for a use-case-centric metric. While their implementation and set of all quality criteria $Q_i (1 \leq i \leq n)$ will hence differ from the one presented in this paper as a proof of concept, the general approach to calculate the metric will still apply.

¹²<https://www.base-search.net>

Table IV
CASE STUDY COVERAGE OF FAIR PRINCIPLES

Principle	Coverage
F1 "(meta)data are assigned a globally unique and persistent identifier"	✓
F2 "data are described with rich metadata"	Q_{geo}, Q_{chrono}
F3 "metadata clearly and explicitly include the identifier of the data it describes"	✓
F4 "(meta)data are registered or indexed in a searchable resource"	✓
A1 "(meta)data are retrievable by their identifier using a standardized communications protocol"	Q_{ret}
A1.1 "the protocol is open, free, and universally implementable"	Q_{ret}
A1.2 "the protocol allows for an authentication and authorization procedure, where necessary"	Q_{ret}
A2 "metadata are accessible, even when the data are no longer available"	✗
I1 "(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation."	✓
I2 "(meta)data use vocabularies that follow FAIR principles"	✗
I3 "(meta)data include qualified references to other (meta)data"	✗
R1 "(meta)data are richly described with a plurality of accurate and relevant attributes"	Q_{geo}, Q_{chrono}
R1.1 "(meta)data are released with a clear and accessible data usage license"	Q_{lic}
R1.2 "(meta)data are associated with detailed provenance"	✗
R1.3 "(meta)data meet domain-relevant community standards"	✗

Although the greatest care was taken in its implementation, the benchmark is a proof-of-concept in the form of software and hence bugs are a possibility. A stricter implementation which has fewer workarounds or an implementation with an additional set of features or another technique on the base level of the implementation might lead to differing metrics for a specific research data product or RDR. Since this is always the case if software is used to produce scientific data, we hence publish the source code of the study with a license, that allows to reuse and modify it, as long as the resulting code is itself made publicly available. This allows for code adaption and reruns of the metric to reproduce or refute our measurements. The idea of use-case-driven FAIR metrics is nevertheless not invalidated by these kinds of shortcomings of prototypical implementations.

A case-study-execution is in principle time-biased. This feature of use-case-centric metrics is independent of the implementation, unlike the aforementioned aspects: Some services could have maintenance activities during the execution. As a first countermeasure we repeated unsuccessful harvests to handle this threat to validity. The execution of a benchmark should always be repeated on a regular basis and the results should always be annotated with the date of the execution. This handling will help to detect and treat shortcomings due to the time bias.

Since the overall process to collect the data and calculate the metric currently takes about two and a half days, it is not feasible to rerun it too often to get a bigger sample size. Another problem in this context is the load step three and five causes on the RDRs, which could be taken for an attack if they run "too efficiently".

B. Lessons learned

In this subsection lessons learned from the use case implementation are given. These are directed primarily at two target groups:

- *RDR providers*: Maintainers of RDRs could be interested in improving their score or in enabling their RDR to be evaluated in a benchmark like the presented one.
- *Researchers*: They could be interested in an implementation of a benchmark to support another use case or they could want to implement a workflow for integration of research data based on the technology used (with or without the intent of measuring FAIRness).

Providers of RDRs should consider registering to re3data.org and offering OAI-PMH as an API to increase the impact of their data. re3data.org is a valuable asset to automate research data integration tasks across RDRs. Some information are inconsistent (e.g. some URLs of APIs registered on re3data.org are not pointing to the actual API endpoint, but to the documentation) or out-of-date. RDR maintainers should regularly check them and request an update. To guarantee interoperability of registry retrievals as in step one and two of our generic workflow, future services

other than re3data.org should be compliant to their schema [30], or at least provide a mapping.

OAI-PMH is a good option for providers of RDRs, if they want to support the generic workflow of research data integration (cf. Figure 1). Nevertheless there are reasons for an update (current version is 2.0) or an alternative protocol supporting the semantics sketched in Section III: Improvements for a retrieval of a complete data catalogue could be achieved if more efficient protocols would be supported (e.g. a compressed metadata dump retrievable by FTP which is in the format of an unchunked ListRecords or ListIdentifiers response). Other missing features are the possibility to retrieve the number of records available via a specific metadata format or a way to retrieve the size of the chunks (how many metadata records are returned per request). Some of these aspects are targeted with the ResourceSync framework [31], but this framework is technically very different from OAI-PMH and it is currently not broadly adopted.

Providers of RDRs which already offer an OAI-PMH interface should review its compliance with the standard. 56 or 34.57% out of 162 RDRs have an unresponsive or uncompliant OAI-PMH endpoint. The usage of resumption tokens to chunk the harvest into manageable packages should be considered if the RDR has more than 1000 items or supports potentially extensive metadata formats, such as DataCite. The size of the chunks provided by the OAI-PMH server should also be chosen with consideration. If the chunk size includes only 10 records per request and the OAI-PMH interface is limited to one request per second, step three of our implementation becomes very time-consuming. Under these circumstances the harvest of a RDR with over two million records would take more than two days (in the best case).

Many image formats support automatic annotation by providing the metadata as a part of the file format. Automatic annotation procedures on ingest or by iterating over the existing research data stock could be implemented to increase the number of temporally (and probably also spatially) annotated data. This will heavily increase both the absolute and the relative score of the research data products.

Researchers and data stewards should always use URIs to identify the license of the research data. Free text does not allow for a doubtless and machine-actionable determination under which conditions a research data item can be reused.

Considering the retrieval of data items more RDRs should follow the example set by PANGAEA and use the Link header to enable direct resource downloading or alternatively honour client-side content negotiation.

To complete the picture with regard to the FAIRness of the research data landscape, additional case studies should be designed, executed and their results published. As already stated in Section II, [9] might be a good option. The research data integration technology presented in [9] is based on

another technology (semantic web and open linked data) and its architecture is compatible with the presented approach to calculate use-case-centric FAIR metrics.

Another use case could include the retrieval of statistical data in a text-based content type (such as CSV files). Finding data sets fitting a specific research question and automatically harmonising and evaluating the full data set is a non-trivial but common task.

A third possibility would be to find a use case that relies on the five FAIR guiding principles which were not covered by our case study.

VII. CONCLUSION

In this paper an approach to design and implement use-case-centric FAIR metrics has been presented. This approach allows to cover research data integration across different data sources in the calculated metric. The use case furthermore provides the necessary context to disambiguate two of the FAIR guiding principles, which have not been measurable so far. The prototypical implementation shows the viability of this approach. The method and parts of the source code can be reused by other case studies, covering other disciplines or technologies. If benchmark executions are scheduled regularly, the analysis of the results will allow a re-evaluation of the first run and an in-depth interpretation of trends can be given.

If more case studies beyond the one presented in this paper will be implemented and executed, two valuable achievements are in sight: On the one hand, the results will help to gain a more complete picture of the research data landscape. On the other hand, getting a higher score in most of the use-case-centric FAIR metrics will motivate providers of RDRs to be more FAIR. It will also help researchers to use RDRs in a way that hopefully results in newly gained knowledge.

ACKNOWLEDGMENTS

We like to thank Stephan Hachinger and Tobias Gugge-mos for insightful feedback to prior versions of this paper.

The paper could not have been written without a lot of open source software. Nevertheless the R package "knitr" (presented in [32]) with its R integration into L^AT_EX deserves specific praise for its outstanding usefulness in dynamically integrating statistical information.

This work was supported by the DFG (German Research Foundation) with the GeRDI project (Grant No. BO818/16-1).

REFERENCES

- [1] P. Ayris, J. Berthou, R. Bruce, S. Lindstaedt, A. Monreale, B. Mons, Y. Murayama, C. Södergård, K. Tochtermann, and R. Wilkinson, "Realising the european open science cloud," *First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud*, 2016.
- [2] L. Federer, "Research data management in the age of big data: Roles and opportunities for librarians," *Information Services & Use*, vol. 36, no. 1-2, pp. 35–43, 2016.
- [3] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific data*, vol. 3, 2016.
- [4] S. Cox and J. Yu, "OzNome 5-start Tool: A Rating System for making data FAIR and Trustable (presentation given at the 2017 eResearch Australasia Conference)," Oct. 2017. [Online]. Available: <https://conference.eresearch.edu.au/wp-content/uploads/2017/07/Simon-Cox.pdf>
- [5] M. D. Wilkinson, S.-A. Sansone, E. Schultes, P. Doorn, L. O. Bonino da Silva Santos, and M. Dumontier, "A design framework and exemplar metrics for fairness," *bioRxiv*, 2017. [Online]. Available: <https://www.biorxiv.org/content/early/2017/12/01/225490>
- [6] M. Wilkinson, L. O. Bonino, N. Nichols, and K. Leinweber, "FAIRMetrics/Metrics: Proposed FAIR Metrics and results of the Metrics evaluation questionnaire," Mar. 2018. [Online]. Available: <https://doi.org/10.5281/zenodo.1205235>
- [7] J. Boehmer, A. Dunning, and M. de Smaele, "Are the fair data principles fair?" in *12th International Digital Curation Conference*, 1 2017.
- [8] —, "Evaluation of data repositories based on the FAIR Principles for IDCC 2017 practice paper," <http://dx.doi.org/10.4121/uuid:5146dd06-98e4-426c-9ae5-dc8fa65c549f>.
- [9] M. D. Wilkinson, R. Verborgh, L. O. Bonino da Silva Santos, T. Clark, M. A. Swertz, F. D. Kelpin, A. J. Gray, E. A. Schultes, E. M. van Mulligen, P. Ciccarese, A. Kuzniar, A. Gavai, M. Thompson, R. Kaliyaperumal, J. T. Bolleman, and M. Dumontier, "Interoperability and fairness through a novel combination of web technologies," *PeerJ Computer Science*, vol. 3, p. e110, Apr. 2017. [Online]. Available: <https://doi.org/10.7717/peerj-cs.110>

- [10] Members of the RDA Research Data Repository Interoperability Working Group, "Research data repository interoperability primer," jun 2017, This document is a Research Data Alliance Supporting Output. [Online]. Available: <https://doi.org/10.15497/RDA00020>
- [11] K. Gregory, P. Groth, H. Cousijn, A. Scharnhorst, and S. Wyatt, "Searching data: A review of observational data retrieval practices," *arXiv preprint arXiv:1707.06937*, 2017.
- [12] R. Devarakonda, G. Palanisamy, J. M. Green, and B. E. Wilson, "Data sharing and retrieval using OAI-PMH," *Earth Science Informatics*, vol. 4, no. 1, pp. 1–5, 2011.
- [13] H. v. d. Sompel, M. L. Nelson, C. Lagoze, and S. Warner, "Resource harvesting within the OAI-PMH framework," *D-Lib Magazine*; 2004 [10] 12, 2004. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0164121214002040>
- [14] M. Kindling, H. Pampel, S. van de Sandt, J. Rücknagel, P. Vierkant, G. Kloska, M. Witt, P. Schirnbacher, R. Bertelmann, and F. Scholze, "The Landscape of Research Data Repositories in 2015: A re3data analysis," *D-Lib Magazine*, vol. 23, no. 3/4, 2017. [Online]. Available: <http://www.dlib.org/dlib/march17/kindling/03kindling.html>
- [15] H. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, "Big data and its technical challenges," *Communications of the ACM*, vol. 57, no. 7, pp. 86–94, 2014.
- [16] C. Lynch, "Big data: How do your data grow?" *Nature*, vol. 455, no. 7209, pp. 28–29, 2008.
- [17] Y. Demchenko, P. Grosso, C. De Laat, and P. Membrey, "Addressing big data issues in scientific data infrastructure," in *Collaboration Technologies and Systems (CTS), 2013 International Conference on*. IEEE, 2013, pp. 48–55.
- [18] J. D. Lusier, W. L. Thompson, J. M. Wilson, B. E. Gorham, and L. D. Dragut, "Using digital photographs and object-based image analysis to estimate percent ground cover in vegetation plots," *Frontiers in Ecology and the Environment*, vol. 4, no. 8, pp. 408–413, 2006. [Online]. Available: [http://dx.doi.org/10.1890/1540-9295\(2006\)4\[408:UDPAOI\]2.0.CO;2](http://dx.doi.org/10.1890/1540-9295(2006)4[408:UDPAOI]2.0.CO;2)
- [19] C. R. Johnson, E. Hendriks, I. J. Berezhnoy, E. Brevdo, S. M. Hughes, I. Daubechies, J. Li, E. Postma, and J. Z. Wang, "Image processing for artist identification," *IEEE Signal Processing Magazine*, vol. 25, no. 4, pp. 37–48, July 2008.
- [20] N. T. de Sousa, W. Hasselbring, T. Weber, and D. Kranzlmüller, "Designing a Generic Research Data Infrastructure Architecture with Continuous Software Engineering," in *Software Engineering Workshops 2018*, ser. CEUR Workshop Proceedings, vol. Vol-2066. CEUR-WS.org, March 2018, pp. 85–88. [Online]. Available: <http://eprints.uni-kiel.de/42215/>
- [21] H. Pampel, P. Vierkant, F. Scholze, R. Bertelmann, M. Kindling, J. Klump, H.-J. Goebelbecker, J. Gundlach, P. Schirnbacher, and U. Dierolf, "Making Research Data Repositories visible: The re3data.org Registry," *PLOS ONE*, vol. 8, no. 11, pp. 1–10, 11 2013. [Online]. Available: <https://doi.org/10.1371/journal.pone.0078080>
- [22] "Databib and re3data.org merge," <https://cms.library.illinois.edu/export/lsdata/news/databibandre3data.html>, Mar 2015, "[Online; Accessed 2018-March-04]". [Online]. Available: <https://cms.library.illinois.edu/export/lsdata/news/databibandre3data.html>
- [23] M. Gudgin, H. F. Nielsen, A. Karmarkar, N. Mendelsohn, J.-J. Moreau, Y. Lafon, and M. Hadley, "SOAP version 1.2 part 1: Messaging framework (second edition)," W3C, W3C Recommendation, Apr. 2007, <http://www.w3.org/TR/2007/REC-soap12-part1-20070427/>.
- [24] S. Harris and A. Seaborne, "SPARQL 1.1 query language," W3C, W3C Recommendation, Mar. 2013, <http://www.w3.org/TR/2013/REC-sparql11-query-20130321/>.
- [25] R. Verborgh and M. Dumontier, "A web api ecosystem through feature-based reuse," *arXiv preprint arXiv:1609.07108*, 2016.
- [26] P. A. Covitz, F. Hartel, C. Schaefer, S. De Coronado, G. Fragoso, H. Sahni, S. Gustafson, and K. H. Buetow, "cacore: A common infrastructure for cancer informatics," *Bioinformatics*, vol. 19, no. 18, pp. 2404–2412, 2003. [Online]. Available: <http://dx.doi.org/10.1093/bioinformatics/btg335>
- [27] R. Fielding and J. Reschke, "Hypertext transfer protocol (http/1.1): Message syntax and routing," Internet Requests for Comments, RFC Editor, RFC 7230, June 2014, <http://www.rfc-editor.org/rfc/rfc7230.txt>. [Online]. Available: <http://www.rfc-editor.org/rfc/rfc7230.txt>
- [28] M. Nottingham, "Web Linking," Internet Requests for Comments, RFC Editor, RFC 5988, October 2010, <http://www.rfc-editor.org/rfc/rfc5988.txt>. [Online]. Available: <http://www.rfc-editor.org/rfc/rfc5988.txt>
- [29] A. Bäcker, C. Pietsch, F. Summann, and S. Wolf, "BASE (Bielefeld Academic Search Engine). Eine Suchmaschinenlösung zur Indexierung wissenschaftlicher Metadaten," *Datenbank-Spektrum*, vol. 17, no. 1, pp. 5–13, 2017.
- [30] P. Vierkant, S. Spier, J. Rücknagel, H. Pampel, F. Fritze, J. Gundlach, D. Fichtmüller, M. Kindling, A. Kirchhoff, H.-J. Goebelbecker, J. Klump, G. Kloska, E. Reuter, A. Semrau, E. Schnepf, M. Skarupianski, R. Bertelmann, P. Schirnbacher, F. Scholze, C. Kramer, R. Ulrich, M. Witt, and C. Fuchs, "Schema for the Description of Research Data Repositories - RFC Version 2.2," Sep. 2014. [Online]. Available: <https://doi.org/10.5281/zenodo.11748>
- [31] M. Klein, R. Sanderson, H. Van de Sompel, S. Warner, B. Haslhofer, M. Nelson, and C. Lagoze, "Resourcesync framework specification," 2016. [Online]. Available: <https://www.openarchives.org/rs/1.1/resourcesync>
- [32] Y. Xie, *Implementing reproducible research*. CRC Press, 2014, vol. 1, ch. knitr: a comprehensive tool for reproducible research in R, p. 20.