Machine-Actionable Assessment of Research Data Products

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität München

eingereicht von
Tobias Weber
München, den 08.12.2020

Erstgutachter: Prof. Dr. Dieter Kranzlmüller
Zweitgutachter: Prof. Dr. Isabella Peters
Tag der mündlichen Prüfung: 19.02.2021

This work is dedicated to Alexandra Elbakyan.

Abstract

Research data management is a relevant topic for academic research which is why many concepts and technologies emerge to face the challenges involved, such as data growth, reproducibility, or heterogeneity of tools, services, and standards. The basic concept of research data management is a research data product; it has three dimensions: the data, the metadata describing them, and the services providing both. Traditionally, the assessment of a research data product has been carried out either manually via peer-review by human experts or automated by counting certain events. We present a novel mechanism to assess research data products.

The current state-of-the-art of machine-actionable assessment of research data products is based on the assumption that its quality, impact, or relevance are linked to the likeliness of peers or others to interact with it: *event-based metrics* include counting citations, social media interactions, or usage statistics. The shortcomings of event-based metrics are systematically discussed in this thesis; they include dependance on the date of publication and the impact of social effects.

In contrast to event-based metrics *benchmarks* for research data products simulate technical interactions with a research data product and check its compliance with best practices. Benchmarks operate on the assumption that the effort invested in producing a research data product increases the chances that its quality, impact, or relevance are high. This idea is translated into a software architecture and a step-by-step approach to create benchmarks based on it.

For a proof-of-concept we use a prototypical benchmark on more than 795,000 research data products deposited at the Zenodo repository to showcase its effectiveness, even with many research data products. A comparison of the benchmark's scores with event-based metrics indicate that benchmarks have the potential to complement event-based metrics and that both weakly correlate under certain circumstances. These findings provide the methodological basis for a new tool to answer scientometric questions and to support decision-making in the distribution of sparse resources. Future research can further explore those aspects of benchmarks that allow to improve the reproducibility of scientific findings.

Kurzfassung

Dass das Management von Forschungsdaten ein relevantes Thema ist, zeigt sich an der Vielzahl an konzeptioneller und technischer Antworten auf die damit einhergehenden Herausforderungen, wie z.B. Datenwachstum, Reproduzierbarkeit oder Heterogenität der genutzten Tools, Dienste und Standards. Das Forschungsdatenprodukt ist in diesem Kontext ein grundlegender, dreiteilig aufgebauter Begriff: Daten, Metadaten und Dienste, die Zugriffe auf die beiden vorgenannten Komponenten ermöglichen. Die Beurteilung eines Forschungsdatenprodukts ist bisher händisch durch den Peer Review oder durch das Zählen von bestimmten Ereignissen realisiert.

Der heutige Stand der Technik, um automatisiert Qualität, Impact oder Relevanz eines Forschungsdatenprodukts zu beurteilen, basiert auf der Annahme, dass diese drei Eigenschaften mit der Wahrscheinlichkeit von Interaktionen korrelieren. *Event-basierte Metriken* umfassen das Zählen von Zitationen, Interaktionen auf sozialen Medien oder technische Zugriffe. Defizite solcher Metriken werden in dieser Arbeit systematisch erörtert; besonderes Augenmerk wird dabei auf deren Zeitabhängigkeit und den Einfluss sozialer Mechanismen gelegt.

Benchmarks sind Programme, die Interaktionen mit einem Forschungsdatenprodukt simulieren und dabei die Einhaltung guter Praxis prüfen. Benchmarks operieren auf der Annahme, dass der Aufwand, der in die Erzeugung und Wartung von Forschungsdatenprodukte investiert wurde, mit deren Qualität, Impact und Relevanz korreliert. Diese Idee wird in dieser Arbeit in eine Software-Architektur gegossen, für deren Implementierung geeignete Hilfsmittel bereitgestellt werden.

Ein prototypischer Benchmark wird auf mehr als 795.000 Datensätzen des Zenodo Repositorys evaluiert, um die Effektivität der Architektur zu demonstrieren. Ein Vergleich zwischen Benchmark Scores und event-basierten Metriken legt nahe, dass beide unter bestimmten Umständen schwach korrelieren. Dieses Ergebnis rechtfertigt den Einsatz von Benchmarks als neues szientrometrisches Tool und als Entscheidungshilfe in der Verteilung knapper Ressourcen. Der Einsatz von Benchmarks in der Sicherstellung von reproduzierbaren wissenschaftlichen Erkenntnissen ist ein vielversprechender Gegenstand zukünftiger Forschung.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich, Tobias Weber, an Eidesstatt, dass die vorliegende Dissertation
Machine-Actionable Assessment of Research Data Products
von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

Heilbronn, 19.7.2021	*Tobias Weber*
Ort, Datum	Unterschrift

Contents

Chapter 1

Introduction

Contents

This chapter provides the necessary background to state and motivate the research question. In Section 1.1 the main concepts are introduced. The research question is presented in Section 1.2 and motivated in Section 1.3 by examples of challenges involved in the context of assessing research data products. These considerations underline the central role of the concept of a research data product. The chapter closes with a tabular overview of the contents of the remaining chapters of this thesis (Section 1.4).

1.1 Definitions of Basic Concepts

This section introduces the main concepts involved in the assessment of a research data product;[1] they are explicated to finally phrase the research question in the next section. The systematic introduction of terms is meant to fix their usage throughout this thesis, especially for terms which are unfortunately not disambiguated in parts of the relevant literature. The following two subsections are ordered from basic concepts up to the concepts which constitute the building blocks of the title and the research question of this thesis.

1.1.1 Research Data Product

Research denotes academic activities in the context of science and the humanities. "Science" has not been chosen to be the basic term, since in English-speaking countries it typically only refers to the natural sciences, whereas the term research is more inclusive (e.g. regarding social sciences and the humanities).

Research data are understood as all forms of digitized content that is input for or output of those activities of researchers, that are necessary to produce or verify knowledge ([WK18]). In our work, this concept has a broader sense compared to the literature, where it is typically used to differentiate supplemented material (e.g. tabular data) from publications in the classical sense (books and articles). Instead, we consider all of the above to be research data.

A *research data service* is a service providing access to either research data, metadata describing these data, or both. A research data service can be uniquely identified by an endpoint and a protocol.

A *research data product* is a composite of four *components*:

- a Persistent Identifier (PID)
- research data
- metadata describing the research data
- research data services hosting research data and their metadata

The PID is a small but essential component, since it provides a mechanism to unambiguously identify a research data product and therefore decide whether two research data products are identical (by comparing their PIDs). This is necessary, since the identity of the other three components are only a necessary, but not a sufficient criterion for identity of two research data products.

[1] The introduced concepts are highlighted in italics.

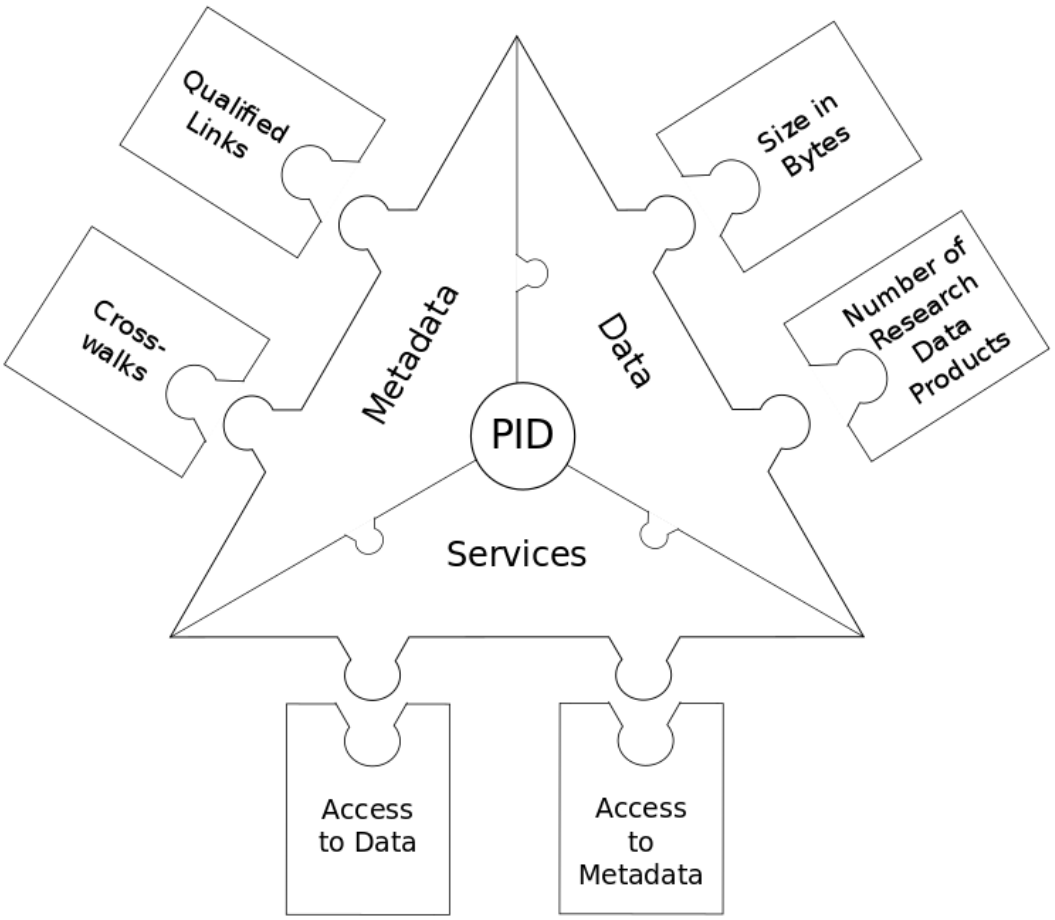

Figure 1.1: Schematic view of a research data product (including challenges for each component)

Figure 1.1 shows a schematic view of a research data product; it is meant to depict the interplay of the components and is to be read inwards to outwards: The circle in the middle depicts the PID of the research data product. The triangular jigsaw pieces arranged around the PID-circle symbolize the three other components of a research data product; they have logical interfaces to each other: the flow of information is represented by tabs (source of information) and blanks (sink of information). The triangle stands for the research data product as a whole. The quadratic jigsaw pieces around the schematic of the research data product exemplify challenges in the handling of research data products:

- **Data** (see Subsection 1.3.1):
  - size of a research data product

  - number of research data products

- **Metadata** (see Subsection 1.3.2):
  - crosswalks between two metadata formats
  - qualified links to other entities (e.g. citations)

- **Services** (see Subsection 1.3.3):
  - access to metadata with a standardized protocol
  - orchestration of access methods to the research data

1.1.2 Machine-Actionable Assessment

Before explaining the concept of an *assessment* of a research data product, the dimensions need to be specified along which such an assessment is carried out: these dimensions are captured by the terms "quality", "impact", and "relevance" of a research data product; these concepts unfortunately have no commonly agreed upon meaning throughout the literature.[2] We thus propose a working definition of the terms to enable consistent usage in the following:

- The *quality of a research data product* is understood as the chance of a research data product to be (re-)used for tasks similar to the one for which the research data product was originally created.[3] The aptness of the research data product for this original task is an edge case included in this definition.

- The *impact of a research data product* is understood as the chance of a research data product to influence the direction of research of a peer in the same or a similar field.

- The *relevance of a research data product* is understood as the chance of a research data product to influence an audience beyond the field of the creator(s). This includes outreach outside the scope of academic research.

An *assessment of a research data product* is the task to map a research data product to $\mathbb{R}^+$ according to its quality, impact, and/or relevance; low numbers indicate a lower quality, impact, and/or relevance. If unspecified, our considerations apply to all three contexts. If only one of the three contexts is of interest, it is specified accordingly. The value a research data product is mapped onto is

[2] Take as an example, how [ALW19] and [SD19] treat impact and relevance to be dimensions of quality, whereas [Gam+20] and [FW17] treat quality and relevance as independent and distinct features.

[3] Use and reuse cannot be easily separated on a conceptual level, see [San+19]

called a *score of a research data product*. Assessments of research data products are not functions in the strict mathematical meaning of the term, since the score of a research data product can vary depending on the time and date the assessment is carried out.

A task is *machine-actionable* if a machine can correctly process the task without human interaction [Wil+16]. An assessment of a research data product is therefore machine-actionable if a machine only needs an initial configuration to map the specified research data products to their scores.

Our work concentrates on two ways to assess research data products:

1. Event-based metrics (state-of-the-art)

2. Benchmarks (as developed by this thesis)

An *event-based metric* is understood to be an assessment of a research data product based on the documentation of events of interactions with the research data product. The frequency of citations, mentions, or downloads are examples for event-based metrics' scores. Event-based metrics work on the assumption that the frequency of certain events correlates with the quality, impact, and/or relevance of a research data product.[4]

Event-based metrics are state-of-the-art assessments of research data products in the context of scientific governance (in the time of writing). Event-based metrics are discussed in-depth in Chapter 3.

A *benchmark for a research data product* is understood to be an assessment of a research data product based on simulated interactions with the research data product: a computer program is used to check those characteristics of the research data product which are taken as signals for the effort put into its creation and curation. Examples for these characteristics are the compliance to data standards, the completeness of metadata, or the accessibility of the research data product via research data services. The benchmark's score of a research data product is determined by a combination of the checks' outcomes. Benchmarks follow the idea that the effort put into the creation and curation of a research data product correlates with its quality, impact, and/or relevance. This is a central assumption of this thesis and it is justified by related findings in the literature, e.g. the fact that publishing primary data as a supplement to a publication increases the impact of the publication ([PV13]), and that improving infrastructure services facilitate data reuse ([BK17]).

[4]The term article-level metric is a subclass of event-based metrics, it was not used in this thesis, since it is used only for articles in journals or conference proceedings ([LF13], [DM14], [BH18b]) and not for other research data

Benchmarks for research data products are a novel approach in the time of writing and their role is yet to be determined (see Section 2.2 for a discussion of current approaches to benchmark research data products). A conceptual architecture for benchmarks for research data products is discussed in Chapter 4.

1.2 Research Question and Contributions

With the basic concepts defined in the previous section, the research question of this thesis can be formulated:

RQ Why and how should research data products be benchmarked?

This question is answered along the discussion of the state-of-the art for assessments of research data products, namely event-based metrics for research data products: it is systematically determined, whether benchmarks for research data products have the potential to complement event-based metrics. These considerations entail a couple of Sub-Questions (SQ):

SQ-1 What are the common shortcomings of event-based metrics in use?

SQ-2 What architecture for benchmarks is induced by these shortcomings?

SQ-3 How can an architecture for benchmarks be instantiated?

SQ-4 Do the scores of a prototypical benchmark correlate with event-based metrics?

SQ-5 Can benchmarks for research data products complement event-based metrics?

The answers to these questions form the Main Contributions (MC) of this thesis:

MC-1 Systematic discussion of shortcomings of event-based metrics (Chapter 3): the discussion is based on a reproducible analysis of commonly reported problems with event-based metrics for research data product. These problems are classified, most importantly with regard to viable mitigations. The main result is the insight which shortcomings are in principle non-mitigatable.

MC-2 A common architecture for benchmarks for research data products (Chapter 4): this architecture is motivated by the challenges discussed below and the non-mitigatable shortcomings of event-based metrics. The architecture

is a blueprint for benchmarks; this blueprint needs a specific use case to fully identify all requirements for an implementation.

MC-3 A recipe to implement benchmarks for research data products and a prototype (Chapter 5): the recipe offers a procedure to realize the architecture presented in MC-2. The prototype is based on the use case to explore the contents of the Zenodo repository for research data products.

MC-4 Empirical evidence for a weak correlation of the scores of a benchmark for a research data product and event-based metrics (Section 6.2): This correlation underlines the assumption, that event-based metrics and benchmarks measure similar features of research data products but are affected by different types of noise.

MC-5 Empirical evidence for the ability of a benchmark for a research data product to complement event-based metrics (Section 6.3): This final finding justifies the continued development of benchmarks for research data products to complement event-based metrics in different usage scenarios, such as scientometric research or service proliferation.

The following section sketches the challenges involved in answering the research question and its sub-questions, and gives a rationale why they cannot be answered component-wise (PID, data, metadata, services), but must be tackled from a holistic stance, that is, with the concept of a research data product at heart.

1.3 Challenges in the Assessment of Research Data Products

This section emphasizes why the importance of the research question and states which challenges need to be addressed in answering it. Assessments of research data products are necessary to decide on the distribution of resources, such as access to funds, to services of infrastructure providers, and to publication opportunities; they are therefore one of the tasks involved in research data management; research data management is considered a high-priority topic in scientific and political discourse alike ([Pat16], [Ayr+16]); these are three of the challenges involved:

- The *growth* of research data is too fast for important resources (experts, funding, etc.) to keep pace. ([Lyn08], [BHS09]).
- The *heterogeneity* of technical solutions developed to manage research data is a challenge [WS18], especially across borders of fields of study [Gru+17].

- The *absence of conventions* that prescribe the orchestration of the technologies used to integrate research data across different sources.[5]

Each of the following three subsections discusses one of these challenges, and its implications for the assessment of research data products; the discussions motivate the research question presented above. These challenges (among others) are also discussed in international associations of experts, such as the Research Data Alliance (RDA)[6] or the Committee on Data for Science and Technology (CODATA).[7] For an in-depth discussion of the challenges and related topics, the above-mentioned organizations and their output (e.g. recommendations or proceedings of conferences) should be consulted. Especially the human-related aspect of the challenges is *not* the focus of this thesis; there are excellent contributions discussing social aspects of research data management (e.g. [San+19] on the semantics of "(re)using data" and the study carried out in [Gre+20] which proposes the evidence-based concept of data communities, called "community of use"). Our work is a contribution to the intersection of computer science and scientometrics and therefore concentrates on machine-related challenges, but societal aspects are discussed where they directly influence said challenges (see Subsection 7.2.4 for an example).

1.3.1 Growth of Research Data

The expression "growth of research data" can have several meanings, among them the growth of the total amount of data measured in *Bytes[8] and the growth of the number of research data products. Both define challenges for assessments of research data products and will be discussed in the following.

Growth in *Bytes

Although there is common agreement, that the growth of research data in *Bytes is both positive and very dynamic, (e.g. [SG06], [Coo+15], [RS16]) there is no easy way to precisely quantify this growth, let alone characterize it as exponential. Evidence for said growth of research data in *Bytes has to be collected over distributed sources:

[5] While the first two challenges are described in the literature, the third is a practical problem which, to our knowledge, has only recently been identified; on the one hand by the experience gained during research projects such as the DFG-funded project GeRDI (Generic Research Data Infrastructure), on the other hand by discussions in the context of the RDA.

[6] `https://rd-alliance.org`

[7] `https://codata.org`

[8] Since storage systems' capacities grow so fast that replacing the star with the "largest" Greek prefix used in the time of writing might make the following considerations look obsolete, the star is used to stress the fact that these considerations apply regardless of the current scale.

- Extrapolations and/or estimations can be found in *literature*:
  - 1 petabyte of data per year in particle physics and astronomy [BHS09]
  - 24 petabyte of data per year in particle physics and biology [Gru+15]

  Ignoring the partial mismatch in fields of study, both point estimates from 2009 and 2015 can be modelled as exponential growth in the 6 years with a growth rate of approximately 1.7.

- *Technical reports* of infrastructure providers sometimes include historical data or estimations of the expected demand:
  - The Royal Society ([Bou+12]), reports growth rates of approximately 250 gigabytes per year for a large institutional research data repository such as the DSpace installation at the Massachusetts Institute of Technology (MIT). This would indicate linear growth.
  - The annual report of the Leibniz Supercomputing Centre (LRZ) ([Lei17]) reports more than 50 million gigabytes of used archive space in 2017, which is more than 7,229 times the amount used in 1997. Figure 1.2 displays the overall development of the storage capacity of the LRZ; the y-axis displays the capacity in gigabytes, the x-axis shows the development over time in years and the name of the storage solution if a new system is purchased that year. The meaning of the three colors is described in the figure. The numbers reported can be modelled as exponential growth with a growth rate of about 1.56.

- It is also an option to infer growth rates of research data from *estimations of global data growth*:
  - ScienceDaily claimed that 90 percent of all data in 2013 were produced in the last two years [SIN13].
  - The International Data Corporation predicts a global data growth from 33 zettabytes (2018) to 175 zettabytes (2025) [RGR18], which can be taken as an indication for a modest exponential growth of research data.[9]

Even if we assume that there is no conclusive evidence for *exponential* growth of all research data in *Bytes, the provided sources indicate that the growth of research data in *Bytes is positive, and a linear model of growth probably underestimates it.

[9] $175 = 33(1.2690185818^7)$, hence a growth rate of about 1.27, which would result in approximately 225 zettabytes in 2026.

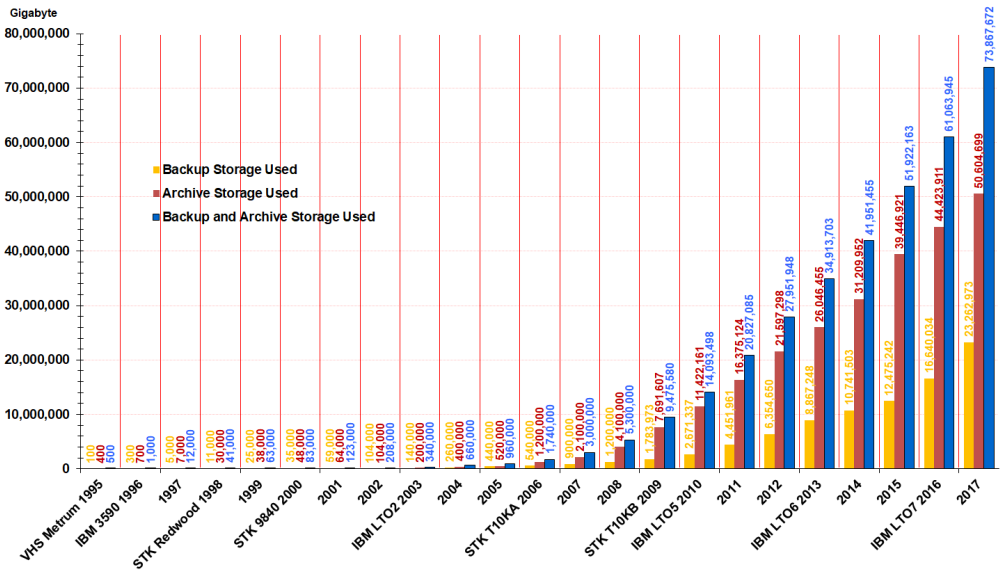

Figure 1.2: Growth of used archive and backup storage at the Leibniz Supercomputing Centre 1995–2017, adapted version from [Lei17]

The following challenge for both research data management and assessments of research data products is a result of the above articulated insights in data growth: the expected upper bound for the size of a research data product is moving beyond the capacities of the technical infrastructure currently available. To manage this issue, it is not sufficient to point to the ever-growing availability of storage space — these systems need proper maintenance by human experts and the curation of large data depositions are also bound to manual work (e.g. to annotate or formally publish the research data). Additionally, all these systems are in the need of funding, which is also a finite resource. Furthermore, it is not sufficient to only store and curate the data to guarantee the reproducibility of results: if researchers wish to analyze the depositions, they need sufficient bandwidth, memory, and computing power to reproduce the claims based on the data; these resources need to scale with the amount of research data in *Bytes.

While this is a known challenge for research data management in general (e.g. [BHS09]), the implication for assessments of research data products needs to be spelled out: machine-actionable solutions to assess research data products have to scale well with the size of the data to be assessed; this means that the time and the resources necessary for the assessment of a research data product should not grow at the same pace as research data in *Bytes, but instead substantially slower. To illustrate this point to the extreme: An assessment should be possible

on standard desktop hardware in a reasonable amount of time, even if the research data product fills an entire tape library.[10]

While these considerations apply to those fields of study which are traditionally counted amongst the "big sciences" (such as astronomy, genomics, or particle physics), they might be less interesting for the fields of the "long tail of science", meaning fields which are not known to depend on the newest IT infrastructure and its performance. Below, another growth-related challenge will be characterized which applies to *all* fields of study.

Growth in Number of Research Data Products

As with the growth in *Bytes, there is no canonical and all-embracing method to quantify the growth in number of research data products; but the literature indicates, that the growth has comparable dynamics:

- In [Pet+17] it is reported that "all research data repositories have witnessed an exponential growth of data deposits".[11]
- [Jin10] presents methods to quantify and estimate the growth of scientific publications and estimated that the 50 millionth article has been published by the end of 2008. The assumed doubling rate is under 24 years for articles.
- The growth rates of the Software Heritage archive is reported to be exponential over a period of over 40 years [RCZ19].

Publicly available numbers of infrastructure providers allow to add further evidence: The number of records available via the interface for the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) of the DataCite index changed from more than 7.44 million records in April 2016 [Rob+17] to more than 18.75 million records in September 2019 [Web+20]; this growth can be modeled with an exponential function with a growth rate of almost 2.[12]

The growth in numbers of research data products creates a similar but not identical challenge than the one sketched in the subsection above. One major difference is the context of the growth: there is no evidence that the growth in research data products is limited to fields of study typically clustered as "big science" (astronomy, life sciences, particle physics). Traditional means to guarantee quality in the bigger part of science and the humanities include manual reviews [WC14].

[10]The current (2020) used space for archive and backup in the tape library of the Leibniz Supercomputing Centre is about 102 million Gigabytes as of May 2020.

[11]This statement does not hold for all repositories today, confirm growth difference between 2018 and 2019 in Table 6.2.

[12]Inference from the global growth as in the previous paragraph is not feasible, since evidence for growth of global data is typically specified in *Bytes, not in discrete data products.

Therefore, the central challenge from this type of growth is that it is not to be expected that the number of persons qualified to assess a research data product grows at the same pace as the number of research data products [Cro14].

1.3.2 Heterogeneity of Metadata for Research Data

After the discussion of growth in terms of *Bytes and number of research data products as a challenge for assessments of research data products, this section presents another challenge: heterogeneity in the context of metadata. One way to deal with the growth discussed in the previous subsection is to concentrate on the metadata component of a research data product. Metadata are typically much more lightweight in terms of *Bytes and content than the whole research data product. The argument of this subsection describes two exemplary services based on metadata and continues by sketching the metadata-related challenge for research data management and assessments of research data products.

The following two services are examples for services with a large stock of data,[13] or services with a very dynamic growth rate,[14] respectively:

- the **Bielefeld Academic Search Engine (BASE)**[15] offers an index to more than 165 million documents (called research data products in this thesis) from 7,888 sources as of April 2020.
- the **DataCite index of research data**[16] includes almost 20 million records of research data products in January 2020 (growth rates are discussed in the previous section).

Both services are built upon a general-purpose metadata standard that enables the service in the first place:

- **Dublin Core**[17] is the basis for BASE and a common standard.[18] Ambiguities and semantic overlaps are the greatest weaknesses of this standard [PC09].
- The **DataCite** metadata standard[19] is the basis for the service with the same name. Although this standard is less criticized for ambiguities and

[13] The COnnecting REpositories (CORE) search engine (`https://core.ac.uk`) is another example

[14] Crossref is a comparable service, though more specialized on publications in the classical sense (`https://crossref.org`)

[15] `https://base-search.net`

[16] `https://datacite.org`

[17] `https://www.dublincore.org`

[18] It is mandatory to distribute Dublin Core on a compliant OAI-PMH server.

[19] `https://schema.datacite.org/`

semantic overlaps than Dublin Core, it still needs additional clarification and specification [Küm+19].

Both services operate on mechanisms to guarantee a minimum of metadata quality,[20] e.g. they manage ambiguities or guarantee minimal provenance-related information. BASE developed a harvesting, cleaning, and enrichment process over years [Bäc+17]. DataCite uses a strategy of incentives: A valid set of mandatory metadata is a requirement to use the DataCite DOI assignment service.

Another shared feature of both services is that they target a general scientific audience and are not designed to foster discipline-specific searches. Although both underlying metadata standards allow to add domain-specific information by using specialized controlled vocabularies, the heterogeneity of the available solutions define the limits of such interdisciplinary services and are therefore good examples of challenges caused by the heterogeneity of the metadata ecosystem:

Table 1.1: Selection of metadata standards in OAI-PMH compliant repositories listed in re3data.org in March 2018 (n=2093, selection criteria: at least 3 repositories, multiple standards can be supported by one repository) [Web18a]

Standard	# of repositories	Standard	# of repositories
DublinCore	90	METS	32
RDF	28	DIDL	26
UKETDC	24	OLAC	19
MODS	17	DCMI-Terms	16
SlimMARC21	15	ORE	10
Datacite	10	ETDMS	10
DIM	8	XOAI	6
DDI	4	Epicur	4
GMD	4	XmetaDissPlus	4
RIOXX	4	Dif	3
ISO2146	3	–	–

- **The number of metadata standards** is a known problem in the metadata ecoystem [Lan11]. Crosswalks from one standard to another are possible solution for this challenge, but crosswalks are rarely canonized, seldom machine-actionable, and often in principle impossible without the loss of information. The challenge becomes clear when it is considered how many

[20] Metadata quality is a much-discussed topic — we roughly follow [BH04], who name completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness and accessibility as the main characteristics of quality metadata.

standards a service provider aggregating research data across different repositories needs to support if full coverage of the landscape is the goal. Table 1.1 shows a selection of the fifty standards that could be found in March 2018 at re3data.org [Web18a], a registry for research data repositories.[21] The columns labeled as "Standard" show the name of the metadata standard and the columns labeled as "# of repositories" show the number of repositories providing metadata compliant to this standard. The table is ordered by number of repositories from top to bottom and left to right (in that order).

- **The number of different and inconsistent link targets** when metadata are enriched with qualified links is another challenge; links typically direct to curated content such as controlled vocabularies or ontologies. The diversity and number of available options makes aggregation over different sources a challenge, since the metadata ecosystem is in a status of "creolization" [WS18]. The following examples add evidence to this claim:
  - In May 2019 five different schemes to specify the field(s) of study of a research data product[22] were in broad use in the DataCite index [Web+20].[23] These schemes are not consistent with each other and there are no canonical and machine-actionable crosswalks between the schemes. More than 1,700 lines of code were necessary to build crosswalks to map them to a minimal scheme of fields of study [WF19].
  - The affiliation of a creator of a research data product can be specified by the ISNI number[24], the ROR-ID[25], the workInfoHomepage attribute of the FOAF-ontology[26], the identifier provided by wikidata[27], or the identifier provided by the German National Library[28]. There are no simple means to automatically translate one of these formats into another and the list is certainly not exhaustive. Similar problems arise when individual researchers or research projects need to be identified.

The challenge for research data management is to cope with this heterogeneity, meaning to pick the best fit for a use case out of the diversity of candidates available. The challenge for aggregating research data across repositories and for

[21] `http://re3data.org`

[22] Dewey Decimal Scheme, Australian and New Zealand Standard Research Classification, Narcis Classification, Basisklassifikation, Linsearch Classification, Bepress Classification

[23] "broad" is defined as used by more than 4,000 records

[24] `https://www.isni.org`

[25] `https://ror.org`

[26] Friend Of A Friend, `http://xmlns.com/foaf/spec/`

[27] `https://www.wikidata.org`

[28] `https://www.dnb.de`

assessing research data products alike is to deal with these choices. Assessment mechanisms that directly evaluate the metadata of a research data product must be based on detailed knowledge of both general-purpose and discipline-specific metadata best practices.

1.3.3 Missing Conventions in Using Research Data Services

After discussing challenges related to the data and metadata component of a research data product, we now focus on a third challenge that is related to the service component.

Research data products having a service component is not a wide-spread idea in literature and hence not broadly discussed. This is probably due to the fact, that resolving citations to books is the paradigm in information retrieval: it begins with a search based on metadata (from the citation) and ends with the manual retrieval of the cited resource.[29] Giving machine-actionable access to information resources is not a central part of this paradigm. The previous considerations about growth suggests that this manual "book paradigm" becomes less feasible in a digital world and the considerations about heterogeneity suggests the challenges involved in such a paradigm shift.

Scientific information retrieval as a machine-actionable task presupposes a service component of a research data product: information retrieval in the scientific context boils down to navigating linked research data products and retrieving the data and metadata components of these research data products. While interoperable metadata are necessary for the former task, research data services are necessary for the latter: to retrieve data and metadata without human interaction is only possible if querying and retrieval of these components is specified via a protocol, each component has a dedicated (technical) endpoint,[30] and if it is clear, under which circumstances which protocol should be used.

The main challenge for assessing research data products related to research data services is missing awareness of this dimension. This challenge is part social, part technical. The social part is not in the scope of this thesis, but nevertheless an important issue.[31] The technical part is to uniformly specify how a machine can query and access research data products.

[29] Clicking on a link is considered to be a manual retrieval as well, as it is often not machine-actionable (the clicking might be machine-actionable, but the identification of the right link normally is not.)

[30] Strictly speaking, not each component needs a dedicated endpoint, but each content presentation of it, e.g. metadata can be presented in different formats.

[31] A central question would be whether a person, team, or institution is responsible for the quality, impact, and/or relevance of a research data product.

In general there is no need for additional protocols or standards to access research data and their metadata, but *uniform* orchestration of the available options, among them OAI-PMH or ResourceSync for metadata and HTTP or FTP for data.[32] The challenge of missing conventions to uniformly access data is greater compared to the same challenge for metadata; OAI-PMH[33] captures the semantics of retrieving all or some of the metadata of a repository in a machine-actionable manner and is in wide use, despite its conceptual and technical shortcomings [Bor14b].

1.3.4 Research Data Product as the Object of Investigation

Above, 3 challenges for assessments of research data products have been presented and discussed: growth, heterogeneity, and missing conventions. These challenges make assessments of research data products a mission with a moving target in a shifting landscape. While previous considerations were theoretical in nature and could therefore address these challenges individually, they are often met all at once in practical tasks of research data management [Gru+17]. One way to mirror this experience for our RQ, is to bundle all relevant aspects into a single object of investigation: the research data product. All three components of a research data product affect the mechanisms to control and to assure quality, impact, and relevance of digitized research output. Discussing them in isolation would be an unwarranted simplification.

The idea of a research data product is found in the literature, but under different names, e.g. "research object"[34] or "digital object" [Cla+19]. The usage of "product" instead of "object" in our approach stresses the service component of a research data product: a product needs a distribution channel.

Additionally, the term "product" underlines the different roles involved in the creation, curation and eventual deprovision of research data products. The perspective of a researcher as the originator and primary user of research data is typically central to descriptions of research data management, although there are more than one role responsible for the quality, relevance, and impact of a research data product (see e.g. [Pen+16] for different "stewards" in the context of research data management) and many sources for research data product do not have their origin in scientific activities in the narrow sense (e.g. statistical data of demographic agencies or measurement and simulation data provided by weather and

[32] At the 13th Research Data Alliance plenary in Philadelphia an international working group to tackle this challenge has been founded: `https://www.rd-alliance.org/repository-interfaces-data-analytics-rida`

[33] `http://www.openarchives.org/OAI/openarchivesprotocol.html`

[34] `http://dgarijo.github.io/ResearchObjects/`

forecasting services).

An important differentiation is our last argument to conceptualize problems and solutions with a composite rather than with its constituent parts: the separation between the characteristics and features of a research data product on the one hand and machine-actionable tasks (such as assessment of a research data product) on the other hand. Complex tasks, such as the research of available resources, the integration of different data sources to reproducibly answer a specific scientific question, or the assessment of a research data product, have one commonality: they are successfully realized semi-automatically or entirely manually by human actors and the involved procedures are customized to the use case at hand (e.g. for a research a specific search engine can be used and only titles of research data products are of interest). Automation on the other hand is only feasible, when a concise model allows to generalize over single instances of interactions. While successful interactions of humans with a research data product are facilitated by the human capacity to sub-consciously bridge conceptual gaps (e.g. a title can occur more than once and a PID is probably unique) machines need a built-in model of the entities involved, their capacities, dependencies, and limitations. This built-in model has to cover all eventualities. The concept of a research data product fulfills this "holistic" requirement.

1.4 Outline of the Thesis

The outline of the thesis is described in Table 1.2. The table follows the order of the subsequent chapters, each row contains a reference to a chapter and a short summary of its content. It allows to follow the scientific narrative of this thesis from a bird's eye view.

Table 1.2: Outline of thesis

Chapter	*Title*	Content
Chapter 1	*Introduction*	Introduction of the necessary background to state and motivate the research question
Chapter 2	*Methodological Approach and Related Work*	Presentation of the methodology to answer the research question in the context of related and preliminary work
Chapter 3	*Event-Based Metrics*	Classification of event-based metrics and discussion of their common shortcomings
Chapter 4	*An Architecture for Benchmarks for Research Data Products*	Discussion of the concept of benchmarks for research data products along the common features of such benchmarks and their description in the form of a software architecture
Chapter 5	*Implementing Benchmarks for Research Data Products*	Presentation of a recipe to realize the architecture for benchmarks for research data products and the prototypical implementation of a benchmark for a research data product
Chapter 6	*Evaluation based on the Prototypical Benchmark*	Empirical evaluation of the prototype in comparison with event-based metrics
Chapter 7	*Conclusions and Future Work*	Summary, discussion, and suggestions for further steps

Chapter 2

Methodological Approach and Related Work

Contents

This chapter introduces the methodology applied throughout this thesis to answer RQ and its sub-questions (SQ1 to SQ5). These considerations are embedded in the context of related and preliminary work. Section 2.1 presents the systematic review to identify and examine shortcomings of existing event-based metrics. How to design benchmarks for research data products is the subject of Section 2.2. Section 2.3 discusses the empirical evaluation of the relation between event-based metrics and benchmarks for research data products. These first three sections include considerations of both methodological issues and related work. In Section 2.4, the author's own preliminary work is presented and set into relation to the content of this thesis along with an assessment of the author's contributions to these publications.

The first three sections of this chapter discuss methodological issues which occurr in the course of answering one or two sub-questions of the research question. The order of these sections is aligned with both the order of the sub-questions and the order of later chapters of this thesis as indicated by Table 2.1. This table shows a row for each of the following three sections and maps each section to the sub-questions of the research question, the contributions achieved by answering the question, and the chapters of this thesis, where said contributions are presented. The abbreviations for sub-questions and main contributions are defined in Section 1.2. The table is meant to ease the navigation through this chapter and to give an overview of the relations of its content to other chapters.

Table 2.1: Mapping of sections of Chapter 2 to sub-questions of the research question (see Section 1.2), main contributions and chapters of this thesis

Section	**Sub-question(s)**	**Main Contribution(s)**	**Chapter(s)**
Section 2.1	SQ-1	MC-1	Chapter 3
Section 2.2	SQ-2 & SQ-3	MC-2 & MC-3	Chapter 4 & 5
Section 2.3	SQ-4 & SQ-5	MC-4 & MC-5	Chapter 6

The following scheme is applied to the sections 2.1 to 2.3:

(1) What is the *desired outcome* produced by answering the sub-question(s)?

(2) What is our *methodological approach* to produce this outcome?

(3) How is the outcome and the approach linked to *related work*?

Step (1) includes a discussion, what quality criteria must be fulfilled by a method producing the desired outcome. The proposed method in Step (2) will be evaluated in Chapter 7. The following General Quality Criteria (GQC) are derived from the considerations in Section 1.3 and can be applied to all sub-questions and approaches, respectively. Step (3) includes an evaluation of existing work to assess the contribution of this thesis along these GQC and quality criteria specific for the sub-question(s) at hand:

GQC-1 Does the approach concern *research* data as opposed to data in general?

GQC-2 Does the approach concern publications, code, other data, or a combination thereof?

2.1 Categorize Event-Based Metrics and their Shortcomings

Desired Outcome

This section presents our approach to answer the first sub-question (SQ-1) of the research question: What are common shortcomings of event-based metrics in use? The desired outcome in answering this sub-question is the systematic discussion of shortcomings of event-based metrics (MC-1). This objective can be achieved in three steps:

1. A *classificatory scheme* to describe and structure different event-based metrics
2. The *systematic identification of shortcomings* of event-based metrics
3. The *discussion* which shortcomings cannot be technically mitigated

This outcome is necessary to motivate benchmarks for research data products: their design should tackle those shortcomings, that cannot be met by event-based metrics. Point 3 of the outcome thus becomes input for the architectural design for benchmarks for research data products (see Section 2.2 and Chapter 4).

Methodological approaches producing the desired outcome should fulfill the following, additional Shortcomings-Discussion Quality Criteria (SQC):

SQC-1 They discuss all types of event-based metrics, not only one group.

SQC-2 The set of shortcomings can be reproduced and is systematically extendable.

These quality criteria are stipulated by the introduction of the concept of event-based metrics and justified by an analysis of related work (see below). Examples for groups of event-based metrics are citation-based metrics, social media metrics or usage metrics (see Section 3.1).

Methodological Approach

The methodological approach chosen to achieve the desired outcome consists of a review of available literature and the discussion of intermediate results with experts.
The review of available literature is carried out along these steps:

1. A set of 21 publications is compiled, based on prior research. With this *base set*,

9 types of re-occurring shortcomings of event-based metrics are identified.

2. The base set is extended by a literature research using the following search engines: The Bielefeld Academic Search Engine (BASE), a general-purpose scientific from more than 7000 sources,[1] and google scholar[2], another general purpose search engine in wide use which has been described as the largest academic search engine with an estimation of 389 million records [Gus19].

   The search term used to query the search engines above is a Boolean combination of the following classes, which have been compiled based on the titles of the base set of publications:

   - A class of terms for **research data products** and their context:
     academia, academic, article, code, data, humanities, paper, publication, research, scholar, science, scientific
   - A class of terms indicating a **critical discussion**:
     assess, discussion, evaluation, evaluate, examine, meaning, shortcoming
   - A class of terms for **event-based metrics**:
     altmetrics, citation, download, impact, metric, ranking, usage, view

   The three topics (in bold above) were present in almost all titles of the base set. The final search term consists of a Boolean conjunction of each class, which itself is represented as a disjunction of all members of the class, i.e. from all three classes, at least one term must match.

   The search excludes publications which are not in English or which are too old to realistically discuss modern approaches: all publications prior to 2000 have been excluded. The first 250 search results ranked by relevance have been evaluated; after scanning the titles of these 250 search results, no new additions to the corpus were found, so this limit was set to keep the results relevant.

   This results in an *extended corpus of 62 publications* (41 more compared to the base set).

3. The publications have been checked for the following, further exclusion criteria:

[1] `https://base-search.net`
[2] `https://scholar.google.com`

- The publication is excluded if it does not discuss shortcomings of event-based metrics.
- The publication is excluded if it discusses mainly metrics applied to journals, institutions or researchers and not research data products.

> 21 of the publications have been excluded based on these criteria, *the extended and cleaned corpus therefore consisted of 41 papers after this step.*

4. Publications are furthermore added by the "snowball method": follow a reference of an already selected publication if the referenced publication promised to add previously not discussed shortcomings, or such shortcomings with only a few references (less than 2).

> Ten publications have been added by this method. *the final corpus of selected literature therefore consists of 51 publications.*

5. The corpus of selected literature is analyzed and passages describing the shortcomings of event-based metrics are extracted. Only those shortcomings are considered, that can be generalized to all event-based metrics, excluding for example those shortcomings that only apply to social media metrics.

> The initial model of 9 shortcomings of event-based metrics has been extended by a tenth shortcoming (coverage, COV) during the analysis of the final corpus.

6. In a conclusive analysis, those shortcomings are identified, that cannot be mitigated technically *in principle*.

The list of publications and a table with the extracted passages and the mapping to the shortcomings is published along this thesis (see Section I, in the appendix).

Intermediate results of the outcome have been discussed on two occasions with international experts:

- 13th Research Data Alliance (RDA) plenary in Philadelphia, USA (April 2019)[3]

[3] `https://www.rd-alliance.org/wg-data-usage-metrics-rda-13th-plenary-meeting`

- 14th RDA plenary in Helsinki, Finland (October 2019)[4]

Feedback from these events has been considered while the final presentation of shortcomings of event-based metrics has been compiled in Chapter 3.

Related Work

In these paragraphs below, four publications are discussed which produced an outcome comparable to the desired outcome described above. After this discussion, some work is presented that proposes classificatory schemes for event-based metrics.

[**Gru14**] is a commentary (SQC-2) on shortcomings of metrics in the context of assessing the impact of research (GQC-1), focusing on publications (GQC-2). It is not limited to a certain type of metric (SQC-1) and is therefore in parts broader, than the considerations of Chapter 3. Beside the general skepticism regarding any quantitative approach to assess research, the paper includes a detailed discussion of citation-based metrics (a subset of event-based metrics).

[**Cro14**] shares this skeptical stance concerning (alt)metrics (SQC-1) as an evaluative tool for scholarly output (GQC-1), focusing on publications (GQC-2). In contrast to [Gru14], this work points out possible beneficial use cases (faster results compared to citations and measuring societal relevance). It is also an opinion piece without a dedicated section which explains the methodological approach (which would allow its reproduction (SQC-2)).

[**Bor14a**] discusses advantages and disadvantages of altmetrics (SQC-1) in the context of assessment of publications (GQC-1, GQC-2) and identifies commercialization, data quality, missing evidence, and manipulation as major shortcomings. How this list is methodically compiled is not explained in the paper (SQC-2), but each claimed shortcoming is supported by a reference (it is also part of the final corpus used in this thesis).

[**Hau16**] discusses heterogeneity, data quality and dependencies as challenges for altmetrics (SQC-1). The context of this publication is scholarly communication (GQC-1), the authors include all type of research data into their considerations (GQC-2). The approach to collect and cluster the challenges seems to be based on prior experience, which is our interpretation — the approach to compile the above-named challenges is not discussed in-depth (SQC-2). The authors furthermore differentiate between bibliometrics (as the application of citation-based metrics) and altmetrics (other event-based metrics); this differentiation is not as fundamental for this thesis as for [Hau16]. This publication names eight shortcomings, but not the entire set of shortcomings found by us; compared to the other publications it was the most extensive.

[4] `https://vimeo.com/367997861`

An overview of how the four discussed publications fulfill the quality criteria is given in Table 2.2. Each row stands for one of the above discussed publications; the columns indicate whether they fulfill the quality criteria (checkmark) or not (cross).

None of the list of shortcomings compiled by the above publications allows to be systematically enhanced (SQC-2, only crosses in the last column of the table); this is the main reason why Chapter 3 fills a gap in the literature. In the course of the review ten types of shortcomings of event-based metrics were identified and none of the reviewed publications discusses all of them; this gap is also filled in Chapter 3. The approach proposed in this thesis is evaluated in Chapter 7 using the same criteria.

Table 2.2: Evaluation of related work for SQ-1

Related Work	GQC-1	GQC-2	SQC-1	SQC-2
[Gru14]	✓	×	✓	×
[Cro14]	✓	×	×	×
[Bor14a]	✓	×	×	×
[Hau16]	✓	✓	×	×

[LF13] and [HBC16] propose a taxonomy for event-based metrics or a subset thereof. Their focus lies on the application of the metrics, not their (technical) origin. The approach chosen in Chapter 3 focuses on the technical creation of the scores of research data products and therefore leads to a similar, but slightly different taxonomy. Where possible, the proposed nomenclature of both publications have been used.

2.2 Design Benchmarks for Research Data Products

In the previous section the desired outcome of SQ-1, the methodological approach to achieve this outcome, and related work is discussed; this chapter follows the same three-fold structure for SQ-2 and SQ-3.

Desired Outcome

This section discusses approaches to answer the second and third sub-questions of the research question: What architecture for benchmarks is induced by the shortcomings of event-based metrics (SQ-2)? How can an architecture for benchmarks of research data products be instantiated (SQ-3)? The desired outcome is a common architecture for these benchmarks (MC-2), a recipe to implement

a benchmark, and a prototype as a proof-of-concept (MC-3). This outcome is achieved by following these steps:

1. Compiling a list of *features* a benchmark of research data products must have.
2. Designing an *architecture* for benchmarks based on these requirements.
3. A *recipe* how to create a benchmark based on this architecture.
4. A *prototype* to proof the viability of the design and the recipe.

These steps are necessary to show how benchmarks for research data products can be realized *in principle* and how a specific benchmark can be created for empirical evaluation.

Approaches producing this outcome should fulfill the following, additional Quality Criteria (BQC):

BQC-1 Do the requirements mirror the concept of machine-actionability?

BQC-2 Is the design flexible enough to support different assessment frameworks?

BQC-3 Does the design include all components of a research data product?

BQC-4 Has a prototype been implemented and is its source code available?

BQC-5 Is the prototype evaluated against shortcomings of event-based metrics?

These quality criteria are either derived from considerations about research data products (see Section 1.3), adapted from existing work (see next subsection), or a consequence of the discussion of shortcomings of event-based metrics (see Chapter 3).

While the bigger part of the criteria (BQC-1, BQC-3, BQC-4, BQC-5) should be straightforward, BQC-2 needs additional context: A prominent example for a assessment framework, as phrased in BQC-2, are the FAIR guiding principles for scientific data management and stewardship (Findability, Accessibility, Interoperability, and Reusability, [Wil+16], again discussed in [Mon+17]). In general, an assessment framework stipulates or even defines what characteristics of a research data product are considered essential for its quality, impact, or relevance. If the design of a benchmark is too heavily tight to one assessment framework, a change of it might break its design. Benchmarks for research data products should support such frameworks, but the design should separate the chosen framework from the technical aspects of benchmarking, thus making the chosen assessment framework explicit, but interchangeable.

Methodological Approach

The methodological approach chosen in this thesis to achieve the desired outcome is aligned with the following scheme:

1. A requirement analysis based on Chapter 1, Chapter 3, and the literature.
2. Presentation of a resulting architecture based on these requirements.
3. A recipe to implement this architecture.
4. A prototypical implementation following the recipe.
5. Discussion of the prototype and publication of its source code.

Related Work

Three related projects/publications are discussed in the following, all of which produced an outcome similar to the desired outcome described above.

The authors of [Wil+18b][5] introduce both a **framework for measurable FAIRness** (BQC-2) of meta(data) (BQC-3) in the context of research data (GQC-1, GQC-2) and tools for semi-automatic assessment; the requirement analysis has been executed through seven interviews (BQC-1). Since the FAIR principles do explicitly target data and metadata components of a research data product, but the service component only in an implicit way, BQC-3 is only partially fulfilled. The framework allows to provide additional, possibly community-specific metrics. Currently 14 examples for such metrics are described in [Wil+18a] and the source code is available (BQC-4).[6] The approach is not evaluated against shortcomings of event-based metrics (BQC-5) since this was not the focus of the authors.

The project **MetaData Improvements and Guidance - MetaDIG** [Hab19] also suggests a benchmarking system. Unfortunately, there we could not find a requirement analysis (BQC-1), nor an in-depth description of the architecture of the benchmark. Some hints can be obtained from the source code, which is openly available (BQC-4).[7] The MetaDig project targets solely the FAIRness (BQC-2) of metadata (BQC-3) of research data (GQC-1), also called research objects (to stress the variability of data formats (GQC-2)). There is currently no evaluation available which assesses the approach against the shortcomings of event-based metrics (BQC-5).

[5] Again discussed in [Wil+19]

[6] `https://github.com/FAIRMetrics/Metrics`

[7] `https://github.com/NCEAS/metadig`

[Cla+19] presents the work of the **FAIRshake** project, which also targets the FAIR principles as a framework for assessments of research data products (GQC-1, BQC-2), The quality of a research data product is explicitly separated from its FAIRness, but the projects and its deliverables concentrate only on the latter. The proposed solution is not only a methodology to assess FAIRness of "digital resources" (GQC-2), with a prototype, but also web services, colored insignia indicating the score of the assessment and technical documentation. It was developed to meet the requirements of the biomedical research community (BQC-2). The software allows to manually or automatically (BQC-1) execute the assessment of data and metadata (BQC-3) and is openly available (BQC-4).[8] There was no evaluation against the shortcomings of event-based metrics (BQC-5).

An overview of how the discussed related work fulfills the defined criteria is given in Table 2.3. The rows represent the three discussed approaches, the columns the general quality criteria and the quality criteria for benchmarks described above. A checkmark depicts the fulfillment of a criterion, a cross its opposite. The tilde stands for partial fulfillment.

All three approaches share the characteristic, that their target audience are mainly repository providers and creators of research data products; the scientometric perspective is not their focus, hence the concentration on the FAIR guiding principles. It is therefore not surprising that none of the related publications evaluate their benchmarks against considerations of shortcomings of event-based metrics. But the main gap identified in the available systems is another one: A design that is not solely targeted at the fulfillment of the FAIR principles. Those two conceptual gaps are bridged by this thesis. The approach of this thesis will be evaluated in Chapter 7 using the same criteria.

Table 2.3: Evaluation of related work for SQ-2 and SQ-3

Related Work	GQC-1	GQC-2	BQC-1	BQC-2	BQC-3	BQC-4	BQC-5
[Wil+18b]	✓	✓	✓	×	∼	✓	×
[Hab19]	✓	✓	∼	×	×	✓	×
[Cla+19]	✓	✓	✓	×	∼	✓	×

2.3 Evaluate Correlation and Complementarity

In the previous section the desired outcome of SQ-2 and SQ-3, the methodological approach to achieve this outcome, and related work is discussed; this chapter

[8] `https://github.com/MaayanLab/FAIRshake`

follows the same three-fold structure for SQ-4 and SQ-5.

Desired Outcome

This section discusses approaches to answer the fourth and fifth sub-questions of the research question: Do the scores of a prototypical benchmark correlate with event-based metrics (SQ-4)? Can benchmark for a research data product complement event-based metrics (SQ-5)? The desired outcome is evidence for the (non-)existence of a correlation between benchmarks for research data products and event-based metrics (MC-4) and for the complementarity of the two (MC-5). The desired outcome of this chapter can be split into three components:

- A *sample* drawn from a collection of research data products.
- An *analysis of the correlation* between the scores of research data products for event-based metrics and benchmarks for research data products (using the sample).
- *Indicators for the complementarity* of benchmarks for research data products to event-based metrics (in the sample), that is, whether they can mitigate some or all of the shortcomings identified in Chapter 3.

If evidence for a weak correlation and complementarity between event-based metrics and benchmarks for research data products is available, further research is motivated, e.g. how benchmarks for research data products can be used to systematically mitigate shortcomings of event-based metrics in other contexts as the context of the prototype and the sample. Approaches producing this desired outcome concerning this evaluation should fulfill the following, additional Evaluation Quality Criteria (EQC):

EQC-1 The sample is drawn reproducibly from a large collection.

EQC-2 The sample is statistically described to manage the effect of hidden variables.

EQC-3 The correlation should be measured by a statistic that does not assume a linear transformation between the spaces of the compared metrics.

EQC-4 The indicators of complementarity are rooted in one or several of the shortcomings described in chapter 3.

These quality criteria have been compiled based on discussions of the literature (see above).

Methodological Approach

The methodological approach to achieve the desired outcome follows this scheme:

1. A sample of research data products is drawn at-large out of a snapshot of the Zenodo repository (February 2020) comprised of more than 1.5 million research data products. "At-large" means that all research data products are added to the sample, that fulfill the following conditions:

   - They have a DOI assigned by zenodo.
   - They are de-duplicated.
   - The prototypical benchmark presented in Section 5.3 can run successfully on them.

   The sample is comprised of 795,363 research data products

   More details to the sampling at-large is given in Subsection 6.1.2. The sampling at-large has been chosen to demonstrate the scalability of benchmark performance with the number of research data products. It furthermore allows to apply descriptive methods of statistics instead of using inferential methods.

2. Event-based metrics are retrieved:

   - *Usage metrics*, that is counts of views and downloads, are part of the snapshot. They correspond to the state when the snapshot was made (March 2020).
   - *Social media metrics* as provided by altmetrics:[9]
     - Tweeters: number of twitter accounts that have tweeted about the research data product.
     - Readers: number of accounts in citeulike, Mendeley, or Connotea that marked the research data product as read.
     - Facebook Walls: number of pages that have shared information about the research data product on Facebook.
     - Feeds: number of blogs that have mentioned the research data product.
     - Posts: number of online documents with one or more links or mentions to the research data product

[9] `https://www.altmetric.com`

– Altmetric Score: the altmetric score is an aggregate value calculated by all social media metrics available to altmetric. While the altmetric score includes all previous social-media metrics it is still not redundant, since it furthermore includes fields that were not exploited for the evaluation, such as videos, reddit threads, or news sources[10] and has the highest coverage.

The score, i.e. the tweeters count, readers count, etc. correspond to the state when the altmetric API was called (April and June 2020)[11]

3. The sample and the population is described statistically with the following variables:

   - type of resource (such as publication, image, data set, software etc.), and
   - year of publication
   - field of study[12]
   - scores of event-based metrics: usage metrics (views and downloads) and social media metrics (see above).[13] These types of event-based metrics are introduced in Chapter 3.

4. The Spearman rank correlation coefficient is calculated for all combinations of scores from the retrieved event-based metrics and those obtained by the prototype, respectively.[14] This coefficient quantifies the correlation of ranks and is defined as the covariance of the ranks of the two variables divided through the multiplication of both standard deviations of the variables. Its value range is $[-1, 1]$, with -1 denoting perfect negative correlation, 0 no correlation and 1 perfect positive correlation.

   Since only the ranks of the variables are used, the comparison is only ordinal (not metric), i.e. the distances of the values of the variables are not used, only the order induced by them. There is no proof, nor a reasonable argument for the claim that there is a linear transformation between the metric

[10] These have been excluded since they could be retrieved for but a small number of Zenodo depositions.

[11] Social media metrics for 19 research data products were retrieved in June, since they were missing in the original set, the rest was retrieved in April.

[12] The model and our approach assume that a research data product can belong to more than one discipline of research.

[13] Citation-based metrics currently have too low of a coverage of research data products deposited in Zenodo (see e.g. [Pet+17]) which does not allow to meaningful compare them to the scores of benchmarks for research data products; they are therefore excluded from the evaluation.

[14] The Spearman rank correlation coefficient is a standard measure for correlation of variables [Spe04], in fact "one of the oldest statistics based on ranks" [Zar05].

spaces of event-based metrics and benchmarks; the spearman rank correlation coefficient is thus the best fit for a correlation quantification, since the comparison of ranks only uses the induced order, not the distances between the scores. Another correlation measure is Kendall's correlation [Ken38] which also solely uses ranks; it is computationally more complex without a corresponding gain in information and therefore not used in this thesis.

5. The list of shortcomings of event-based metrics which cannot be mitigated *in principle* is taken as a basis to evaluate the complementarity of event-based metrics and benchmarks for research data products. This is either done by *a priori* reasoning or by a quantitative analysis of the distributions of scores of event-based metrics and benchmarks. There are two edge cases of complementarity, when we take benchmarks for research data products to be to a corrective factor of event-based metrics:[15]

   - Low scores in event-based metrics and high scores in benchmarks for research data products. In this case the score of the benchmark can be taken to be an indicator for a *"sleeping beauty"* [Raa04], i.e. a research data product that may not have received the attention it deserved due to its quality, impact, or relevance.[16]
   - High scores in event-based metrics and low scores in benchmarks for research data products. In this case the score of the benchmark can be taken to be an indicator for a *"bad example"*, i.e. a research data product that had received more attention than can be justified by its quality, impact, or relevance.

Related Work

No publications were found that discuss the comparison between event-based metrics and benchmarks for research data products. This can be explained by the rather recent upcoming of benchmarks for research data products. Chapter 6 closes this gap.

Unlike the previous section, the related work of this section will hence *not* be clustered along their fulfillment of requirements EQC-1 to EQC-4 (these quality criteria will be used in Chapter 7). Instead, this subsection is organized as follows: After an overview, we will shortly present papers who either are exemplary for

[15] These two edge cases are similar to Case II and Case III of [Hau+14b], which makes a similar complementarity analysis between twitter coverage and citations of papers.

[16] There might, of course, be other interpretations, such as the benchmark's result not corresponding to quality, impact, or relevance in a specific instance; but it still seems more probable to find "sleeping beauties" in the set of low event-based scores and high benchmark scores than in randomized samples of the whole population.

their sample technique, their selection of correlation indicators, or their discussion of complementarity.

Overview: Related work discussed in the following consists mainly of papers comparing different types of event-based metrics. There are two approaches to explain the weak or missing correlation of different types of event-based metrics [Sug+17]: On the one hand is the assumption that different types of event-based metrics indicate different features of research data products such as quality, impact, or relevance of a research data product ([LF13], [HBC16]). On the other hand, event-based metrics can also be taken to be receivers of *one* signal (i.e. one feature) and the imperfect correlation between their scores can be explained by their ability to filter out different types of noise ([AR13], [CZW15], [HCL15]). According to this view, event-based metrics (and benchmarks for research data products) are imperfect measurement devices, that could complement each other in the task to assess research data products. A very similar conclusion in the context of journals is drawn in [Hau12] and a blueprint for general assessements is sketched in [MH15]. This view is also the underlying view of this thesis and Chapter 6 will provide evidence for such a perspective.

Sampling Methods: In [TJR14] and [Pet+16], the Thomson Reuters Data Citation Index (DCI) is used as a data source for data citations. Although it offers a single point of entry, we did not use this service: it is part of the service "Web of Science" which is not open for researchers without a paid subscription; this hinders reproducibility. Open alternatives (e.g. the event-API of DataCite) have too low of a coverage and are only in an experiential state as of the time of writing. Citation-based metrics are therefore excluded in the evaluation.

Correlation Measures: An overview of comparisons of event-based metrics is given in [Sug+17] (paragraph "Social Media Metrics"). [Pet+17] uses the Pearson correlation coefficient to state a correlation between event-based metrics based on Twitter and Mendeley, respectively. This coefficient is not used in this thesis, since it presupposes linear transformability between the different metric spaces. The Spearman rank correlation coefficient is also used in the literature ([Kra+15a], [HCL15]) and is only based on the order induced by the metric spaces (not the distances). It is therefore a good choice to answer SQ-4.

Indicators of Complementarity: [Ke+15] proposes a method to identify "sleeping beauties" without the need of specifying arbitrary parameters, such as the necessary time for a research data product to be discovered. Unfortunately, this method works only after the fact, i.e. it only enables the identification of sleeping

beauties after a period of time has passed and enough citation data are available. The proposed indicators in Chapter 6 are not tied to this time-dependence problem.

2.4 Own Preliminary Work

In this section, the author's own publications and their relation to this thesis are presented. This section also includes software and other data published. The two main objectives are to indicate which parts of this thesis are directly or indirectly based on previous work, and to identify the author's contributions to this prior work in comparison with co-authors.

Publications

In the following all peer-reviewed publications until today of the author of this thesis are listed in order of date of publication. Each publication is briefly summarized by its abstract, copied from its original publication. A short paragraph after the abstract comments on the relation of the publication with this thesis. The author's contribution is assessed on par with the Contributor Roles Taxonomy (CRediT).[17] This standard is designed to make contributions of authors transparent and comparable beyond the writing, revision or editing of manuscripts [Hol19]. The taxonomy includes 14 roles that are typically contributing to a publication, and short descriptions of these roles to foster uniform and consistent application of the taxonomy.

[17] `https://casrai.org/credit`

Challenges in Creating a Sustainable Generic Research Data Infrastructure ([Gru+17])

Authors: Richard Grunzke, Tobias Adolph, Christoph Biardzki, Arndt Bode, Timo Borst, Hans-Joachim Bungartz, Anja Busch, Anton Frank, Christian Grimm, Wilhelm Hasselbring, Anastasia Kazakova, Atif Latif, Fidan Limani, Mathis Neumann, Nelson Tavares de Sousa, Jakob Tendel, Ingo Thomsen, Klaus Tochtermann, Ralph Müller-Pfefferkorn, and Wolfgang E. Nagel

Abstract: Research data management is of the utmost importance in a world where research data is created with an ever increasing amount and rate and with a high variety across all scientific disciplines. This paper especially discusses software engineering challenges stemming from creating a long-living software system. It aims at providing a reference implementation for a federated research data infrastructure including inter-connected individual repositories for communities and an overarching search based on metadata. The challenges involve a high variety of evolving requirements, the management and development of the distributed and federated infrastructure that are based on existing components, the piloting within the use cases, the efficient training of users, and how to enable the future sustainable operation.

Relation to thesis: The ideas of this paper influenced the description of challenges in Section 1.3 and are remotely connected to the discussion of features in Chapter 4.

Author's contribution: Conceptualization, Writing - original draft (this publication was published under the birth name of the author: Adolph)

How FAIR Can you Get? Image Retrieval as a Use Case to Calculate FAIR Metrics ([WK18])

Authors: Tobias Weber and Dieter Kranzlmüller

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.

Relation to thesis: This paper is a conceptual predecessor of the architecture presented in Chapter 4.

Author's contribution: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing

Addressing knowledge and know-how biases in the environmental sciences with modern data and compute services ([Wei+18])

Authors: Jens Weismüller, Stephan Hachinger, Hai Nguyen, and Tobias Weber

Abstract: In their daily work, environmental scientists typically need to access data from a variety of sources, analyze and process them with different tools, and model the data using heterogeneous IT systems. Gathering all the necessary knowledge and executing the corresponding workflows repeatedly consumes a lot of the researcher's time, which leads to a problem we call the "knowledge and know-how bias": Scientists will generally prefer data from sources they are familiar with, and focus on computational methods and tools they know. This undesirable situation can be improved by services that help them with their core workflows in data-driven research. Typically, these have to focus on creating interfaces between heterogeneous data sources and heterogeneous computational tools and methods. We believe that optimizing scientific workflows – which in the environmental sciences typically involve data and metadata in diverse formats, as well as a vast variety of software stacks and libraries for data analysis – should not be the primary task of a scientist, but rather a central service of modern scientific data and computing centers. A key prerequisite to efficiently use heterogeneous computational tools on a variety of data is interoperability: technical aspects such as data repositories and file formats need to be considered as well as metadata and organizational aspects. With their expertise in this area, scientific computing centers can provide scientists with specifically tailored, yet flexible solutions. With this aim in mind, we exemplarily discuss efforts to set up closer collaborations between scientists and the Leibniz Supercomputing Centre (LRZ, Garching, Germany). High-level IT services developed in such contexts should enable environmental scientists to reduce the knowledge and know-how biases in their research.

Relation to thesis: This paper includes thoughts loosely related to the challenges discussed in Section 1.3.

Author's contribution: Conceptualization, Methodology, Writing - original draft, Writing - review & editing

Designing a Generic Research Data Infrastructure Architecture with Continuous Software Engineering ([Sou+18]

Authors: Nelson Tavares de Sousa, Wilhelm Hasselbring, Tobias Weber, and Dieter Kranzlmüller

Abstract: Long-living software systems undergo a continuous development including adaptions due to altering requirements or the addition of new features. This is an even greater challenge if neither all users nor requirements are known at an initial design phase. In such a context, complex restructuring activities are much more probable, if the challenges are not taken into account from the beginning. We introduce a combination of the concepts of domain-driven design and self-contained systems to meet these challenges within the system's architecture design. We show the merits of this approach by designing an architecture for a generic research data infrastructure, a use case where the mentioned challenges can be found. Embedding this approach within continuous software engineering, allows to implement and integrate changes continuously, without neglecting other crucial properties such as maintainability and scalability.

Relation to thesis: The proposed design in this paper influenced the scalability disucssion in Chapter 4.

Author's contribution: Writing - original draft, Writing - review & editing

Methods to Evaluate Lifecycle Models for Research Data Management ([WK19])

Authors: Tobias Weber and Dieter Kranzlmüller

Abstract: Lifecycle models for research data are often abstract and simple. This comes at the danger of oversimplifying the complex concepts of research data management. The analyses of 90 different lifecycle models lead to two approaches to assess the quality of these models. While terminological issues make direct comparisons of models hard, an empirical evaluation seems possible.

Relation to thesis: The relevance of this paper for the thesis is indirect: in this publication evidence is collected to sustain the claim that the concept of a data-lifecycle is currently neither mature, nor consistent enough to be a scientific object of investigation. As a consequence, the concept of a research data product is the main object of investigation for this thesis.

Author's contribution: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing

Using supervised learning to classify metadata of research data by field of study ([Web+20])

Authors: Tobias Weber, Dieter Kranzlmüller, Michael Fromm, and Nelson Tavares de Sousa

Abstract: Many interesting use cases of research data classifiers presuppose that a research data item can be mapped to more than one field of study, but for such classification mechanisms, reproducible evaluations are lacking. This paper

closes this gap: It describes the creation of a training and evaluation set comprised of labeled metadata, evaluates several supervised classification approaches, and comments on their application in scientometric research. The metadata were retrieved from the DataCite index of research data, pre processed, and compiled into a set of 613,585 records. According to our experiments with 20 general fields of study, multi layer perceptron models perform best, followed by long short-term memory models. The models can be used in scientometric research, for example to analyze interdisciplinary trends of digital scholarly output or to characterize growth patterns of research data, stratified by field of study. Our findings allow us to estimate errors in applying the models. The best performing models and the data used for their training are available for re use.

Relation to thesis: The results of this paper influenced the discussion of challenges in Section 1.3 (especially with regard to heterogenity). The described model is furthermore used to describe the distribution of fields of study in the sample in Chapter 6.

Author's contribution: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing - original draft, Writing - review & editing

Standardizing a Standard: Why and how a Best Practice Guide for the DataCite Metadata Schema was created ([Küm+20])

Authors: Sonja Kümmet, Stephan Lücke, Julian Schulz, Martin Spenger, and Tobias Weber

Abstract: In order to promote the FAIRness of research data, the use of a widespread metadata schema is recommended to describe the data. The DataCite Metadata Schema published by the consortium of the same name has meanwhile established itself as a model used worldwide. However, the evaluation of DataCite XML files created by project managers at the IT Group Humanities of the LMU Munich and at the Leibniz Supercomputing Centre revealed the need to extend the standard. Against this background, representatives of data creators, data curators and data aggregators participated in the development of a best practice guide for DataCite in order to increase the interoperability of (meta-)data through a stronger standardization. This paper describes the development process towards the now published Best Practice Guide, discusses the reasons for its development, and presents the main features of the guide and the potential of its future application.

Relation to thesis: This publication and the described best practice guide are the sources for the creation of the prototype along the recipe presented in Chapter 5.

Author's contribution: Conceptualization, Investigation, Methodology, Writing - original draft, Writing - review & editing

Shortcomings of Usage Metrics for Research Data ([Web20])

Authors: Tobias Weber

Abstract: The FAIR assessment of research data is a hard task, both for technical and social reasons. Usage metrics are a central pillar in solving these problems. However, usage statistics are not without shortcomings, which are presented and classified in this paper; technical and organizational mitigations exist or are in development, but there are still issues demanding further research. A list of such open challenges is another contribution of this paper. Facing these challenges will contribute to open and transparent scientometric research and fair assessment of digital scholarly output.

Relation to thesis: This publication is a subset of the considerations in Section 3.2 for usage metrics. It is planned as an RDA recommendation and output of the working group "Data Usage Metrics" — and is currently in preparation.

Author's contribution: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft, Writing - review & editing

Data, Code and Technical Reports

The following data and code publications are supplementary material to articles presented in the previous subsection or grey literature (non-peer reviewed articles). They are listed here, since they are produced by the author of this thesis:

- [WK18]
  - Data Publication accompanying the paper "Methods to Evaluate Lifecycle Models for Research Data Management" ([Web18a])
  - Software Publication accompanying the paper "How FAIR can you get? Image Retrieval as a Use Case to calculate FAIR Metrics" ([Web18c])
- [WK19]
  - Data Publication accompanying the paper "Methods to Evaluate Lifecycle Models for Research Data Management" ([Web18b])
- [Web+20]
  - Source Code and Configurations for Publication "Using Supervised Learning to Classify Metadata of Research Data by Discipline of Research" ([WF19])
  - s-sized Training and Evaluation Data for Publication "Using Supervised Learning to Classify Metadata of Research Data by Discipline of Research" ([Web19d])

  - m-sized Training and Evaluation Data for Publication "Using Supervised Learning to Classify Metadata of Research Data by Discipline of Research" ([Web19b])
  - l-sized Training and Evaluation Data for Publication "Using Supervised Learning to Classify Metadata of Research Data by Discipline of Research" ([Web19a])
  - Raw Data for Publication "Using Supervised Learning to Classify Metadata of Research Data by Discipline of Research" ([Web19c])
  - Statistics and Evaluation Data for Publication "Using Supervised Learning to Classify Metadata of Research Data by Discipline of Research" ([WFS19])

- [Küm+20]
  - DataCite Best Practice Guide ([Küm+19])

Data and code published along this thesis is listed in the appendix (see Section I).

Chapter 3

Event-Based Metrics

Contents

This chapter presents a classification of event-based metrics and discusses their shortcomings. In Section 3.1 a short overview of event-based metrics is given, including examples for these metrics. The result of a literature review concerning shortcomings of event-based metrics is presented in Section 3.2. Based on these findings, principal shortcomings are identified in Section 3.3, which cannot be technically mitigated.

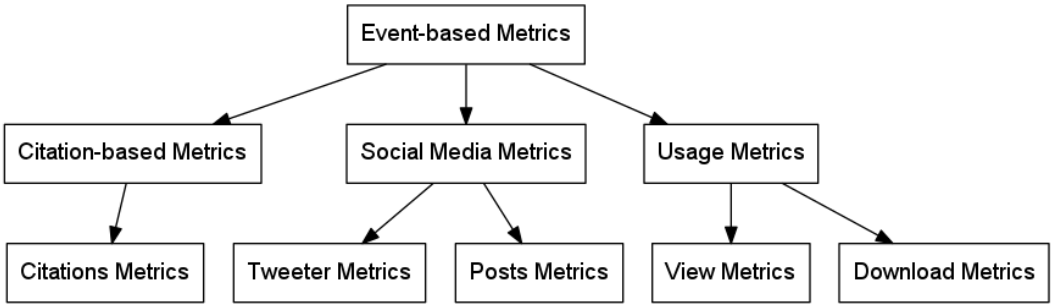

Figure 3.1: A Taxonomy for Event-based Metrics

3.1 Overview of Event-based Metrics

An event-based metric is understood to be an assessment of a research data product based on the documentation of events of interactions with the research data product. The frequency of citations, mentions, or downloads are examples for event-based metrics' scores. Although the concept "event-based metric" is to our knowledge coined by this thesis, the idea to point out and discuss similarities between citation-based metrics, social media metrics, and usage-based metrics is not new [WC12], [Ham14], [Hau16], [May+17], [Rav+17], [Sug+17].[1] All these publications propose or presuppose a collective term for a set of metrics beyond simple citation-based metrics, while none of them sets this collective into relation to another type of metric, as is done in our work with the comparison of event-based metrics with benchmarks for research data products.

We concentrate on three main types of event-based metrics:

- Citation-based Metrics (Subsection 3.1.1)
- Social Media Metrics (Subsection 3.1.2)
- Usage Metrics (Subsection 3.1.3)

Figure 3.1 shows the taxonomy induced by these three types: The figure shows a taxonomic tree and is to be read from top to bottom: The root stands for all

[1] The most similar term to event-based metrics is "scholarly metrics" in [Sug+17]: *"scholarly metrics are thus defined as indicators based on recorded events of acts (e.g., viewing, reading, saving, diffusing, mentioning, citing, reusing, modifying) related to scholarly documents (e.g., papers, books, blog posts, datasets, code) or scholarly agents (e.g., researchers, universities, funders, journals)."* This term has not been chosen for this thesis for two reasons: First, it is more extensive than event-based metrics: It also includes metrics solely based on artificially produced events (as in the context of benchmarks), while event-based metrics is derived from events which have already happened and are not produced for the sake of the assessment of a research data product. Secondly, the term "scholarly metrics" focuses on the context of the metrics (scholarly communication) and not on their origin (events).

event-based metrics, the first layer below the root depicts the three types discussed in this thesis and the last layer exemplifies the events that are captured by each metric. The figure gives an overview of the types of event-based metrics relevant to this thesis without any claim to completeness.

From a conceptual point of view, all presented event-based metrics share a common anatomy:

1. The basic building blocks are *events*, meaning real-world phenomena (i.e. not artificially created) happening at a discrete point in time in relation to a research data product. Examples for events include somebody citing a research data product or tweeting about it.

2. The occurrences of a selection of events are documented in a *log*; the log can be actually realized (by a file or a database), but it could also be just a conceptual ensemble of all events of a certain type.

3. The *score* for a given research data product is then the mapping from the log to a number, e.g. by counting how often a certain type of event occurred in the log.

4. The last layer of this scheme must be clearly separated from the other three: the *application* of the metric. The application adds an interpretation to the score, such as "many clicks mean high quality".

The event, the log, the score, and the application can be understood as layers of a scheme that abstracts from events by *documentation*, then from the log by *quantification* and finally from the score by *evaluation*. The layered scheme of event-based metrics is depicted by Figure 3.2. In the first row, screenshots give examples for events: references in an article, tweets about a code base and downloads of data sets. The second row shows conceptual logs (for citations and tweets) and a logfile recording access to a research data product. The third row shows how the logs are mapped to a score of a research data product and the last row shows an application. The direction from top to bottom stands for the stepwise abstraction from events via documentation, quantification, and evaluation.

One of the most discussed use cases for event-based metrics is their application to measure some sort of "scientific success", especially in situations in which decisions are made regarding funding or hiring. Before we discuss the three main types of event-based metrics in the remainder of the section, a word of caution is due applying to *all* quantitative models used in these contexts: Assessing individuals or deciding which research project is worth funding is always more than a mere calculation — or to rephrase a passage from the Leiden Manifesto for research metrics: evaluation should be led by *judgements*, not *data* [Hic+15]. This does

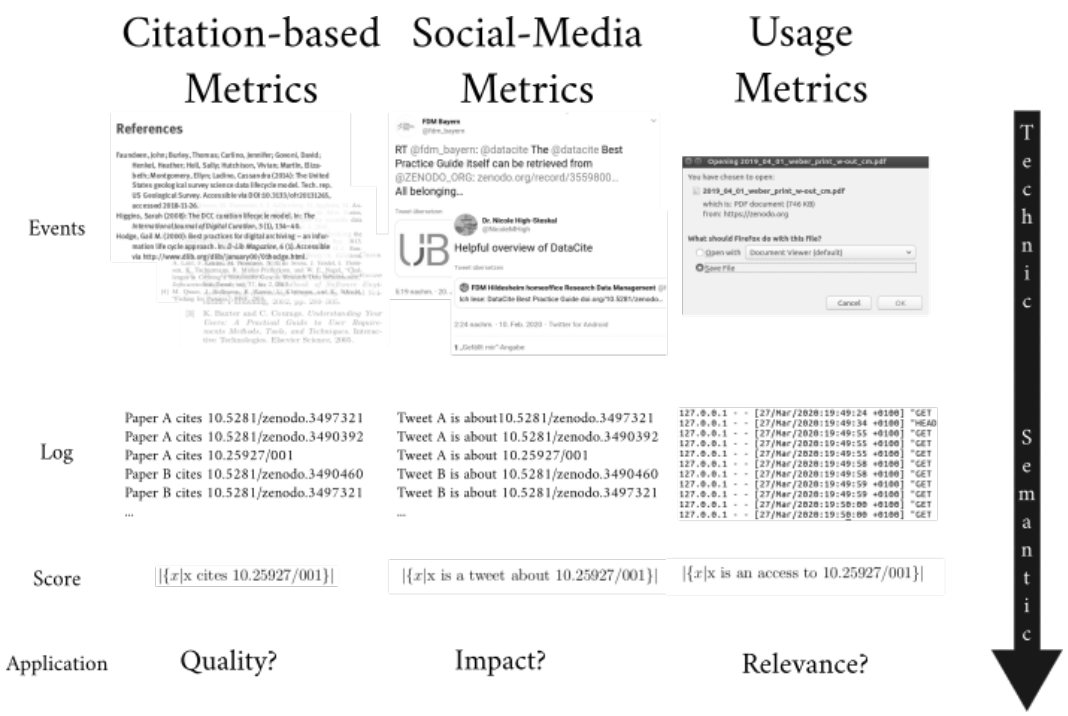

Figure 3.2: Examples of event-based metrics and their different stages

not entail that metrics cannot support decisions, but all metrics have limitations *and* are in the need of contextualization.

To our knowledge, this analysis of identical features of event-based metrics has not been carried out systematically in related work, although conceptual predecessors include:

- [Cro01] discusses the application of the same methodology used at citation analysis to resources of the "web" and focus on the "science in action", i.e. science as a system of interconnected events.
- [BI04] introduces a classification of different events, but not based on their origin but based on the fields scrutinizing the events (*metrics, including infometrics, cybermetrics and webometrics).
- [HBC16] discusses efforts to find a synthesis of altmetrics (see Subsection 3.1.2), i.e. a conceptual interpretation of their meaning and application.

The layered model can also be considered as a contribution of this thesis to scientometrics, although its objective is external to scientometrics: to describe the commonalities of the state-of-the-art way to assess research data products to identify common shortcomings that can be (partially) remedied by a new approach, i.e. by benchmarks for research data products.

3.1.1 Citation-based Metrics

Citations are references to other research data products in the data or metadata component of a research data product. A citation can also be considered as an event, since the research data product is published at a discrete point in time, which is identical with the point in time the citation comes into existence. Citation-based metrics are thus assessments of research data products based on an aggregation of citations.

Citation-based metrics are often described with economical metaphors: they are called the "gold standard" of scholarly metrics [Sug+17], a "currency" [Opp15], or "central to the incentive system" by constituting "an underlying sense of distributive justice" [Mer88]. Their importance in the assessment of publications is the reason other forms of research data strive to be part of this system of distributive justice: there are principles for "data citation"[2] [Dat14] and a roadmap for repositories which strive for technically compliance to these principles [Fen+19]; the same endeavor is discussed for citations of software [SKN16]. The meaning and limitations of citation-based metrics have been discussed extensively in the 1970s and 1980s (see [ALW19] for an overview and primary sources). *Avant la lettre*, similarities in referencing relations across different types of research data products have been described as early as 2001 (see e.g. [Cro01]).

The simplest citation-based metric is built on counting as an aggregation function: the sum of all available citations of a research data product determines the score of a research data product; but there are other metrics such as the journal impact factor [Gar55] or the h-index [Hir05]. Both are also based on counting citations. These types of metrics have been criticized, especially when a researcher is evaluated by a metric designed to evaluate journals[3] or when decisions in academia are only supported by quantitative evaluations, without any qualitative considerations (see above).[4] Since both metrics do not assess research data products, but journals and researchers respectively, they are out of the scope of this thesis; but the critique for these types of metrics influenced this work both technically and ethically: the claim that assessments of research data products have to be both transparent and adaptable is a technical design principle for the architecture proposed in Chapter 4 as well as a dimension along which our results are critically appraised.

[2] meaning research data that are not code nor publications

[3] e.g. the San Francisco Declaration on Research Assessment, initialized on the Annual Meeting of The American Society for Cell Biology (ASCB) in San Francisco in December 2012, `https://sfdora.org/read`, see also [Gam+20]

[4] e.g. The Leiden Manifesto for research metrics [Hic+15]

3.1.2 Social Media Metrics

One reaction to many of the critical discussions of citation-based metrics was the development of alternative metrics, or altmetrics for short [PGT12].[5] Social media metrics is a subset of these altmetrics; the term "social media metrics" is used in this thesis as it is described in [Sug+17]: *"[...] metrics, which have a clear focus on social functions; that is, [which are derived from] those platforms which allow users to connect and interact with each other; create and reuse content; and comment on, like, and share user-provided content."* Exploiting logs of social interactions to assess research data items (beyond the formal communication of peer-reviewed findings) promises to include societal relevance to the assessment of a research data product [BH18b].

Since the events needed to derive social media metrics are mainly happening on platforms which are run by private companies, the logs are also in private hands. APIs to services such as the microblogging service Twitter[6] or the reference manager Mendeley[7], allow for *ad hoc* analyses; but typically, social media metrics are provided by private companies as a service.[8]

As with citations, metaphors to make social media metrics understandable often involve economic imagery; In [HBC16], the authors propose three theories to better understand altmetrics, two of which bear clear economical reference: social capital and attention economics.[9] Despite the theoretical issues with the meaning of these metrics, it was disputed whether altmetrics in general and social media metrics in special could replace citation-based metrics. Several studies could only find weak ([Zuc+15], [CZW15]) or no correlation ([Pet+16]) between citation-based metrics and altmetrics. The dominant conclusion in the scientometric literature is therefore that citation-based metrics and social media metrics measure different characteristics of research data products.

3.1.3 Usage Metrics

Usage metrics are often considered a subclass of altmetrics ([Dor13],[MT14]). Metaphorically speaking, they are differentiated from social media metrics by the events on which they are based: while social media metrics are based on *conversations about* research data products, usage metrics are based on *interactions with* research data products. Any technical and discrete interaction with the research

[5] See also: `http://altmetrics.org/manifesto/`

[6] `https://www.twitter.com`

[7] `https://www.mendeley.com`

[8] The same applies to citation-based metrics, with the prominent exception of the Initiative for Open Citations (I4OC, `https://i4oc.org/`).

[9] The third theory, impression management, has its roots rather in dramaturgy.

data product is considered usage in this thesis.

Canonical examples for such interactions are "views", which are visits on the landing page of a research data product, and "downloads" which is the retrieval of the data or metadata of a research data product via its services. Usage metrics therefore differ in a crucial point from the other event-based metrics since they potentially honor all three components of a research data product.

Providers of repositories for research data products collect these interactions with a research data product on their system; metrics can be calculated by applying web analytics. The COUNTER[10] standard for publishers of articles has been introduced to make these metrics more comparable and to offer a unified approach to tackle technical problems, such as "double clicks" or denial-of-service attacks [She04]. A similar standard for data sets has also been proposed [Fen+18].

3.1.4 Other Event-Based Metrics

The list of event-based metrics presented so far is most probably not exhaustive, instead citation-based metrics, social media metrics and usage metrics are the dominant examples in the relevant literature. In the future, more event types might provide the basis for other event-based metrics: e.g. the performance of research data products on reproducibility platforms such as PopperCI [Jim+17][11] or decentralized recommendation systems such as plaudit.[12]

Since the role of these less frequently used event-based metrics is yet to be determined, the empirical evaluation in Chapter 6 exclusively takes two out of the three aforementioned types into account, but all theoretical considerations, especially the following concerning common shortcomings of event-based metrics, should apply to all other types as well.

3.2 Shortcomings of Event-based Metrics

This section summarizes the results of a systematic literature review to find and cluster common shortcomings of event-based metrics (as mentioned in Section 2.1). Some shortcomings referenced in literature only apply to one type of event-based metrics[13] or one type of research data product.[14] These were excluded from the following results, since they cannot be used to answer the question, how benchmarks

[10] `https://www.projectcounter.org`

[11] see also `https://falsifiable.us`

[12] `https://plaudit.pub`

[13] e.g. the problem how to count events in usage-based metrics (double-clicks, sessions), see [Fen+18]

[14] e.g. it is currently an unsolved problem how to aggregate event counts in the case of research data that are composites of other research data.

for research data products could contribute to mitigate the *common* shortcomings of all event-based metrics.

Ten types of shortcomings of event-based metrics have been identified. Table 3.1 shows each of these shortcomings in one row; the columns contain information about the shortcoming (from left to right): an identifier for the shortcoming, its name, a short description, the number of sources referencing the shortcoming and a reference to a table with details for that shortcoming. The tables displaying the details (Table A.1 to Table A.10) can be found in the appendix; they provide additional background, classification information and a list of the publications referencing the shortcoming. The rows are displayed in decreasing order of number of sources referencing the shortcomings.

"Data quality" is a shortcoming that has been discussed by some of the publications in the corpus, namely [ZC18], [Hau16], and [Hic+15]. It was not added to this list as an own type of shortcoming for two reasons: first, it is doubtful that publications using the term "data quality" share a common understanding of that term. Second, the term is too coarse-grained: several shortcomings presented in this chapter can be subsumed under this concept, such as COV (the log of events has low quality since it does cover too many or too few events) or TRST (the log of events includes providers of event data that are not trustworthy).

Despite the heading of this section, a quick remark about the *advantages* of event-based metrics might be warranted: they are easy and cost-efficient to implement, scrutinized by scientometric research and agnostic to any technology applied in the creation of a research data product. The empirical nature of collecting and counting events to assess the output of research might be a reason why event-based metrics are so wide-spread in current evidence-based fields of research.[15] Another advantage of event-based metrics is that they can be used for machine-actionable assessments of research data products: If it is specified which scores are considered thresholds or indicators of quality, impact, or relevance of a research data product, no human interaction is necessary to carry out an assessment.

3.3 Ways to Mitigate Shortcomings of Event-based Metrics

The shortcomings presented in the previous section are classified below to identify shortcomings which can be most informative in the creation of an architectural design of benchmarks for research data products: the requirements presented in Chapter 4 are based on these flaws of event-based metrics which cannot be easily mitigated. The objective of this section is not to discuss all aspects of mitigation

[15] A word of caution is phrased in the fifth main finding of [Wil+17]: "measure what matters" instead of measuring those features that are easy to collect.

Table 3.1: Shortcomings of event-based metrics

Identifier	Name	Description	#	Details
COV	Missing coverage	A score of an event-based metric is too low, because not all events are in the log.	29	Table A.1
COR	Doubtful correlation	Correlation with quality, impact, or relevance of a research data product is doubtful.	27	Table A.2
NORM	Normalization	The scores of two research data products are not comparable, since they must be normalized.	23	Table A.3
GAME	Gaming	The score of a research data product is too high, since it has been artificially and intentionally increased.	16	Table A.4
BAND	Sensivity to Social Effects (bandwagon)	The score of a research data product is too high (to correlate with quality, impact, relevance of a research data product) since social effects, such as the Matthews effect or manipulation, lead to skewed distributions.	13	Table A.5
TIME	Dependence on Time	The score of a research data product is only available after a certain period passed (in order for the events to occur).	9	Table A.6
TRST	Missing Trustworthiness	The scores of two research data products cannot be compared, since the one source adding to the log of the metric is trusted while the other is not.	9	Table A.7
CTXT	Missing Context	The score of a research data product cannot be used to assess the quality, impact, or relevance since necessary context is missing.	8	Table A.8
DUP	Duplication	The score of a research data product is too low since the research data product is duplicated over different service providers.	2	Table A.9
VER	Versioning	The score of a research data product is too low, since its predecessors or successors are not accounted for.	2	Table A.10

in-depth, but to give the right amount of context to justify the classification of the shortcomings of event-based metrics along the following classes:

1. *Simple shortcomings* can be mitigated easily, i.e. the technical solutions are available, they can be implemented from an organizational point of view, and their mitigation does not increase the negative impact of other shortcomings.

2. *Normal shortcomings* can be mitigated, but this mitigation comes at a cost: the technical solutions need (further) research and development, their implementation is hard from an organizational point of view, or they can only be mitigated at the cost of increasing the negative impact of other shortcomings.

3. *Principal shortcomings* cannot be mitigated at all, i.e. there is in principle no technical solution available other than complementing event-based metrics with metrics that are not based on events.

The remainder of this section is aligned with these three categories: firstly, mitigations for three simple shortcomings will briefly be sketched. Secondly, five normal shortcomings will be discussed, i.e. it will be argued why these shortcomings are *not* simple and what available technical or organizational options for mitigations exist. Lastly, it will be argued that TIME and BAND are principal shortcomings.

3.3.1 Mitigations of Simple Shortcomings

Missing Context (CTXT)

The mitigation of CTXT is straightforward: providing the missing context. This shortcoming applies to the format of the log of event, i.e. what is recorded. Candidates to add to a log can be directly read from the presentation of the shortcoming in table A.8:

- Engagement with the research data product could be provided by qualified references in the case of citations (*how* or *why* the research data product is cited), context of the social interaction on a platform (e.g. along the three categories proposed in [HBC16]: access, appraise, apply), or details about the usage (e.g. *who* used it, or to what *end*).

- Differentiating technical processing from human interaction is addressed in the COUNTER code of practice [She04], [Fen+18].[16]

[16] The proposed solution might be in the need of an overhaul, when technical processing becomes more sophisticated than today. At the moment it consists of evaluating the user agent field of an HTTP request.

Duplication (DUP)

From a technical point of view, DUP is a solved problem awaiting implementation: persistent identifiers like Digitial Object Identifiers (DOI)[17] allow to unambiguously identify digital resources such as research data products. Persistent identifiers for other resources, such as researchers (e.g. ORCID IDs[18]) or academic institutions (e.g. the Research Organization Registry (ROR)[19]) furthermore allow to link solutions of this problem to solutions of CTXT, as proposed in projects like FREYA.[20]

If every research data product is unambiguously identifiable, event-based metrics can be aggregated accordingly and the score of a research data product is not skewed by duplicates.

Versioning (VER)

As with DUP, VER is a technically solved issue, that awaits implementation: Versioning semantics has been standardized in the software community and its results can be re-used in the context of research (e.g. semantic versioning[21]). Metadata standards like DataCite allow to create machine-actionable links from one version of a research data product to another, thus documenting the version history.[22]

Usage of these machine-actionable solutions allows users of event-based metric to implement applications of these metrics that take a version history into account when appropriate or aggregate over all versions in other cases.

3.3.2 Mitigation Strategies for Normal Shortcomings

This subsection discusses normal shortcomings. While GAME and NORM are each discussed in a dedicated paragraph, the mitigation of COV and COR on the one hand, and COV and TRST on the other hand are intertwined and therefore discussed together in the last two paragraphs of this subsection.

Gaming (GAME)

"Goodheart's law" as phrased by Marilyn Strathern, says that "[w]hen a measure becomes a target, it ceases to be a good measure." [Str97] While we do not share

[17] `https://www.doi.org/`
[18] `https://orcid.org`
[19] `https://ror.org`
[20] `https://www.project-freya.eu`
[21] `https://semver.org/`
[22] see the relation types "isVersionOf", "isNewVersionOf", and "isPreviousVersionOf" in standard 4.3 [Dat19]

this stance, at least not in its radical form, that a targeted measure cannot be a suitable tool for scientometricians, it provides the motivation to constantly check incentive systems for false play.

Some authors suggest that altmetrics are easier to game than citation-based metrics [Bor14a], [23] but at the same time the incentive for a malevolent actor is much greater for citation-based metrics at the time of writing.[24] Work to mitigate the effects of self-citation [BK11] and citation cartels [FP16] has been published. Citation "theft" is also an issue, considering e.g. that review article receive more citation than the work reviewed [Gru14].

Since social media metrics and usage-based metrics are not (yet) a considerable part of the academic incentive system, detection methods for unethical *scientific* behavior on social platforms, or countermeasures thereof have not received the same attention in research as compared to gaming of citation-based metrics. Many of the countermeasures in other fields, such as the study of political discourse on social media platforms, is adaptable to the context of assessment of a research data product, e.g. the detection of social bots [Wan10].

In general, the strategy to harden event-based metrics against gaming is to constantly invest in research and technology to identify occurrences of misbehavior and find systematic countermeasures. While this policy is possible in principle, it still requires additional resources and intellectual work, which is why GAME has been classified as a normal shortcoming.

Normalization (NORM)

[BH18b] suggests applying "established normalization processes" to altmetrics. The review carried out in [Wal16] references at least two types of normalization procedures (average citation count and highly cited publications) to achieve normalization. This is evidence that there actually is *not* a commonly agreed upon or standardized way to normalize citations as indicated by the authors of [BH18b].[25] Further technical issues with proposed normalization procedures are yet unsolved, including but not limited to: missing robustness,[26] the arbitrary choice of dimen-

[23] Unsurprisingly, the altmetric manifesto claims the opposite, see [Pri+10].

[24] Academic hiring and tenure decisions are often based on citation-based metrics

[25] See [ALW19] for an overview of citation-based normalization procedures and their discussion

[26] Normalizations are often based on the mean (see e.g. [CLB10]), not the median of distributions of citations (or other types of events). This can be criticized for only softening, but not erasing the effect of skewed distributions, see [ALW19].

sions to normalize against,[27] and the open scale of the normalized scores.[28]

These technical issues need to be addressed to mitigate NORM. The further need for research and development is why NORM is classified as a normal shortcoming.

Missing Coverage vs Missing Correlation (COV/COR)

Considered in isolation, COV has a straightforward solution: Make the log recording the events as inclusive as possible. A target conflict arises, when mitigation strategies for COR are considered: If only those events are included in the log which are considered to be indicators of quality, impact, or relevance of a research data product, the question whether event-based metrics correlate with these would be solved.[29] COV and COR thus cannot be mitigated without increasing their respective negative impact: if the log is compiled in a manner that is very inclusive, quality issues might arise and vice versa, if it is handled too restrictive, relevant events might be excluded. These considerations justify the classification of COV and COR as normal shortcomings.

Missing Coverage vs Missing Trustworthiness (COV/TRST)

A similar situation to COV and COR can be described, when considering the effects of mitigations of COV and TRST: The straightforward solution to mitigate trustworthiness issues is to restrict the event log to those sources for which trust could be established (e.g. by means of certification), which directly effects the negative impact of COV. Therefore, TRST is also classified as a normal shortcoming.

3.3.3 Principal Shortcomings

This subsection finally discusses those shortcomings which should be most informative for the architectural design of benchmarks for research data products, since there is no technical mitigation.

[27] Canonical example for such a dimension is the field of study, for which a big number of schemes is available (see [Web+20] for a selection). The choice of the classification scheme has a crucial impact to the results of normalization, but is often dictated by the tools and services available, e.g. the classification of the service "Web of Science", see e.g. [Wal+11b].

[28] Even normalization against "highly-cited publications" (see [Wal16]) does not allow to define an upper bound, since there is (currently) no reasonable and robust candidate for an upper bound for the number of events.

[29] This mitigation of COR might deprive event-based metrics of their possibility to be the basis of machine-actionable assessments of research data products, since these selections are often manual.

Social Effects (BAND)

All types of events listed in Section 3.1 mirror social interaction, since the basic intuition behind event-based metrics is that the frequency with which humans interact with a research data product correlates with its quality, impact, or relevance. But human behavior is much more complicated than suggested by this intuition: *"People do behave in the same manner towards things, not because these things are identical in a physical sense, but because they have learnt to classify them as belonging to the same group, because they can put them to the same use or expect from them what to the people concerned is an equivalent effect."* [Hay73] Two consequences can be drawn from this statement, which are of importance for the usage of event-based metrics for assessments of research data products:

1. What is measured by a score of an event-based metric is heavily dependent on the "measurement device", that is the specific event-based metric: actors in research will adapt their behavior to meet the expectations.[30] As an example, if many tweets about a research data product are desirable for a researcher, his or her behavior will change, thus leading to a situation in which a scientometrician faces methodological "dilemmatics": the scientometrician cannot optimize generalizability, precision, and realistic conditions of empricial observance at the same time (see [McG81]).

2. The (social) events are not mutually independent: almost all reported distributions of citations, social interactions or usage are skewed towards high scores (see e.g. [Sug+17], [Pet+16]); research data products with higher scores are more likely to collect even higher scores.[31] This phenomenon has multiple names, among them Matthew effect [Hau16], power law [Zuc+15], or bandwagon effect.[32]

Since social norms effect event-based metrics and there is no known way to mitigate them without begging the question, BAND is classified as a principal shortcoming.

Timeliness (TIME)

While some types of event-based metrics, like social media metrics, allow to assess research data products "soon" after their publications [BH18b], [Gam+20], for others, like citations, years must pass [SD19]. Although event-based metrics differ in the length of this "time gap" between publication and assessment [WHH15],

[30] This is related to GAME, since social motivation to comply to perceived standards might in border cases be indistinguishable from attempts to game event-based metrics.

[31] [BD08] is even bold enough to describe this effect mathematically: *"[...] the expected number of future citations is a linear function of the current number [...]*

[32] `https://en.wikipedia.org/wiki/Bandwagon_effect`

they all have one characteristic in common: a research data product cannot be assessed by event-based metrics in a machine-actionable way *before* it has been published. Measurable interactions with a research data product can only occur after the world had been granted access to it. These considerations justify the classification of TIME as a principal shortcoming.

Chapter 4

An Architecture for Benchmarks for Research Data Products

Contents

This chapter discusses the concept of benchmarks for research data products in depth. Section 4.1 derives the features every benchmark must meet. The architecture derived from these features is presented in Section 4.2.

4.1 Feature Extraction

The considerations of this section do *not* constitute a requirement analysis in the traditional sense: they do not (fully) specify what a concrete implementation of a benchmark must fulfill. Instead, they specify which features must be shared by *all* possible implementations to be subsumed under the concept of a benchmark for a research data product. The discussion of the features in this section and the design presented in the next section are thus mainly an explication of the term "benchmark".

Although the following subsections are not meant to carry out requirement engineering, they include some of the activities of "requirement elicitation" and "requirement analyses" as defined in [BF+14]: identification of the sources for features and their discussion (characterization, classification, discussion of conflicts). But these steps are taken without specific values for the typical bounding parameters of a software project (project's scope and resources, customers, and other stakeholders, etc.).

The result of this section is thus a blueprint that can be completed with a full-fledged elicitation and analysis of requirements with further additions, as proposed in Section 5.1. The primary goal of this thesis is *not* an implemented system that a specific stakeholder needs for a certain purpose, but the specification of a scientific measurement device anybody can rebuild.

The features are extracted from the following sources:

1. The *general domain of this thesis*, i.e. the challenges involved in assessments of research data products — this domain and the challenges have been introduced in Section 1.3. This source is discussed in Subsection 4.1.1.

2. The methodological quality criteria for the design of benchmarks for research data products, as discussed in the context of *related work* (Section 2.2) — this source is discussed in Subsection 4.1.2.

3. *Shortcomings of event-based metrics* which are hard or impossible to mitigate with the means of event-based metrics alone — these have been identified in Section 3.2. This source is discussed in Subsection 4.1.3

4. *Best practices for benchmarks in other domains*, such as performance measurements in High-Performance Computing (HPC), or validation and verification benchmarks for computational simulations — this source is discussed in Subsection 4.1.4

These four types of sources define the structure of the following subsections, in which the derived features are discussed. Each feature is summarized in a dedicated

table since they are typically motivated by more than one source. Additionally, these tables include an identifier for each feature, to reference them unambiguously. The section closes with a discussion of conflicting features.

4.1.1 Based on the General Domain of the Thesis

There is a basic intuition why benchmarks are suited to assess research data products: the effort put into the creation and curation of a research data product correlates with its quality, impact, and relevance. If benchmarks provide a means to quantify this effort they can be used as a quantitative indicator for these three qualities.

The main feature for an architecture of benchmarks is thus that its building blocks allow to model specific interactions with research data products — namely those interactions that are informative of the effort put into their creation and curation. These building blocks form the body of the architecture, since the main features can be derived from them:

F-D1 a common and generic interface for research data products[1]

F-D2 a model for interactions with a research data product

F-D3 the mapping of the behavior of a research data product to $\mathbb{R}^+$

F-D4 the orchestration of all components of the architecture

The following paragraphs comment on each of these features.

F-D1: An Interface for Research Data Products

This feature targets two of the three challenges discussed in Section 1.3, namely heterogeneity and missing conventions:

- The interface should *abstract from the heterogeneity* if this is possible without arbitrary choices (F-D1/1).

  An example for such an abstraction is the mapping of the fields "originator" of metadata standard A and "author" of metadata standard B to a common attribute "creator" of the interface.

- If neither abstraction nor commonly accepted conventions allow to define the interface — that is, an "homogenization" is not possible — the selection of a standard or technique should be *configurable* (F-D1/2); this means that the

[1] The identifier F-D1 is to be read as "1st Feature derived from the general Domain".

necessary choices can be delayed until the context of the interaction with the research data products is given (see next paragraph).

An example for such a configuration would be the specification of a field of study of the research data product, for which a lot of different, partly inconsistent standards exist: the interface for research data products should offer the possibility to check for one or several standards, without favoring a specific standard; this includes the possibility that none of the tested standards is supported.

Table 4.1: F-D1: An Interface for Research Data Products

Details	**F-D1/1** The interface should abstract from heterogeneity
	F-D1/2 The interface should support the configuration of different technical solution in cases in which a choice for a specific technical solution is arbitrary.
	F-D1/3 The interface should support all four components of a research data product (PID, data, metadata, and services).
	F-D1/4 The interface should support the option to access the research data product through an authorization and authentication layer.
Motivation	• Heterogeneity (Section 1.3) • Missing conventions (Section 1.3) • BQC-2 (Section 2.2) • BQC-3 (Section 2.2) • TIME (Section 3.2) • CTXT (Section 3.2)

These considerations are not limited to the metadata component (as might be inferred by the examples given) — PIDs, data, and services should be handled in an analogous manner. Table 4.1 offers a tabular overview over feature F-D1, including sub-features motivated by following considerations. The table lists the components of a feature (Details) and the justifications for this feature (Motivation). It summarizes the considerations distributed all over Section 4.1 in one place and are meant to provide an overview of the displayed feature. The following ta-

bles (Table 4.2, Table 4.3, Table 4.4, Table 4.5) have the same structure and serve the same purpose for the other features.

FD-2: A Model for Interactions with a Research Data Product

Heterogeneity and missing conventions are also motivations for this feature, but in this instance the interface should allow to model best practices, thus setting them apart from the multitude of other interactions with equal or similar outcome:

- The model for interactions with research data products should allow to *encapsulate best practices* (F-D2/1). If a research data product is compliant to best practices, it probably was created and is curated with a certain effort.

  For example [Küm+19] prescribes to use geonames[2] to uniquely identify geolocations, thus identifying this curation style of geolocations as a best practice, although other services offer the same or similar functionalities. The model for interactions with a research data product should provide the means to encapsulate the check whether a geoname was used to specify a geolocation.

- Only those interactions should be supported by the model that can be *fully automated* (F-D2/2). This feature of the model for interactions with a research data product is motivated by the growth dynamics of research data products: This ensures that no manual steps are necessary to carry out the assessment of a research data product.

  Re-using the example above, whether a geolocation is specified by a valid geoname is an automatable test.

- The possibility to model interactions *as independent as possible* is necessary to facilitate parallel execution (F-D2/3); this sub-feature is also motivated by the growth dynamics of research data products, since parallel execution of benchmarks allows to score a larger number of research data products in the same time.

  While checking geonames is an example that is easily parallelized on the client side, limitations and bottlenecks have to be considered regarding the requested service.

Table 4.2 offers a tabular overview over feature F-D2.

[2] `https://www.geonames.org`

Table 4.2: F-D2: A Model for Interactions with Research Data Products

Details	**F-D2/1** The model allows to encapsulate best practices.
	F-D2/2 The model should demand that the interactions are automated, i.e. they should be machine-actionable tasks.
	F-D2/3 The specific interactions should be as independent as possible to enable parallel execution.
Motivation	• Heterogeneity (Section 1.3)
	• Missing conventions (Section 1.3)
	• BQC-1 (Section 2.2)
	• BAND (Section 3.2)
	• CTXT (Section 3.2)

F-D3: Mapping of the behavior of a Research Data Product onto $\mathbb{R}^+$

The architecture should include the logic to map the behavior of a research data product onto values in $\mathbb{R}^+$(F-D3) The main motivation for F-D3 is comparability to event-based metrics, which also have metric output. Such a mapping to $\mathbb{R}^+$ does not entail a linear transformation from the metric space of benchmark scores to the metric space of scores of event-based metrics.[3] From the general domain of this thesis, only one feature can be derived:

- The mapping must be realized with uniform and clear semantics (F-D3/1).

  An example for such a mapping is to define 0 to signal the worst possible quality, impact, or relevance, while 1 is the best grade. All values in between would indicate a tendency corresponding to their distance to 1 (or 0).

Table 4.3 offers a tabular overview over feature F-D3, including sub-features motivated by following considerations.

FD-4: Orchestration of Execution, Components, and Values

The fourth and last feature derived from the general domain of the thesis is that the architecture should allow to orchestrate the execution, the composition of its components and the aggregation of the values onto which the behavior of the research data products is mapped. (F-D4):

[3]It is not assumed that the distances in both spaces have the same "meaning". The purely ordinal nature of the comparison is discussed in Section 2.3 and exemplified in Section 6.2.

Table 4.3: F-D3: Mapping the behavior of a research data product onto $\mathbb{R}^+$

Details	**F-D3/1** The mapping should have uniform and clear semantics. **F-D3/2** The mapping should only be loosely coupled to the model for interactions with research data products. **F-D3/3** The range of possible values should be bounded.
Motivation	• Comparability to event-based metrics (Section 1.3) • BQC-2 (Section 2.2) • NORM (Section 3.2) • CTXT (Section 3.2)

- Since the interactions should be as independent as possible (F-D2/3), the architecture should allow to *orchestrate the execution of interactions* with the research data products (F-D4/1) to enable parallelization and an asynchronous communication paradigm.

  An example for an orchestration of the execution is to run all interactions of the benchmarks in parallel for one research data product or to run the benchmark in parallel for multiple research data products.

- The architecture should allow to *orchestrate its components*, i.e. interactions with research data products and mappings from behavior to $\mathbb{R}^+$, (F-D4/2); this should include the possibility to skip certain combinations if they are not suitable for a research data product (F-D4/3). The motivation for this sub-feature is to cope with another aspect of the heterogeneity of research data products, since it allows to adapt the benchmark to a specific research data product, as well as to the use case at hand.

As an example, an interaction that checks whether the language of the research data product is specified, makes no sense, if the language of the research data product is irrelevant, as in the case of images (which are not figures at the same time).

- The architecture should also allow to *orchestrate the mapped values*, so that a single, weighted score of a research data product is the result of the benchmark (F-D4/4). This weighted score should mirror the holistic stance regarding research data management.

  An example is to calculate the arithmetic mean of all mappings as a result of the benchmark.

Table 4.4 offers a tabular overview over F-D4, including sub-features motivated by following considerations.

Table 4.4: F-D4: Orchestration of Execution, Components, and Values

Details	**F-D4/1** The orchestration should provide enable parallel execution.
	F-D4/2 The orchestration should provide means to compose the components of the benchmark.
	F-D4/3 The orchestration should allow to skip interactions if suitable.
	F-D4/4 The orchestration should provide variing functionality to aggregate the values onto which the behavior of research data products is mapped.
Motivation	• Growth dynamics (Section 1.3)
	• Heterogeneity (Section 1.3)
	• Holistic stance towards research data management (Section 1.3)
	• BQC-3 (Section 2.2)
	• COV (Section 3.2)

4.1.2 Based on the Evaluation of Related Work

The second source for features is the discussion of the methodology of and related work to this thesis in Section 2.2. Three out of the five quality criteria are relevant:[4]

- BQC-1: The features should mirror the concept of machine-actionability, which is already captured by F-D2/2.
- BQC-2: The architectural design should be flexible enough to support different assessment frameworks (such as the FAIR principles). This requirement can be honored by extending F-D3 with the additional sub-feature that the mapping of the behavior of research data products to $\mathbb{R}^+$ should be only loosely coupled to the model for interactions with research data products (F-D3/2). This means to separate the concern of modeling interactions from the concern to "grade" a certain behavior

  Another sub-feature affected by BQC-2 is F-D4/2 since it prescribes the possibility to (re)-arrange checks and evaluations which facilitates a flexibility towards different assessment frameworks.

  The last sub-feature affected by BQC-2 is F-D4/4, the weighting of different evaluations into one score of a research data product; this weighting must be configurable to support a broad bandwidth of assessment frameworks.
- BQC-3: All components of a research data product should be included in the architectural design; this sub-feature is partially captured by F-D4. The additional sub-feature F-D1/3 applies BQC-3 to the interface of research data products, namely, to support all four components of an research data product (PID, data, metadata, and services).

4.1.3 Based on the Shortcomings of Event-Based Metrics

The third source for features is the discussion of shortcomings of event-based metrics in Section 3.2: since sub-question SQ-5 of the research question asks whether benchmarks can complement event-based metrics, the architecture has to be aligned to major shortcomings of event-based metrics to deliver promising answers to that question. From a system architect's point of view, the discussion of shortcomings in this context is to learn from the problems of other solutions.

The following shortcomings from Section 3.2 are sources for features:

- BAND: as argued in Subsection 3.3.3, the effects of social dynamics on event-based metrics is a principal shortcoming, i.e. there is no possible mitigation

[4] BQC-4 and BQC-5 rather concern the implementation of the prototype and its evaluation, not its architecture.

strategy with the means of event-based metrics alone. It is an additional motivation for F-D2/2, since fully automated workflows typically are independent of human activity, ergo social activity.[5]

- TIME: this shortcoming of event-based metrics is also classified as a principal shortcoming. The problem that event-based metrics are only available *after* a certain time is passed is solved in benchmarks by design: the moment a research data product is created, it can be scored — even in its initial and pre-published state. The problem of scoring before publication can be solved with the feature, that the interface of research data products supports restricted access to its components, via its service components (F-D1/4).

- NORM: the problem how event-based metrics can (best) be normalized is categorized as a normal shortcoming, i.e. it can be mitigated in principle given the resources to develop the necessary solutions. This shortcoming is instructive for the architecture of benchmarks for research data products, since its main problem, the unbounded value range of scores of event-based metrics can be mitigated for benchmarks with a simple design decision: to define a upper and lower bound for scores of benchmarks (F-D3/3).

- CTXT: the shortcoming of missing context is classified as a simple shortcoming, i.e. there are measures available to mitigate its negative effects: providing the missing context. It is noteworthy that F-D1, F-D2, and F-D3 already demand that the context of the calculation of the score is specified in a way that allows to comprehend the final mapping of the behavior of the research data product to $\mathbb{R}^+$. The feature F-A1 discusses contextual information necessary to comprehend the final score of a benchmark (see below).

- COV: The shortcoming of deficient coverage (of events) applies to the log layer of event-based metrics (see Figure 3.2). Since the "events" of benchmarks, i.e. the interactions with a research data products are artificially created, the coverage (which research data products or interactions are included?) can be configured at will (F-D4/2). This shortcoming is therefore mitigated by design.

- TRST: The reasons to doubt reported scores of benchmarks will have a substantial overlap with the reason to doubt event-based metrics. A possible remedy in both cases in transparency. The consequences for benchmarks for research data products is spelled out in the following sub-section.

[5]It would be premature to declare benchmarks free from social effects of any kind, but they are certainly less affected by social dynamics than scores of event-based metrics. See also Subsection 7.2.4 for a discussion.

- DUP: The shortcoming of missed events due to duplicated data sets is handled in the proposed architecture by the PID component: if two research data products have no common identifier, they are treated as separate entities.
- VER: The problem of different versions of the same research data product is another aspect that must be taken care in the context of reporting (see next sub-section).

Two shortcomings are not discussed here, but in Chapter 6 (COR) and in Subsection 7.2.4 (GAME). The reason for that is, that they cannot be mitigated in the architecture, but only in the context of implementation, with a specific benchmark at hand. The discussion of the effect of these shortcomings to benchmarks for research data products is therefore postponed.

4.1.4 Based on Benchmarking Best Practices

To include lessons learned from benchmarking experiences in other fields, two additional sources were analyzed:

- [HB15] discusses best practices in reporting measurement results of performance experiments in HPC benchmarks.
- [OT08] presents best practices in creating verification and validation benchmarks for computational simulations.

We chose these two sources since they present the state-of-the-art of two fields in which benchmark scores and the reporting of computer-generated metrics play a major role. Although both sources deal with quality criteria for research data products — measurement data and simulation software — these quality criteria were not the focus of our analysis; in the current context the quality criteria for benchmarks were analyzed, not for the benchmarked object. Both documents could serve as a source to implement a specific benchmark though, as outlined in Section 5.1.

The following considerations derived and abstracted from the two sources lead to additional features for an architecture of benchmarks for research data products:

- The transfer of the main theme of [HB15], *comparability*, to the present discussion results in the requirement that the scores of two research data products calculated by a benchmark have to bring the two research data products in an informative relation (F-A1). This mainly affects the reporting of a benchmark score, which should include all necessary information to comprehend the result and determine whether two scores can in fact be compared (F-A4/1).

An example of missing comparability is the case, when only the final scores of the benchmark are reported, but not the values over which the score was aggregated.

- Both papers discuss issues of *reproducibility*.[6] The lesson learned from this discussion to the question at hand is to demand that all results are included in the report of each component in order to be able to recalculate the score if necessary (F-A4/2).

  An example for an un-reproducible score of a benchmark is the report of the score and all un-aggregated values without the specification of the aggregation function (candidates are: arithmetic mean, harmonic mean, weighted mean, winsorized mean, median — to name but a few).

- Another aspect mentioned in both papers is the replicability of scores. It is necessary to state whether the interactions are *random or deterministic* (F-A4/3). If the interaction is random, an expected window of outcomes (including a confidence statement) should be specified. In general scores will not be replicated easily, since they depend on volatile assumptions (such as the availability of services), replicability is thus *not* a feature demanded by the architecture.

  An example for a random interaction is to check the specific value of a measurement device. Another example highlights unwanted "randomness" in an interaction with a research data product: if the version of the research data product is not specified, the outcome of the interaction appears random (but is in fact deterministic given the version of the research data product).

- Another issue raised by both papers is the *precision* of the reported results: computers can only approximate certain real numbers due to roundoff errors, that is the difference between the real number and its representation in hardware. The reporting should always specify the used precision, report the actual, un-rounded floating-point values of all intermediate steps, and specify if and how the final score of the benchmark has been rounded (F-A4/4).

- [OT08] demands conceptual descriptions of the components of a benchmark that contains enough information to *understand its purpose* (F-A4/5).

Table 4.5 offers a tabular overview over feature F-A1.

[6] Following the suggestions of [Bar18] this thesis uses the term "reproduce" (and its derived forms, such as "reproducibility") to denote "same data+same methods=same results", as opposed to "replicate", which denotes "new data and/or new methods in an independent study = same findings".

Table 4.5: F-A1: Reports to make Scores Comprehensible and Reproducible

Details	These points apply to reports of each major component of the architecture: **F-A4/1** A report should provide all information necessary to compare the scores of two research data products. **F-A4/2** A report should allow to reproduce the scores. **F-A4/3** A report should specify whether the involved interactions were random or deterministic. **F-A4/4** A report should specify the floating point precision used for metric values, report non-rounded values as intermediate results, and if the final result is rounded, specify the decimal position up to which the value was rounded. **F-A4/5** A report should provide enough information to comprehend the purpose of the component.
Motivation	• CTXT • TRST • VER • [HB15] • [OT08]

4.1.5 Conflicting Features

A possible conflict might arise between the feature to encapsulate best practices (F-D2/1) with the required independence of these encapsulations (F-D2/3) and the exploitation of this independence in designing parallelized and asynchronous communication patterns (F-D4/1). Two types of dependencies are problematic in this sense:

1. Interactions might be dependent since they request the same resources. An example for this type of dependency are interactions with the same (network-bound) service component. Such a dependency is called *concurrent* dependency below: they decrease the potential of parallel execution.

2. Interactions might be dependent since they need to be carried out in a specific order, or their serial execution might just be more performant than their parallel execution. One example for this type of dependency is the relation between the check whether a research data product has a valid PID and the

check whether the PID resolves to a landing page. The second is conceptually dependent on the first (an invalid PID cannot resolve), therefore the second check can be skipped. This type of dependency is called *serial* dependency in the following: the interaction must be executed in a specific order, or the serial execution might outperform parallelized processing.

Outside of the context of a specific use case and the implementation of a benchmark, these dependencies are impossible to resolve *a priori*, but possible mitigation strategies are sketched in Section 5.1.

4.2 An Architectural Design for Benchmarks

This section presents the architecture of benchmarks for research data products which adheres to the features discussed in the previous section. The architecture is independent from its possible implementation (although one implementation is sketched in Section 5.3).

The benchmark architecture is modularized along the principle of the separation of concerns [Dij82]: there are five main components of the architecture which are aligned with the five main requirements discussed in the previous section. Table 4.6 provides the mapping of features to components of the architecture: Each row displays the feature, the conceptual question raised by the feature, the component responsible to implement the feature, and the subsection in which the component is discussed in-depth.

Table 4.6: Overview of Main Components of the Architecture

Feature	Conceptual Question	Component	Subsec.
F-D1	What interface is shared by all research data products?	Interface for research data products	4.2.1
F-D2	How can the interactions with a research data product be modeled to probe it for a certain quality?	Check	4.2.2
F-D3	How can the behavior of an research data product be mapped to $\mathbb{R}^+$?	Evaluation	4.2.3
F-D4	How should the components of the architecture be orchestrated?	Benchmark	4.2.4
F-A1	How can reproducibility and comprehension of the scores be facilitated?	Report	4.2.5

This summary of the architecture helps to navigate the following subsections, by highlighting the relations between the components: a *benchmark* is the orches-

tration of *checks* (probing of a research data product) and *evaluations* (assessment of the probe's result); this orchestration results in a score for a research data product compliant to a common *interface*; all these components contribute a part of the *report* that allows to reproduce and comprehend the score of the research data product.

4.2.1 An Interface for Research Data Products

The first component of the benchmark architecture is the interface for research data products. A common interface for all research data products is a necessary precondition to define all following steps in this subsection. Without such an interface, the proposed architecture would be under-specified, especially in the light of heterogeneities discussed in Section 1.3. The conceptual structure of research data products is sketched in Figure 4.1 (a modified version of Figure 1.1 in Subsection 1.1.1), and a template for an interface is depicted in an UML class diagram in Figure 4.2. Before the class diagram and the depicted interfaces are discussed, it is motivated by a few conceptual remarks along Figure 4.1.

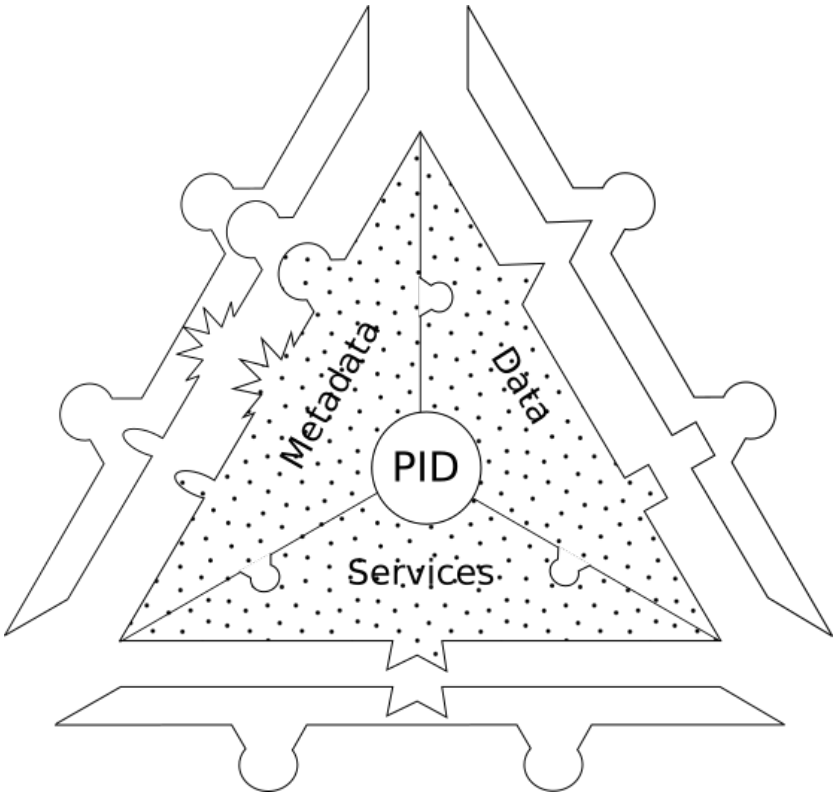

Figure 4.1: A research data product with facade layers

In a "creolized" world ([WS18]), a common interface for research data products can be realized by using two common design patterns[7] namely the *adapter* and the *facade* pattern [Gam+94]. This idea is depicted in Figure 4.1, in which the inner, dotted triangle stands for a specific research data product with its custom interfaces and the pieces surrounding the triangle stand for the adapters. The tabs inside the dotted triangle represent internal connections between the components and the tabs pointing outside of the spotted triangle represent the different interfaces, standards, and implementations for interacting with the data, metadata, and service component.[8]

Each of the pieces around the dotted triangle represents an adapter for one of the components: the interface(s) of each component is (are) converted into the interface(s) the benchmark implementation expects, thus creating the generic interface demanded by F-D1/1. It is still possible to use and probe a research data product by its "original interfaces"(F-D1/2).

Combined, the three adapters form a facade, i.e. they limit and simplify the functionalities of the original, lower-level interfaces to those which are common among research data products. In the following, a triangle with six identical looking tabs is used as the graphical representation of the "variable" standing for this facade, i.e. the interface the benchmark's implementation expects.

Figure 4.2 displays the mentioned ideas in UML notation: the main classes are depicted inside the "Framework" package (the enclosing box): the ResearchDataProduct is a composite ("has-a"-relationship) of Metadata, Data, and Service objects (the last one via the ServiceBundle object that encapsulates the management of credentials and the selection of services for a given purpose) and the pid attribute. The classes depicted outside of the enclosing box give examples for realizations of the Metadata, Data, and Service interfaces. These realizations are there to show how F-D1/2 can be satisified, since they can provide functionality that could not be homogenized without arbitrary choices (e.g. getting the number of cells makes sense for CSVData, but not for PDFData).

Figure 4.2 is an inversion of Figure 4.1: the inner triangle in Figure 4.1 corresponds to a combination of the specific realizations of the interfaces (everything outside the enclosing package box). The components inside the package box symbol correspond to the adapter parts outside the dotted triangle — the two figures

[7] "*A design pattern names, abstracts, and identifies the key aspects of a common design structure that make it useful for creating a reusable object-oriented design.*" [Gam+94] A design pattern is defined by its name, a problem description, a solution and the consequences of the solution. The canonical description of most design patterns can be found in [Gam+94].

[8] The PID component is left out, since it is rather a simple component (typically a string that is a persistent and globally unique identifier by convention). Furthermore, three dimensions are easier to represent than four; it is to be noted though, that in principle all considerations also apply to PIDs.

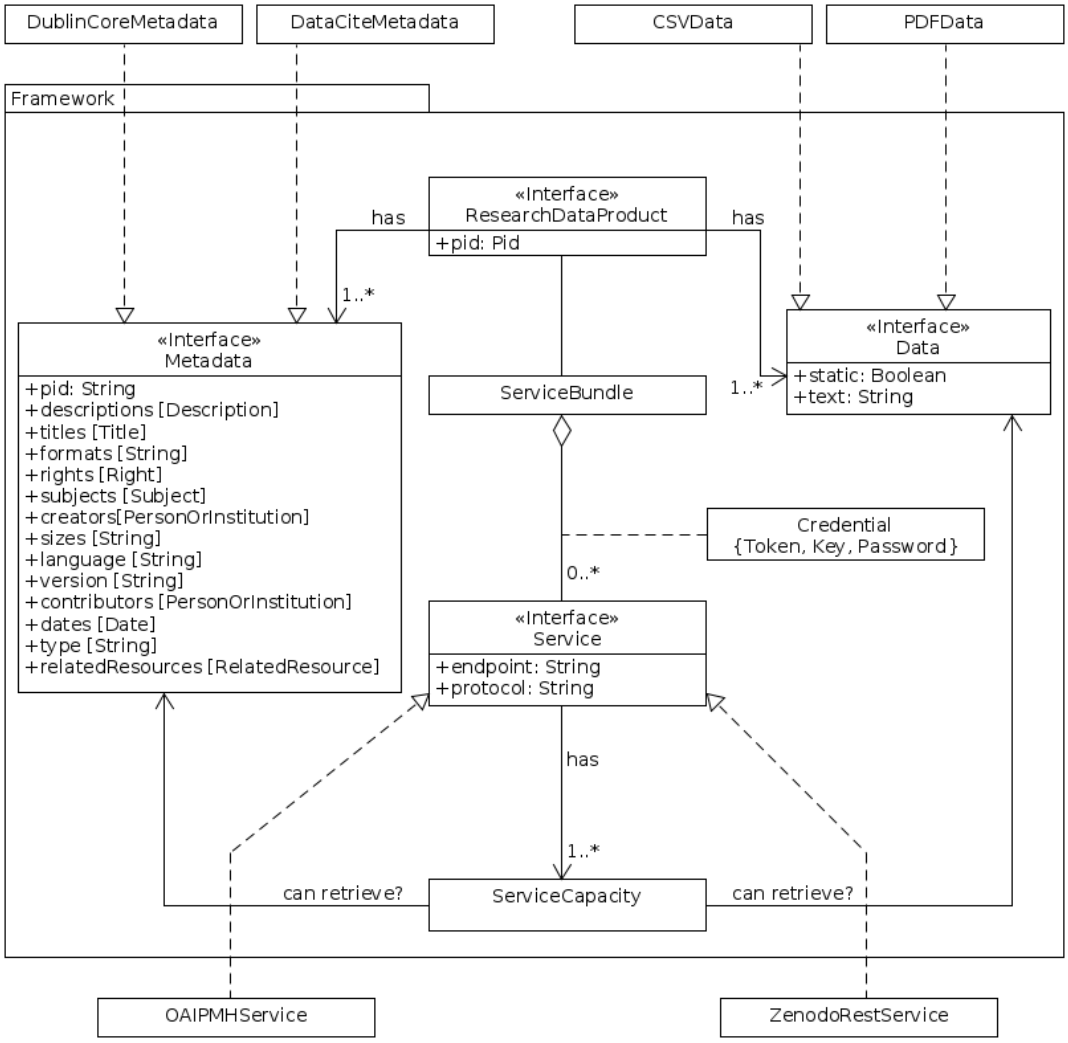

Figure 4.2: UML class diagram of a research data product

are thus inverted views of the same model: while Figure 4.1 depicts how to abstract from the heterogeneities to a common interface, Figure 4.2 shows how it can be realized.[9]

[9]Since each of the specific classes realizes an interface, they are *not* adapters in the strict (object-oriented) sense of that pattern — there is no dedicated class which provides the adaption logic, nor is there a class corresponding to the "conceptual" facade. But this is only an implementation detail, since it could be implemented that way, if it would bring advantages (code reusage, etc.). The class diagram shown in Figure 4.2 concentrates on the specification of the interface for research data products and the explication of the underlying model and is therefore kept as simple as possible.

The class diagram specifies the four interfaces ResearchDataProduct, Metadata, Data, and Service to satisfy F-D1/1 and F-D1/3. It specifies a very minimalistic set of common attributes for each interface, which is more extensive only in the case of the Metadata interface.[10] It is possible that the selected attributes and methods need extension or revision (see Subsection 4.2.5), so the proposed version in Figure 4.2 should be considered as a template in which the specific versions can be replaced, given more context. Section 5.1 provides a recipe of how the interface for research data products can be extended in the context of a specific use case.

The class ServiceBundle is a collection of services that offers the typical get-and-store logic of collections, but provides two additional features:

- The ServiceBundle allows to select a service for a task based on its capacities. This entails the capacity model being a part of the conceptual interface for the service components of research data products and therefore contributing to the satisfaction of F-D1/1 and F-D1/3.[11]

- The ServiceBundle stores Credentials necessary to use the services and maps them to the services, thus satisfying F-D1/4.

The enclosing box defines the boundaries of package "Framework", beyond which the components of the architecture are specific to one or several use cases and are not part of the template. An implementation of such a framework is sketched in Subsection 5.3.2.

4.2.2 Checks

The problem of modelling interactions with a research data product to probe it for a certain quality, is solved by the concept of a *check*. The intuition behind a check is best captured when considering it as an example for the *command* pattern [Gam+94]: it encapsulates a request (to a research data product) in an object. A check fulfills two objectives:

1. Automation of interactions with the research data product.

[10] This interface was mainly influenced by the DataCite standard [Dat19], but contains a few simplifications and some original features were skipped.

[11] The displayed capacity model only models the capacity to retrieve data or metadata. This is a rather simplistic approach for the sake of brevity and will probably be in need of extension in an implementation to capture the full width of research data services, such as services providing streaming access to research data products to sensor data or filter/selection functionality to chunk the data into smaller pieces. This model can be extended by following the steps layed out in Section 5.1.

2. Mapping the behavior of the research data product to a well-defined and limited set of result types.

The following list shows four result types, which can also be found in the prototype (see Section 5.3):[12]

1. Boolean results, indicating whether one or several criteria is met by a research data product

2. Metric results, quantifying a feature of a research data product

3. Categorical results, are strings selected out of a finite and pre-defined set of strings indicating the detected category of a research data product or one or several of its components.

4. List results, which are lists of values of any type (can also include strings, etc.)

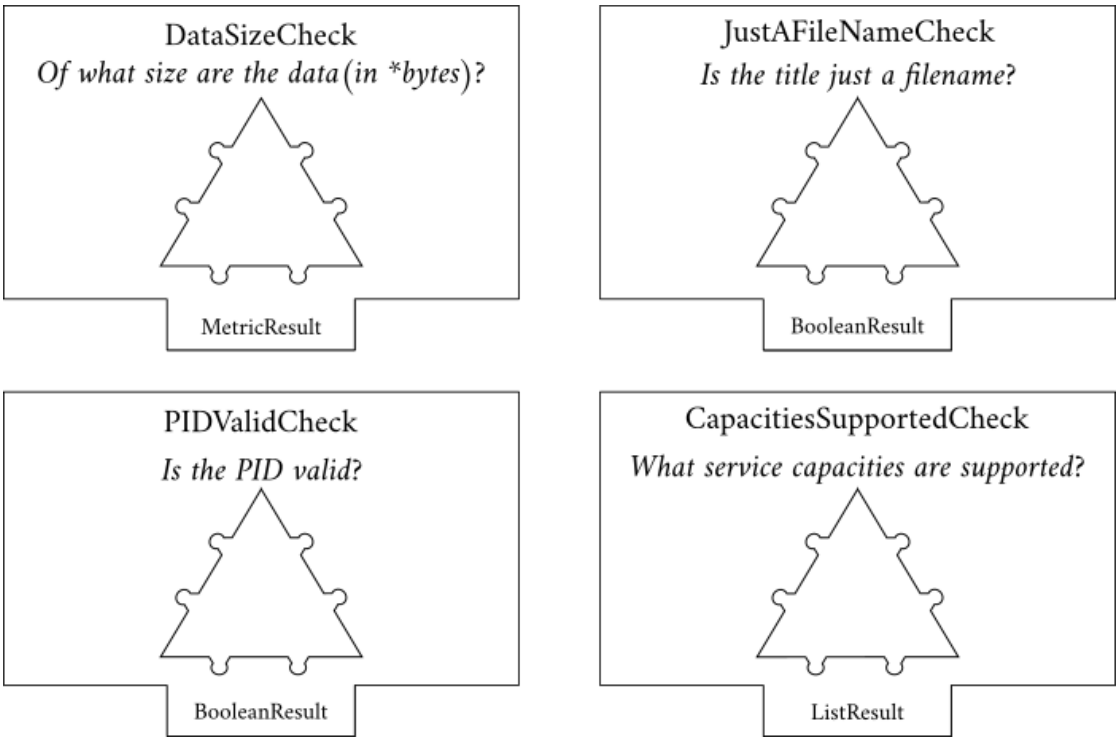

Figure 4.3: Examples for Checks with different result types

[12]This list can be extended if necessary. It was compiled based on the typical types of variables of most programming languages.

Figure 4.4 depicts four examples of checks: each check is represented by a box with a blank in the middle which stands for the expected interface the research data product must provide (see previous subsection). The plug-like extension at the bottom of each check represents the type of the result(s) of the check:

- The DataSizeCheck concerns the data component of the research data product. Its result is metric.
- The JustAFileNameCheck concerns the metadata component of the research data product (e.g. the title of a research data product). Its result is of Boolean type.
- The PIDValidCheck concerns the PID component of the research data product. Its result is of Boolean type.
- The CapacitiesSupportedCheck concerns the service component of the research data product. Its result is a list type, comprised of categorical results.

A check can be a simple access to features of the research data product, such as "give me the title(s)" — but it can also involve more complex processing, e.g. "does any title include numbers?", or "calculate the arithmetic mean of all metric columns of the csv". A check can probe one or several components of the research data product at the same time, but in consideration of F-D2/3, separate aspects should not be put in one check.

All technical interactions with a research data product have to be encapsulated into the checks (F-D2/1, F-D2/2), i.e. the check needs to take care of all expected eventualities or return a specific failure signal in the event of unexpected events. Since checks might include probing unstable resources such as network-bound features of the research data product it must be guaranteed that the checks are robust enough to return a (valid or invalid) result.

A check is non-evaluative, i.e. it just returns a result, without evaluating the quality, impact or relevance of the research data product as "good" or "bad"; a good example to illustrate this separation is the BooleanResult type - it is a different matter to determine whether a title is just a filename and whether it is good or bad practice for a title to be just a filename. Checks are thus the glue between the interface of research data products and evaluations (see next subsection).

Figure 4.4 depicts a UML class diagram to formally specify the previous conceptual considerations, the examples for Checks and Results correspond to the previous given examples:

- The abstract Check class offers the basic functionality, especially considering F-A1. The check-method offers the functionality to satisfy F-D2/1. The

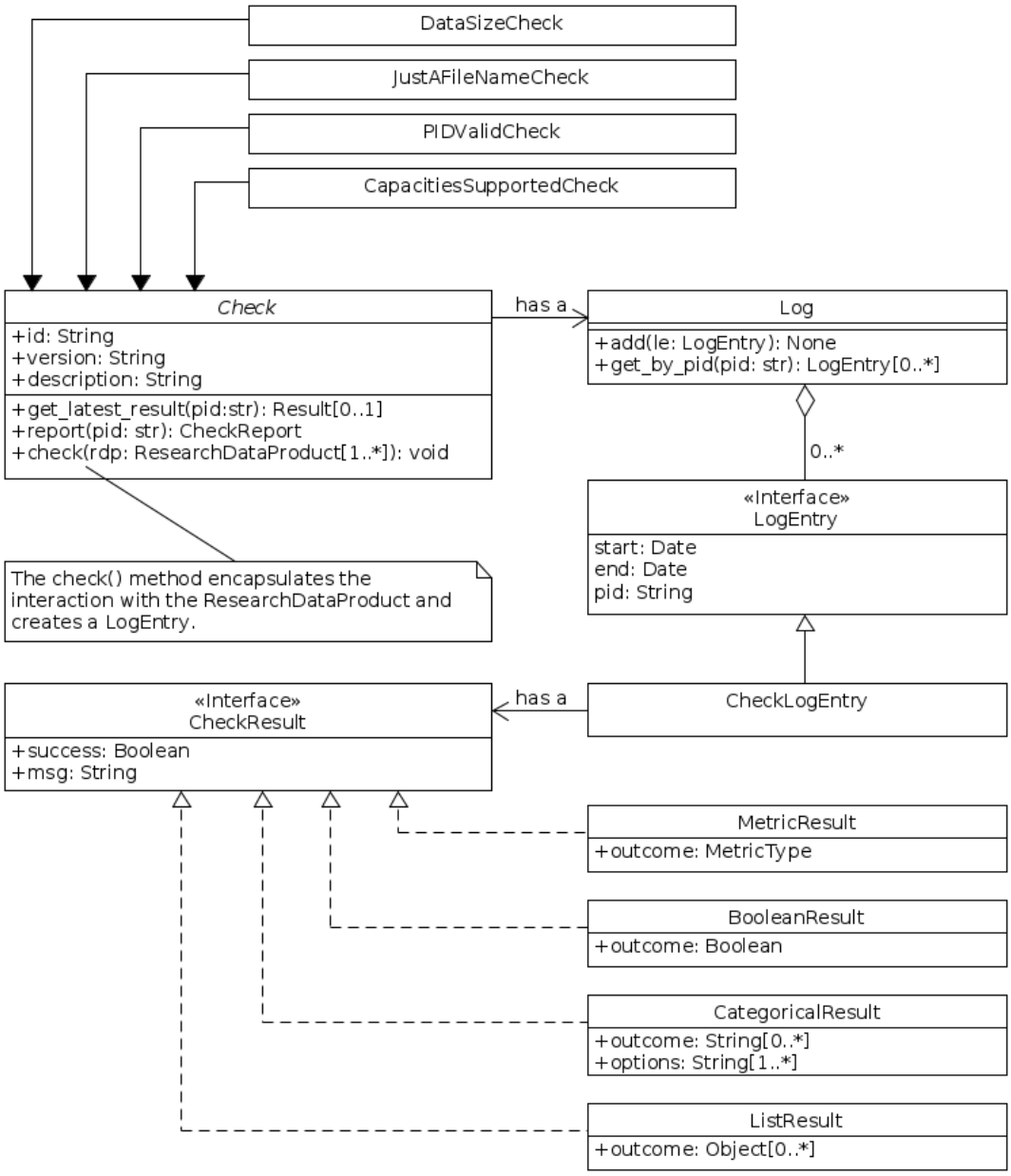

Figure 4.4: UML class diagram of Checks and related Objects

check-method and the getter for the results allow an asynchronous communication pattern.

- Each check has a log in which the CheckResult is stored together with some metadata (including when and for which research data product the check has been run).
- The types of results listed are represented in the lower part of the diagram. The CheckResult interface offers functionality to make checks run robust and machine-actionable (F-D2/2).

The only requirement left undiscussed, F-D2/3, i.e. that checks should be as independent from each other as possible, cannot be answered on the architectural level, but must be faced when specific checks are implemented or added to a benchmark.

4.2.3 Evaluations

The concept of an *evaluation* is necessary to model the mapping of the result of a check to $\mathbb{R}^+$. Its role is to enable two features of the architecture:

- Decoupling the evaluation of a research data product from checking it (F-D3/2)
- Reusability regarding re-occurring evaluation patterns (code) and runtime results (n:m-mappings of checks to evaluations)

To map "good" results of checks to 1 and "bad" results to 0, and interpreting intermediate values to indicate a corresponding tendency is a solution that satisfies both F-D3/1 and F-D3/3.

Figure 4.5 depicts the interplay between research data products, checks and evaluations in the conceptual notation introduced above: An evaluation is represented by an hexagon that has blanks for the result type of a check. The upper part of Figure 4.5 displays an example for an evaluation that is re-usable: the evaluation just returns 1 if the result is true (in case the result type is Boolean) or 0 if it is false. Further examples for such evaluations include:

- Evaluate to 1 if a result is identical to a specific value, 0 otherwise.
- Evaluate to 1 if a result lies between two values, 0 otherwise.
- Evaluate to 1 if the result is false, 0 otherwise.
- Evaluate to $\frac{i}{n}$, with i as the number of Booleans in the result with value "false".

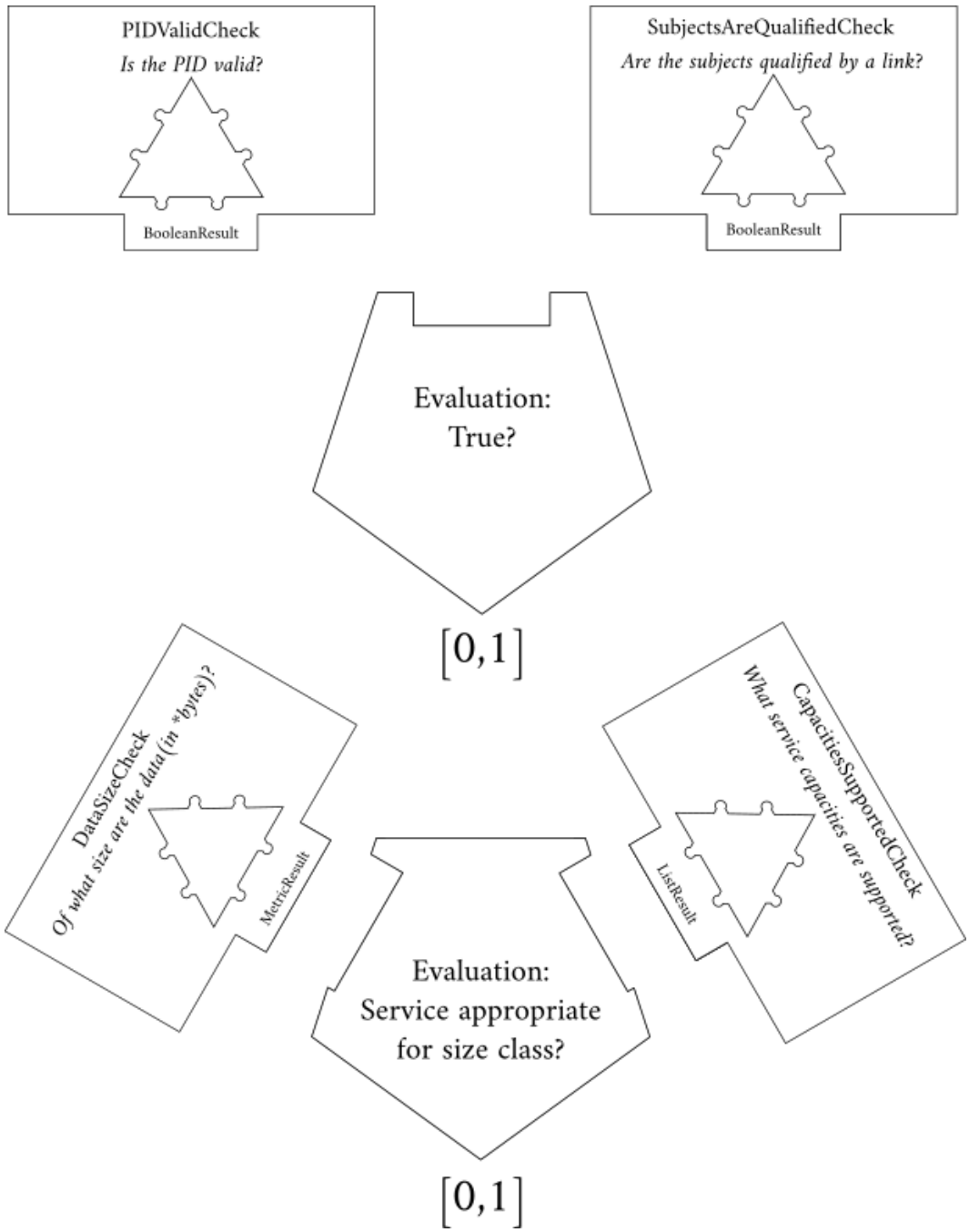

Figure 4.5: Examples for Evaluations and their relation to checks

- Evaluate to a return value of a user-provided function (customization).

The lower part of Figure 4.5 depicts the last type of evaluation; a function encapsulates the knowhow which service capacity is appropriate for a certain data size.[13] It is also an example for an evaluation that needs more than one check to run its evaluation. Since the evaluation has also access to the success flag of the result, it can decide whether a non-successful test should be evaluated (e.g. as 0) or whether an exception is raised.

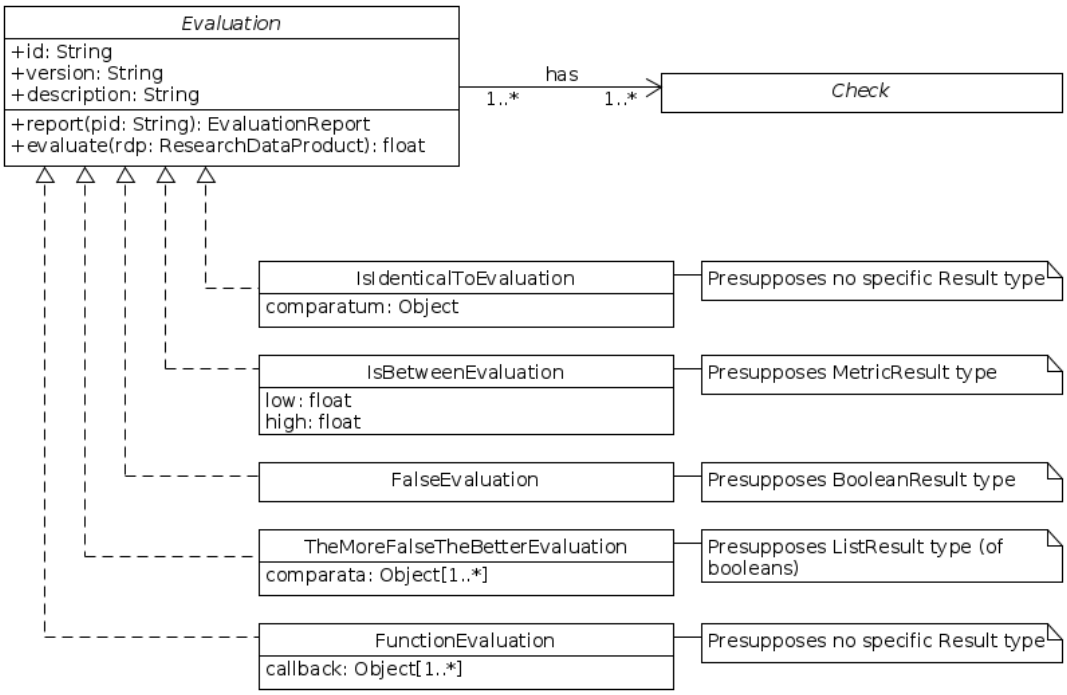

Figure 4.6: UML class diagram for evaluations

Figure 4.6 is a UML class diagram to formalize the previous comments. The abstract base class provides functionality to satisfy F-A1 (the report method). The list of examples above is also specified together with the type of result expected by the inheriting class; it corresponds to the examples for evaluations listed above. The n:m-cardinality between Evaluations and Checks is the notational equivalent of the feature to re-use the result of a check in different evaluations and to use multiple checks in a single evaluation.

[13] An example would be to prescribe the use of the gridFTP protocol to transfer a large amount of data; this is more efficient than the transfer via HTTP/S. See `https://www.ogf.org/documents/GFD.20.pdf` for further details.

4.2.4 Benchmarks

Benchmarks are the architectural component that orchestrates all other components. The term "orchestrates" is a fitting term, since the three aspects that are solved with the concept of a benchmark can all be enlightened by orchestral metaphors:

1. A benchmark is the conductor of the checks, which means that a benchmark schedules start and stops of checks and decides when evaluations are carried out. Together with the independence of checks (F-D2/3) this concerns requirements F-D4/1 and F-D4/3.

2. A benchmark is also the composer of the interplay between checks and evaluations, i.e. the benchmarks manages the n:m mappings between checks and evaluations. This aspect touches F-D4/2.

3. A benchmark is, last but not least, in charge of the concert review, i.e. aggregating the output of all evaluations and condense it into a single value between 0 and 1. This aspect is important for F-D4/4.

Figure 4.7 is meant to show the interplay between the three orchestration tasks. It is to be read top to bottom, at the left side the components are depicted which are central to each step:

1. selecting a *research data product*

2. running the *checks* on it

3. *evaluating* the checks' results

4. weighting these evaluations into a final *score*.

The three responsibilities of the benchmark are depicted by little numbers in Figure 4.7:

(1) **conducting the assessment**: the first task of the benchmark is to execute all checks on a given research data product

(2) **composing the assessment**: the checks' results are mapped to one or many evaluations

(3) **finalizing the assessment**: there is an aggregation of all evaluations, depicted by the scale at the layer Score

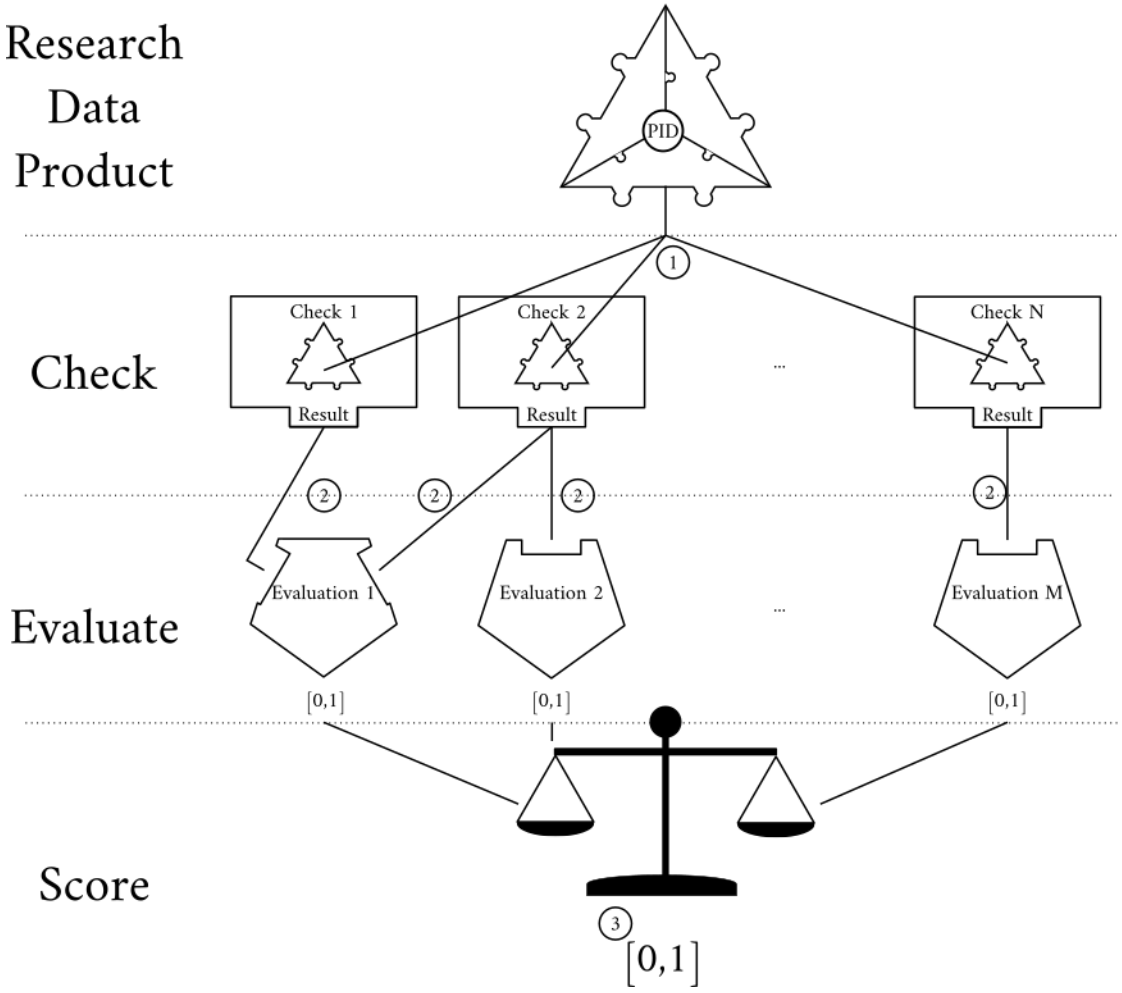

Figure 4.7: Overall schematic for the architecture with the focus on benchmarks

This last layer Score has another variable element, which again stresses the template-like nature of the proposed benchmark architecture: the only requirement is that the benchmark maps the result of the different evaluations to [0,1], it does not specify *how*. One reason for this variability is F-D4/4 (via BQC-2), the possibility to support different assessment frameworks. The possibilities to weigh the different evaluations can be categorized in two main classes:

- *Static weights*: The weights of each evaluation for a research data product are determined independently of the outcome of the checks/evaluations of other research data products. The simplest example would be the arithmetic mean of all evaluations.
- *Dynamic weights*: The weights are a function of the outcome of the checks/e-

valuations of other research data products. An example would be a weighting function that is biased towards checks/evaluations whose outcome's distribution is skewed in a population of research data products, hence giving those research data products more weight, which excel (by fulfilling a quality requirement that is only met by a few research data products, see also [WK18]).

Static weights are easier to reproduce since no additional state must be reported. Dynamic weights might be more discriminative: scores calculated with dynamic weights will not only carry information about each research data product, but also about the use case mirrored in the collection of checks and evaluations. For dynamic weights additional information must be reported, such as the distribution of the evaluations and the function to map them to weights ensuring a final score between the defined bounds 0 and 1.

As in the previous subsections, these conceptual considerations need to be formalized into a design that can be implemented. Structural diagrams (such as UML class diagrams) cannot honor the "active" role of benchmarks in the proposed architecture, which is why Figure 4.8 shows a UML sequence diagram. It displays three basic activities:

1. Setup: The yellow activity boxes and the messages attached to them show the setup phase. It corresponds to the second responsibility of benchmarks in the conceptual Figure 4.7, which is the coordination of checks and evaluations.

2. Run: The orange activity boxes and the messages attached to them show the coordination of the asynchronous checks by the benchmark object. It corresponds to the first responsibility of benchmarks in the conceptual Figure 4.7.

3. Report: The red activity boxes and the messages attached to them show the calculation of evaluations and score coordinated by the benchmark object. The coordination is encapsulated into the creation of the reports. It corresponds to the third responsibility of benchmarks in the conceptual Figure 4.7.

Both the Run and the Report activities are shown with only one research data product as argument, which is a simplification: they support batch processing (e.g. by looping of a given set of research data products). The following subsection will give more details on reports, the last major component of the architecture to be presented. This will provide missing context to point 3 above.

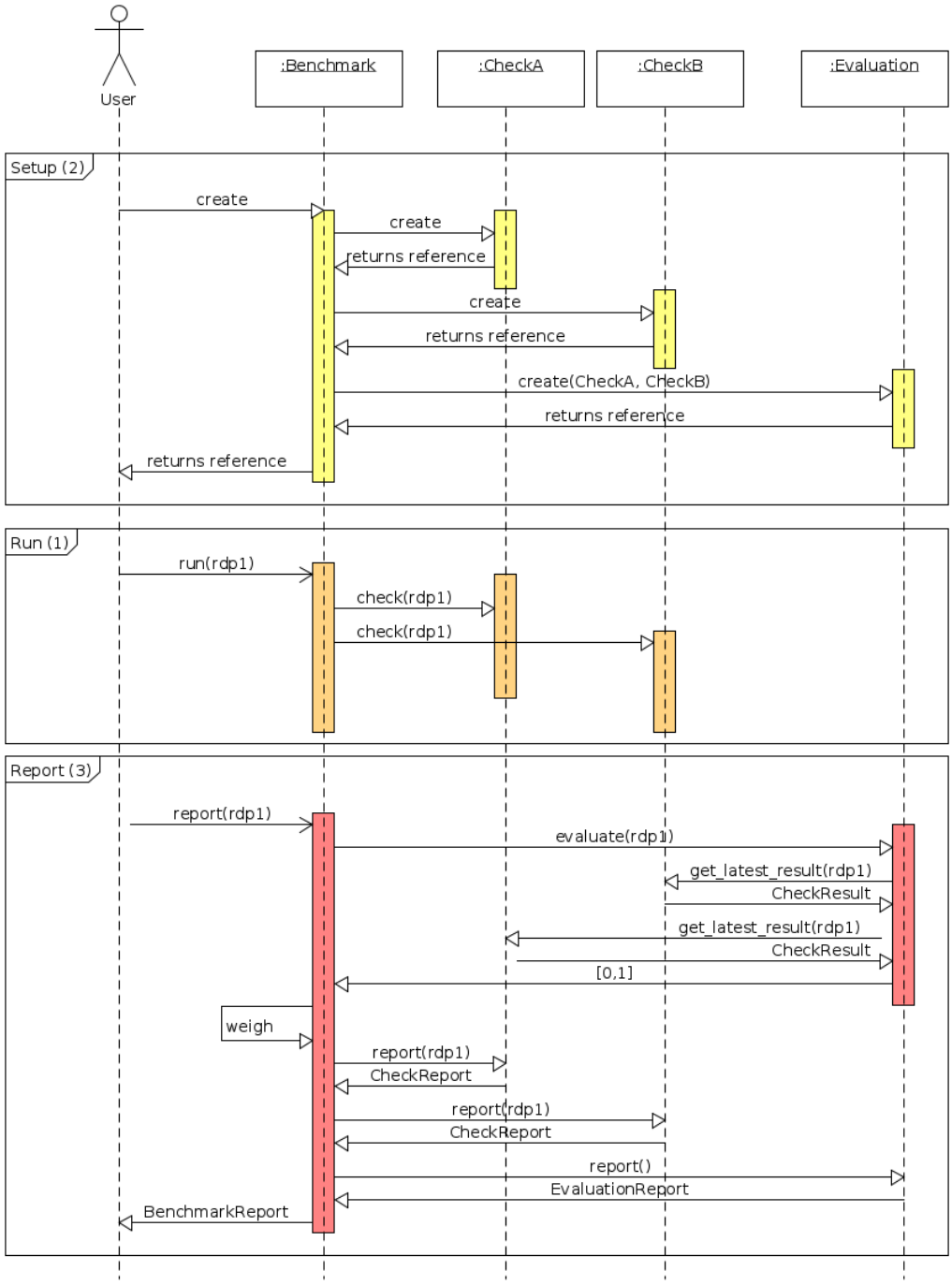

Figure 4.8: UML sequence diagram for the setup, run and report activities

4.2.5 Reports

The role of the report object is to satisfy requirement RA-A1, that is to make sure the score of a benchmark is reproducible and comprehensible. Figure 4.9 provides an UML class diagram with the proposed design for reports. The inheritance of the three Report classes ensures a unified interface, whereas the composite structure of the BenchmarkReport allows a concise reporting (all reports of checks and evaluations are bundled into the BenchmarkReport).

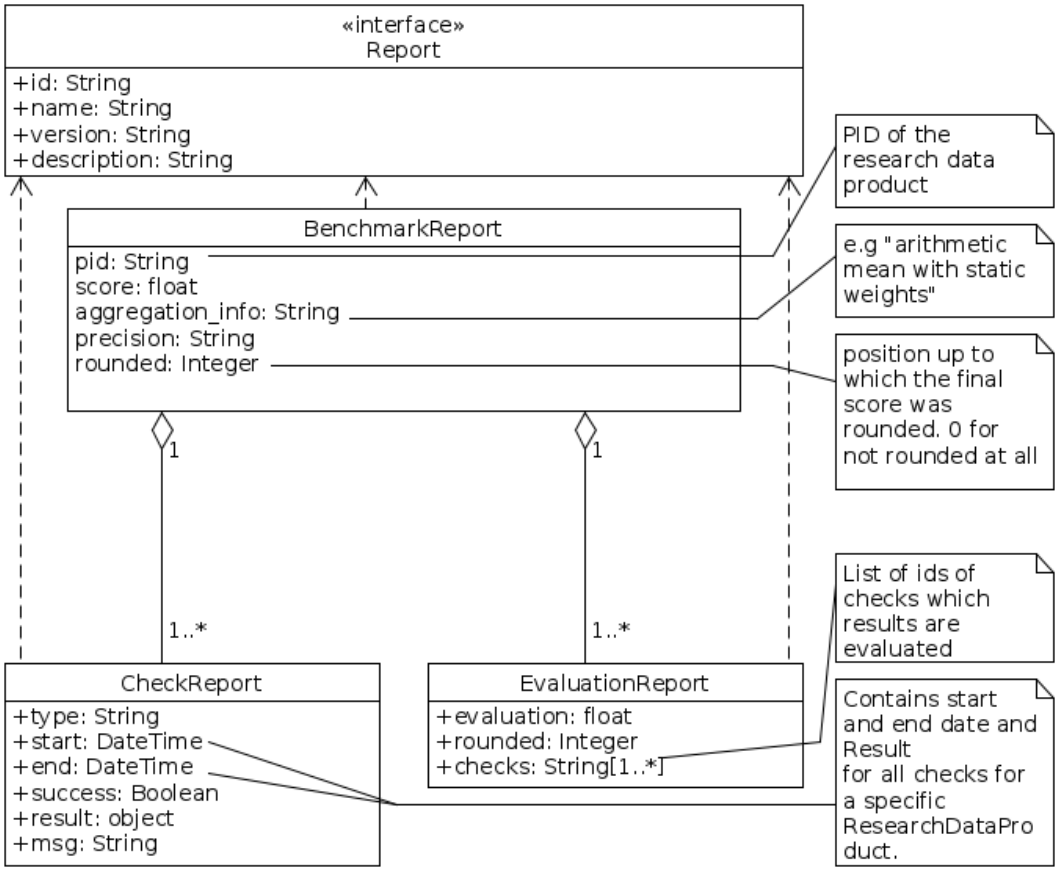

Figure 4.9: UML class diagram for reports

This design satisfies the requirements discussed in Section 4.1:

- The base interface of reports ensures that each component is identified and described (F-A4/5) due to the attributes “id” and “description”.
- The version attribute is of specific importance since it allows to judge to

which extent two scores are comparable (F-A4/1). It is discussed in-depth in the paragraph below.

- The reporting of the checks' result (in its log) in the CheckReport, the evaluations' value in the EvaluationReport, and the "weighting" and the "rounded" attributes of the BenchmarkReport allow to reproduce the final score even manually (F-A4/2).

- Whether or not a check encapsulates a random event (F-A4/3) is reported with the "type" attribute of the BenchmarkReport class.

- The issues related to floating point precision (F-A4/5) are handled by the "precision" attribute of the BenchmarkReport class. It is to be read as the default precision throughout the report.

Versioning

In typical software projects, versioning is a detail of implementation and deployment, but no architectural concern; this does not apply to the benchmark architecture presented in this chapter: four of the five components presented so far — the interface of research data products, checks, evaluations, and benchmarks — constitute a measurement device, whose output is not only a function of the interplay between the input (a research data product) and their constituent components, but also of the version of their components:

- Fixing software bugs might change the behavior of a component.

- Newer versions might add new interactions to a component.

- Backward-incompatible updates might (partially) break interfaces.

These considerations imply one response to the requirement of comparability (F-A4/1) and reproducibility (F-A4/2): to achieve comparable and reproducible benchmark scores, implementations must provide consistent versioning as a core attribute of the implemented components. Re-running the benchmark on the same state with the same version must result in the same scores (in case all checks are deterministic checks).[14] There are three possibilities how changes to a component affect the scores of the different versions of a benchmark:

[14] Checks encapsulating random behavior must include measures of variations to judge whether a result can be reproduced.

- Two versions of a component are considered **inconsistent**, if the score based on the new version is different from the score based on the old version and the old score *cannot* be derived from the new score. Scores based on inconsistent versions can never be compared.
- Two versions of a component are considered **consistent**, if the score based on the new version is different from the score based on the old version and the old score *can* be derived from the new score. Scores based on consistent versions can be compared with limitations.
- Two versions of a component are considered **equivalent**, if the score based on the new version is identical to the score based on the old version. Scores based on equivalent versions can be compared without limitations.

The following maps these considerations to the components of the architecture and gives examples for each (applicable) type of version relation:

- **Interface for research data products**: Different versions can only be inconsistent or equivalent, depending whether the new interface breaks the old interface (an extension of the interface is indifferent regarding comparability).
- **Checks**:
  - A newer version of a check that changes the result type is inconsistent with its older versions.
  - A newer version of a check with a categorical result type is consistent with its older versions, if the set of possible values of the newer version is a superset of the possible values of the older versions.
  - A newer version of a check is equivalent with older versions if the newer version is faster but produces the same results.
- **Evaluations**: Different versions can only be inconsistent or equivalent, depending whether the evaluation logic changed between the versions.
- **Benchmarks**:
  - A newer version of a benchmark that changes the weighting function is inconsistent with an older version.
  - A newer version of a benchmark that includes inconsistent changes of its constituent evaluations or checks is itself consistent if the changed components can be skipped in the run of the benchmark. The same applies to added or re-arranged combinations of checks and evaluations.

– A newer version of a benchmark that changes solely the orchestration of runs of checks is equivalent to its predecessor.

Each component therefore needs to include its own version as a core attribute, as depicted in Figure 4.9. One example for specifying the version compliant with the consideration above, is to use an adaptation of semantic versioning:[15] Each version of a component is specified in the format MAJOR.MINOR.PATCH:

1. Increase MAJOR, if the changes lead to inconsistent versions.
2. Increase MINOR, if the change leads to consistent versions.
3. Increase PATCH, if the change leads to equivalent versions

The five components of the benchmark architecture define a common conceptual framework for all benchmarks for research data products. The next chapter will give hints how such an architecture can be realized and what resources have been used to implement the prototypical benchmark used in Chapter 6.

[15] `https://semver.org`

Chapter 5

Implementing Benchmarks for Research Data Products

Contents

This chapter discusses the implementation of benchmarks for research data products. A recipe to transform the benchmark architecture presented in Section 4.2 into an operational benchmark prototype is presented in Section 5.1. Different use cases for benchmarks are discussed in Section 5.2. The chapter concludes with an overview of the software published alongside this thesis, including the prototype used for the evaluation in Chapter 6.

5.1 A Recipe to Build Benchmarks

In this section a step-by-step approach is provided to realize the benchmark architecture described in the previous chapter. This recipe is an aggregation of the practical insights gained during the implementation of the prototype.

The recipe is built based on the assumption, that the following resources are used, which are provided alongside this thesis:

- A library for research data products (Subsection 5.3.1)
- A framework to build the components of a benchmark. (Subsection 5.3.2)
- A prototypical benchmark (Subsection 5.3.3)

The library and the framework are extended to realize the prototype that is used for the evaluation carried out in Chapter 6; this prototype can additionally serve as an example for other implementations. This section provides the means to get from a use case to the identification and analysis of requirements for a specific benchmark and finally to its implementation.

5.1.1 A Step-by-step Approach

The recipe contains several steps, which can be traversed in different permutations (that is there are different possible paths through the recipe):

- Specify use case(s)
- Review existing benchmarks
- Reuse an existing benchmark
- Analyze evidence
- Identify implementable components
- Handle dependencies between checks
- Implement components
- Orchestrate the implemented components

The steps are not numbered, since there is no specific order in which theses steps are carried out (there is more than one path through the recipe, depending on the use cases at hand). Each of the steps is described in-depth in the following

paragraphs; Figure 5.1 visualizes the different permutations of the steps (the different paths in the figure). The figure works like a state machine, with the states depicted by the boxes which correspond to the above mentioned steps.

Diamonds and branching arrows depict decisions by which the implementation is steered. The criteria for choosing a branch are given by short texts. The specification of the use case(s) is the initial step. The recipe ends with all use cases covered either by existing benchmarks or by newly implemented benchmarks in compliance with the architecture laid out in the previous chapter. An implementation is finished when all components have been implemented (loop on the bottom right, the state includes substates) and orchestrated. Each of the boxes corresponds to a paragraph in this subsection.

Specify Use Case(s)

The most important step in creating a new benchmark is to collect the basic facts concerning its objectives, users, the targeted workflows, and the actors in this workflow. The following questions can serve as a guideline to specify the use cases:

1. *What should the benchmark measure?* Obvious candidates are quality, relevance, or impact of research data products. Additionally, it should be specified, whether all types of research data products are measured, or only a subclass (such as papers, software or tabular data).

2. *Who is the target audience of the benchmark?* Should the benchmark be informative to a certain community, such as researchers of a specific domain (e.g. archaeology, life sciences etc.), meta-scientists (scientometricians, philosophers of science), funding bodies, journalists, non-academic audience, etc.? Are there multiple groups involved and if so, how are their objectives to be prioritized?

3. *Is the benchmark specific to a certain workflow?* Especially important is the information whether there are one or several workflows, since multiple workflows should result in separate components or even separate benchmarks. Can they be condensed into a short description, such as image processing, simulations, qualitative research (e.g. digitizing interviews), textual criticism, machine-learning pipelines, publishing, reviewing, programming, etc.? Is there written evidence for best practices concerning these workflows (publications, blogs, interviews, etc.)? This evidence is used as the basis to specify the requirements for the components.

4. *Is the benchmark specific to a certain type of actor?* Actors of typical research

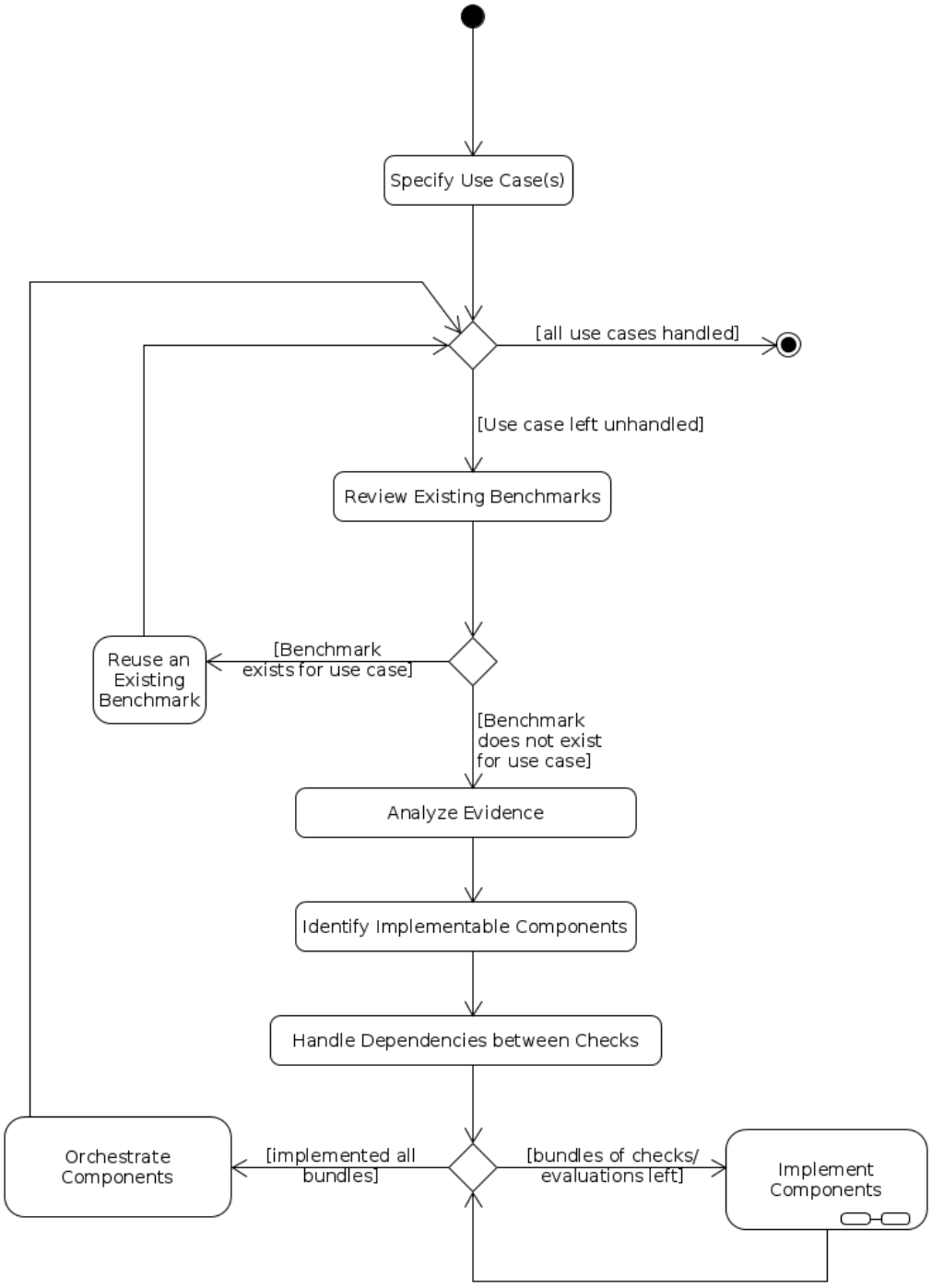

Figure 5.1: Overview of the step-by-step approach to implement customized benchmarks for research data products

data workflows include, but are not limited to, data producers, researchers, reviewers, data stewards, curators, funders, etc.

The answers to these questions specify a use case of a benchmark for a research data product. Three examples for such use cases defined along these questions are presented in Section 5.2. The information included in the description of the use case should be mirrored to the description attribute of the benchmark and its components. The use case might even suffice to specify the first bundle of evaluations and checks of the benchmark.

One possible outcome of this step is the insight that instead of one benchmark, several independent benchmarks are needed since the described use cases overlap only partially.

Review Existing Benchmarks

The next step includes the exploration of the landscape around the use case(s) defined above. This allows to identify re-usable benchmarks, to research best practices, and to specify the relation of the new benchmarks to already existing resources. The following questions can serve as a guideline:

- Are there already benchmarks (partially) covering the use case(s)?
- Can existing benchmarks be adapted to cover the use case(s)?
- How would a new benchmark be compared to existing benchmarks (extension, complement, replacement)?

References to relevant benchmarks should be put into the documentation. This can guide users of new benchmarks in their decision which benchmark to use for their use case and to attribute credits to the creators.

This step has two possible outcomes for each use case: either the use case can be covered by re-using an already existing benchmark or the requirement analysis initiated by the description of use case(s) has to be completed to implement a new benchmark.

Reuse an Existing Benchmark

It is possible that benchmarks for research data products are available, but not in a version compliant to the specifications of Chapter 4. In this case it must be considered whether the effort to adapt an existing benchmark to make it compliant (especially to guarantee comparability and reproducibility, cf. RQ-A1) or the effort to implement the benchmark from scratch is smaller (Subsection 5.1.2 provides a method to estimate the effort of the second option). Each option is discussed

below. The following aspects help to make a final decision whether to re-use the benchmark or not:

- Is the benchmark published under a license that allows to adapt it?
- Is the benchmark actively maintained?
- Can the benchmark be integrated into the parameters of the use case (e.g. programming language, requirements for the runtime environment, etc.)?

Analyze Evidence

Together with the specification of use cases for the benchmark (see above), this step "Analyze Evidence" concludes a full requirement analysis: it identifies the requirements the interface for research data products and the other components (checks, evaluations, and benchmark) have to fulfill to meet the use case's objectives. Possible sources to find evidence include but are not limited to:

- Resources with a description of the data workflows identical or similar to the workflows in the use case description. An example for this type of source is the method section of a scientific article, such as section IV in [WK18].
- Resources created for educational or instructional purposes, such as [Wic+14], [Küm+19], [HB15], or [OT08] (see also Subsection 4.1.4).
- Interviews with stakeholders explaining their methodological approach, such as the interviews with researchers described in [BH18a].

Resources in this context include, but are not limited to articles, books, book chapters, white papers, technical reports, and blog posts. Any such resource should describe interactions with research data products and evaluate these interactions, i.e. characterizing certain workflows as good or bad practice. Each of the resources is analyzed towards the following scheme:

- Does it describe features of research data products which are considered good or bad practice (useful for the extension of the interface of research data products or for implementations of these interfaces)?
- Does it contain descriptions of interactions with research data products that can be implemented (useful for implementations of checks)?
- Are standards, tools, or workflows evaluated — e.g. is compliance to a standard described as desirable or superfluous (useful for implementations of evaluations)?

- Does the resource contain statements which can be used to weigh certain features of research data products against others, e.g. being downloadable by PID is described as a "nice to have", while standardized metadata are characterized as "of the essence" (useful for the weighting function of a benchmark)?

If the resource contains one or several of these options, the information should be excerpted and then clustered into common patterns; e.g. some activities such as simulating with a certain model, or statistical analyses with a specific script can be clustered, if they are similarly discussed across resources. The list of clustered evidence should be compiled in a way that documents the original sources and that allows to extension for updated versions of the benchmark. This list of excerpted evidence is the output of this step.

Identify Implementable Components

One can only consider those aspects of the research data products, interactions, evaluation, or weighting which are machine-actionable, i.e. which can be implemented. The compliance to a standard for example should only be considered, if this compliance is testable (e.g. via a schema language like XSL [Kay17]).

This step results in three outputs:

1. a list containing checks and evaluations to-be-implemented
2. a list containing implementation taks with regard to the interface of research data products
3. a document describing the weighting logic of benchmarks (with references to the first list).

The first list should be extended with the information necessary to comply to the interfaces sketched in Section 4.2:

- An *ID* of the component
- The initial *version* (see Figure 4.2.5)
- A *description*, which includes the references to the resources by which this component was motivated.
- For checks, the information should be added whether the outcome of the check is *deterministic or random*.

IDs are especially helpful to keep consistent references between the three documents.

Handle Dependencies between Checks

Our framework for benchmarks offers simple base classes for benchmarks which operate on the assumption, that checks are independent of each other. This allows to schedule them for parallel execution only considering the resources available.

In the current step, the checks listed and specified in step "Identify Implementable Components" are analyzed to determine whether there are *concurrent* or *serial* dependencies (see Section 4.1). Both types of dependencies need to be handled on the level of the benchmark object, i.e. on the mode of scheduling provided by the benchmark class. A heuristic approach might lead to an easy extension of the benchmark's base class:

- Certain types of checks might be scheduled with fixed intermediate waiting times, to mitigate bottlenecks involved in concurrent dependencies.
- Certain types of serial dependencies can be ignored, if the performance penalty is acceptable: An example would be a check that becomes irrelevant based on the outcome of another check — the effort to execute the superfluous step nevertheless might be smaller compared to the effort to implement a benchmark class that enforces a certain scheduling order.

If these heuristic approaches are not viable, a new benchmark class must be implemented including a scheduling algorithm that can handle the detected dependencies. The implementation should take advantage of best practices developed in the context of compiler theory (see e.g. the section on data dependence in [PW86]).

Another resort would be to eliminate checks from the list if their implementation is too costly and other candidates are available to achieve the benchmark's objective.

Implement Components

In this step the list of of checks and evaluations is converted into a list of bundles; a bundle is a set of evaluations and checks that will be connected by the benchmark. An example would be a bundle containing the check for available data transmission protocols, the check for the size of a research data product, and the evaluation that assesses the retrieved results of these two checks for their suitability (i.e. is the transmission protocol a state of the art solution for the size of the research data product, see Figure 4.5).

To guarantee the accuracy of the benchmark, each of the bundles should be accompanied by a set of research data products which cover the bandwidth of possible outputs of the evaluation. Such a set of examples enables a test-driven approach, that detects whether the checks and evaluations work as specified and how the version of these components must be changed on updates (see Figure 4.2.5)

With the bundles and the according test cases available the components can be implemented along the following steps:

1. Determine whether the interface for research data products suffices to implement the check. If this is not the case, either extend the interface (if the missing functionality can be used for all or almost all research data products), or add a new class realizing the Data, Metadata or Service interfaces. This step might include the implementation of new ServiceCapability classes.

2. For each check in the bundle:

   (a) Determine which type of result the check has. Implement this result type if necessary.

   (b) Implement the check.

3. Implement the evaluation.

Orchestrate the Implemented Components

If the implemented checks and evaluations inherit from the base class provided by the framework, they comply with the interface necessary to coordinate all components of a benchmark. This orchestration thus only needs to be configured bundle-by-bundle in analogy of the example provided by the prototypical benchmark.

Two of the three responsibilities of benchmarks remain, the scheduling of checks and the weighting of evaluations. The former has been discussed in the context of dependent checks: Either the simple orchestration functionality of the framework (Subsection 5.3.2) suffices or a new scheduling logic must be implemented.

The weighting of the bundles is a function that determines the final score of a research data product. Implementation-wise three different ways to combine bundles can be separated:

- All evaluations' outcomes count equal, i.e. each bundle has a weight of $\frac{1}{n}$ with n being the number of bundles. The framework offers the arithmetic mean as the standard weighting function. If another dispersion measure is more suitable (e.g. median), it must be implemented accordingly. The arithmetic mean over equally weighted bundles is the recommended weighting function if the bundles cannot be prioritized based on the analyzed evidence. This functionality is provided by the Benchmark class of the framework.

- The evaluations' outcomes have different, but fixed weights between 0 and 1 that sum up to 1. A configuration is necessary to specify the weight of each bundle. This functionality is provided by the WeightedBenchmark class

of the framework. As with equally weighted bundles, a dispersion measure other than arithmetic mean must be implemented.

- The evaluations' outcomes have different and dynamic weights. The weight of an evaluation is e.g. determined by its adoption of a larger set of research data products. This functionality is provided by the DynamicBenchmark class of the framework.

5.1.2 Estimation of Effort

In the previous subsection a step-by-step approach to build a benchmark (or re-use an existing one) has been presented. This subsection's objective is to map the different types of implementation tasks mentioned in the previous subsection to a simple scheme to categorize the necessary effort:[1]

- *Small*: The task can be realized by an experienced programmer in less than a day.
- *Medium*: The task can be realized in 1-3 days by an experienced programmer.
- *Large*: The task needs more than 3 days of work of an experienced programmer.

The following tasks are included in the step-by-step approach (see above):

- *Extending the Interface for Research Data Products*: The task to extend the interface of Metadata, Data, and Service in the framework itself is not very complex, but the effort to support all types of heterogeneity hidden behind these interfaces is. The effort is therefore categorized as *large*.
- *Adding Realizations for Interfaces for Research Data Products*: In contrast to the previous task, this only means to implement one class that is compliant to the interfaces defined in the framework. The effort is thus categorized as *small*.
- *Adding a new check* heavily depends on the interaction encapsulated by the check. On average it is assumed to be rather a *small* task.
- *Adding new types of results* is a *small* task.
- *Adding new types of evaluations* can be a *small* task, if the evaluation logic is not specific to a certain domain but can become a *medium* taks in cases in which customized functions and mappings must be implemented to encapsulate the complex knowhow.

[1] The categorization is motivated by the experiences made in developing the prototype. They should be considered as rules of thumb rather than precise estimations.

- *The implementation of a new scheduling logic for benchmarks* is — as discussed previously — a complex and thus a *large* task.

5.2 Exemplary Use Cases for Benchmarks

This section gives three example use cases of benchmarks for research data products. The first example is the basis for the prototypical benchmark used in the next chapter and sketched in Subsection 5.3.3. The presentation of the use cases is aligned with the first step of the recipe (see Subsection 5.1.1). The use cases are selected out of the set of possibilities since they cover the bandwidth of scenarios for which benchmarks for research data products can be used.

5.2.1 Exploring the Contents of a Repository

1. **What should the benchmark measure?**
   The benchmark should measure the overall quality of research data products stored in the Zenodo repository.[2] This repository hosts more than 1.5 million records of different type, including but not limited to data sets, figures, images, source code, conference posters, and slides for academic talks.[3]

2. **Who is the target audience of the benchmark?**
   There are two main users for the intended benchmark:

   - The most important perspective of potential users of the benchmark is the perspective of the *service provider* of the repository; the benchmark should serve as a means to gain overview of the quality of the depositions to adapt existing terms and policies of the service or to develop new or extend existing service components.
   - The second group is comprised of *researchers* interested in depositing their research data products in the Zenodo repository. This group would use the benchmark to decide whether to upload their output to Zenodo (is the repository upholding a high standard?).

3. **Is the benchmark specific to a certain workflow?**
   In general, there is no specific workflow relevant to *all* depositions, other than uploading and describing the research data product; but certain clusters in the depositions can be detected (see Section 6.1):

[2] `https://zenodo.org`
[3] For a more detailed introduction see Subsection 6.1.1.

- Zenodo includes a lot of source code depositions, since it provides an integration with github.[4] This aspect can be called "software publishing".
- Many depositions in Zenodo are conference posters; another cluster of workflows can thus be subsumed under the term "poster depositions".
- There are also many white papers, best practice reports or preprints in Zenodo; a third group of workflows can therefore be labeled "publishing grey literature".

4. **Is the benchmark specific to a certain type of actor?**
   Because of the heterogeneity of workflows, there is also no specific role that is central for this benchmark — submitters e.g. include researchers, research software engineers, data curators, information specialists and policy makers.

This use case is the basis for the prototypical benchmark presented in Subsection 5.3.3 which is used for the empirical evaluation in Chapter 6.

5.2.2 Scientometric Research

1. **What should the benchmark measure?**
   The benchmark should measure the overall impact of research data products of a large set of published research to identify variables correlating with high benchmark scores. Candidates for such variables include but are not limited to: language of the publication, native language, education, financial background, and affiliation of the creators, number and diversity of coauthors, presentation on high-profile conferences or publication in high-profile journals. An additional question to ask would be whether there are identifiable trends in the correlation over time (i.e. are some variables getting more importance since the rise of digital methods in academic research).

2. **Who is the target audience of the benchmark?**
   The target audience is mainly comprised of scientometricians. Their goals include the identification of models that can explain scientific success or failure, or the detection of common patterns of practices of research and their role in academic careers.

3. **Is the benchmark specific to a certain workflow?**
   Although publications are typically structured in a similar way across fields with re-occurring elements, such as a section discussing related work, workflows vary depending on the field and the language of publication. To refine

[4] `https://github.com`

the scope of the benchmark certain fields of study or languages of publications can be defined: the concentration on publications in the environmental sciences should define a typical structure from which common workflows can be derived such as presenting the model parameters (for computational simulations) or discussing sensor calibration and sample sizes.

4. **Is the benchmark specific to a certain type of actor?**
   The main actor mirrored in the benchmark is the researcher publishing their work. A concentration on certain fields of study might be beneficial (see previous question).

5.2.3 Continuous Integration for Research Data Products

1. **What should the benchmark measure?**
   The benchmark should measure the quality of research data products which are submitted for peer review to a publishing body (such as a journal or a repository). The submission should only be processed by a human reviewer if the submitted research data product passes a quality threshold.

2. **Who is the target audience of the benchmark?**
   There are two main users of this benchmark:

   - The *submitters* of research data products. This group wants its submission to be tested in favor for further manual processing. Submitters therefore need to get transparent access to the criteria that lead to the benchmark score, which define the basic quality criteria of the publishing body.[5]
   - The *editors* of the publishing body, organizing the peer review. Their interest is to reduce the false positive rate (ratio of submissions with a low quality which have a score above the threshold) an the false negative rate (ratio of submissions with a high quality which have a score below the threshold).

3. **Is the benchmark specific to a certain workflow?**
   In general, there is no special context of a workflow, other than defined by the submission guidelines. Given a specific publishing body, workflows can be identified analogous to the previous use case.

[5] A *de facto* benchmark is the check of latex submissions to ArXiv (`https://arxiv.org`), that is the check whether it can be built by the platform and whether all resources to build the document are provided in the correct format. Since this a non-trivial task, it can be argued that this is already filter mechanism similar to the one sketched in this use case.

4. **Is the benchmark specific to a certain type of actor?**
   The most important actor is the reviewer who has an implicit model of quality and a method to apply it to the submission in a way that is most time-efficient. This perspective should lead the implementation of the benchmark.

This use case is connected to the third question listed in Section 7.4. Possible future work might find ways to exploit benchmarks to save time and to make quality of research data products more transparent.

These three examples indicate the variety of useful applications of benchmarks to help organize the vast amount of research data products. They also stipulate a rule concerning the specificity of benchmarks: The more specific a use case is, the more informative are its results for the defined use case and the easier are requirements identifiable — but the less informative are the scores of the benchmark beyond the use case. Furthermore, the more abstract a use case is, the better it can be generalized beyond its original use case, but the harder it is to avoid an implementation of checks which arbitrarily favor research cultures and practices of one field over others.

Section 7.3 will give further remarks on the applications of benchmarks, primarily meant as recommendations to different stakeholders in the research community (researchers, service providers and funding bodies).

5.3 Components of the Prototypical Benchmark

All code described below is written in Python3[6] and published under an Apache 2.0 license. The appendix includes for instructions to retrieve and test the version of the code used by us (see Chapter 6). The publication of the source code should facilitate the reproducibility of our findings.

5.3.1 A Library for Research Data Products

The library provides the common interface for research data products (as depicted in Figure 4.2), as well as a selection of concrete implementations of these interfaces. Although the realizations are mainly motivated by the use case behind the prototypical implementation (see previous section) it can in principle be used for other use cases, even beyond the assessment of a research data product. This is the main reason, why this functionality has been published in a separate project.

Currently the following features are supported:

- Support for Metadata in DataCite format

[6] `https://python.org`

- Support for several types of Data, including CSV, PDF, and zipped source code
- Support for Services as offered by Zenodo (OAI-PMH, REST, HTTP)
- A simple set of service capacities

The publication of the source code is accompanied by documentation to ease the first steps.[7]. At present, 27 tests are part of the code. Both documentation and tests complement the content provided in this thesis.

The current version of the library is available at `https://github.com/tgweber/rdp`.

5.3.2 A Framework for Checks, Evaluations and Benchmarks

The framework enables the customized creation of benchmarks along the step-by-step approach presented in Subsection 5.1.1: it provides the base classes for Checks, Evaluations and Benchmarks along different types of Results (see also Figure 4.4, Figure 4.6, and Figure 4.8).

Additionally, it includes 38 implementations of checks, 12 implementations of evaluations and the base implementation of a Benchmark.

The framework is documented[8] and at present is distributed with 60 tests. Both resources can be consulted in addition to the documentation provided in this thesis.

5.3.3 The Prototypical Benchmark

The prototypical benchmark used for the evaluation in the next chapter is published along the framework (in the same repository and under the same license).[9] It uses at present 38 of the checks and 11 of the evaluations of the framework and adds 7 customized evaluations. It is a benchmark with static weights. A list of used pairs of evaluation and checks can be found in the appendix (Section D).

Using the modularized library and framework presented above allows to specify the prototype in a single file and concentrate on the domain aspects sketched in the use case description (Section 5.2).

The next chapter will show an evaluation of the prototype to answer SQ-4 and SQ-5.

[7] `https://github.com/tgweber/rdp/blob/master/README.md#quick-start`
[8] `https://github.com/tgweber/breadp/blob/master/README.md#quick-start`
[9] `https://github.com/tgweber/breadp/blob/master/breadp/benchmarks/example.py`

Chapter 6

Evaluation based on the Prototypical Benchmark

Contents

The empirical part of this thesis is presented in this chapter. The first section introduces Zenodo, the data source for the evaluation. Section 6.2 and Section 6.3 discuss SQ-4 and SQ-5, respectively.

6.1 Population and Samples

In this section the population — the research data products stored in the Zenodo repository – and the samples drawn from this population are characterized. The samples are not drawn at random, but at large, i.e. if a research data product hosted in Zenodo fulfilled a set of technical requirements necessary for the evaluation, it is included.

Subsection 6.1.1 provides a short introduction of the Zenodo repository, i.e. the data source of the population. The sampling mechanism is shortly discussed in Subsection 6.1.2 (adding to Section 2.3). The population and the samples are characterized along the following variables:

- the types of research data products (Subsection 6.1.3)
- the age of research data products (Subsection 6.1.4)
- the field of study of a research data product (Subsection 6.1.5)
- the different scores of event-based metrics and the distribution of scores of the benchmark run (Subsection 6.1.6)

The analysis of these variables is carried out to manage the effect of hidden variables, i.e. to have a broad understanding of the main dimensions of the analyzed samples to assess the effects of these dimensions on the results presented. The analysis of types and field of study is specifically relevant for the discussion in Section 6.2, whereas the age is relevant for Section 6.3.

6.1.1 The Zenodo Repository

Zenodo[1] is an open data repository hosted at CERN in Geneva, Switzerland. Realized in the context of the OpenAIRE project it was launched in May 2013 and was intended as a "catch-all repository for EC funded research" [NS14].[2] Zenodo accepts all kinds of research data products and can be used with an ORCiD identifier[3] to upload up to 50 GB of data per deposition (as of April 2020).

Zenodo has been discussed in scientometric literature in general ([SGS17]) and in the context of metrics ([Pet+17]). Both publications date back to 2017, a time when Zenodo's stock of data was about a tenth of its amount in April 2020.

[1] `https://zenodo.org` — the name is derived from Zenodotus, the first librarian of the ancient library of Alexandria.

[2] see also `https://about.zenodo.org`

[3] Open Research Contributor Identification Initiative, a non-profit organization handing out identifications for researchers to uniquely identify and reference their outputs, see `https://orcid.org`.

Two specific features of Zenodo need further explanation, as they affect the sampling procedure explained below: DOI versioning and DOI allocation.

DOI Versioning

Depositions in Zenodo cannot be deleted, once they are published — but they can be superseded by a newer version. Each version is identified by its own DOI, while the whole research data product has a DOI to denote the "concept" of the deposition. The metadata and the services offered by Zenodo are managed mainly by the concept-DOI which points at the latest version. The management of these versions and the conceptual representation of all versions is realized by DOI references, hence the name "DOI versioning".

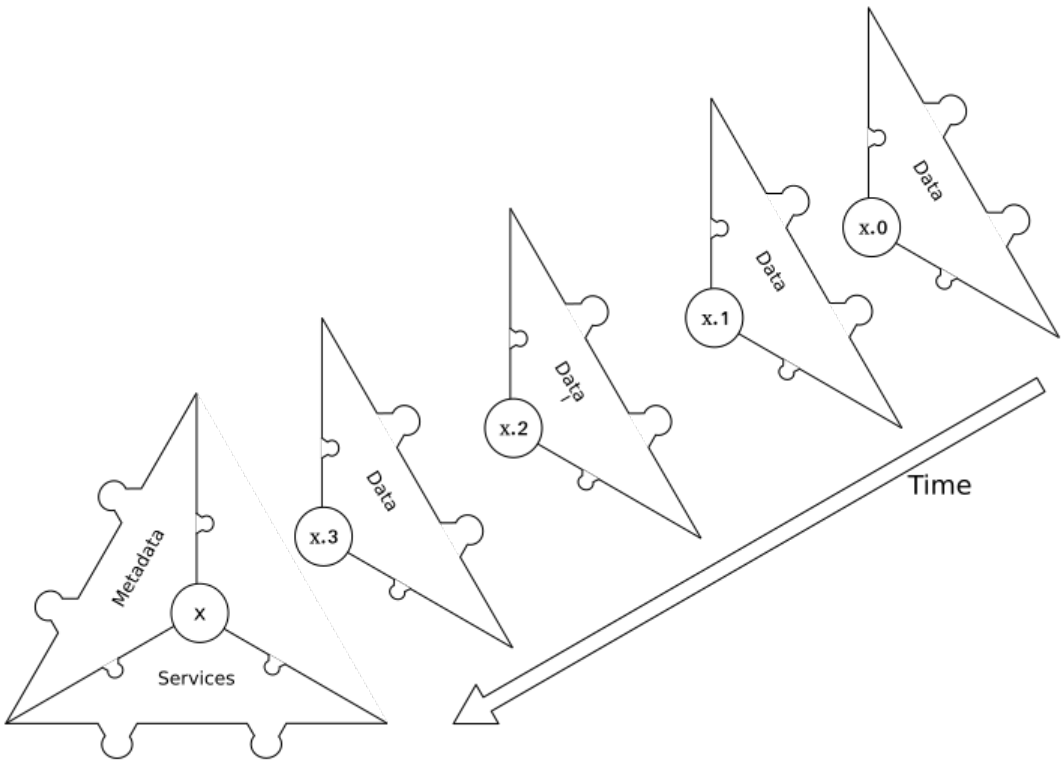

Figure 6.1: Zenodo's DOI versioning

Figure 6.1 depicts the management of versions: In the lower left corner an incomplete research data product is depicted with the PID "x" and the data component is left blank; this incomplete research data product corresponds to the concept-DOI. Closest to the incomplete research data product is the data component with the PID "x.3" depicting the latest version of the data published on

Zenodo. Combined with this latest data component the research data product is complete. Behind the latest version, prior versions of the data component are aligned in a row in the right upper direction, their PIDs are predecessors of the latest PID "x.3". Following the arrow pointing from the first to the latest versions in reverse order, means to reverse back in time to prior versions of the research data product, eventually arriving at the originally published data component. Figure 6.1 provides an imagery to understand the following considerations.

The implications of DOI-versioning need to be scrutinized, before the sampling and evaluation can be discussed:

- Only concept-DOIs are considered PIDs in the sense presented in Subsection 1.1.1.
- The prototypical benchmarks only uses the concept DOI and therefore runs on the latest version of a deposition of a research data product — the other DOIs are considered duplicates.
- The collection of social media scores and usage scores accounts for all versions, i.e. scores are summed up over versions. Usage statistics derived from Zenodo already contain the total scores, if the concept-DOI is used. The collection of social media metrics is carried out by exploiting the different DOIs (altmetrics use this as the primary identification mechanism of research data products, which means there is no threat of duplicating scores).

DOI Allocation

If a deposition uses the DOI allocation mechanism of Zenodo, the resulting DOI corresponds to this format: 10.5281/zenodo.[zenodo_id] This gives a programmatic way to derive the zenodo_id from the DOI. This identifier can be utilized to use other services of Zenodo, e.g. the API or OAI-PMH. At the time of our evaluation, 861,158 research data products have a DOI allocated by Zenodo.

Zenodo furthermore allows to deposit a research data product with an already existing DOI outside of the Zenodo namespace and therefore non-compliant to the format specified above. 664,094 research data products deposited in Zenodo have such a DOI outside of the Zenodo namespace at the time of our evaluation. There is no programmatic way to determine the zenodo_id from such a DOI, short of calling the custom API of Zenodo. In our evaluation we decided to exclude these research data products, since it means to build too much custom code for Zenodo. The main reason for this decision is not the nececssary effort itself, but the comparability to other repositories.[4]

[4]These issues are discussed in an RDA working group: `https://www.rd-alliance.org/repository-interfaces-data-analytics-rida`.

6.1.2 The Samples

The samples are not drawn at random but at large — all available research data products in Zenodo which comply to the following criteria have been included:

- The research data product has a DOI; the DOI is necessary to detect duplicates and retrieve additional event-based metrics for the research data product. 339 research data products have no DOI.

  1,524,913 research data products remain with DOI.

- There are no duplicates of the research data product (DOI-versioned research data products are considered to be duplicates of each other, see above). 65,208 have been excluded as duplicates.

  1,459,705 remain after deduplication.

- The DOI of the research data product is assigned by Zenodo (see above for a rationale). 663,753 research data products have been excluded for this reason.

  795,952 research data products remain with a DOI assigned by Zenodo.

- The benchmark score could be calculated without error. 589 research data products have been excluded for this reason.[5]

  795,363 research data products remain with successfully calulated scores. (dotted circle in Figure 6.2)

The set of research data products with successfully calculated scores consists of 52.15 percent of the 1,525,252 research data products stored in the Zenodo repository in March 2020. To answer SQ-4 (Do the scores of a prototypical benchmark correlate with event-based metrics?) and SQ-5 (Can benchmarks for research data products complement event-based metrics?) two samples are drawn at large:

- **Sample U** is comprised of those research data products which have a benchmark score *and* for which Usage metrics (therefore the U) are retrievable (see Section 2.3 for details on the retrieval). 15,555 research data products do not have view or download metrics.

  779,081 research data products have a view score *and* a benchmark score

[5]The main reason for such an error is that the specified id could not be used to successfully instantiate the metadata component of a research data product (HTTP status code 422).

779,020 research data products have a download score *and* a benchmark score.

Since both types of usage metrics are conceptually very close, they are treated as one set in the descriptive comparison between sample U and the population in the remainder of this section.

In general, when it is not necessary to differentiate between view and download metrics, sample U denotes all those research data products for which download metrics, view metrics, or both are available. They are discussed separately in Section 6.2 and Section 6.3, respectively.

- **Sample S** is comprised of those research data products which have a benchmark score *and* for which Social media metrics (therefore the S) are retrievable via the altmetrics API (see Item 2.3 for details on the retrieval and the description of each of the retrieved social media metrics). For 786,248 research data products no Social media metric was retrievable via altmetric.

  9,115 research data products have an altmteric score *and* a benchmark score.

  In sample S only the altmetric score is available for all research data products. Other social media metrics, such as tweeters or readers, are only partially available. Since these different social media metrics (including the altmetric score) are conceptually close, they are treated as one set in the comparison between sample S and the population in the remainder of this section.

  In general, when it is not necessary to differentiate between social media metrics, sample S denotes all those research data products for which altmetric scores are available. All social media metrics are discussed separately in Section 6.2 and Section 6.3.

The Venn-diagram depicted in Figure 6.2 schematically shows the set theoretical relations of the samples and the population.[6] The rectangular box depicts the population, whereas the circle with the dots stands for the scored research data products. The circle with the horizontal lines stands for all research data products for which usage metrics could be retrieved and the circle with the wavy lines stands for all research data products for which social media metrics could be retrieved. The figure should help to understand how the samples, the scored research data products and the population overlap: Sample U corresponds to the

[6]The figure does not display its elements proportionally to the size of their counterparts.

blue area with horizontal lines and dots (which partially includes some wavy lines), whereas sample S corresponds to the grey area with wavy lines and dots (which partially includes some horizontal lines.

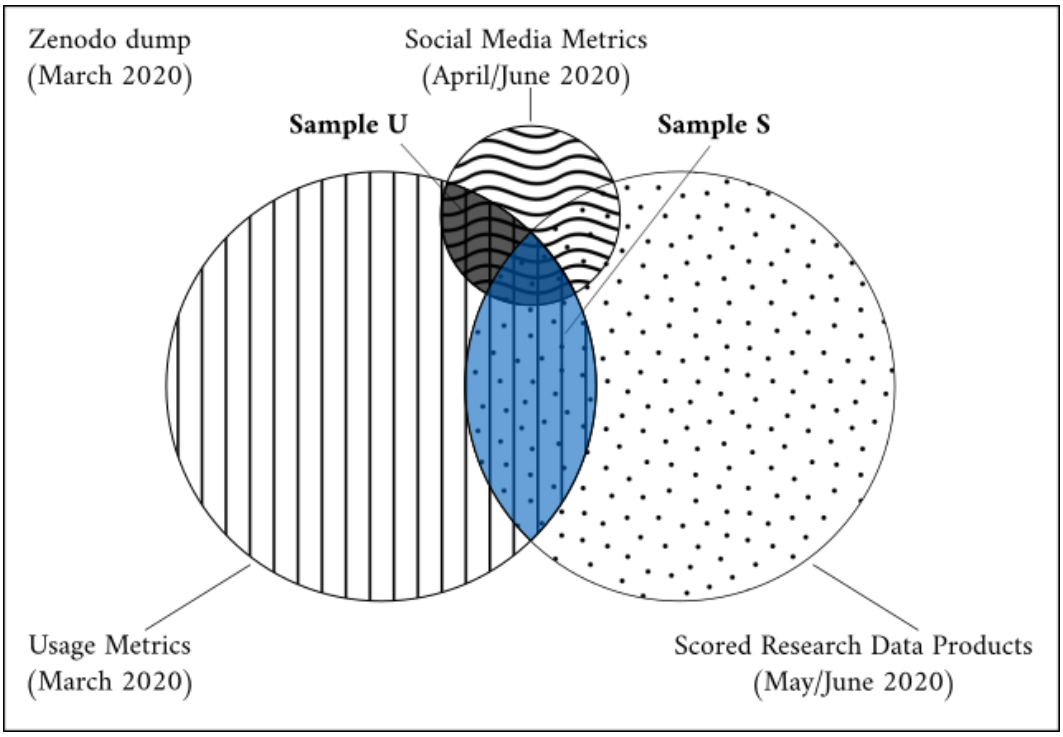

Figure 6.2: Venn diagram of the Zenodo population and the drawn samples

The next paragraphs describe the distribution of types, age, field of study, and scores of the research data products in the samples and the differences of the samples compared to the whole set of research data products in Zenodo.

6.1.3 Types of Research Data Products

Table 6.1 shows the distribution of the different types of research data products in Zenodo (in the population and both samples). Each row corresponds to one of the 9 types supported by Zenodo, columns with a heading starting with a # show the absolute number of research data products of a type, whereas columns with a heading starting with "Pct." show the percentage of research data products of a type. The numbers allow us to discuss the differences between the samples (colored areas in Figure 6.2) and the population (box in Figure 6.2).

The most noteworthy discrepancies between population and sample U are images and publications: while images are overrepresented in comparison to their

Table 6.1: Types of research data products in the population and the sub-samples

Type of research data product	# Pop.	Pct. Pop.	# U	Pct. U	# S	Pct. S
dataset	62,189	4.08 %	44,925	5.77 %	1,374	15.07 %
image	523,847	34.34 %	440,597	56.61 %	174	1.91 %
lesson	1,617	0.11 %	1,377	0.18 %	91	1.00 %
other	3,856	0.25 %	3,155	0.41 %	160	1.76 %
poster	5,509	0.36 %	5,187	0.67 %	523	5.74 %
presentation	13,558	0.89 %	12,448	1.60 %	1,274	13.98 %
publication	833,095	54.62 %	233,518	30.00 %	4,719	51.77 %
software	79,747	5.23 %	35,415	4.55 %	717	7.87 %
video	1,834	0.12 %	1,671	0.21 %	83	0.91 %

Research data products in zenodo: 1,525,252 — in sample U: 778,293 — in sample S: 9,115

occurrence in the population, publications are underrepresented. Aside from these two deviations the distributions are mostly similar. In the evaluation of the research question, the effect of images and publications on the results should be scrutinized.

In sample S images are underrepresented compared to the population, while datasets, posters, presentations are overrepresented. The distribution of types of research data products of sample S therefore deviates from the population more clearly than the distribution of sample U. In Section 6.2 the effect of publications on the correlation are discussed.

A detailed statistical summary with regard to each type can be found in the appendix (Section B, in Table B.1 to Table B.18).

6.1.4 Age of Research Data Products

Table 6.2 shows the growth of depositions of research data products at Zenodo from 2014 to 2020, each row corresponding to a year. The columns share their semantics with the columns of Table 6.1 (see above). As with the previously discussed table, deviations of the samples from the population are discussed to contextualize the assessment of our evaluation. As discussed in Section 3.2, TIME — i.e. time-dependency of event-based metrics — is a crucial shortcoming of such assessments of research data products.

In sample U the year 2018 is underrepresented, whereas 2019 is overrepresented. Sample S overrepresents 2015 while 2018 is underrepresented. According to [FC20] the half-life of social media metrics vary across their type with Twitter having an "altmetric half-life" (i.e. time span between publication and occurrence of half the tweets) of 13 days, while altmetric scores of wikipedia half-life is specified as 515 days. Since only 2019 lies in this longer time span (measured from the retrieval

Table 6.2: Years of publication for research data products

Year of deposition	# Population	Pct. Population	# U	Pct. U	# S	Pct. S
2014	4,272	0.28 %	3,399	0.44 %	38	0.42 %
2015	16,637	1.09 %	12,803	1.65 %	784	8.60 %
2016	97,542	6.40 %	83,099	10.68 %	999	10.96 %
2017	212,113	13.91 %	126,724	16.28 %	1,520	16.68 %
2018	676,860	44.38 %	117,426	15.09 %	2,342	25.69 %
2019	464,365	30.45 %	393,605	50.57 %	2,792	30.63 %
2020	53,463	3.51 %	41,237	5.30 %	640	7.02 %

Research data products in zenodo: 1,525,252 — in sampleU: 778,293 — in sampleS: 9,115

of the altmetric data), the discussion of "younger" research data products is an important part in the discussions below. In contrast to the year of publication, a correlation between age in days and the obtained score of a research data product allows a more fine-grained analysis (see Table 6.6).

6.1.5 Field of Study

Table 6.3 is similar to Table 6.1 and Table 6.2: each row stands for a field of study, and the columns characterize the absolute and relative distribution among the field of study. The table displays the distribution of research data products among fields of study for those subsets of the population (33.24 %), sample U (44.63 %), and sample S (57.83 %), that allow to be classified.[7] The number of research data products for each field add to more than the total in each set and the percentage to more than 100 %; this is due to the multi-label classification used, which allows to label a research data product with more than one field of study.[8] Like in the previous subsections, the purpose of Table 6.3 is to identify deviations between population and samples, and biases (in this case towards single fields); some research data products will occur more than once (especially in Section C), but only when field-specific distributions are discussed.

The distribution of research data products over fields of studies in the population is heavily biased (most prominently towards biological sciences),[9] Sample U and the classified research data products roughly follow the same distribution.

[7] A research data product needs to have a title and/or description in English of at least 10 words to be classified.

[8] For the classification the "mlp_l" classifier was used from `https://github.com/tgweber/fosc` in version 0.0.1. An appraisal of the estimated errors of the classifier is given in Section C — especially Table C.1 — see also [Web+20].

[9] This is in accordance with many other findings in the literature see [Wal+11a] [Kra+15b] [Pet+16] [GHA18] [Web+20].

Table 6.3: Fields of study of research data products in population and in the sub-samples

Field of Study	# Pop.	Pct. Pop.	# U	Pct. U	# S	Pct. S
Mathematical Sciences	8,655	1.71 %	6,754	1.94 %	83	1.57 %
Physical Sciences	22,670	4.47 %	11,621	3.35 %	251	4.76 %
Chemical Sciences	9,918	1.96 %	6,372	1.83 %	123	2.33 %
Earth and Environmental Sciences	57,755	11.39 %	35,214	10.14 %	485	9.20 %
Biological Sciences	237,332	46.81 %	158,045	45.50 %	1,032	19.58 %
Agricultural and Veterinary Sciences	1,443	0.28 %	1,057	0.30 %	30	0.57 %
Information and Computing Sciences	61,057	12.04 %	42,410	12.21 %	1,794	34.04 %
Engineering and Technology	61,305	12.09 %	46,391	13.36 %	509	9.66 %
Medical and Health Sciences	74,394	14.67 %	48,669	14.01 %	674	12.79 %
Built Environment and Design	2,083	0.41 %	1,709	0.49 %	27	0.51 %
Education	4,483	0.88 %	3,900	1.12 %	121	2.30 %
Economics	4,784	0.94 %	3,641	1.05 %	58	1.10 %
Commerce, Management, Tourism and Services	8,087	1.59 %	6,555	1.89 %	74	1.40 %
Studies in Human Society	8,934	1.76 %	7,580	2.18 %	199	3.78 %
Psychology and Cognitive Sciences	5,864	1.16 %	4,163	1.20 %	116	2.20 %
Law and Legal Studies	1,486	0.29 %	1,130	0.33 %	25	0.47 %
Studies in Creative Arts and Writing	626	0.12 %	572	0.16 %	12	0.23 %
Language, Communication and Culture	7,282	1.44 %	5,929	1.71 %	183	3.47 %
History and Archaeology	1,962	0.39 %	1,662	0.48 %	55	1.04 %
Philosophy and Religious Studies	672	0.13 %	576	0.17 %	29	0.55 %

Classifiable research data products in zenodo: 507,049 — in sampleU: 347,355 — in sampleS: 5,271

Sample S underestimates Biological Sciences and overestimates Information and Computing Sciences.

A detailed statistical summary with regard to each field of study can be found in the appendix (Section C, Table C.2 to Table C.61). These summaries allow to interpret the results in the light of cultural patterns of each field of study.

6.1.6 Scores of Benchmark and Event-Based Metrics

In the previous subsections the distribution of variables in the population and the samples have been discussed. Different event-based metrics are clustered above, that are scrutinized in detail in this section.

Table 6.4 offers common statistics of the distribution for the scores of research data products. Each row stands for a event-based metric except the last one which shows the distribution of the scores of the prototypical benchmark. "Views" and "Downloads" are subsets of sample U, while "Tweeters, "Readers", "Facebook Walls", "Feeds", "Posts" and "Altmetric Score" are subsets of sample S. The following list explains each column for a event-based metric e (examples for e are the number of views or the number of readers):

- #: Number of research data products with at least 100 values of e, for social

media metrics, each row includes only those research data products out of SampleS with a score higher than 0.

- **Min**: Minimum value of e, e.g. lowest count of views
- **1st**: First quartile, i.e. smallest value which is larger or equal than 25 % of the values in e
- **Med**: Median or second quartile, i.e. smallest value which is larger or equal to 50 % of the values in e
- **Mean**: Arithmetical mean for all e
- **3rd**: Third quartile, i.e. smallest value which is larger or equal than 75 % of the values in e
- **Max**: Maximum value of e, e.g. highest count of views
- **SD**: The standard deviation is the root of the arithemtical mean of squared distances to the arithmetical mean of e. It indicates the grade of dispersion of e
- **# Char**: Number of characteristic values of e, i.e. the number of all values e includes.

Table 6.4: Scores of research data products in the sample

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	779,081	0.000	2.000	4.000	20.119	13.000	216482.000	379.052	2,088
Downloads	779,020	0.000	1.000	2.000	13.676	8.000	81994.000	226.117	1,664
Tweeters	7,887	1.000	1.000	2.000	6.933	6.000	874.000	24.141	130
Readers	1,725	1.000	1.000	2.000	7.729	7.000	925.000	29.136	76
Facebook Walls	580	1.000	1.000	1.000	1.302	1.000	11.000	0.980	8
Feeds	681	1.000	1.000	1.000	1.203	1.000	18.000	0.859	6
Posts	9,115	1.000	1.000	2.000	7.593	6.000	1477.000	30.488	160
Altmetric Score	9,115	0.250	1.000	1.850	5.307	4.750	632.400	17.927	836
Benchmark Score	795,363	0.400	0.586	0.607	0.601	0.621	0.759	0.040	2,460

Only displaying those event-based metrics with at least 100 instances — n = 795,363

All distributions of event-based metrics show a skewness to a few small values (e.g. 50 % of research data products have been downloaded at most twice) and a high rate of dispersity with extreme outliers.[10] The skewness is important for the

[10]For these reasons the distributions are not graphically displayed, since box plots or cumulative frequency distribution diagrams are rather perverted to un-informative clusters of pixels.

discussion of correlation, since the order of research data products induced by the scores do not exhaust the full space of the characteristic values: e.g. 75% of the views are distributed in only 0.62 % of the possible values of the distribution.

The benchmark scores on the other hand are distributed symmetrically (median and mean are almost identical) with a low rate of dispersion, but equally distributed over the values.

6.2 Correlation between Event-based Metrics and Benchmarks

In this section SQ-4 will be answered: "Do the scores of a prototypical benchmark correlate with event-based metrics?" With a weak correlation (between 0.2 and 0.4) this would indicate that the prototypical benchmark measures features similar to those typically associated with event-based metrics (e.g. quality, relevance, or impact of a research data product). If the correlation is too low or even negative the evidence would suggest to reject such an hypothesis. A stronger correlation would counter-indicate the assumed complementariness of the two types of assessment of a research data product (see next section below).

Table 6.5 shows the measured correlation between benchmark scores and the event-based metrics also displayed in Table 6.4.[11] Each row corresponds to one of the event-based metrics. The columns starting with "S" display the spearman rank correlation (value range -1 to 1), while the columns starting with "#" show the size of the sub-sample. The two columns ending with "all" show the values for the whole sub-sample, while "w/o images" and "w/o Bio" denote the values for the samples without images and those without research data products classified as "Biological Sciences". The samples are skewed towards these two values (see Subsection 6.1.3 and Subsection 6.1.5).

The displayed values indicate that in general, only for Tweeters, Posts and Altmetric Scores the question can be answered positively (which are all social media metrics). Views are correlated weakly in negative direction.

The impact of images on the reported correlation is very strong: without images a weak correlation is indicated also for Downloads (a usage metric) and the Views are not weakly correlated negatively. There is a statistical approach to explain this: The distribution of all event-based metrics for images is more skewed compared to other types of research data products (see Table B.3 in the appendix). The correlation of benchmark scores with the number of downloads is larger than or equal than 0.08 for all other types (see Section B).

[11] The row "Benchmark Score" is missing, since this is the value the correlation is calculated against.

Table 6.5: Correlation with benchmark scores for all research data products

Event-Based Metric	S all	# all	S w/o images	# w/o images	S w/o Bio	# w/o Bio
Views	-0.248	779,081	0.069	337,696	-0.241	621,034
Downloads	-0.099	779,020	0.270	337,699	-0.080	620,964
Tweeters	0.253	7,887	0.250	7,725	0.242	7,041
Readers	-0.063	1,725	-0.061	1,718	-0.063	1,551
Facebook Walls	0.025	580	0.035	575	0.037	508
Feeds	0.079	681	0.076	677	0.069	641
Posts	0.247	9,115	0.245	8,941	0.236	8,083
Altmetric Score	0.197	9,115	0.193	8,941	0.185	8,083

Only displaying those event-based metrics with at least 100 instances

The impact on the score of event-based metrics of research data products classified as Biological Sciences is negligible, since the distribution of the samples "w/o Bio" is very similar to the complete sample. In general the reported correlation varies between fields of study, even inside fields typically clustered together (such as sciences or humanities): While Chemical Sciences, Computer and Information Sciences, and Studies in Human Society show a weak correlation for Downloads with the benchmark scores, values in Physical Sciences and Studies and Creative Arts and Writing indicate no correlation, and Values in Mathematical Sciences, Economics, and Law and Legal Studies show negative numbers. There is no straightforward, i.e. statistical explanation for this variation (see Section C for an detailed display of correlation measures by field of study). The majority of fields support a weak correlation between Tweeters and benchmark scores.

Answer to **SQ-4**:
There is a weak correlation between the scores of the prototypical benchmark and some social media metrics (Tweeters, Posts); if all types of research data products except images are considered, there is evidence for a weak correlation between scores of the prototypical benchmark and download metrics.

6.3 Complementariness of Event-based Metrics and Benchmarks

This section discusses the answer to SQ-5: "Can benchmarks for research data products complement event-based metrics?" If benchmarks for research data products can complement event-based metrics, their continued development is justified

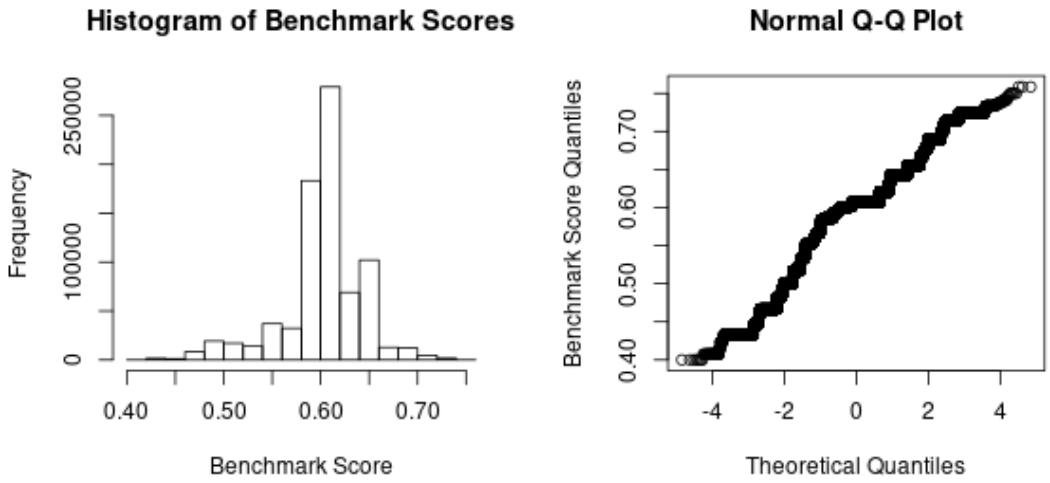

Figure 6.3: Normal distribution of Benchmark scores

to improve applications currently relying only on event-based metrics.

The answers to SQ-4 and SQ-5 are not independent of each other: The stronger the correlation between benchmarks for research data products and event-based metrics is, the smaller is the capacity of benchmarks to complement event-based metrics; vice versa, the higher is the robustness of the assessment. The stated weak correlation above leaves an open window for this complementarity. Since a difference in information content that does not suffice to support a strong correlation on the one hand, can be used to improve the applications of event-based metrics on the other hand; an example would be the improvement of an ordering of research data products with regard to quality, impact or relevance. The complementarity of event-based metrics and benchmark scores can be exploited in a multidimensional assessment context, as sketched in [MH15].

As discussed in Subsection 3.3.3, the most informative shortcomings of event-based metrics for an architecture of benchmarks for research data products are BAND and TIME. The two following subsections provide empirical evidence that scores of benchmarks are not susceptible or not as susceptible to them as scores of event-based metrics and suggest ways to exploit these insights.

6.3.1 No BANDwagon effect for Benchmark SCORes

BAND, i.e. the sensitivity to social effects, such as the bandwagon or Matthews effect can be identified by few extreme outliers of the distribution of scores. Table 6.4 provides evidence for this characteristic for all displayed event-based metrics: the

mean is clearly right of the median and the standard deviation indicates a high dispersion. The statistical values for benchmark scores rather indicate a symmetrical distribution (its mean is almost identical to its median and the standard deviation is rather small).

It is obvious to test whether the distribution of benchmark scores resembles a normal distribution. Figure 6.3 provides two figures to support the assumption of normality: On the left there is a histogram (the x-axis shows the benchmark scores in clustered values, whereas the height on the y-axis corresponds to the cardinality of the values), which can be interpreted as having an approximate bell shape. On the right, a qq-plot approximates a line (a scatter plot of pairs of quantiles of the benchmark scores with the quantiles of a artificially created normal distribution).

Due to their approximate normal distribution, benchmarks for research data products do not show the symptoms of BAND. This fact can be exploited when said extreme outliers are to be assessed: does the score of the benchmark justify its exposition? If the benchmark is customized for the type of research data product assessed a mathematical model for correction is in reach (e.g. multiplying the score of the event-based metric with the score of the benchmark).

The different distribution patterns of benchmarks for research data products and event-based metrics can be exploited further: another shortcoming of event-based metrics is COR, which is the doubtful correlation of scores of event-based metrics with quality, impact, or relevance of a research data product. Statistical indication for this shortcoming includes clusters around a small set of small values, since it is a symptom of missing differentiation between research data products. Scores of benchmarks are approximately distributed normally in the prototypical evaluation *and* they are weakly correlated with some event-based metrics; they can therefore be used as a correction factor in the context of COR.

Figure 6.4 shows four scatter plots of benchmark scores with downloads, views, Tweeters and Posts. All four are clustered on the x-axis, i.e. on the scores of event-based metrics, but approximately distributed normally on the y-axis. This depicts the new information introduced by the benchmark scores: research data products which are hardly distinguishable or even indistinguishable with their event-based metrics' score, can now be separated with the help of the second dimension. If we assume that the scores of a benchmark for a research data product is an (imperfect) signal for quality, impact, or relevance of a research data product, a two-dimensional analysis of research data products will provide better results (e.g. when research data products should be ordered in a search result, or filtered by a "quality threshold"). Again, the simplest mathematical model to achieve this is the multiplication of scores.

The upper left quadrant of the four plots in Figure 6.4 is a promising place to look for so-called "sleeping beauties", i.e. research data products which are of

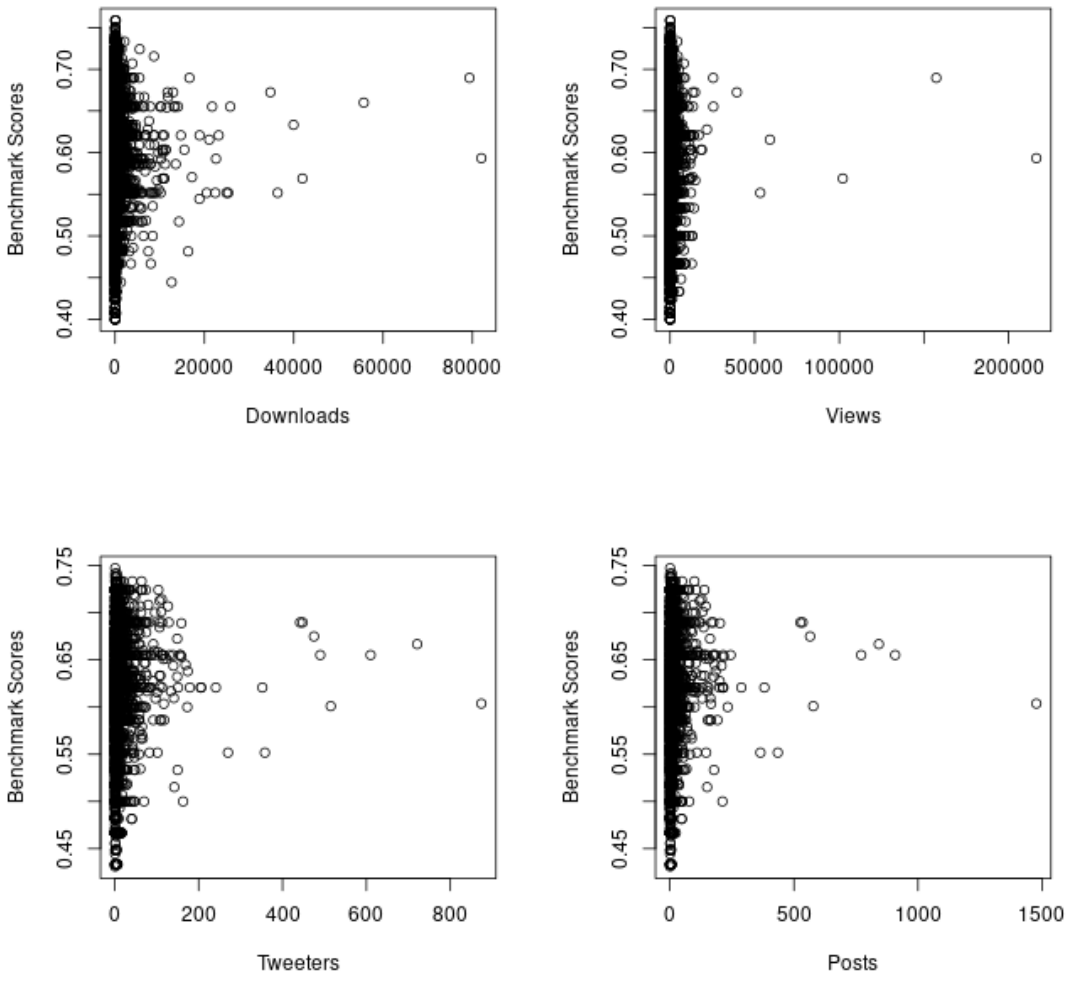

Figure 6.4: Scatter Plots of Benchmark Scores and Scores of Selected Event-Based Metrics

high quality, impact, or relevance, but are not singled out by event-based metrics. The lower right quadrant is not populated in the four examples, but this area is a reasonable starting point to identify "bad examples", i.e. research data products which have a high attention score not because but despite their low quality, impact, or relevance.

6.3.2 TIME-Independence of Benchmark Scores

TIME is the second shortcoming of event-based metrics that was crucial for the architectural design of benchmarks for research data products presented in Chapter 4. Its main impact on scores of event-based metrics concerns the availability of the scores. The concept of a "half-life" of an event-based metric as e.g. discussed in [FC20] for social media metrics shows the expected latency time between publication and the point in time when half of the expected events have occurred. By design benchmarks for research data products do not share this shortcoming, since they can be executed any time. Benchmarks for research data products can therefore not only complement event-based metrics with regard to data availability, but offer a context of application which is unreachable to event-based metrics: assessment of a research data product before or shortly after its publication.

The TIME-dependence of event-based metrics has additional consequences that exceed simple matters of practicability: event-based metrics' scores are a function of the age of a research data product, since they are all based on the occurrence of events which cannot be forced to happen without begging the purpose of the assessment of a research data product.

Table 6.6: Correlation of score and age in days

Metric	Spearman	n
Views	0.462	779,081
Downloads	0.348	779,020
Tweeters	-0.134	34,312
Readers	0.113	1,725
Videos	0.173	85
Facebook Walls	0.016	580
Feeds	0.024	681
Posts	-0.148	9,115
Altmetric Score	-0.055	9,115
Benchmark Score	-0.345	795,363

Table 6.6 shows the Spearman correlation between the scores of assessments and the age of the research data product in days (measured from at the day the

Zenodo dump was created, n shows the size of the sample).[12] The half-life of a social media metrics is suggested by [FC20] to model the time-dependant behavior of these metrics; especially tweets show an extreme behavior, since they tend to occur either around the time of publication of the research data product or not at all (e.g. [SPB12] provides evidence for this claim for preprints on arXiv, as does the correlation depicted in Table 6.6). Usage metrics seem to develop more continuously, hence a larger correlation of scores for views and downloads compared to social media metrics. In total, scores of benchmarks seem to show the reverse behavior compared to usage metrics and the negative correlation between age and score is even larger than for all social media metrics. A hypothetical reason for this phenomenon might be that the younger a research data product is, the more likely it is to comply to the evaluation framework that underlies the benchmark. Raised awareness for data literacy, stricter requirements of publishers and increasing support of service providers for researchers are possible explanations.

Finally, Table 6.7 shows location and a dispersion measure stratified by year (not accumulative) for selected event-based metrics and scores of benchmarks for research data products: each row corresponds to a year. The first letter of a column corresponds to the type of score:

- v stands for views
- d stands for downloads
- p stands for posts
- t stands for tweeters
- s stands for score of a benchmark

All but the first letter indicate the type of statistic displayed:

- m stands for the median, a robust location measure (robust with regard to outliers)
- sd stands for standard deviation, a dispersion measure

Table 6.7 shows that the scores of event-based metrics heavily vary with age: while views and — with one outlying exception — downloads are accumulated continuously, social media metrics work the other way around, the more recent the more uptake. All event-based metrics show volatility with regard to their dispersion, i.e. the standard deviation shows that there is typically a large influence of extreme outliers in every year. This makes the time-normalization of a score a challenge.

[12] 2020 is omitted since its the values would obviously biased due to its incompleteness.

Table 6.7: Location and Dispersion Measures for Scores by Year

year	vm	vsd	dm	dsd	pm	psd	tm	tsd	sm	ssd
2014	13	66.49	13	276.95	1.00	10.95	1	10.27	0.59	0.05
2015	19	404.35	9	145.18	2.00	18.03	2	13.70	0.56	0.05
2016	10	268.74	5	159.73	2.00	21.33	1	15.37	0.59	0.04
2017	8	277.06	5	227.64	2.00	56.89	2	41.74	0.60	0.04
2018	7	225.65	22	219.41	2.00	38.51	2	38.01	0.60	0.05
2019	2	318.24	2	258.81	3.00	98.25	2	80.75	0.61	0.03

The last two columns show that the benchmark for a research data product does not suffer from a comparable time-dependence (although it is weakly correlated with the age of a research data product). Median and stand deviation are almost constant.

This fact can be exploited, when event-based metrics are normalized with regard to time. In this sense benchmarks for research data products complement event-based metrics also in the context of NORM, which is the challenge to normalize scores of event-based metrics for the sake of comparability. As a side note, the design of the architecture for benchmarks for research data products also helps to mitigate TRST (missing trustworthiness of reported scores) and CTXT (missing context to interpret the scores): the reporting component of the architecture allows to reproduce the scores and offers a lot of context to assess the suitability of the benchmarks for a given use case.

Answer to **SQ-5**:
Benchmarks for research data products can complement event-based metrics with regard to BAND, COR, TIME, NORM, TRST, and CTXT.

In this chapter, the empirical evaluation necessary to answer SQ-4 and SQ-5 have been given. The results of this chapter and the previous chapters will be summarized in the next chapter and critically appraised.

Chapter 7

Conclusions and Future Work

Contents

This final chapter concludes by summarizing the contributions and discussing the main findings; it closes by identitfying the next steps that can be taken from here. The main findings are summarized in Section 7.1. and critically appraised in Section 7.2. Recommendations for different stakeholders of research data management are presented and discussed in Section 7.3. Section 7.4 identifies open questions and possible next steps. The chapter closes with a final outlook.

7.1 Summary

This thesis answers the question why research data products should be benchmarked and how. Chapter 3 and Chapter 6 answer the first part ("why"), while Chapter 4 and Chapter 5 answer the second part ("how"). This subsection briefly summarizes the main findings of those chapters.

In Chapter 3 the current state-of-the-art to assess research data products is introduced, namely event-based metrics, with 3 major sub-classes: citation-based metrics, social media metrics and usage metrics. 10 common shortcomings of event-based metrics are identified in an extendable and reproducible manner by an analysis of 62 publications. These shortcomings are classified concerning possible mitigations. Two shortcomings, BAND (social sensitivity of event-based metrics) and TIME (dependence on the temporal occurrence of events) are identified as *principal* shortcomings, i.e. shortcomings that cannot be mitigated for event-based metrics with the technical means available.

Benchmarks are introduced as an alternative to event-based metrics in Chapter 4. This chapter includes the methodic derivation of an architecture for these benchmarks for research data products by defining a framework of five main features and 19 sub-features which are satisfied by 5 main components: research data products, checks, evaluations, benchmarks, and reports. The architecture and components are motivated by the shortcomings discussed in Chapter 3, the discussion of related work in Chapter 2, literature on benchmarking in other fields, and the general challenges in research data management (see Section 1.3). Chapter 5 provides the means to realize this architecture, namely a step-by-step recipe and the presentation of a prototype.

Chapter 6 provides empirical evidence for the hypothesis that benchmarks for research data products have the potential to complement event-based metrics. A prototypical benchmark is used to score more than 795,000 research data products from Zenodo, an open repository hosting more than 1.5 million depositions. The resulting scores are compared to retrieved social media metrics and usage metrics. There is a weak correlation between certain types of social media metrics (tweeters and posts) and usage metrics (downloads, if images are excluded). The statistical analysis of scores of event-based metrics and benchmarks suggest that benchmarks can complement event-based metrics with regard to BAND, TIME and other shortcomings, namely COR (the doubtful correlation of scores of event-based metrics with quality, impact, or relevance of a research data product), TRST (missing trustworthiness of reported scores) and CTXT (missing context to evaluate the aptness of a score for a given use case).

In the light of this summary, the research question of this thesis can be answered:

Our results suggest that research data products should be benchmarked, because benchmarks have the potential to complement event-based metrics to navigate the creolized and fast-developing landscape of research data management.

Along with the thesis all resources to build customized benchmarks are published. Benchmarks are a new tool to assess research data products ready to use.

7.2 Discussion

Benchmarks for research data products offer machine-actionable and reproducible ways to assess research data products. Their reporting component allows users to recalculate, understand, and compare the score of a research data product. Additionally, the evaluation carried out in Chapter 6 shows that they are equipped to scale with large numbers of research data products. As indicated in [WK18], benchmarks allow to translate claims about machine-actionability into code and thus to specify vague descriptions with concrete implementations.

In Chapter 2, 13 quality criteria are identified for the three methodological approaches that produce the outcome summarized above. The approaches are presented and contextualized in Section 2.1, Section 2.2, and Section 2.3. The three following subsections discuss the approaches, the quality criteria and the findings of Chapter 3 to Chapter 6 in the same order; each subsection maps to the identically named subsections of Chapter 2.

Table 7.1 to Table 7.3 offer a tabular overview of the compliance of the work presented with the quality criteria. Each row corresponds to a criteria and the first column lists the identifier of the criterion assigned in Chapter 2. The remaining columns provide the description of the quality critieria ("Description"), the compliance of this thesis with the criteria ("Compliance", indicated by a ✓) and a column with references to the parts of this document that justify the appraisal ("References"). The tables show that the results of this thesis close gaps in the literature as discussed in Chapter 2.

7.2.1 Categorize Event-Based Metrics and Their Shortcomings

The rationale of Chapter 3 is to analyze the conceptual "gaps" of event-based metrics to motivate the effort necessary to implement benchmarks for research data products. Main objections to our approach might include doubts about the completeness of the enumeration of shortcomings and the way they are clustered together as types.

The first objection can be substantiated by evidence for a shortcoming that is not listed in Section 3.2. The two possible implications of such an objection do not harm the findings of this thesis: Either benchmarks for research data products can also handle the additional shortcomings, in which case the motivation to use benchmarks to assess research data products will only be increased; or benchmarks for research data products cannot handle them, in which case the motivation justified by the already described shortcomings will not cease to exist.

The second objection, namely criticizing how the shortcomings are clustered together based on the evidence given in the appendix (Section A), can be substantiated by another clustering. But even if the evidence is reordered and presented in a different manner, the motivation and discussions of Section 3.2 is not dependent on the presentations of the shortcomings. TIME, the dependence of event-based metrics on the discrete event of publication, will prevail as a principal shortcoming, even if this shortcoming is clustered together with other shortcomings, or split into separate shortcomings. The same line of thought applies to the other shortcomings: their mode of presentation does not affect the argument for benchmarks for research data products.

Objections substantiated with the evidence described above can extend the results of Chapter 3, since this thesis complies to SQC-1 and SQC-2 (see Table 7.1).

Table 7.1: Compliance of this thesis with the quality criteria from Chapter 2.1

Identifier	Description	Compliance	References
GQC-1	Does the approach concern *research* data as opposed to data in general?	✓	Section 3.1
GQC-2	Does the approach concern publications, code, other data or a combination thereof?	✓	Section 3.1
SQC-1	The literature review discusses all types of event-based metrics, not only one group.	✓	Section 2.1
SQC-2	The set of shortcomings of event-based metrics can be reproduced and is systematically extendable.	✓	Section 2.1, Section A + Section I

7.2.2 Design Benchmarks for Research Data Products

Chapter 4 provides a methodic derivation of an architecture for benchmarks, while Chapter 5 shows how such an architecture can be realized and subsequently be implemented. The main rationale behind these chapters is to show that benchmarks

can be realized based on the methodic discussions of shortcomings of event-based metrics.

A major objection to the results of Chapter 4 and Chapter 5 might consist of pointing out that there is a better way to realize benchmarks for research data products than presented in Section 4.2, or a better recipe to realize them than presented in Section 5.1. Such an objection can be substantiated with an alternative architecture or recipe and an empirically testable claim how these superseded the solutions presented here.

We expect that evolving technical possibilities and stricter scientific requirements produce such an architecture (or recipe) eventually; but it is doubtful that the improved versions are built on weaker requirements than presented in this thesis. Newer versions will most likely include the main requirements and considerations laid out by us and build upon them, which in turn does not invalidate the findings of Chapter 4 and Chapter 5, but incorporates them.

Table 7.2: Compliance of this thesis with the quality criteria from Chapter 2.2

Identifier	Description	Compliance	References
GQC-1	Does the approach concern *research* data as opposed to data in general?	✓	Subsection 4.1.1
GQC-2	Does the approach concern publications, code, other data or a combination thereof?	✓	Subsection 4.1.1
BQC-1	Do the requirements for an architecture of benchmarks for research data products mirror the concept of machine-actionability?	✓	F-D2/2
BQC-2	Is the design flexible enough to support different assessment frameworks?	✓	Section 5.1
BQC-3	Does the design include all components of a research data product?	✓	Figure 4.2
BQC-4	Has a prototype been implemented and is its source code available?	✓	Subsection 5.3.3, Section I
BQC-5	Is the prototype evaluated against shortcomings of event-based metrics?	✓	Section 6.3, Subsection 7.2.2 + Subsection 7.2.4

BQC-5 includes an evaluation against the shortcomings of event-based metrics that is only carried out partially in Section 6.3. The remainder of this subsec-

tion completes this discussion with one exception (GAME is discussed in Subsection 7.2.4).

Since the architecture for benchmarks for research data products presented in Chapter 4 was informed by a systematic discussion of the shortcomings, benchmarks by design do not share many of the shortcomings of event-based metrics, most prominently there is no dependence on the occurrence of events (TIME) and the effects of social dynamics are nowhere near the extreme of event-based metrics (BAND). As argued in Section 6.3, there is evidence that benchmarks for research data products have the potential to mitigate COR, NORM, TRST, and CTXT.

COV, the coverage of all events by an event-based metric, does not apply to benchmarks for research data products for conceptual reasons (benchmarks are not based on events). DUP and VER have been avoided by simple design decisions, namely the holistic stance of the concept of a research data product (hosting on different services imply a different research data product entity and the reporting of timestamps in reports manages different scores due to different versions).

The only remaining shortcoming that is not discussed above is GAME, which is discussed in Subsection 7.2.4.

7.2.3 Evaluate Correlation and Complementarity

Chapter 6 completes the findings of Chapter 3 to answer the question why research data products should be assessed by benchmarks. It does so by providing empirical evidence that there are circumstances under which event-based metrics weakly correlate with benchmarks for research data products and that the two types of assessments can complement each other.

Table 7.3 shows the compliance of the approach taken in Chapter 6 with the quality criteria defined in Section 2.3.

Discussion of Low and Negative Spearman Correlation

There are three major objections against the conclusion that benchmarks for research data products measure similar features compared to event-based metrics:

1. There is a negative Spearman correlation of usage metrics for images and scores of the prototypical benchmarks.

2. The Spearman correlation between view counts and scores of the prototypical benchmark are smaller than 0.1 (even with images excluded)

3. The correlation between some Social Metrics (Readers, Facebook Walls and Feeds) and scores of the prototypical benchmark is between -0.1 and 0.1.

Table 7.3: Compliance of this thesis with the quality criteria from Chapter 2.3

Identifier	Description	Compliance	References
GQC-1	Does the approach concern *research* data as opposed to data in general?	✓	Section 6.1
GQC-2	Does the approach concern publications, code, other data or a combination thereof?	✓	Subsection 6.1.3, Section B
EQC-1	The sample is drawn reproducibly from a large collection.	✓	Section 6.1, Section I
EQC-2	The sample is statistically described to manage the effect of hidden variables.	✓	Section 6.1
EQC-3	The correlation should be measured by a statistic that does not assume a linear transformation between the spaces of the compared metrics.	✓	Section 6.2
EQC-4	The indicators of complementarity are rooted in one or several of the shortcomings described in Chapter 3.	✓	Section 6.3

All three are counter-indicative to the claim that benchmarks for research data products and event-based metrics measure similar features. The next paragraphs discuss the impact of this counter-indication.

As already pointed out in Subsection 6.1.3, *images* are overrepresented in the sample, and (as pointed out in Section 6.2), the distribution of images are more skewed than the scores of other event-based metrics. There are recommendations to deposit images in open repositories such as Zenodo under an open license, even the containing publication is closed access — this way the images can be used in presentations without copyright infringement.[1]

This might explain the low numbers of views and downloads of images, since the bigger part of the accesses to the content of the images might happen via the containing publication, which is also prominently advertised and cited.

An explanation for the missing correlation of scores of the prototypical benchmark for non-images and *view* counts is the ease to game views compared to all other event-based metrics. Creating a high number of views for a research data

[1] `https://medium.com/@malte.elson/retaining-copyright-for-figures-in-academic-publications-to-allow-easy-citation-and-reuse-77c6e2b511fe` [retrieved 2020-06-15]

products requires less resources than creating high number of downloads, especially for research data products of large size. Other shortcomings (other than GAME) which can also be affected by the ease to produce high view counts are COR and BAND. It is noteworthy that even in the sample presented in Chapter 6, download counts and view counts "only" correlate with a Spearman correlation of 0.56.

The last "problematic" class of event-based metrics without at least a weak correlation (*Readers, Facebook Walls* and *Feeds*) all share a relatively small sample size (see Table 6.5). On the one hand this is evidence for COV, the shortcoming of event-based metrics to not cover all or enough research data products of interest; on the other hand it has to be noted that the Spearman correlation for readers and facebook walls are not negative and none of the three event-based metrics have a correlation that would contradict the complementarity of event-based metrics and benchmarks for research data products. The evidence presented in Chapter 6 justifies the claim that with Zenodo and the prototypical benchmark a part of the research question remains unanswered due to the lack of data.

Selection of Prototype, Event-Based Metrics and Population/Sample for the Evaluation

Another objection against the findings in Chapter 6 concerns the selection of the DataCite best practice guide as source for the prototypical benchmark, of the Zenodo repository as a population to draw a sample for the evaluation, and of the event-based metrics to evaluate correlation and complementarity. The objection is similar in all three cases: The evaluation does not sustain the claim that benchmarks for research data products correlate with event-based metrics outside of the chosen context for evaluation.

The objection in principle is justified, but we do not claim to provide evidence that benchmarks for research data products and event-based metrics relate to each other in all possible circumstances. The claim is weaker: there are circumstances in which the two modes of assessment weakly correlate and there are circumstances in which they can complement each other. This is an existential statement, it is not universal.

A benchmark moves on a spectrum between being *general*, i.e. applying to many research data products with the limitation that its result might be less informative for each research data product, and being *specific*, i.e. applying to few research data products with more informative scores. The prototypical benchmark is rather general, it is therefore possible that future benchmarks differ in their relation to event-based metrics. To measure the extent to which modes of assessments of research data products indeed measure quality, impact, or relevance additional qualitative research is necessary (e.g. interviews). This thesis qualifies benchmarks

for research data products as promising evaluation candidates for such a research project.

7.2.4 Threats to Validity

In this subsection, principal shortcomings of benchmarks for research data products are discussed. This discussion is necessary to evaluate how these shortcomings affect the ability of benchmarks to indicate quality, impact, or relevance of a research data product.

Gaming with Benchmarks

As cited in Section 3.3, Goodheart's law says, that when a measure becomes a target, it ceases to be a good measure [Str97]. While the radical form of this stance invites doubts concerning its validity, the danger of gaming for the informative value of scores of benchmarks for research data products cannot be denied. This danger is not immanent in our evaluation, since there was no specific incentive for research data products to achieve a high score in the prototypical benchmark. But given that benchmarks for research data products may become a widely used tool for the assessment of a research data product, the threat of gaming becomes a challenge.

Two types of "meddling with scores" are possible: improving the effort put into the creation and curation of research data products is one of them, which is indeed the good type of gaming: an increased score of a research data product then mirrors improved quality, impact, or relevance. The attempt to "hack" a benchmark, i.e. modify a research data product with little effort so that checks and evaluations lead to an increased score *without* improving the research data product, is a challenge. The same mitigation strategy compared to event-based metrics is applicable in the context of benchmarks for research data products: constantly improve and re-run the benchmarks, to detect and mitigate specific gaming strategies.

Another alternative is to do entirely without machine-actionable assessment of a research data product. Considering the growth patterns in digital output (Section 1.3), the alternative entails to provide manual resources in scale with the growth dynamics or to refrain from assessing research data products at all.

Social Assumptions in Best Practices

The discussion of *social effects* in Section 6.3 might suggest that we support the claim that benchmarks for research data products are completely unaffected by social influences and measure "objectively". This impression might be based on

the fallacy, that "computers always get it right" ([Bro18]) — a point of view that ignores the way human concepts and models influence software and its output.

The architecture presented in Section 4.2 and the recipe to realize it (Section 5.1) provide the means to manage this "conceptual overload", i.e. to specify and document the social givens, such as best practices. Especially the chosen evidence for a benchmark (step "Analyze Evidence") will heavily impact the implementation of a benchmark with all assumptions and stereotypes carried over from requirement analysis to a running benchmark. As with the previous discussion, the most promising solution is to accept that the development of assessing methods is never finished and should always be ready for review and improvement, especially when it comes to socially produced prejudices.

Artificiality and Dependence on Specific Technological Solutions

Benchmarks artificially produce events of interaction with research data products. For these events to be reproducible, i.e. implementable, they need to be based on simplification assumptions (e.g. the order in which two versions of a research data product are published corresponds to the order of their creation/collection) and on a finite set of technical solutions, such as communication protocols or metadata standards. These assumptions necessitate a conceptual distance between the "reality" of research data management and the automated checks of a benchmark for a research data product.

This shortcoming is not shared by event-based metrics, since they abstract from an individual use case and the used technology. This complementarity is a direct consequence of CTXT of event-based metric. A good mitigation strategy is to carry out an assessment of a research data product with both event-based metrics and benchmarks for research data products.

Implementation effort

Finally, the effort necessary to implement a benchmark for a research data product compliant to the recipe described in Section 5.1 is substantial: Even if the library and the framework published together with this thesis are used, the ingredients necessary to realize a new benchmark for a research data product include access to suitable sources for best practices, programming and deployment know-how and the corresponding computing resources to test and run the benchmark.

A possible mitigation strategy for this shortcoming is to re-use generic benchmarks and accept that they might not entirely capture the requirements of the given use case.

7.3 Recommendations

This sections lists a set of recommendations for different actors in the field of research data management concerning the development and application of benchmarks for research data products.

- **Researchers and Research Data Specialists** are the roles who definitely determine the effort put into the creation and curation of a research data product. The advent of benchmarks for research data products should make submission and curation tasks easier, since the reports produced by a benchmark can be used before publication or dissemination to give hints on how to improve the research data product.

- **Scientometricians** measure the output and describe the communication of researchers. Benchmarks for research data products offer a tool to understand trends and make statements about the dissemination of certain best practices or techniques. It allows to create measurement data, whereas event-based metrics can only be used "after the fact".

- **Research Software Engineers** are the crucial actors in developing new benchmarks for research data products. The recipe in Section 5.1 and the resources presented in Section 5.3 are available for re-use. Sharing new benchmarks will hopefully contribute to the acknowledgment of the crucial role research software engineers play in the creation of scientific progress.

- **Research Data Service Providers** are mainly responsible for the service component of research data products: the performance of a repository heavily influences the score of a benchmark. An active role in the development of benchmarks for research data products ensures a responsible usage of computing resources (e.g. by rate-limiting), but also helps providers to update their services to facilitate higher scores for the depositions of their clients.

- **Policy Makers and Research Funders** are essential in determining if and how the scores of benchmarks for research data products become targets for researchers and other roles. As with all metrics, it is important to allow for contextualization and complementation (see [Hic+15]). The development and maintenance of benchmarks for research data products need appropriate resources.

Finally, it must be stressed, that the best assessment of a research data product always includes a human verdict. Machine-actionable assessments of research data products are no substitutes for a human assessment, they are a tool to support it.

7.4 Future Work

This section provides a list of open questions which can be tackled by future work:

- *Do specific benchmarks for research data products have a tighter Spearman correlation to event-based metrics than the generic prototype of this thesis?* — "specific" and "generic" denote two extremes on a spectrum: a specific benchmark is informative for a comparatively small number of research data products, but mirrors a certain data workflow more closely, while a generic benchmark applies to many research data products but is less close to model a workflow in its entirety.

- *How can event-based metrics and benchmarks be objectively tested to measure quality, impact, or relevance of a research data product?* A possible approach is to interview researchers and ask them to "grade" a set of research data products. The order induced by the grades can then be compared by the order induced by the machine-actionable modes of assessment of a research data product.

- *How can benchmarks for research data products be integrated into reproducibility platforms?* Platforms such as PopperCI [Jim+17], initiatives like ReScience [Rou+17], or journals like the Journal for Open Source Software (JOSS)[2] include steps normally taken in continuous integration and continuous delivery systems in the context of software development. The checks of a benchmark can play the role of tests in quality control and quality assurance workflows. See also the third use case of the implementation chapter in Subsection 5.2.3.

- *Which mathematical models are suitable to combine the scores of event-based metrics and benchmarks for research data products.* In Section 6.3 a simple multiplicative model was suggested, but this suggestion ignores several issues, such as the problem of zero values. Additionally, simple composite indicators (such as the altmetric score), hide information necessary to contextualize the score (see especially the 2nd, 6th, and 8th principle of the Leiden Manifesto [Hic+15]), which would suggest using multidimensional models (see e.g. [Hau12] for a model used for journals), i.e. vectors of different scores.

[2] `https://joss.theoj.org/` [accessed 2020-06-16]

7.5 Outlook

Benchmarks for research data products are a new tool for a challenging task: divide the vast amount of available research data products into two sets - those products that will be appraised by human actors and those that are not. This division is necessary, since human beings are not equipped to keep pace with the growth of data and computing capacities that we saw in past times. Benchmarks have the potential to complement event-based metrics and improve their usage in identifying those research data products which are worthy to be scrutinized by human experts. They are thus an important utility to face the challenge of data-driven research.

If we compare the ever-changing landscape of research data management with the high seas, assessing research data products corresponds to navigational tasks. The role of benchmarks and event-based metrics then closely resembles the role of compasses and coast-bound navigation in naval history: compasses improved navigation, but did not make a look to the shorelines redundant. Both were necessary means to avoid being wrecked and both helped human beings to find their way through stormy seas.

Appendix

Contents

A Tabular Overviews for Shortcomings

The following tables provide details to the 10 shortcomings discussed in Section 3.2. Each table is structured identically:

- **Description** gives a short definition of the shortcoming.
- **Identifier** is an abbreviation for the shortcoming for later reference.
- **Background** adds further information, such as sub-classes of shortcomings of the presented type.
- **Scheme** classifies the shortcoming with regard to the conceptual scheme of event-based metrics presented in Section 3.1.
- **Evidence** lists all publications out of the corpus refer to this type of shortcoming.

The tables are presented in decreasing order of the number of publications discussing the shortcoming.

A.1 Missing Coverage

Table A.1: Coverage as a shortcoming of event-based metrics

Missing coverage	
Description	A score of an event-based metric is too low, because not all events are in the log.
Identifier	COV
Background	Reasons why events are not in the log: • The assessed research data product is not in English or is "non-western" [AT14] [Ham14] [Hic+15] [Zuc+15] [Mac+18] • The events are not frequent enough [The+13] [Bor14a] [JA15] [Pet+16] [Obe17] • References to a research data product are not detectable [Fea14] [May+17] [BH18b] • Platform providers do not grant access to the information [BD08] [PPH12] [MT14] • The log is curated and some events are excluded [Pet+14] [Wal16] [SD19]
Scheme	Coverage concerns the completeness of the *log* of events.
Class	Coverage is a normal shortcoming; it has a target conflict with COR and with TRST.
Evidence	[AT14] [JA15] [BH18b] [Mac+18] [MT14] [Zuc+15] [Obe17] [SD19] [Pet+14] [Gru14] [Ham14] [BD08] [WC12] [CZW15] [Rod15] [HCL15] [May+17] [Cha13] [Pet+16] [Kra+15a] [ZC18] [Hau16] [Wal16] [Bor14a] [Hic+15] [PPH12] [The+13] [Fea14]

A.2 Doubtful Correlation

Table A.2: Coverage as a shortcoming of event-based metrics

Doubtful correlation with quality, impact, or relevance	
Description	Scores of event-based metrics do not correlate with quality, impact, or relevance
Identifier	COR
Background	This shortcoming is characterized in different ways: • event-based metrics do *not* correlate with or imply quality, impact, or relevance e.g.: [Dor13] [Gam+20] [WHH15]; or there is no evidence for such a correlation [The+13] • it is unclear what a high score exactly means [BH18b] [BD08] [Cha13]. • event-based metrics actually measure something different like promotion [Ham14] or attention/popularity [Sug+17] • event-based metrics are even associated with bad practices [She+19]
Scheme	Doubts concerning correlation applies to the *application* layer.
Class	Coverage is a normal shortcoming; it has a target conflict with COR.
Evidence	[AT14] [JA15] [BH18b] [Mac+18] [MT14] [Zuc+15] [Obe17] [SD19] [Pet+14] [Gru14] [Ham14] [BD08] [WC12] [CZW15] [Rod15] [HCL15] [May+17] [Cha13] [Pet+16] [Kra+15a] [ZC18] [Hau16] [Wal16] [Bor14a] [Hic+15] [PPH12] [The+13] [Fea14]

A.3 Normalization

Table A.3: Normalization as a shortcoming of event-based metrics

Normalization	
Description	The scores of two research data products are not comparable, since they must be normalized.
Identifier	NORM
Background	Candidates to normalize against • field of study e.g. in [Har09] [JA15] [BH18b] • age of the research data product e.g. in [Rav+17] [Gam+20] [Ke+15] • number of authors/creators of the research data product e.g. in [CLB10]
Scheme	Normalization issues arise in the calculation of the *score*.
Class	Normalization is a normal shortcoming.
Evidence	[Har09] [Rav+17] [JA15] [BH18b] [Gam+20] [FW17] [Gru14] [Ham14] [Sug+17] [BD08] [Ham14] [WC12] [Rod15] [May+17] [Cha13] [CLB10] [Wal+11b] [Hau16] [Ke+15] [Wal16] [Bor14a] [Bor16] [Hic+15] [The+13]

A.4 Gaming

Table A.4: Gaming as a shortcoming of event-based metrics

Gaming	
Description	The score of a research data product is too high, since it has been artificially and intentionally increased.
Identifier	GAME
Background	Some authors consider altmetrics to be easier to game than citation-based metrics [Bor14a].
Scheme	Gaming applies to the artificial creation of *events*
Class	Gaming is a normal shortcoming.
Evidence	[Dor13] [Rav+17] [BH18b] [Gam+20] [FW17] [Gru14] [Sug+17] [WC14] [Hau+14a] [WC12] [May+17] [DRT14] [Hau16] [Bor14a] [PPH12] [Fea14]

A.5 Sensitivity to social effects (bandwagon)

Table A.5: Social effects as a shortcoming of event-based metrics

Sensitivity to social effects (bandwagon)	
Description	The score of a research data product is too high (to correlate with quality, impact, relevance of a research data product) since social effects lead to skewed distributions.
Identifier	BAND
Background	An example for such effects is the Matthews effect.
Scheme	This sensitivity occurs on the *event* layer.
Class	Sensitivity to social effects is a principal shortcoming.
Evidence	[Rav+17] [Gru14] [Sug+17] [LSJ15] [RF13] [BD08] [Cro14] [CZW15] [CLB10] [Kra+15a] [Hau16] [Ke+15] [Bor14a]

A.6 Timeliness

Table A.6: Timeliness as a shortcoming of event-based metrics

Dependence on Time	
Description	The score of a research data product is only available after a certain period of time passed (in order for the events to occur).
Identifier	TIME
Background	Time also plays a role in NORM, when two research data products of different age are compared. But TIME is *not* identical with this issue. Some authors see altmetrics better suited to mitigate this shortcoming, since citations take longer to accumulate than social media events or usage statistics [BH18b], [Gam+20].
Scheme	Timeliness applies on the *event* layer.
Class	Timeliness is a principal shortcoming.
Evidence	[TC11] [BH18b] [Gam+20] [WHH15] [SD19] [Gru14] [RF13] [CZW15] [ZCW14]

A.7 Missing Trustworthiness

Table A.7: Missing trustworthiness as a shortcoming of event-based metrics

Missing trustworthiness	
Description	The scores of two research data products cannot be compared, since the one source adding to the log of the metric is trusted while another is not.
Identifier	TRST
Background	Reasons to be skeptical: • Economic interest [Har09], [JPH18], [Gam+20], [MT14], [Hau16] • Missing transparency [Cha13], [ZC18], [ZCW14]
Scheme	Trustworthiness is an issue when events are aggregated in a *log*.
Class	Missing trustworthiness is a normal shortcoming; it has a target conflict with COV
Evidence	[Har09] [JPH18] [Gam+20] [MT14] [Cha13] [ZC18] [Hau16] [Hic+15] [ZCW14]

A.8 Missing Context

Table A.8: Missing context as a shortcoming of event-based metrics

Missing context	
Description	The score of a research data product cannot be used to assess the quality, impact, or relevance since necessary context is missing.
Identifier	CTXT
Background	Examples for missing context: • Engagement with the research data product [BH18b] • Role of the person engaged with the research data product [FW17], [Sug+17] ,[Hau16], [Hic+15] • Technical processing as opposed to human interaction (difference to GAME: the increased amount of events is not intended to game the metric) [BH18b], [Sug+17], [Hau+14a]
Scheme	Missing context is a problem of the information stored in the *log*.
Class	Missing context is a simple shortcoming.
Evidence	[BH18b] [FW17] [Mac+18] [Sug+17] [Hau+14a] [Hau16] [Bor14a] [Hic+15]

A.9 Duplication

Table A.9: Duplication as a shortcoming of event-based metrics

Duplication	
Description	The score of a research data product is too low, since the research data product is duplicated over different service providers.
Identifier	DUP
Background	The availability through different platforms is only given if both instances are identified with the same PID. If different identifiers are used, the research data products are different by definition.
Scheme	Duplication is a problem of the *log*.
Class	Duplication is a simple shortcoming.
Evidence	[MT14], [ZC18]

A.10 Versioning

Table A.10: Versioning as a shortcoming of event-based metrics

Versioning	
Description	The score of a research data product is too low, since its predecessors or successors are not accounted for.
Identifier	VER
Background	Publications can be considered to have a version history (preprint, publication, postprint) [Bor14a] Other types of research data product obviously can be versioned.
Scheme	The incorrect aggregation of versions is a problem of the *score*.
Class	Versioning is a simple shortcoming.
Evidence	[Hau16], [Bor14a]

B Evaluation Data by Type of Research Data Product

This section of the appendix shows detailed information parallel to Chapter 6, but stratified by the type of the research data product. Each of its subsection shows information to one of the 9 types supported by zenodo:

1. A table showing the distribution of scores of research data products for the type. The columns are on par with Table 6.4 and are described in Subsection 6.1.6.

2. A table showing correlation of event-based metrics with benchmark scores for the type. The columns are on par with Table 6.5 and are described in Section 6.2

B.1 Research Data Products of Type Publication

Table B.1: Scores of research data products of type publication

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	832,205	0.000	5.000	7.000	16.387	10.000	216482.000	368.462	1,457
Downloads	832,210	0.000	10.000	21.000	24.925	26.000	81994.000	200.628	1,391
Tweeters	29,441	1.000	1.000	2.000	8.774	6.000	7097.000	70.268	277
Readers	33,688	1.000	3.000	9.000	33.779	26.000	9409.000	123.943	643
Videos	367	1.000	1.000	1.000	1.458	1.000	13.000	1.240	9
Facebook Walls	9,909	1.000	1.000	1.000	1.988	2.000	415.000	4.982	39
Feeds	6,782	1.000	1.000	1.000	1.903	2.000	68.000	2.216	30
Posts	44,188	1.000	1.000	2.000	9.361	6.000	8322.000	71.810	359
Altmetric Score	44,188	0.250	1.000	3.000	10.918	6.200	6704.998	68.512	3,801
Benchmark Score	234,351	0.407	0.583	0.593	0.595	0.621	0.759	0.045	1,560

Only displaying those event-based metrics with at least 100 instances — n = 833,095

Table B.2: Correlation with benchmark scores for type publication

Name	Spearman	n
Views	0.049	233,518
Downloads	0.201	233,520
Tweeters	0.322	29,441
Readers	-0.068	1,267
Facebook Walls	0.052	470
Feeds	0.073	309
Posts	0.283	4,719
Altmetric Score	0.228	4,719

Only displaying those event-based metrics with at least 100 instances

B.2 Research Data Products of Type Image

Table B.3: Scores of research data products of type image

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	506,505	0.000	2.000	3.000	5.076	5.000	8928.000	17.919	262
Downloads	506,437	0.000	1.000	1.000	4.097	3.000	8421.000	20.505	350
Tweeters	185	1.000	1.000	2.000	5.524	4.000	169.000	15.688	24
Posts	244	1.000	1.000	1.000	4.922	3.000	181.000	15.164	27
Altmetric Score	244	0.250	1.000	1.850	4.763	7.850	120.450	9.935	62
Benchmark Score	456,732	0.423	0.601	0.607	0.609	0.607	0.724	0.024	530

Only displaying those event-based metrics with at least 100 instances — n = 523,847

Table B.4: Correlation with benchmark scores for type image

Name	Spearman	n
Views	-0.513	441,385
Downloads	-0.435	441,321
Tweeters	0.314	185
Posts	0.271	174
Altmetric Score	0.308	174

Only displaying those event-based metrics with at least 100 instances

B.3 Research Data Products of Type Software

Table B.5: Scores of research data products of type software

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	79,739	0.000	12.000	47.000	222.705	161.000	14834.000	697.099	1,018
Downloads	79,739	0.000	2.000	7.000	42.462	23.000	10700.000	245.484	363
Tweeters	918	1.000	1.000	2.000	4.460	4.000	163.000	9.827	43
Readers	166	1.000	1.000	2.000	9.494	6.000	329.000	33.414	27
Feeds	127	1.000	1.000	1.000	1.039	1.000	3.000	0.232	3
Posts	1,070	1.000	1.000	2.000	4.495	4.000	212.000	11.260	40
Altmetric Score	1,070	0.250	1.000	1.600	3.740	4.700	105.100	6.766	187
Benchmark Score	35,416	0.400	0.467	0.500	0.518	0.552	0.747	0.054	424

Only displaying those event-based metrics with at least 100 instances — n = 79,747

Table B.6: Correlation with benchmark scores for type software

Name	Spearman	n
Views	0.068	35,415
Downloads	0.117	35,415
Tweeters	0.008	918
Readers	-0.092	110
Posts	0.023	717
Altmetric Score	-0.021	717

Only displaying those event-based metrics with at least 100 instances

B.4 Research Data Products of Type Dataset

Table B.7: Scores of research data products of type dataset

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	62,189	0.000	4.000	14.000	110.241	35.000	59008.000	699.987	1,114
Downloads	62,189	0.000	4.000	7.000	123.023	36.000	55668.000	963.719	855
Tweeters	1,595	1.000	1.000	3.000	8.937	7.000	874.000	30.587	85
Readers	374	1.000	1.000	4.000	59.971	20.000	9926.000	523.375	86
Feeds	179	1.000	1.000	1.000	1.251	1.000	12.000	1.005	6
Posts	1,849	1.000	1.000	3.000	9.628	7.000	1477.000	42.989	92
Altmetric Score	1,849	0.250	1.000	2.250	7.356	5.080	748.738	28.224	379
Benchmark Score	44,925	0.407	0.586	0.587	0.602	0.621	0.759	0.044	750

Only displaying those event-based metrics with at least 100 instances — n = 62,189

Table B.8: Correlation with benchmark scores for type dataset

Name	Spearman	n
Views	0.225	44,925
Downloads	0.080	44,925
Tweeters	0.076	1,595
Readers	-0.155	194
Feeds	-0.003	115
Posts	0.099	1,374
Altmetric Score	0.110	1,374

Only displaying those event-based metrics with at least 100 instances

B.5 Research Data Products of Type Presentation

Table B.9: Scores of research data products of type presentation

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	13,556	0.000	7.000	16.000	51.777	42.000	5316.000	159.220	543
Downloads	13,556	0.000	6.000	13.000	33.676	30.000	6120.000	107.072	376
Tweeters	1,319	1.000	1.000	3.000	7.086	8.000	448.000	16.926	55
Readers	100	1.000	1.000	2.000	4.640	4.000	93.000	10.330	19
Posts	1,378	1.000	1.000	4.000	7.853	8.000	535.000	19.598	61
Altmetric Score	1,378	0.250	1.000	2.700	5.275	5.650	323.110	12.054	277
Benchmark Score	12,449	0.407	0.586	0.621	0.630	0.655	0.759	0.047	320

Only displaying those event-based metrics with at least 100 instances — n = 13,558

Table B.10: Correlation with benchmark scores for type presentation

Name	Spearman	n
Views	0.119	12,448
Downloads	0.122	12,448
Tweeters	0.076	1,319
Posts	0.076	1,274
Altmetric Score	0.082	1,274

Only displaying those event-based metrics with at least 100 instances

B.6 Research Data Products of Type Poster

Table B.11: Scores of research data products of type poster

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	5,505	0.000	8.000	16.000	38.901	38.000	1918.000	85.178	292
Downloads	5,505	0.000	7.000	13.000	26.402	26.000	4031.000	76.120	211
Tweeters	549	1.000	1.000	3.000	6.078	7.000	149.000	10.651	38
Posts	563	1.000	1.000	3.000	6.973	8.000	162.000	12.342	42
Altmetric Score	563	0.250	1.000	2.350	4.563	5.250	95.700	7.583	163
Benchmark Score	5,191	0.444	0.593	0.621	0.625	0.655	0.733	0.043	271

Only displaying those event-based metrics with at least 100 instances — n = 5,509

Table B.12: Correlation with benchmark scores for type poster

Name	Spearman	n
Views	0.107	5,187
Downloads	0.142	5,187
Tweeters	0.140	549
Posts	0.119	523
Altmetric Score	0.166	523

Only displaying those event-based metrics with at least 100 instances

B.7 Research Data Products of Type Video

Table B.13: Scores of research data products of type video

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	1,791	0.000	7.000	14.000	39.607	30.000	2316.000	119.079	173
Downloads	1,791	0.000	2.000	4.000	18.534	10.000	1067.000	70.298	110
Benchmark Score	1,714	0.407	0.552	0.621	0.602	0.655	0.733	0.056	141

Only displaying those event-based metrics with at least 100 instances — n = 1,834

Table B.14: Correlation with benchmark scores for type video

Name	Spearman	n
Views	0.058	1,671
Downloads	0.119	1,671

Only displaying those event-based metrics with at least 100 instances

B.8 Research Data Products of Type Lesson

Table B.15: Scores of research data products of type lesson

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	1,595	0.000	7.000	19.000	130.579	65.500	9750.000	571.876	241
Downloads	1,595	0.000	6.000	14.000	164.915	48.000	14290.000	735.257	244
Posts	106	1.000	1.000	2.000	12.943	8.000	381.000	41.239	27
Altmetric Score	106	0.250	1.600	6.472	10.167	7.675	185.700	21.070	54
Benchmark Score	1,399	0.407	0.552	0.607	0.601	0.633	0.724	0.056	120

Only displaying those event-based metrics with at least 100 instances — n = 1,617

Table B.16: Correlation with benchmark scores for type lesson

Name	Spearman	n
Views	0.176	1,377
Downloads	0.137	1,377

Only displaying those event-based metrics with at least 100 instances

B.9 Research Data Products of Type Other

Table B.17: Scores of research data products of type other

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	3,825	0.000	5.000	13.000	87.223	46.000	8665.000	499.777	309
Downloads	3,826	0.000	4.000	10.000	61.160	28.000	5446.000	317.482	243
Tweeters	185	1.000	1.000	4.000	11.546	9.000	442.000	35.903	33
Posts	199	1.000	1.000	4.000	14.020	12.000	526.000	41.408	44
Altmetric Score	199	0.250	1.050	3.700	8.724	7.825	314.112	24.391	99
Benchmark Score	3,186	0.433	0.586	0.600	0.606	0.655	0.733	0.052	181

Only displaying those event-based metrics with at least 100 instances — n = 3,856

Table B.18: Correlation with benchmark scores for type other

Name	Spearman	n
Views	0.278	3,155
Downloads	0.318	3,156
Tweeters	0.110	185
Posts	0.206	160
Altmetric Score	0.142	160

Only displaying those event-based metrics with at least 100 instances

C Evaluation Data by Field of Study

This section of the appendix displays detailed information parallel to Chapter 6, but stratified by field of study. Table C.1 displays in each row a field of study, a lower estimate of the number of research data products, the actual number of classification, an upper estimate, precision, and recall of the classifier for the field of study. Further information on the training and application of the classifier can be found in [Web+20].

Table C.1: Classification with estimated errors, precision and recall of the classifier

Field of Study	Lower Estimate	# Classified	Upper Estimate	Precision	Recall
Mathematical Sciences	6,924	8,655	11,078	0.80	0.72
Physical Sciences	21,990	22,670	24,257	0.97	0.93
Chemical Sciences	8,133	9,918	11,703	0.82	0.82
Earth and Environmental Sciences	46,204	57,755	70,461	0.80	0.78
Biological Sciences	208,852	237,332	261,065	0.88	0.90
Agricultural and Veterinary Sciences	1,241	1,443	2,309	0.86	0.40
Information and Computing Sciences	50,067	61,057	74,490	0.82	0.78
Engineering and Technology	48,431	61,305	78,470	0.79	0.72
Medical and Health Sciences	63,235	74,394	87,041	0.85	0.83
Built Environment and Design	1,729	2,083	2,958	0.83	0.58
Education	3,766	4,483	6,231	0.84	0.61
Economics	3,779	4,784	6,698	0.79	0.60
Commerce, Management, Tourism and Services	6,470	8,087	12,050	0.80	0.51
Studies in Human Society	7,505	8,934	12,240	0.84	0.63
Psychology and Cognitive Sciences	5,102	5,864	7,213	0.87	0.77
Law and Legal Studies	1,367	1,486	2,184	0.92	0.53
Studies in Creative Arts and Writing	582	626	926	0.93	0.52
Language, Communication and Culture	6,117	7,282	10,049	0.84	0.62
History and Archaeology	1,727	1,962	2,727	0.88	0.61
Philosophy and Religious Studies	591	672	1,048	0.88	0.44

Classifiable research data products in zenodo: 507,049

The remainder of this section includes a subsection for each of the fields in order of appearance in Table C.1. In each subsection three tables can be found:

1. A table showing the distribution of types of research data products in the field of study in absolute and relative numbers.

2. A table showing the distribution of scores of research data products in the field of study. The columns are on par with Table 6.4 and are described in Subsection 6.1.6.

3. A table showing correlation of event-based metrics with benchmark scores in the field of study. The columns are on par with Table 6.5 and are described in Section 6.2

C.1 Mathematical Sciences

Table C.2: Types of research data products in Mathematical Sciences

dataset	image	lesson	other	poster	presentation	publication	software	video
208 (2.40%)	458 (5.29%)	15 (0.17%)	37 (0.43%)	28 (0.32%)	66 (0.76%)	6,507 (75.18%)	1,329 (15.36%)	7 (0.08%)

Table C.3: Assessment scores in Mathematical Sciences

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	6,754	0.000	6.000	10.000	25.294	20.000	6061.000	134.447	245
Downloads	6,754	0.000	3.000	6.000	12.343	10.000	2484.000	55.764	153
Benchmark Score	6,788	0.433	0.586	0.600	0.598	0.621	0.747	0.038	163

Only displaying those event-based metrics with at least 100 instances — n = 6,788

Table C.4: Correlation in Mathematical Sciences

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	-0.075	6,754	-0.077	6,471
Downloads	0.084	6,754	0.096	6,471

Only displaying those event-based metrics with at least 100 instances

C.2 Physical Sciences

Table C.5: Types of research data products in Physical Sciences

dataset	image	lesson	other	poster	presentation	publication	software	video
10,898 (48.07%)	707 (3.12%)	11 (0.05%)	63 (0.28%)	892 (3.93%)	1,999 (8.82%)	6,601 (29.12%)	1,409 (6.22%)	90 (0.40%)

Table C.6: Assessment scores in Physical Sciences

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	11,621	0.000	8.000	16.000	36.904	29.000	6090.000	153.096	384
Downloads	11,621	0.000	5.000	11.000	28.574	37.000	4020.000	97.490	301
Tweeters	207	1.000	1.000	2.000	3.691	3.000	118.000	8.956	19
Posts	251	1.000	1.000	2.000	4.056	3.000	162.000	11.268	23
Altmetric Score	251	0.250	1.000	1.500	5.375	5.080	334.176	23.347	69
Benchmark Score	11,672	0.433	0.586	0.600	0.600	0.621	0.747	0.048	393

Only displaying those event-based metrics with at least 100 instances — n = 11,672

Table C.7: Correlation in Physical Sciences

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	0.081	11,621	0.083	11,215
Downloads	0.025	11,621	0.026	11,215
Tweeters	0.251	773	0.237	769
Posts	0.194	251	0.185	247
Altmetric Score	0.140	251	0.129	247

Only displaying those event-based metrics with at least 100 instances

C.3 Chemical Sciences

Table C.8: Types of research data products in Chemical Sciences

dataset	image	lesson	other	poster	presentation	publication	software	video
862 (8.69%)	1,679 (16.93%)	7 (0.07%)	22 (0.22%)	98 (0.99%)	125 (1.26%)	6,492 (65.46%)	610 (6.15%)	23 (0.23%)

Table C.9: Assessment scores in Chemical Sciences

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	6,372	0.000	6.000	12.000	32.511	22.000	11735.000	211.979	262
Downloads	6,373	0.000	3.000	7.000	17.264	15.000	1738.000	60.649	205
Tweeters	107	1.000	1.000	2.000	3.374	5.000	17.000	3.372	14
Posts	123	1.000	1.000	2.000	3.707	5.000	22.000	3.946	16
Altmetric Score	123	0.250	0.750	1.500	2.842	3.650	15.980	3.199	44
Benchmark Score	6,457	0.433	0.586	0.600	0.607	0.621	0.733	0.040	289

Only displaying those event-based metrics with at least 100 instances — n = 6,457

Table C.10: Correlation in Chemical Sciences

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	0.034	6,372	0.038	5,297
Downloads	0.163	6,373	0.215	5,297
Tweeters	0.339	989	0.346	985
Posts	0.140	123	0.139	121
Altmetric Score	0.266	123	0.267	121

Only displaying those event-based metrics with at least 100 instances

C.4 Earth and Environmental Sciences

Table C.11: Types of research data products in Earth and Environmental Sciences

dataset	image	lesson	other	poster	presentation	publication	software	video
4,411 (7.64%)	29,771 (51.55%)	26 (0.05%)	105 (0.18%)	344 (0.60%)	615 (1.06%)	20,085 (34.78%)	2,330 (4.03%)	68 (0.12%)

Table C.12: Assessment scores in Earth and Environmental Sciences

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	35,215	0.000	4.000	9.000	21.463	17.000	12460.000	139.419	474
Downloads	35,217	0.000	1.000	5.000	14.551	12.000	25312.000	166.561	361
Tweeters	404	1.000	1.000	2.000	8.119	6.000	874.000	45.239	38
Readers	107	1.000	1.000	2.000	9.551	8.500	99.000	17.000	30
Posts	485	1.000	1.000	2.000	9.115	5.000	1477.000	68.730	38
Altmetric Score	485	0.250	1.000	2.000	5.863	5.080	632.400	29.866	130
Benchmark Score	36,873	0.433	0.588	0.601	0.598	0.607	0.733	0.039	840

Only displaying those event-based metrics with at least 100 instances — n = 36,873

Table C.13: Correlation in Earth and Environmental Sciences

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	0.124	35,215	0.246	15,716
Downloads	0.179	35,217	0.454	15,716
Tweeters	0.180	5,317	0.174	5,299
Readers	-0.135	107	-0.135	107
Posts	0.123	485	0.112	466
Altmetric Score	0.125	485	0.117	466

Only displaying those event-based metrics with at least 100 instances

C.5 Biological Sciences

Table C.14: Types of research data products in Biological Sciences

dataset	image	lesson	other	poster	presentation	publication	software	video
21,511 (9.06%)	162,491 (68.47%)	51 (0.02%)	231 (0.10%)	428 (0.18%)	421 (0.18%)	46,592 (19.63%)	5,495 (2.32%)	112 (0.05%)

Table C.15: Assessment scores in Biological Sciences

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	158,047	0.000	3.000	6.000	14.591	14.000	14017.000	136.440	641
Downloads	158,056	0.000	1.000	4.000	10.148	9.000	23184.000	123.893	462
Tweeters	846	1.000	1.000	1.000	5.090	4.000	515.000	22.132	36
Readers	174	1.000	1.000	3.000	5.971	6.750	74.000	10.239	25
Posts	1,032	1.000	1.000	1.000	5.003	4.000	578.000	24.618	39
Altmetric Score	1,032	0.250	0.500	1.000	3.256	3.000	402.600	16.039	151
Benchmark Score	168,780	0.427	0.587	0.600	0.594	0.602	0.733	0.030	1,340

Only displaying those event-based metrics with at least 100 instances — n = 168,780

Table C.16: Correlation in Biological Sciences

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	0.076	158,047	0.128	44,749
Downloads	0.090	158,056	0.566	44,750
Tweeters	0.335	12,928	0.331	12,853
Readers	-0.084	174	-0.084	174
Posts	0.321	1,032	0.319	959
Altmetric Score	0.270	1,032	0.270	959

Only displaying those event-based metrics with at least 100 instances

C.6 Agricultural and Veterinary Sciences

Table C.17: Types of research data products in Agricultural and Veterinary Sciences

dataset	image	lesson	other	poster	presentation	publication	software	video
33 (2.29%)	110 (7.62%)	2 (0.14%)	3 (0.21%)	28 (1.94%)	24 (1.66%)	1,217 (84.34%)	24 (1.66%)	2 (0.14%)

Table C.18: Assessment scores in Agricultural and Veterinary Sciences

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	1,057	0.000	6.000	10.000	21.476	20.000	2495.000	83.705	100
Downloads	1,057	0.000	5.000	8.000	24.174	14.000	3211.000	132.108	101
Benchmark Score	1,061	0.433	0.586	0.621	0.611	0.626	0.724	0.036	93

Only displaying those event-based metrics with at least 100 instances — n = 1,061

Table C.19: Correlation in Agricultural and Veterinary Sciences

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	-0.052	1,057	-0.056	1,026
Downloads	0.122	1,057	0.128	1,026

Only displaying those event-based metrics with at least 100 instances

C.7 Information and Computing Sciences

Table C.20: Types of research data products in Information and Computing Sciences

dataset	image	lesson	other	poster	presentation	publication	software	video
4,606 (7.54%)	2,811 (4.60%)	166 (0.27%)	480 (0.79%)	929 (1.52%)	2,723 (4.46%)	30,310 (49.64%)	18,826 (30.83%)	206 (0.34%)

Table C.21: Assessment scores in Information and Computing Sciences

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	42,410	0.000	5.000	10.000	53.988	28.000	25453.000	326.646	991
Downloads	42,410	0.000	3.000	6.000	31.236	14.000	55668.000	393.857	715
Tweeters	1,683	1.000	1.000	3.000	9.084	8.000	610.000	23.202	83
Readers	322	1.000	1.000	3.000	8.233	7.000	329.000	21.569	41
Facebook Walls	107	1.000	1.000	1.000	1.336	1.000	6.000	0.931	6
Feeds	187	1.000	1.000	1.000	1.241	1.000	4.000	0.614	4
Posts	1,794	1.000	1.000	3.000	10.403	9.000	908.000	30.328	98
Altmetric Score	1,794	0.250	1.000	2.700	7.240	6.787	442.360	17.211	417
Benchmark Score	42,559	0.433	0.583	0.600	0.602	0.633	0.747	0.054	701

Only displaying those event-based metrics with at least 100 instances — n = 42,559

Table C.22: Correlation in Information and Computing Sciences

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	0.045	42,410	0.120	40,315
Downloads	0.348	42,410	0.362	40,315
Tweeters	0.196	2,965	0.194	2,950
Readers	-0.102	322	-0.100	320
Facebook Walls	-0.051	107	-0.038	104
Feeds	0.037	187	0.029	185
Posts	0.200	1,794	0.198	1,782
Altmetric Score	0.184	1,794	0.181	1,782

Only displaying those event-based metrics with at least 100 instances

C.8 Engineering and Technology

Table C.23: Types of research data products in Engineering and Technology

dataset	image	lesson	other	poster	presentation	publication	software	video
2,381 (3.88%)	2,979 (4.86%)	99 (0.16%)	237 (0.39%)	461 (0.75%)	1,222 (1.99%)	49,934 (81.45%)	3,842 (6.27%)	150 (0.24%)

Table C.24: Assessment scores in Engineering and Technology

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	46,391	0.000	6.000	10.000	27.104	20.000	21732.000	179.644	632
Downloads	46,392	0.000	4.000	7.000	18.871	13.000	7983.000	119.574	496
Tweeters	466	1.000	1.000	1.000	4.571	3.000	352.000	18.959	30
Posts	509	1.000	1.000	2.000	5.010	3.000	381.000	20.443	35
Altmetric Score	509	0.250	0.750	1.000	3.782	2.850	334.176	17.722	93
Benchmark Score	46,603	0.433	0.586	0.600	0.605	0.621	0.733	0.038	524

Only displaying those event-based metrics with at least 100 instances — n = 46,603

Table C.25: Correlation in Engineering and Technology

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	-0.066	46,391	-0.067	44,749
Downloads	0.122	46,392	0.133	44,750
Tweeters	0.174	1,835	0.172	1,832
Posts	0.115	509	0.113	507
Altmetric Score	0.087	509	0.085	507

Only displaying those event-based metrics with at least 100 instances

C.9 Medical and Health Sciences

Table C.26: Types of research data products in Medical and Health Sciences

dataset	image	lesson	other	poster	presentation	publication	software	video
3,035 (4.08%)	22,307 (29.98%)	62 (0.08%)	310 (0.42%)	491 (0.66%)	496 (0.67%)	40,287 (54.15%)	7,277 (9.78%)	129 (0.17%)

Table C.27: Assessment scores in Medical and Health Sciences

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	48,671	0.000	5.000	10.000	27.915	19.000	12178.000	193.804	634
Downloads	48,669	0.000	3.000	6.000	18.185	12.000	55668.000	285.864	467
Tweeters	615	1.000	1.000	2.000	7.208	5.000	721.000	33.427	42
Readers	113	1.000	1.000	2.000	2.938	3.000	30.000	3.910	11
Posts	674	1.000	1.000	2.000	8.159	5.000	842.000	38.077	50
Altmetric Score	674	0.250	1.000	1.500	5.331	3.588	610.430	26.614	140
Benchmark Score	49,840	0.433	0.586	0.600	0.599	0.621	0.737	0.045	762

Only displaying those event-based metrics with at least 100 instances — n = 49,840

Table C.28: Correlation in Medical and Health Sciences

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	0.049	48,671	0.062	36,721
Downloads	0.173	48,669	0.234	36,721
Tweeters	0.305	3,777	0.296	3,762
Readers	-0.170	113	-0.143	110
Posts	0.264	674	0.254	659
Altmetric Score	0.262	674	0.253	659

Only displaying those event-based metrics with at least 100 instances

C.10 Built Environment and Design

Table C.29: Types of research data products in Built Environment and Design

dataset	image	lesson	other	poster	presentation	publication	software	video
42 (2.02%)	52 (2.50%)	2 (0.10%)	8 (0.38%)	23 (1.10%)	113 (5.42%)	1,811 (86.94%)	30 (1.44%)	2 (0.10%)

Table C.30: Assessment scores in Built Environment and Design

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	1,709	0.000	6.000	11.000	31.315	22.000	4350.000	149.655	136
Downloads	1,709	0.000	5.000	8.000	19.958	16.000	1594.000	75.107	122
Benchmark Score	1,709	0.463	0.586	0.600	0.612	0.633	0.724	0.037	74

Only displaying those event-based metrics with at least 100 instances — n = 1,709

Table C.31: Correlation in Built Environment and Design

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	-0.117	1,709	-0.112	1,666
Downloads	0.031	1,709	0.037	1,666

Only displaying those event-based metrics with at least 100 instances

C.11 Education

Table C.32: Types of research data products in Education

dataset	image	lesson	other	poster	presentation	publication	software	video
41 (0.91%)	24 (0.54%)	20 (0.45%)	17 (0.38%)	17 (0.38%)	91 (2.03%)	4,155 (92.68%)	107 (2.39%)	11 (0.25%)

Table C.33: Assessment scores in Education

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	3,900	0.000	5.000	11.000	31.902	23.000	5316.000	141.468	236
Downloads	3,900	0.000	5.000	8.000	27.666	17.000	5296.000	138.423	213
Tweeters	114	1.000	1.000	4.000	8.535	12.000	73.000	12.049	27
Posts	121	1.000	1.000	5.000	10.636	14.000	104.000	16.117	33
Altmetric Score	121	0.250	1.000	3.700	6.579	8.400	53.800	8.516	74
Benchmark Score	3,907	0.444	0.586	0.600	0.606	0.621	0.724	0.041	108

Only displaying those event-based metrics with at least 100 instances — n = 3,907

Table C.34: Correlation in Education

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	-0.048	3,900	-0.049	3,883
Downloads	0.021	3,900	0.020	3,883
Tweeters	0.300	170	0.300	167
Posts	0.198	121	0.198	118
Altmetric Score	0.221	121	0.220	118

Only displaying those event-based metrics with at least 100 instances

C.12 Economics

Table C.35: Types of research data products in Economics

dataset	image	lesson	other	poster	presentation	publication	software	video
100 (2.09%)	179 (3.74%)	5 (0.10%)	16 (0.33%)	8 (0.17%)	40 (0.84%)	4,250 (88.84%)	179 (3.74%)	7 (0.15%)

Table C.36: Assessment scores in Economics

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	3,641	0.000	5.000	10.000	26.436	19.000	5329.000	123.066	194
Downloads	3,641	0.000	4.000	7.000	27.757	14.000	16351.000	332.858	172
Benchmark Score	3,657	0.467	0.586	0.600	0.609	0.633	0.724	0.039	133

Only displaying those event-based metrics with at least 100 instances — n = 3,657

Table C.37: Correlation in Economics

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	-0.162	3,641	-0.167	3,529
Downloads	-0.002	3,641	-0.004	3,529

Only displaying those event-based metrics with at least 100 instances

C.13 Commerce, Management, Tourism and Services

Table C.38: Types of research data products in Commerce, Management, Tourism and Services

dataset	image	lesson	other	poster	presentation	publication	software	video
40 (0.49%)	12 (0.15%)	8 (0.10%)	28 (0.35%)	11 (0.14%)	46 (0.57%)	7,814 (96.62%)	116 (1.43%)	12 (0.15%)

Table C.39: Assessment scores in Commerce, Management, Tourism and Services

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	6,555	0.000	5.000	11.000	31.626	22.000	8234.000	185.964	277
Downloads	6,555	0.000	4.000	8.000	29.672	16.000	7571.000	193.863	266
Benchmark Score	6,557	0.433	0.586	0.616	0.612	0.633	0.745	0.037	99

Only displaying those event-based metrics with at least 100 instances — n = 6,557

Table C.40: Correlation in Commerce, Management, Tourism and Services

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	-0.235	6,555	-0.236	6,547
Downloads	-0.134	6,555	-0.135	6,547

Only displaying those event-based metrics with at least 100 instances

C.14 Studies in Human Society

Table C.41: Types of research data products in Studies in Human Society

dataset	image	lesson	other	poster	presentation	publication	software	video
220 (2.46%)	52 (0.58%)	17 (0.19%)	73 (0.82%)	37 (0.41%)	155 (1.73%)	8,182 (91.58%)	189 (2.12%)	9 (0.10%)

Table C.42: Assessment scores in Studies in Human Society

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	7,580	0.000	5.000	8.000	26.337	17.000	5186.000	135.218	278
Downloads	7,580	0.000	3.000	6.000	22.140	12.000	10970.000	184.235	246
Tweeters	184	1.000	1.000	2.000	6.793	6.000	205.000	19.919	26
Posts	199	1.000	1.000	2.000	7.583	6.000	215.000	22.074	31
Altmetric Score	199	0.250	1.000	1.850	5.133	5.200	141.750	13.831	75
Benchmark Score	7,585	0.467	0.593	0.593	0.607	0.621	0.733	0.036	161

Only displaying those event-based metrics with at least 100 instances — n = 7,585

Table C.43: Correlation in Studies in Human Society

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	0.117	7,580	0.116	7,552
Downloads	0.199	7,580	0.199	7,552
Tweeters	0.411	453	0.411	453
Posts	0.277	199	0.286	198
Altmetric Score	0.316	199	0.309	198

Only displaying those event-based metrics with at least 100 instances

C.15 Psychology and Cognitive Sciences

Table C.44: Types of research data products in Psychology and Cognitive Sciences

dataset	image	lesson	other	poster	presentation	publication	software	video
576 (9.82%)	253 (4.31%)	9 (0.15%)	50 (0.85%)	66 (1.13%)	83 (1.42%)	3,938 (67.16%)	848 (14.46%)	41 (0.70%)

Table C.45: Assessment scores in Psychology and Cognitive Sciences

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	4,163	0.000	7.000	14.000	61.072	31.000	59008.000	975.639	268
Downloads	4,163	0.000	4.000	8.000	33.353	16.000	21076.000	404.345	206
Posts	116	1.000	1.000	2.000	6.448	5.000	108.000	13.567	23
Altmetric Score	116	0.250	0.938	1.500	4.365	5.020	70.600	8.265	46
Benchmark Score	4,188	0.400	0.583	0.600	0.601	0.621	0.733	0.048	172

Only displaying those event-based metrics with at least 100 instances — n = 4,188

Table C.46: Correlation in Psychology and Cognitive Sciences

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	-0.010	4,163	-0.011	4,031
Downloads	0.150	4,163	0.157	4,031
Posts	0.090	116	0.090	116
Altmetric Score	0.088	116	0.088	116

Only displaying those event-based metrics with at least 100 instances

C.16 Law and Legal Studies

Table C.47: Types of research data products in Law and Legal Studies

dataset	image	lesson	other	poster	presentation	publication	software
5 (0.34%)	2 (0.13%)	3 (0.20%)	12 (0.81%)	2 (0.13%)	25 (1.68%)	1,434 (96.50%)	3 (0.20%)

Table C.48: Assessment scores in Law and Legal Studies

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	1,130	0.000	4.000	9.000	25.998	18.000	1888.000	105.325	112
Downloads	1,130	0.000	5.000	8.000	36.451	16.000	5482.000	254.936	120
Benchmark Score	1,131	0.467	0.586	0.621	0.617	0.655	0.724	0.044	47

Only displaying those event-based metrics with at least 100 instances — n = 1,131

Table C.49: Correlation in Law and Legal Studies

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	-0.201	1,130	-0.202	1,128
Downloads	-0.088	1,130	-0.088	1,128

Only displaying those event-based metrics with at least 100 instances

C.17 Studies in Creative Arts and Writing

Table C.50: Types of research data products in Studies in Creative Arts and Writing

dataset	image	lesson	poster	presentation	publication	software	video
29 (4.63%)	6 (0.96%)	2 (0.32%)	1 (0.16%)	6 (0.96%)	556 (88.82%)	17 (2.72%)	9 (1.44%)

Table C.51: Assessment scores in Studies in Creative Arts and Writing

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	572	0.000	5.000	9.000	40.026	17.250	4369.000	251.745	84
Downloads	572	0.000	4.000	7.000	28.921	14.000	3192.000	150.592	89
Benchmark Score	574	0.457	0.586	0.621	0.608	0.621	0.724	0.037	43

Only displaying those event-based metrics with at least 100 instances — n = 574

Table C.52: Correlation in Studies in Creative Arts and Writing

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	0.046	572	0.042	566
Downloads	0.031	572	0.037	566

Only displaying those event-based metrics with at least 100 instances

C.18 Language, Communication and Culture

Table C.53: Types of research data products in Language, Communication and Culture

dataset	image	lesson	other	poster	presentation	publication	software	video
455 (6.25%)	273 (3.75%)	24 (0.33%)	87 (1.19%)	51 (0.70%)	180 (2.47%)	5,716 (78.49%)	433 (5.95%)	63 (0.87%)

Table C.54: Assessment scores in Language, Communication and Culture

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	5,929	0.000	6.000	12.000	42.910	28.000	13088.000	284.163	306
Downloads	5,929	0.000	5.000	9.000	37.574	21.000	22577.000	354.134	293
Tweeters	120	1.000	1.000	1.500	3.467	4.000	63.000	6.861	15
Posts	183	1.000	1.000	1.000	3.301	3.000	85.000	7.660	18
Altmetric Score	183	0.250	1.000	2.700	3.416	3.350	45.900	4.899	50
Benchmark Score	5,951	0.444	0.586	0.621	0.612	0.638	0.759	0.044	150

Only displaying those event-based metrics with at least 100 instances — n = 5,951

Table C.55: Correlation in Language, Communication and Culture

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	-0.031	5,929	-0.032	5,785
Downloads	0.044	5,929	0.046	5,785
Tweeters	0.152	200	0.162	197
Posts	0.093	183	0.095	181
Altmetric Score	0.052	183	0.055	181

Only displaying those event-based metrics with at least 100 instances

C.19 History and Archaeology

Table C.56: Types of research data products in History and Archaeology

dataset	image	other	poster	presentation	publication	software	video
303 (15.44%)	360 (18.35%)	37 (1.89%)	19 (0.97%)	28 (1.43%)	1,162 (59.23%)	37 (1.89%)	16 (0.82%)

Table C.57: Assessment scores in History and Archaeology

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	1,662	0.000	6.000	12.000	25.733	27.000	696.000	50.536	142
Downloads	1,662	0.000	4.000	9.000	27.734	19.000	2626.000	102.584	163
Benchmark Score	1,692	0.467	0.597	0.621	0.621	0.655	0.726	0.042	96

Only displaying those event-based metrics with at least 100 instances — n = 1,692

Table C.58: Correlation in History and Archaeology

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	-0.012	1,662	0.010	1,352
Downloads	0.026	1,662	0.013	1,352

Only displaying those event-based metrics with at least 100 instances

C.20 Philosophy and Religious Studies

Table C.59: Types of research data products in Philosophy and Religious Studies

dataset	image	other	poster	presentation	publication	software
13 (1.93%)	10 (1.49%)	2 (0.30%)	2 (0.30%)	11 (1.64%)	614 (91.37%)	20 (2.98%)

Table C.60: Assessment scores in Philosophy and Religious Studies

Name	#	Min	1st	Med	Mean	3rd	Max	SD	# Char
Views	576	1.000	6.000	12.000	30.502	27.250	1224.000	81.630	103
Downloads	576	0.000	6.000	9.000	35.366	23.000	3785.000	169.510	106
Benchmark Score	576	0.466	0.583	0.613	0.608	0.655	0.737	0.047	64

Only displaying those event-based metrics with at least 100 instances — n = 576

Table C.61: Correlation in Philosophy and Religious Studies

Event-Based Metric	S all	# all	S w/o images	# w/o images
Views	0.095	576	0.093	566
Downloads	0.113	576	0.123	566

Only displaying those event-based metrics with at least 100 instances

D Pairs of Evaluation and Checks of the Benchmark

The following 34 Evaluation-Check-Pairs (ECP) were used in the prototypical benchmark:

ECP-1 evaluates whether a research data product has a valid DOI as PID.

ECP-2 evaluates whether the DOI of the research data product resolves.

ECP-3 evaluates whether the creators of the research data product have valid OrcIDs.

ECP-4 evaluates whether the creators of the research data product have distinguishable family and given names.

ECP-5 evaluates whether the creators include institutions.

ECP-6 evaluates whether the titles of a research data product are (probably) just a file name.

ECP-7 evaluates whether the titles of a research data product include only one un-specified, i.e. one main title.

ECP-8 evaluates whether at least on title is in English.

ECP-9 evaluates whether the subject tags of the research data product are qualified, i.e. whether a scheme is specified by name or URL.

ECP-10 evaluates whether there is at least one subject tag specified for the research data product.

ECP-11 evaluates whether the subject tag of the research data product contain a DDC field of study specification.

ECP-12 evaluates whether the subject tag of the research data product contain a reference to a concept defined in wikidata.

ECP-13 evaluates whether the creation or collection date of the research data product is specified.

ECP-14 evaluates whether there an issuance date is specified that is consistent with the publication year of the research data product.

ECP-15 evaluates whether same-named date specifications of the research data product are separated by additional information.

ECP-16 evaluates whether the size of the research data product is specified exactly once.

ECP-17 evaluates whether the size of the research data product is specified in a parsable *bytes format.

ECP-18 evaluates whether the formats of the research data product are specified in a valid MimeType format.

ECP-19 evaluates whether at least one parsable license statement is specified for the research data product.

ECP-20 evaluates whether the rights of the research data product are specified in a SPDX-compliant manner.

ECP-21 evaluates whether the research data product has at least one description.

ECP-22 evaluates whether the description of the research data product is less than 300 words long.

ECP-23 evaluates whether at least one of the descriptions of the research data product is in English.

ECP-24 evaluates whether at least one description of type “Abstract” is specified for the research data product.

ECP-25 evaluates whether the types of the descriptions of the research data product are consistent with the types prescribed by the best practice guide.

ECP-26 evaluates whether the rightsholder for the research data product is specified if the research data product is not openly licensed.

ECP-27 evaluates whether the version of the research data product is specified in semantic versioning (this pair is optional).

ECP-28 evaluates whether the language of the research data product is specified as an ISO-639-1 code (this pair is optional).

ECP-29 evaluates whether the contributors of the research data product have valid OrcIDs (this pair is optional).

ECP-30 evaluates whether the contributors of the research data product have distinguishable family and given names (this pair is optional).

ECP-31 evaluates whether the contributors of the research data product include institutions (this pair is optional).

ECP-32 evaluates whether the types of contributor are consistent with the prescription of the best practice guide (this pair is optional).

ECP-33 evaluates if the research data product specifies a relation to another resource of type 'HasMetadata', that this relation also contains a resolvable schemeURI and schemeType (this pair is optional).

ECP-34 evaluates whether the specified related resources of the research data product are consistent with the recommended types of the best practice guide (this pair is optional).

E List of Figures

F List of Tables

G Glossary

assessment of a research data product is the task to map a research data product to $\mathbb{R}^+$ according to its quality, impact, and/or relevance; low numbers indicate a lower quality, impact, and/or relevance. If unspecified, our considerations apply to all three contexts. If only one of the three contexts is of interest, it is specified accordingly. The value a research data product is mapped onto is called a *score of a research data product* v, 1, 2, 4–8, 10, 12, 15–17, 28, 33, 42, 45, 46, 48, 52–55, 58, 59, 61, 102, 112, 116, 118, 119, 121, 126–128, 130, 132–136, 188, 190

benchmark for a research data product is understood to be an assessment of a research data product based on simulated interactions with the research data product: a computer program is used to check those characteristics of the research data product which are taken as signals for the effort put into its creation and curation. Examples for these characteristics are the compliance to data standards, the completeness of metadata, or the accessibility of the research data product via research data services. The benchmark's score of a research data product is determined by a combination of the checks' outcomes. v, 5–7, 18, 19, 21, 26, 28–33, 42, 44, 47, 48, 53, 57–59, 65–68, 70, 88, 89, 93, 99, 102, 117–119, 121–123, 126–137

controlled vocabulary is a list of terms with well-defined rules to add or (in rare occasions) delete terms. Typically, an institution or individual expert is in charge to oversee the curation of the vocabulary. 13, 14, 189

event-based metric is understood to be an assessment of a research data product based on the documentation of events of interactions with the research data product. The frequency of citations, mentions, or downloads are examples for event-based metrics' scores. v, 5–7, 18, 19, 21–33, 41–44, 47–55, 58, 62, 63, 65, 66, 106, 109, 112, 114–119, 121–123, 126–137, 140–149, 159, 183–185, 209

impact of a research data product is understood as the chance of a research data product to influence the direction of research of a peer in the same or a similar field v, 4, 5, 16, 26, 32, 33, 48, 49, 53, 54, 59, 62, 76, 91, 100, 116, 118, 119, 121, 126, 132, 133, 136, 142, 144, 147, 188

information retrieval is finding digital material of an unspecified nature that satisfies an information need from within large unstructured and potentially distributed collections (adapted definition of [SMR07]) 15

machine-actionable is a characteristic of tasks, meaning that machines can correctly process it without human interaction [Wil+16] v, 5, 10, 14–17, 26, 48, 51, 53, 55, 62, 65, 78, 95, 127, 133, 135, 136

ontology denotes a collection of formalized statements, typically about a specific domain. The statements are in formalized in the style of predicate logic, which allows to represent concepts and relations between these concepts, and to infer new statements. An ontology can be seen as a structured controlled vocabulary. 14

provenance is the context of creation of a research data product; it includes, but is not limited to: creators, contributors, sources, methodological approaches. 13

quality of a research data product is understood as the chance of a research data product to be (re-)used for tasks similar to the one for which the research data product was originally created v, 4, 5, 16, 26, 32, 33, 48, 49, 53, 54, 59, 62, 76, 91, 99, 101, 102, 116, 118, 119, 121, 126, 132, 133, 136, 142, 144, 147, 188

relevance of a research data product is understood as the chance of a research data product to influence an audience beyond the field of the creator(s). This includes outreach outside the scope of academic research v, 4, 5, 16, 26, 32, 33, 46, 48, 49, 53, 54, 59, 62, 76, 91, 116, 118, 119, 121, 126, 132, 133, 136, 142, 144, 147, 188

research denotes academic activities in the context of science and the humanities 2, 48, 54

research data are understood as all forms of digitized content that is input for or output of those activities of researchers, that are necessary to produce or verify knowledge ([WK18]). In our work, this concept has a broader sense compared to the literature, where it is typically used to differentiate supplemented material (e.g. tabular data) from publications in the classical sense (books and articles). Instead, we consider all of the above to be research data 2, 4, 5, 7–10, 12, 14, 16, 20, 24, 27, 45, 47, 91, 190

research data management is understood as the set of all tasks related to the creation, publication, and curation of research data products v, 7, 8, 10, 12, 14, 16, 64, 126, 127, 134, 135, 137

research data product is the combination of research data, metadata describing them and services hosting both v, 1–8, 10–12, 14–19, 21–23, 25–33, 37, 42–55, 57–76, 78, 81–83, 86–91, 93–102, 106–119, 121–123, 126–137, 141, 143–149, 159, 180–184, 188–190

research data service is a service providing access to either research data, metadata describing these data, or both. A research data service can be uniquely identified by an endpoint and a protocol 2, 5, 15, 74, 188, 190

scientometrics is a field studying the quantifiable aspects of academic output. It is often considered a sub-field of bibliometrics (the study of quantifiable aspect of books, articles and other publications), but in our work we use this term to include the study of any scientific digital output, including data which cannot (easily) be printed in a human-readable format. v, 7, 46, 52, 91, 106

score of a research data product is the value a research data product is mapped onto by an assessment of a research data product 5, 6, 25, 29, 31–33, 42, 43, 45, 48, 49, 51, 53, 54, 61, 62, 64–68, 71, 85, 86, 97, 106, 111–114, 117–119, 121–123, 126, 127, 130, 131, 133, 135, 136, 141–149, 159, 188

H Bibliography

[ALW19] Dag W. Aksnes, Liv Langfeldt, and Paul Wouters. "Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories". In: *SAGE Open* 9.1 (2019), p. 2158244019829575. DOI: 10.1177/2158244019829575.

[AR13] Euan Adie and William Roe. "Altmetric: enriching scholarly content with article-level discussion and metrics". In: *Learned Publishing* 26.1 (2013), pp. 11–17. DOI: 10.1087/20130103.

[AT14] A. Abrizah and Mike Thelwall. "Can the impact of non-Western academic books be measured? An investigation of Google Books and Google Scholar for Malaysia". In: *Journal of the Association for Information Science and Technology* 65.12 (2014), pp. 2498–2508. DOI: 10.1002/asi.23145.

[Ayr+16] P Ayris, JY Berthou, R Bruce, S Lindstaedt, A Monreale, B Mons, Y Murayama, C Södergård, K Tochtermann, and R Wilkinson. "Realising the European Open Science Cloud". In: *First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud* (2016).

[Bäc+17] Amelie Bäcker, Christian Pietsch, Friedrich Summann, and Sebastian Wolf. "BASE (Bielefeld Academic Search Engine). Eine Suchmaschinenlösung zur Indexierung wissenschaftlicher Metadaten". In: *Datenbank-Spektrum* 17.1 (2017), pp. 5–13.

[Bar18] Lorena A. Barba. *Terminologies for Reproducible Research*. 2018. arXiv: 1802.03311 [cs.DL].

[BD08] Lutz Bornmann and Hans-Dieter Daniel. "What do citation counts measure?: A review of studies on citing behavior". In: *Journal of Documentation* 64.1 (Jan. 2008), pp. 45–80. ISSN: 0022-0418. DOI: 10.1108/00220410810844150.

[BF+14] Pierre Bourque, Richard E Fairley, et al. *Guide to the software engineering body of knowledge (SWEBOK (R)): Version 3.0*. Ed. by Pierre Bourque and Richard E. (Dick) Fairley. IEEE Computer Society Press, 2014.

[BH04] Thomas R. Bruce and Diane I. Hillmann. "Metadata in Practice". In: ed. by Diane I. Hillmann and Elaine L. Westbrooks. ALA Editions, 2004. Chap. The Continuum of Metadata Quality: Defining, Expressing, Exploiting, pp. 238–256.

[BH18a] Bradley Wade Bishop and Carolyn Hank. "Measuring FAIR Principles to Inform Fitness for Use." In: *IJDC* 13.1 (2018), pp. 35–46.

[BH18b] Lutz Bornmann and Robin Haunschild. "Alternative article-level metrics: The use of alternative metrics in research evaluation". In: *EMBO reports* 19.12 (Dec. 2018), e47260. ISSN: 1469-3178.

[BHS09] Gordon Bell, Tony Hey, and Alex Szalay. "Beyond the Data Deluge". In: *Science* 323.5919 (Mar. 2009), pp. 1297–1298. ISSN: 1095-9203. DOI: 10.1126/science.1170411.

[BI04] Lennart Björneborn and Peter Ingwersen. "Toward a basic framework for webometrics". In: *Journal of the American Society for Information Science and Technology* 55.14 (2004), pp. 1216–1227. DOI: 10.1002/asi.20077.

[BK11] Christoph Bartneck and Servaas Kokkelmans. "Detecting h-index manipulation through self-citation analysis". In: *Scientometrics* 87.1 (2011), pp. 85–98. ISSN: 1588-2861. DOI: 10.1007/s11192-010-0306-5.

[BK17] Libby Bishop and Arja Kuula-Luumi. "Revisiting Qualitative Data Reuse: A Decade On". In: *SAGE Open* 7.1 (2017), p. 2158244016685136. DOI: 10.1177/2158244016685136.

[Bor14a] Lutz Bornmann. "Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics". In: *Journal of Informetrics* 8.4 (2014), pp. 895–903. ISSN: 1751-1577. DOI: 10.1016/j.joi.2014.09.005.

[Bor14b] Timo Borst. "Repositorien auf ihrem Weg in das Semantic Web: semantisch hergeleitete Interoperabilität als Zielstellung für künftige Repository-Entwicklungen". In: *Bibliothek Forschung und Praxis* 38.2 (2014), pp. 257–265.

[Bor16] Lutz Bornmann. "To what extent does the Leiden manifesto also apply to altmetrics?: A discussion of the manifesto against the background of research into altmetrics". In: 40.4 (Jan. 2016), pp. 529–543. ISSN: 1468-4527. DOI: 10.1108/OIR-09-2015-0314.

[Bou+12] Geoffrey Boulton, Philip Campbell, Brian Collins, Peter Elias, Wendy Hall, Graeme Laurie, Onora O'Neill, Michael Rawlins, J Thornton, Patrick Vallance, et al. *Science as an open enterprise*. Tech. rep. The Royal Society, 2012.

[Bro18] Meredith Broussard. *Artificial Unintelligence: How Computers Misunderstand the World*. MIT Press, Dec. 2018.

[Cha13] Scott Chamberlain. "Consuming Article-Level Metrics: Observations and Lessons". In: *Information Standards Quarterly* 2 (2013), pp. 4–13. URL: https://www.niso.org/sites/default/files/stories/2017-08/FE_Chamberlain_Consuming_ALMs_isq_v25no2.pdf.

[Cla+19] Daniel J.B. Clarke, Lily Wang, Alex Jones, Megan L. Wojciechowicz, Denis Torre, Kathleen M. Jagodnik, Sherry L. Jenkins, Peter McQuilton, Zachary Flamholz, Moshe C. Silverstein, Brian M. Schilder, Kimberly Robasky, Claris Castillo, Ray Idaszak, Stanley C. Ahalt, Jason Williams, Stephan Schurer, Daniel J. Cooper, Ricardo de Miranda Azevedo, Juergen A. Klenk, Melissa A. Haendel, Jared Nedzel, Paul Avillach, Mary E. Shimoyama, Rayna M. Harris, Meredith Gamble, Rudy Poten, Amanda L. Charbonneau, Jennie Larkin, C. Titus Brown, Vivien R. Bonazzi, Michel J. Dumontier, Susanna-Assunta Sansone, and Avi Ma'ayan. "FAIRshake: Toolkit to Evaluate the FAIRness of Research Digital Resources". In: *Cell Systems* 9.5 (2019), pp. 417–421. ISSN: 2405-4712. DOI: 10.1016/j.cels.2019.09.011.

[CLB10] Rodrigo Costas, Thed N. van Leeuwen, and María Bordons. "A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age on productivity and impact". In: *Journal of the American Society for Information Science and Technology* 61.8 (2010), pp. 1564–1581. DOI: 10.1002/asi.21348.

[Coo+15] Charles E. Cook, Mary Todd Bergman, Robert D. Finn, Guy Cochrane, Ewan Birney, and Rolf Apweiler. "The European Bioinformatics Institute in 2016: Data growth and integration". In: *Nucleic Acids Research* 44.D1 (Dec. 2015), pp. D20–D26. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1352.

[Cro01] Blaise Cronin. "Bibliometrics and beyond: some thoughts on web-based citation analysis". In: *Journal of Information Science* 27.1 (2001), pp. 1–7. DOI: 10.1177/016555150102700101.

[Cro14] David Crotty. "Altmetrics: Finding Meaningful Needles in the Data Haystack". In: *Serials Review* 40.3 (2014), pp. 141–146. DOI: 10.1080/00987913.2014.947839.

[CZW15] Rodrigo Costas, Zohreh Zahedi, and Paul Wouters. "Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective". In: *Journal of the Association for Information Science and Technology* 66.10 (2015), pp. 2003–2019. DOI: 10.1002/asi.23309.

[Dat14] Data Citation Synthesis Group. "Joint Declaration of Data Citation Principles". In: (2014). DOI: 10.25490/A97F-EGYK.

[Dat19] DataCite Metadata Working Group. *DataCite Metadata Schema for the Publication and Citation of Research Data v4.3*. Tech. rep. DataCite, 2019. DOI: 10.14454/F2WP-S162.

[Dij82] Edsger W. Dijkstra. "Selected Writings on Computing: A personal Perspective". In: *Selected Writings on Computing: A personal Perspective*. Ed. by David Gries. New York, NY: Springer New York, 1982. Chap. On the Role of Scientific Thought, pp. 60–66. ISBN: 978-1-4612-5695-3. DOI: 10.1007/978-1-4612-5695-3_12.

[DM14] Anup Kumar Das and Sanjaya Mishra. *Genesis of Altmetrics or Article-level Metrics for Measuring Efficacy of Scholarly Communications: Current Perspectives*. en. Tech. rep. 2014. URL: http://eprints.rclis.org/23581/.

[Dor13] Bertil Dorch. *Altmetrics to quantify the impact of scientific research published in open full text repositories*. en. 2013. URL: https://hal-hprints.archives-ouvertes.fr/hprints-00822129.

[DRT14] Emilio Delgado López-Cózar, Nicolás Robinson-García, and Daniel Torres-Salinas. "The Google scholar experiment: How to index false papers and manipulate bibliometric indicators". In: *Journal of the Association for Information Science and Technology* 65.3 (2014), pp. 446–454. DOI: 10.1002/asi.23056.

[FC20] Zhichao Fang and Rodrigo Costas. "Studying the accumulation velocity of altmetric data tracked by Altmetric.com". In: *Scientometrics* 123.2 (May 2020), pp. 1077–1101. ISSN: 1588-2861. DOI: 10.1007/s11192-020-03405-9.

[Fea14] Robin Featherstone. "Scholarly tweets: measuring research impact via altmetrics". In: *Journal of the Canadian Health Libraries Association/Journal de l'Association des bibliothèques de la santé du Canada* 35.2 (2014), pp. 60–63.

[Fen+18] Martin Fenner, Daniella Lowenberg, Matt Jones, Paul Needham, Dave Vieglais, Stephen Abrams, Patricia Cruse, and John Chodacki. *Code of practice for research data usage metrics release 1*. Tech. rep. PeerJ Preprints, 2018.

[Fen+19] Martin Fenner, Mercè Crosas, Jeffrey S. Grethe, David Kennedy, Henning Hermjakob, Phillippe Rocca-Serra, Gustavo Durand, Robin Berjon, Sebastian Karcher, Maryann Martone, and Tim Clark. "A data citation roadmap for scholarly data repositories". In: *Scientific Data* 6.1 (2019), p. 28. ISSN: 2052-4463. DOI: 10.1038/s41597-019-0031-8.

[FP16] Iztok Fister and Matjaž Perc. "Toward the Discovery of Citation Cartels in Citation Networks". In: *Frontiers in Physics* 4 (2016), p. 49. ISSN: 2296-424X. DOI: 10.3389/fphy.2016.00049.

[FW17] Seena Fazel and Achim Wolf. "What is the impact of a research publication?" In: *Evidence-based mental health* 20.2 (May 2017), pp. 33–34. ISSN: 1468-960X. DOI: 10.1136/eb-2017-102668.

[Gam+20] J.M. Gamble, Robyn L. Traynor, Anatoliy Gruzd, Philip Mai, Colin R. Dormuth, and Ingrid S. Sketris. "Measuring the impact of pharmacoepidemiologic research using altmetrics: A case study of a CNODES drug-safety article". In: *Pharmacoepidemiology and Drug Safety* 29.S1 (2020), pp. 93–102. DOI: 10.1002/pds.4401.

[Gam+94] Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley Professional, Nov. 1994. ISBN: 0201633612.

[Gar55] Eugene Garfield. "Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas". In: *Science* 122.3159 (1955), pp. 108–111. ISSN: 0036-8075. DOI: 10.1126/science.122.3159.108.

[GHA18] Koraljka Golub, Johan Hagelbäck, and Anders Ardö. "Automatic classification using DDC on the Swedish union catalogue". eng. In: *CEUR Workshop Proceedings*. Vol. 2200. Porto, Portugal: CEUR, 2018, pp. 4–16.

[Gre+20] Kathleen Gregory, Paul Groth, Andrea Scharnhorst, and Sally Wyatt. "Lost or Found? Discovering Data Needed for Research". In: *Harvard Data Science Review* (Apr. 30, 2020). DOI: 10.1162/99608f92.e38165eb.

[Gru+15] R. Grunzke, A. Aguilera, W. E. Nagel, J. Krüger, S. Herres-Pawlis, A. Hoffmann, and S. Gesing. "Managing Complexity in Distributed Data Life Cycles Enhancing Scientific Discovery". In: *2015 IEEE 11th International Conference on e-Science*. Aug. 2015, pp. 371–380. DOI: 10.1109/eScience.2015.72.

[Gru+17] Richard Grunzke, Tobias Adolph, Christoph Biardzki, Arndt Bode, Timo Borst, Hans-Joachim Bungartz, Anja Busch, Anton Frank, Christian Grimm, Wilhelm Hasselbring, Anastasia Kazakova, Atif Latif, Fidan Limani, Mathis Neumann, Nelson Tavares de Sousa, Jakob Tendel, Ingo Thomsen, Klaus Tochtermann, Ralph Müller-Pfefferkorn, and Wolfgang E. Nagel. "Challenges in Creating a Sustainable Generic Research Data Infrastructure". In: *4th Collaborative Workshop on Evolution and Maintenance of Long-Living Software Systems (EMLS'17)*. Vol. 37. Softwaretechnik Trends 2. Feb. 2017, pp. 74–77.

[Gru14] Thorsten Gruber. "Academic sell-out: how an obsession with metrics and rankings is damaging academia". In: *Journal of Marketing for Higher Education* 24.2 (2014), pp. 165–177. DOI: 10.1080/08841241.2014.970248.

[Gus19] Michael Gusenbauer. "Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases". In: *Scientometrics* 118.1 (2019), pp. 177–214.

[Hab19] Ted Habermann. "MetaDIG recommendations for FAIR DataCite metadata". In: (2019). DOI: 10.5438/2CHG-B074.

[Ham14] Björn Hammarfelt. "Using altmetrics for assessing research impact in the humanities". In: *Scientometrics* 101.2 (2014), pp. 1419–1430. ISSN: 1588-2861. DOI: 10.1007/s11192-014-1261-3.

[Har09] Stevan Harnad. "Open Access Scientometrics and the UK Research Assessment Exercise". en. In: *Scientometrics* 79.1 (2009), pp. 147–156. DOI: 10.1007/s11192-009-0409-z.

[Hau+14a] Stefanie Haustein, Timothy D. Bowman, Kim Holmberg, Andrew Tsou, Cassidy R. Sugimoto, and Vincent Larivière. "Tweets as impact indicators: Examining the implications of automated bot accounts on Twitter". In: *ArXiv* (2014). arXiv: 1410.4139 [cs.DL].

[Hau+14b] Stefanie Haustein, Isabella Peters, Cassidy R. Sugimoto, Mike Thelwall, and Vincent Larivière. "Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature". In: *Journal of the Association for Information Science and Technology* 65.4 (2014), pp. 656–669. DOI: 10.1002/asi.23101.

[Hau12] Stefanie Haustein. *Multidimensional Journal Evaluation*. Berlin, Boston: De Gruyter Saur, 2012. DOI: 10.1515/9783110255553.

[Hau16] Stefanie Haustein. "Grand challenges in altmetrics: heterogeneity, data quality and dependencies". In: *Scientometrics* 108.1 (July 2016), pp. 413–423. ISSN: 1588-2861. DOI: 10.1007/s11192-016-1910-9.

[Hay73] Friedrich August von Hayek. "Modes of Individualism and Collectivism". In: ed. by John O'Neill. Gregg Revivals, 1973. Chap. Scientism and the Study of Society, pp. 27–67.

[HB15] Torsten Hoefler and Roberto Belli. "Scientific Benchmarking of Parallel Computing Systems: Twelve Ways to Tell the Masses When Reporting Performance Results". In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. SC '15. Austin, Texas: Association for Computing Machinery, 2015. ISBN: 9781450337236. DOI: 10.1145/2807591.2807644.

[HBC16] Stefanie Haustein, Timothy D. Bowman, and Rodrigo Costas. "Interpreting "altmetrics": viewing acts on social media through the lens of citation and social theories". In: *ArXiv* abs/1502.05701 (2016).

[HCL15] Stefanie Haustein, Rodrigo Costas, and Vincent Larivière. "Characterizing Social Media Metrics of Scholarly Papers: The Effect of Document Properties and Collaboration Patterns". In: *PLOS ONE* 10.3 (Mar. 2015), pp. 1–21. DOI: 10.1371/journal.pone.0120495.

[Hic+15] Diana Hicks, Paul Wouters, Ludo Waltman, Sarah De Rijcke, and Ismael Rafols. "Bibliometrics: the Leiden Manifesto for research metrics". In: *Nature News* 520.7548 (2015), p. 429.

[Hir05] J. E. Hirsch. "An index to quantify an individual's scientific research output". In: *Proceedings of the National Academy of Sciences* 102.46 (2005), pp. 16569–16572. DOI: 10.1073/pnas.0507655102.

[Hol19] Alex O Holcombe. "Contributorship, not authorship: Use credit to indicate who did what". In: *Publications* 7.3 (2019), p. 48.

[JA15] Hamid R. Jamali and Dariush Alimohammadi. "Blog Citations as Indicators of the Societal Impact of Research : Content Analysis of Social Sciences Blogs". EN. In: *International Journal of Knowledge Content Development & Technology* (2015). DOI: 10.5865/IJKCT.2015.5.1.015.

[Jim+17] I. Jimenez, M. Sevilla, N. Watkins, C. Maltzahn, J. Lofstead, K. Mohror, A. Arpaci-Dusseau, and R. Arpaci-Dusseau. "The Popper Convention: Making Reproducible Systems Evaluation Practical". In: *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. May 2017, pp. 1561–1570. DOI: 10.1109/IPDPSW.2017.157.

[Jin10] Arif E Jinha. "Article 50 million: an estimate of the number of scholarly articles in existence". In: *Learned Publishing* 23.3 (2010), pp. 258–263.

[JPH18] Arlette Jappe, David Pithan, and Thomas Heinze. "Does bibliometric research confer legitimacy to research assessment practice? A sociological study of reputational control, 1972-2016". In: *PLOS ONE* 13.6 (June 2018), pp. 1–28. DOI: 10.1371/journal.pone.0199031.

[Kay17] Michael Kay. *XSL Transformations (XSLT) Version 3.0*. W3C Recommendation. W3C, June 2017. URL: https://www.w3.org/TR/2017/REC-xslt-30-20170608/.

[Ke+15] Qing Ke, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. "Defining and identifying Sleeping Beauties in science". In: *Proceedings of the National Academy of Sciences* 112.24 (2015), pp. 7426–7431. ISSN: 0027-8424. DOI: 10.1073/pnas.1424329112.

[Ken38] Maurice G Kendall. "A new measure of rank correlation". In: *Biometrika* 30.1/2 (1938), pp. 81–93.

[Kra+15a] Peter Kraker, Elisabeth Lex, Juan Gorraiz, Christian Gumpenberger, and Isabella Peters. "Research Data Explored II: the Anatomy and Reception of figshare". English. In: *Proceedings of the 20th International Conference on Science and Technology Indicators (STI 2015)*. 2015.

[Kra+15b] Peter Kraker, Elisabeth Lex, Juan Gorraiz, Christian Gumpenberger, and Isabella Peters. "Research Data Explored II: the Anatomy and Reception of figshare". English. In: *Proceedings of the 20th International Conference on Science and Technology Indicators (STI 2015)*. 2015.

[Küm+19] Sonja Kümmet, Stephan Lücke, Julian Schulz, Martin Spenger, and Tobias Weber. *DataCite Best Practice Guide*. Version Version 1.0. Nov. 2019. DOI: 10.5281/zenodo.3559800.

[Küm+20] Sonja Kümmet, Stephan Lücke, Julian Schulz, Martin Spenger, and Tobias Weber. "Standardizing a Standard: Why and how a Best Practice Guide for the DataCite Metadata Schema was created". In: *Korpus im Text* (2020). URL: http://www.kit.gwi.uni-muenchen.de/?p=51272&v=1.

[Lan11] Betty Landesman. "Seeing Standards: A Visualization of the Metadata Universe http://www.dlib.indiana.edu/ jenlrile/metadatamap/". In: *Technical Services Quarterly* 28.4 (2011), pp. 459–460. DOI: 10.1080/07317131.2011.598072.

[Lei17] Leibniz Rechenzentrum der Bayerischen Akademie der Wissenschaften. *Jahresbericht 2017*. Tech. rep. Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities, 2017. URL: https://www.lrz.de/wir/berichte/JB/JBer2017.pdf.

[LF13] Jennifer Lin and Martin Fenner. "Altmetrics in evolution: Defining and redefining the ontology of article-level metrics". In: *Information standards quarterly* 25.2 (2013), p. 20.

[LSJ15] Shibo Li, Eugene Sivadas, and Mark S. Johnson. "Explaining article influence: capturing article citability and its dynamic effects". In: *Journal of the Academy of Marketing Science* 43.1 (2015), pp. 52–72. ISSN: 1552-7824. DOI: 10.1007/s11747-014-0392-7.

[Lyn08] Clifford Lynch. "Big data: How do your data grow?" In: *Nature* 455.7209 (2008), pp. 28–29.

[Mac+18] Shona Mackinnon, Bogna A. Drozdowska, Michael Hamilton, Anna H. Noel-Storr, Rupert McShane, and Terry Quinn. "Are methodological quality and completeness of reporting associated with citation-based measures of publication impact? A secondary analysis of a systematic review of dementia biomarker studies". In: *BMJ Open* 8.3 (Mar. 2018). DOI: 10.1136/bmjopen-2017-020331.

[May+17] Matthew S. Mayernik, David L. Hart, Keith E. Maull, and Nicholas M. Weber. "Assessing and tracing the outcomes and impact of research infrastructures". In: *Journal of the Association for Information Science and Technology* 68.6 (2017), pp. 1341–1359. DOI: 10.1002/asi.23721.

[McG81] Joseph E McGrath. "Dilemmatics: The study of research choices and dilemmas". In: *American Behavioral Scientist* 25.2 (1981), pp. 179–210.

[Mer88] Robert K. Merton. "The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property". In: *Isis* 79.4 (1988), pp. 606–623. DOI: 10.1086/354848.

[MH15] Henk F. Moed and Gali Halevi. "Multidimensional assessment of scholarly research impact". In: *Journal of the Association for Information Science and Technology* 66.10 (2015), pp. 1988–2002. DOI: 10.1002/asi.23314. eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23314. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23314.

[Mon+17] Barend Mons, Cameron Neylon, Jan Velterop, Michel Dumontier, Luiz Olavo Bonino da Silva Santos, and Mark D Wilkinson. "Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud". In: *Information Services & Use* 37.1 (2017), pp. 49–56.

[MT14] Ehsan Mohammadi and Mike Thelwall. "Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows". In: *Journal of the Association for Information Science and Technology* 65.8 (2014), pp. 1627–1638. DOI: 10.1002/asi.23071.

[NS14] Lars Holm Nielsen and Tim Smith. *Zenodo Overview*. Mar. 2014. DOI: 10.5281/zenodo.8428.

[Obe17] Ursula Oberst. "Measuring the Societal Impact of Research with Altmetrics: An Experiment". In: *027.7 Zeitschrift für Bibliothekskultur / Journal for Library Culture* 5.1 (2017), pp. 16–21. ISSN: 2296-0597. DOI: 10.12685/027.7-5-1-167.

[Opp15] Harmen Oppewal. "Citations as a currency: Every performance measure creates its own behaviour: Commentary on the Soutar, Wilkinson, & Young article". In: *Australasian Marketing Journal (AMJ)* 23.2 (2015). Special Issue on Resource Management in Buyer-seller Relationships, pp. 162–164. ISSN: 1441-3582. DOI: 10.1016/j.ausmj.2015.04.011.

[OT08] William L. Oberkampf and Timothy G. Trucano. "Verification and validation benchmarks". In: *Nuclear Engineering and Design* 238.3 (2008). Benchmarking of CFD Codes for Application to Nuclear Reactor Safety, pp. 716–743. ISSN: 0029-5493. DOI: 10.1016/j.nucengdes.2007.02.032.

[Pat16] Dimple Patel. "Research data management: a conceptual framework". In: *Library Review* 65.4/5 (2016), pp. 226–241. DOI: 10.1108/LR-01-2016-0001.

[PC09] Jung-ran Park and Eric Childress. "Dublin Core metadata semantics: an analysis of the perspectives of information professionals". In: *Journal of Information Science* 35.6 (2009), pp. 727–739. DOI: 10.1177/0165551509337871.

[Pen+16] Ge Peng, Nancy A Ritchey, Kenneth S Casey, Edward J Kearns, Jeffrey L Privette, Drew Saunders, Philip Jones, Tom Maycock, and Steve Ansari. "Scientific stewardship in the Open Data and Big Data era—Roles and responsibilities of stewards and other major product

stakeholders". In: *D-Lib Magazine* 22.5/6 (May 2016). DOI: 10.1045/may2016-peng.

[Pet+14] Isabella Peters, Alexandra Jobmann, Christian P Hoffmann, Sylvia Künne, Jasmin Schmitz, and Gabriele Wollnik-Korn. "Altmetrics for large, multidisciplinary research groups: Comparison of current tools". In: *Bibliometrie — Praxis und Forschung* 3 (2014).

[Pet+16] Isabella Peters, Peter Kraker, Elisabeth Lex, Christian Gumpenberger, and Juan Gorraiz. "Research data explored: an extended analysis of citations and altmetrics". In: *Scientometrics* 107.2 (May 2016), pp. 723–744. ISSN: 1588-2861. DOI: 10.1007/s11192-016-1887-4.

[Pet+17] Isabella Peters, Peter Kraker, Elisabeth Lex, Christian Gumpenberger, and Juan Ignacio Gorraiz. "Zenodo in the Spotlight of Traditional and New Metrics". In: *Frontiers in Research Metrics and Analytics* 2 (2017), p. 13. ISSN: 2504-0537. DOI: 10.3389/frma.2017.00013.

[PGT12] Jason Priem, Paul Groth, and Dario Taraborelli. "The Altmetrics Collection". In: *PLOS ONE* 7.11 (Nov. 2012), pp. 1–2. DOI: 10.1371/journal.pone.0048753.

[PPH12] Jason Priem, Heather A. Piwowar, and Bradley M. Hemminger. *Altmetrics in the wild: Using social media to explore scholarly impact.* 2012. arXiv: 1203.4745 [cs.DL].

[Pri+10] J. Priem, D. Taraborelli, P. Groth, and C. Neylon. *Altmetrics: A manifesto.* http://altmetrics.org/manifesto. [accessed 2020-07-24]. Oct. 2010.

[PV13] Heather A. Piwowar and Todd J. Vision. "Data reuse and the open data citation advantage". In: *PeerJ* 1 (Oct. 2013), e175. ISSN: 2167-8359. DOI: 10.7717/peerj.175.

[PW86] David A Padua and Michael J Wolfe. "Advanced compiler optimizations for supercomputers". In: *Communications of the ACM* 29.12 (1986), pp. 1184–1201.

[Raa04] Anthony F. J. van Raan. "Sleeping Beauties in science". In: *Scientometrics* 59.3 (2004), pp. 467–472. ISSN: 1588-2861. DOI: 10.1023/B:SCIE.0000018543.82441.f1.

[Rav+17] James Ravenscroft, Maria Liakata, Amanda Clare, and Daniel Duma. "Measuring scientific impact beyond academia: An assessment of existing impact metrics and proposed improvements". In: *PLOS ONE* 12.3 (Mar. 2017), pp. 1–21. DOI: 10.1371/journal.pone.0173152.

[RCZ19] Guillaume Rousseau, Roberto Di Cosmo, and Stefano Zacchiroli. *Growth and Duplication of Public Source Code over Time: Provenance Tracking at Scale*. 2019. arXiv: 1906.08076 [cs.SE].

[RF13] Rousseau Ronald and Y. Ye Fred. "A multi-metric approach for research evaluation". In: *Chinese Science Bulletin* 58.26 (Sept. 2013), pp. 3288–3290. ISSN: 1861-9541. DOI: 10.1007/s11434-013-5939-3.

[RGR18] David Reinsel, John Gantz, and John Rydning. "The digitization of the world: from edge to core". In: *Framingham: International Data Corporation* (2018). URL: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.

[Rob+17] Nicolas Robinson-Garcia, Philippe Mongeon, Wei Jeng, and Rodrigo Costas. "DataCite as a novel bibliometric source: Coverage, strengths and limitations". In: *Journal of Informetrics* 11.3 (2017), pp. 841–854. ISSN: 1751-1577. DOI: 10.1016/j.joi.2017.07.003.

[Rod15] Costas Rodrigo. "The thematic orientation of publications mentioned on social media". In: 67.3 (Jan. 2015). Ed. by Dr Cassidy R. Sugimoto Dr Stefanie Haustein and Dr Vincent Larivière, pp. 260–288. ISSN: 2050-3806. DOI: 10.1108/AJIM-12-2014-0173.

[Rou+17] Nicolas P. Rougier, Konrad Hinsen, Frédéric Alexandre, Thomas Arildsen, Lorena A. Barba, Fabien C.Y. Benureau, C. Titus Brown, Pierre de Buyl, Ozan Caglayan, Andrew P. Davison, Marc-André Delsuc, Georgios Detorakis, Alexandra K. Diem, Damien Drix, Pierre Enel, Benoît Girard, Olivia Guest, Matt G. Hall, Rafael N. Henriques, Xavier Hinaut, Kamil S. Jaron, Mehdi Khamassi, Almar Klein, Tiina Manninen, Pietro Marchesi, Daniel McGlinn, Christoph Metzner, Owen Petchey, Hans Ekkehard Plesser, Timothée Poisot, Karthik Ram, Yoav Ram, Etienne Roesch, Cyrille Rossant, Vahid Rostami, Aaron Shifman, Joseph Stachelek, Marcel Stimberg, Frank Stollmeier, Federico Vaggi, Guillaume Viejo, Julien Vitay, Anya E. Vostinar, Roman Yurchak, and Tiziano Zito. "Sustainable computational science: the ReScience initiative". In: *PeerJ Computer Science* 3 (Dec. 2017), e142. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.142.

[RS16] Francesco Ronzano and Horacio Saggion. "An empirical assessment of citation information in scientific summarization". In: *international conference on applications of natural language to information systems*. Springer. 2016, pp. 318–325.

[San+19] Stephanie van de Sandt, Artemis Lavasa, Sünje Dallmeier-Tiessen, and Vivien Petras. "The Definition of Reuse". In: *Data Science Journal* 18.1 (2019), p. 22.

[SD19] Mozhdeh Salajegheh and Sareh Dayari. "Comparing the Citations Counts and Altmetrics of the Top Medical Science Journals in Scopus". In: *International Journal of Information Science and Management (IJISM)* 17.1 (2019), p. 59.

[SG06] Alexander Szalay and Jim Gray. "2020 Computing: Science in an exponential world". In: *Nature* 440.7083 (2006), p. 413.

[SGS17] Miguel-Angel Sicilia, Elena García-Barriocanal, and Salvador Sánchez-Alonso. "Community Curation in Open Dataset Repositories: Insights from Zenodo". In: *Procedia Computer Science* 106 (2017). 13th International Conference on Current Research Information Systems, CRIS2016, Communicating and Measuring Research Responsibly: Profiling, Metrics, Impact,Interoperability, pp. 54–60. ISSN: 1877-0509. DOI: 10.1016/j.procs.2017.03.009.

[She+19] Hadas Shema, Oliver Hahn, Athanasios Mazarakis, and Isabella Peters. "Retractions from altmetric and bibliometric perspectives". In: *Information - Wissenschaft & Praxis* 70 (2019), pp. 110–98.

[She04] Peter T. Shepherd. "COUNTER: towards reliable vendor usage statistics". In: *VINE* 34.4 (Jan. 2004), pp. 184–189. ISSN: 0305-5728. DOI: 10.1108/03055720410570975.

[SIN13] SINTEF. "Big Data for better or worse: 90% of world's data generated over last two years". In: *SCIENCE DAILY* (2013). URL: www.sciencedaily.com/releases/2013/05/130522085217.htm.

[SKN16] Arfon M. Smith, Daniel S. Katz, and Kyle E. and Niemeyer. "Software citation principles". In: *PeerJ Computer Science* 2 (Sept. 2016), e86. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.86.

[SMR07] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *An introduction to information retrieval*. Cambridge University Press, 2007.

[Sou+18] Nelson Tavares de Sousa, Wilhelm Hasselbring, Tobias Weber, and Dieter Kranzlmüller. "Designing a Generic Research Data Infrastructure Architecture with Continuous Software Engineering". In: *Software Engineering Workshops 2018*. Vol. Vol-2066. CEUR Workshop Proceedings. CEUR-WS.org, Mar. 2018, pp. 85–88. URL: http://eprints.uni-kiel.de/42215/.

[SPB12] Xin Shuai, Alberto Pepe, and Johan Bollen. "How the Scientific Community Reacts to Newly Submitted Preprints: Article Downloads, Twitter Mentions, and Citations". In: *PLOS ONE* 7.11 (Nov. 2012), pp. 1–8. DOI: 10.1371/journal.pone.0047523.

[Spe04] Charles Spearman. "The Proof and Measurement of Association Between Two Things". In: *American Journal of Psychology* 15.1 (1904), pp. 72–101.

[Str97] Marilyn Strathern. "'Improving ratings': audit in the British University system". In: *European Review* 5.3 (1997), pp. 305–321. DOI: 10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4.

[Sug+17] Cassidy R. Sugimoto, Sam Work, Vincent Larivière, and Stefanie Haustein. "Scholarly use of social media and altmetrics: A review of the literature". In: *Journal of the Association for Information Science and Technology* 68.9 (2017), pp. 2037–2062. DOI: 10.1002/asi.23833.

[TC11] David Tarrant and Les Carr. "Using the Co-Citation Network to Indicate Article Impact". en. In: *Alt-Metrics Workshop @ Web Science 2011*. June 2011. URL: http://eprints.soton.ac.uk/id/eprint/272684.

[The+13] Mike Thelwall, Stefanie Haustein, Vincent Larivière, and Cassidy R. Sugimoto. "Do Altmetrics Work? Twitter and Ten Other Social Web Services". In: *PLOS ONE* 8.5 (May 2013), pp. 1–7. DOI: 10.1371/journal.pone.0064841.

[TJR14] Daniel Torres-Salinas, Evaristo Jiménez-Contreras, and Nicolas Robinson-García. "How many citations are there in the Data Citation Index?" In: *Proceedings of the STI Conference*. 2014. URL: https://arxiv.org/abs/1409.0753.

[Wal+11a] Ulli Waltinger, Alexander Mehler, Mathias Lösch, and Wolfram Horstmann. "Hierarchical Classification of OAI Metadata Using the DDC Taxonomy". In: *Advanced Language Technologies for Digital Libraries*. Ed. by Raffaella Bernardi, Sally Chambers, Björn Gottfried, Frédérique Segond, and Ilya Zaihrayeu. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 29–40. ISBN: 978-3-642-23160-5.

[Wal+11b] Ludo Waltman, Nees Jan van Eck, Thed N. van Leeuwen, Martijn S. Visser, and Anthony F.J. van Raan. "Towards a new crown indicator: Some theoretical considerations". In: *Journal of Informetrics* 5.1

(2011), pp. 37–47. ISSN: 1751-1577. DOI: 10.1016/j.joi.2010.08.001.

[Wal16] Ludo Waltman. "A review of the literature on citation impact indicators". In: *Journal of Informetrics* 10.2 (2016), pp. 365–391. ISSN: 1751-1577. DOI: 10.1016/j.joi.2016.02.007.

[Wan10] Alex Hai Wang. "Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach". In: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2010, pp. 335–342. DOI: 10.1007/978-3-642-13739-6_25.

[WC12] P. Wouters and R. Costas. "Users, Narcissism and control - Tracking the impact of scholarly publications in the 21st century." In: *Proceedings of 17th International Conference on Science and Technology Indicators*. Vol. 2. 2012, pp. 847–857.

[WC14] Ludo Waltman and Rodrigo Costas. "F1000 Recommendations as a Potential New Data Source for Research Evaluation: A Comparison With Citations". In: *Journal of the Association for Information Science and Technology* 65.3 (2014), pp. 433–445. DOI: 10.1002/asi.23040.

[Web+20] Tobias Weber, Dieter Kranzlmüller, Michael Fromm, and Nelson Tavares de Sousa. "Using supervised learning to classify metadata of research data by field of study". In: *Quantitative Science Studies* 1.2 (2020), pp. 525–550. DOI: 10.1162/qss_a_00049.

[Web18a] Tobias Weber. *Data Publication accompanying the paper "Methods to Evaluate Lifecycle Models for Research Data Management"*. Nov. 2018. DOI: 10.25927/002.

[Web18b] Tobias Weber. *Data Publication accompanying the paper "Methods to Evaluate Lifecycle Models for Research Data Management"*. Nov. 2018. DOI: 10.25927/002.

[Web18c] Tobias Weber. *Software Publication accompanying the paper "How FAIR can you get? Image Retrieval as a Use Case to calculate FAIR Metrics"*. Oct. 2018. DOI: 10.25927/001.

[Web19a] Tobias Weber. *l-sized Training and Evaluation Data for Publication "Using Supervised Learning to Classify Metadata of Research Data by Discipline of Research"*. Oct. 2019. DOI: 10.5281/zenodo.3490460.

[Web19b] Tobias Weber. *m-sized Training and Evaluation Data for Publication "Using Supervised Learning to Classify Metadata of Research Data by Discipline of Research"*. Oct. 2019. DOI: 10.5281/zenodo.3490458.

[Web19c] Tobias Weber. *Raw Data for Publication "Using Supervised Learning to Classify Metadata of Research Data by Discipline of Research"*. Oct. 2019. DOI: 10.5281/zenodo.3490329.

[Web19d] Tobias Weber. *s-sized Training and Evaluation Data for Publication "Using Supervised Learning to Classify Metadata of Research Data by Discipline of Research"*. Oct. 2019. DOI: 10.5281/zenodo.3490396.

[Web20] Tobias Weber. "Shortcomings of Usage Metrics for Research Data". in preparation. 2020.

[Wei+18] Jens Weismüller, Stephan Hachinger, Hai Nguyen, and Tobias Weber. "Addressing knowledge and know-how biases in the environmental sciences with modern data and compute services". In: *EGU General Assembly Conference Abstracts*. Vol. 20. 2018, p. 4399.

[WF19] Tobias Weber and Michael Fromm. *Source Code and Configurations for Publication "Using Supervised Learning to Classify Metadata of Research Data by Discipline of Research"*. Oct. 2019. DOI: 10.5281/zenodo.3490466.

[WFS19] Tobias Weber, Michael Fromm, and Nelson Tavares de Sousa. *Statistics and Evaluation Data for Publication "Using Supervised Learning to Classify Metadata of Research Data by Discipline of Research"*. Oct. 2019. DOI: 10.5281/zenodo.3490468.

[WHH15] Birge Wolf, Anna-Maria Häring, and Jürgen Heß. "Strategies towards Evaluation beyond Scientific Impact. Pathways not only for Agricultural Research". In: *Organic Farming* 1.1 (2015). ISSN: 2297-6485. URL: http://www.librelloph.com/organicfarming/article/view/of-1.1.3.

[Wic+14] Hadley Wickham et al. "Tidy data". In: *Journal of Statistical Software* 59.10 (2014), pp. 1–23.

[Wil+16] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3 (2016).

[Wil+17] JR Wilsdon, Judit Bar-Ilan, Robert Frodeman, Elisabeth Lex, Isabella Peters, and Paul Wouters. *Next-generation metrics: Responsible metrics and evaluation for open science*. 2017. DOI: 10.2777/337729.

[Wil+18a] Mark Wilkinson, Luiz Olavo Bonino, Nolan Nichols, and Katrin Leinweber. *FAIRMetrics/Metrics: Proposed FAIR Metrics and results of the Metrics evaluation questionnaire*. Mar. 2018. DOI: 10.5281/zenodo.1205235. URL: https://doi.org/10.5281/zenodo.1205235.

[Wil+18b] Mark D. Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos, and Michel Dumontier. "A design framework and exemplar metrics for FAIRness". In: *Scientific Data* 5.1 (2018), p. 180118. ISSN: 2052-4463. DOI: 10.1038/sdata.2018.118. URL: https://doi.org/10.1038/sdata.2018.118.

[Wil+19] Mark D. Wilkinson, Michel Dumontier, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Mario Prieto, Dominique Batista, Peter McQuilton, Tobias Kuhn, Philippe Rocca-Serra, Mercie Crosas, and Erik Schultes. "Evaluating FAIR maturity through a scalable, automated, community-governed framework". In: *Scientific Data* 6.1 (2019), p. 174. ISSN: 2052-4463. DOI: 10.1038/s41597-019-0184-5.

[WK18] Tobias Weber and Dieter Kranzlmüller. "How FAIR Can you Get? Image Retrieval as a Use Case to Calculate FAIR Metrics". In: *2018 IEEE 14th International Conference on e-Science (e-Science)*. Oct. 2018, pp. 114–124. DOI: 10.1109/eScience.2018.00027.

[WK19] Tobias Weber and Dieter Kranzlmüller. "Methods to Evaluate Lifecycle Models for Research Data Management". In: *Bibliothek Forschung und Praxis* 43 (2020 2019). 1, p. 75. ISSN: 18657648. DOI: 10.1515/bfp-2019-2016.

[WS18] Peter Wittenburg and George Strawn. *Common Patterns in Revolutionary Infrastructures and Data*. [Online; Accessed 2018-Nov-13]. Feb. 2018. URL: https://www.rd-alliance.org/sites/default/files/Common_Patterns_in_Revolutionising_Infrastructures-final.pdf.

[Zar05] Jerrold H. Zar. "Spearman Rank Correlation". In: *Encyclopedia of Biostatistics*. American Cancer Society, 2005. ISBN: 9780470011812. DOI: 10.1002/0470011815.b2a15150.

[ZC18] Zohreh Zahedi and Rodrigo Costas. "General discussion of data quality challenges in social media metrics: Extensive comparison of four major altmetric data aggregators". In: *PLOS ONE* 13.5 (May 2018), pp. 1–27. DOI: 10.1371/journal.pone.0197326.

[ZCW14] Zohreh Zahedi, Rodrigo Costas, and Paul Wouters. "How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications". In: *Scientometrics* 101.2 (Nov. 2014), pp. 1491–1513. ISSN: 1588-2861. DOI: 10.1007/s11192-014-1264-0.

[Zuc+15] Alesia Ann Zuccala, Frederik Verleysen, Roberto Cornacchia, and Tim Engels. "Altmetrics for the humanities: Comparing Goodreads reader ratings with citations to history books". In: 67.3 (Jan. 2015). Ed. by Dr Cassidy R. Sugimoto Dr Stefanie Haustein and Dr Vincent Larivière, pp. 320–336. ISSN: 2050-3806. DOI: 10.1108/AJIM-11-2014-0152.

I Data and Code Availability Statement

This subsection list additional material for this thesis to reproduce the findings or to re-use part of the developed resources to replicate the findings in a different setting. Subsection I.1 lists all retrieved or calculated data which provides evidence for the claims in this thesis. In Subsection I.2 the code to retrieve, calculate and evaluate the data is listed.

I.1 Data

- Tabular overview of shortcomings of event-based metrics, as used in Chapter 3: DOI: 10.5281/zenodo.4284733
- Raw data (scores of event-based metrics and benchmark runs), as used in Chapter 6: DOI: 10.5281/zenodo.4284737

I.2 Code

- Software used to create the benchmark scores as described in Chapter 5 and used in Chapter 6: DOI: 10.5281/zenodo.4284750
- Statistical scripts for Chapter 6 and the appendix: DOI: 10.5281/zenodo.4284747

J Acknowledgements

First of all, I want to thank Prof. Dr. Dieter Kranzlmüller who has been my supervisor during my thesis. His guidance was crucial for me, not only in creating this thesis, but also in the development of my own views on academia, research, and life in general. I am grateful that he gave me the opportunity to work in exciting settings, to meet interesting people and to face and master technical challenges at the Leibniz Supercomputing Centre and beyond.

I also have to thank Prof. Dr. Isabella Peters for interesting discussions and important insights. The opportunity to present ideas at her chair was a milestone to finish the thoughts that led to this thesis.

During my affiliation with the Leibniz Supercomputing Centre and the chair of Prof. Dr. Kranzlmüller many colleagues provided me with resources and council, discussed aspects of my thesis, or simply cheered me up, when the need arose. In alphabetical order I like to thank Dr. Werner Baur, Dr. Christoph Biardzki, Dr. Michael Brenner, Dr. Vitalian Danciu, Dr. Niels Fallenbeck, Dr. Nils gentschen Felde, Michael Fromm, Dr. Anton Frank, Tobias Fuchs, Sophia Grundner-Culemann, Tobias Guggemos, Pascal Jungblut, Dr. Stephan Hachinger, Dr. Helmut Heller, Maximilian Höb, Annette Kostelezky, Roger Kowalewski, Jan Schmidt, and many more. I also profited from many discussions with colleagues at the Universitätsbibliothek and the IT Gruppe Geisteswissenschaften. The LRZ compute cloud and the support offered by its staff were invaluable for the evaluation part of this thesis.

I am thankful to Nelson Tavares de Sousa with whom I shared the worries and delights of a PhD student's life, although his research took place at the other side of the country. I thank him and Dr. Timo Borst, Anja Busch, Dr. Christian Grimm, Dr. Richard Grunzke, Prof. Dr. Wilhelm Hasselbring, Dr. Jakob Tendel, and Robin Weiss for their collaboration during the GeRDI project, their insightful comments, the collaboratively published work, and last but not least for the fun we had.

I am highly appreciative of the DFG for granting the GeRDI project that supported me over the bigger part of my doctoral research. I thank RDA Europe and the Sloan Foundation for travel grants to Philadelphia and Rome, respectively. Martin Fenner (Datacite), Daniella Loewenberg (Make Data Count), and Stacy Konkiel (Altmetric) were invaluable partners not only for the support their organizations/projects provided me, but also as esteemed discussion partners.

My family and friends supported me throughout my PhD with encouragement and in times with desparetly needed diversion. Last, but most certainly not least, I thank Katja for her support — intellectually, emotionally, and financially.