universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

ViDa: Online Visualization Platform for Dataset Reuse

verfasst von | submitted by

Bc. Hynek Zemanec

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 935
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Medieninformatik
Betreut von | Supervisor:	Univ.-Prof. Torsten Möller PhD
Mitbetreut von | Co-Supervisor:	Laura Koesten MSc Ph.D.

Acknowledgements

Valuable accomplishments in any avenue of life require grit and fortitude. At the same time, we humans are remarkably adept at deluding ourselves that our achievements are entirely of our own making. However, we do not exist in vacuum. I strongly believe that gratitude for the external opportunities must balance our personal contribution. In keeping with that theme, I would like to thank the following individuals.

To my co-supervisors Torsten Möller, Laura Koesten, and Kathleen Gregory: thank you for giving me the chance to work on this topic and for the time and energy you have devoted to this project. Your expertise have continously helped to steer the course of the thesis in the right direction. To Laura in particular, thank you for your consistent feedback during our countless meetings, which continually moved the project forward, even as you patiently endured my occasional challenges.

To my parents, who have tirelessly stood by me both financially and emotionally throughout my years of study. Děkuji vám za vaši stálou podporu ve všech šťastných i těžkých obdobích mého života. Rodiče si člověk nevybírá, ale já bych vás za nic ani za nikoho nevyměnil. Mám vás moc rád.

To my stunning life partner and source of inspiration, Jaroslava, who shows me what it means to persevere. Your fearless pursuit of new experiences and your courage in facing challenges continually inspire me. Thank you for staying by my side, even when I hardly deserve your company.

Finally, I would like to thank my pug Gyros, who consistently reminded me of the importance of energy conservation and the fine art of relaxation.

Abstract

Millions of datasets are freely accessible online, yet their effective reuse remains limited. Data reusers face barriers when assessing whether a dataset is fit for a new purpose. Adopting a design science approach, this thesis combines insights from literature, a survey of 14 dataset repositories and portals, iterative low-fidelity prototyping, and evaluation through semi-structured interviews and user testing. A systematic evaluation against 47 recommendations reveals significant gaps in current platforms, particularly regarding data quality indicators and visualization. Semi-structured interviews with four data producers informed the design of two prototype versions through multiple iterations. The primary contribution is ViDa, an online dataset visualization platform designed for both data producers and reusers. ViDa aims to reduce documentation burden for producers through guided metadata workflows while providing interactive dashboards that combine contextual metadata with automatically generated visualizations through dataset profiling. These visualizations enable potential reusers to quickly gain an overview of datasets and make informed reuse decisions. The platform requires minimal technical expertise in visualization or metadata curation. To our knowledge, ViDa is the first non-commercial solution offering this suite of functionalities. User testing with ten participants achieved a usability score of 92.13%, indicating the platform is intuitive and effective for accomplishing visualization tasks under the tested conditions.

Kurzfassung

Millionen von Datensätzen sind online frei zugänglich, dennoch bleibt ihre effektive Wiederverwendung begrenzt. Datennutzende stoßen auf Hindernisse, wenn sie beurteilen möchten, ob ein Datensatz für einen neuen Zweck geeignet ist. Mit einem Design-Science-Ansatz kombiniert diese Arbeit Erkenntnisse aus der Literatur, einer Erhebung von 14 Datenrepositorien und -portalen, iterativem Low-Fidelity-Prototyping sowie einer Evaluation durch halbstrukturierte Interviews und Nutzertests. Eine systematische Bewertung anhand von 47 Empfehlungen deckt erhebliche Lücken in bestehenden Plattformen auf, insbesondere hinsichtlich Datenqualitätsindikatoren und Visualisierung. Halbstrukturierte Interviews mit vier Datenproduzierenden flossen über mehrere Iterationen in den Entwurf zweier Prototypversionen ein. Der zentrale Beitrag ist ViDa, eine Online-Plattform zur Datensatzvisualisierung, die sowohl für Datenproduzierende als auch für Datennutzende konzipiert ist. ViDa zielt darauf ab, den Dokumentationsaufwand für Produzierende durch geführte Metadaten-Workflows zu reduzieren und gleichzeitig interaktive Dashboards bereitzustellen, die kontextuelle Metadaten mit automatisch generierten Visualisierungen durch Datensatz-Profiling kombinieren. Die Plattform erfordert minimale technische Kenntnisse in Visualisierung und Metadatenkuration. Diese Visualisierungen ermöglichen potenziellen Wiederverwendenden, schnell einen Überblick über Datensätze zu erhalten und fundierte Wiederverwendungsentscheidungen zu treffen. Nach unserem Kenntnisstand ist ViDa die erste nicht-kommerzielle Lösung, die dieses Funktionsspektrum bietet. Nutzertests mit zehn Teilnehmenden ergaben einen Usability-Score von 92,13%, was darauf hindeutet, dass die Plattform unter den getesteten Bedingungen intuitiv und effektiv für die Durchführung von Visualisierungsaufgaben ist.

Contents

List of Tables

List of Figures

1. Introduction

"We must all accept that science is data and that data are science". This is how Hanson and colleagues [2] conclude their article, emphasizing the critical responsibility of publishers and data producers to curate data, enabling them to be as widely accessible as possible. Millions of datasets are now freely accessible on the web, creating what appears to be an unprecedented opportunity for scientific advancement. Reused across different contexts and disciplines, datasets may generate insights well beyond their original purpose. England's NHS Hospital Episode Statistics (HES) dataset illustrates this potential [3]. Originally collected for billing and administration, it later revealed higher weekend mortality and care quality issues, driving national policy reforms at minimal extra cost [4]. In another instance, a 2021 meta-analysis combining data from various studies on malaria during pregnancy revealed diminished effectiveness of sulfadoxine-pyrimethamine [5]. This prompted the WHO to revise its guidelines for malaria treatment in pregnant women [6]. Despite these success stories, the mere availability of data does not guarantee its visibility and usability [4].

Researchers, data practitioners and other audiences attempting to investigate unfamiliar datasets face a variety of barriers. Understanding the methodological choices and context that shaped data publication remains challenging [7]. Interpreting domain-specific variables presents similar difficulties, as does evaluating data quality [8]. These obstacles make it difficult to assess whether a dataset is fit for a new purpose. Recognition of such challenges has prompted various initiatives to improve dataset reusability. While the FAIR data principles initially focused on machine consumption, they are increasingly extended to support human understanding [9]. Alongside these efforts, frameworks such as the Data Nutrition Label aim to standardize dataset documentation and improve reusability [10]. Yet we still lack a clear understanding of what specific information individuals need to confidently reuse datasets and how this information should be presented.

To address these gaps, this thesis pursues three objectives: (1) to identify what contextual information data reusers need to assess dataset fitness for purpose, (2) to examine what features current data repositories provide to support dataset understanding and reuse, and (3) to derive the requirements a data visualization platform must meet to support data reuse. To achieve these objectives, it adopts a design science approach [11, 12], combining literature review, iterative design, and empirical evaluation. These efforts culminated in the implementation of **ViDa**, an online dataset visualization platform.

Grounded in best practices from the literature, ViDa addresses both sides of the reuse equation. For data producers, it reduces documentation burden through guided metadata workflows. For data reusers, it provides an interactive dashboard combining contextual metadata with visualizations that are automatically generated through dataset profiling. This approach requires little to no visualization and metadata curation expertise from

data publishers. While the underlying principles may generalize to other formats, ViDa focuses specifically on tabular datasets.

This thesis makes the following contributions:

- A systematic evaluation of data repositories and portals ($n = 14$) against an explicit list of 47 recommendations as identified by Gregory and Koesten [13]
- The design and implementation of *ViDa*, an automated dataset visualization platform that requires minimal technical expertise from users
- Insights into data producer and reuser needs gathered through prototype evaluation and ViDa user testing

The thesis is structured as a multi-study project concluded by a common discussion. Chapter 3 reports on a survey of 14 dataset repositories and portals, comparing their features against 47 recommendations identified by Gregory and Koesten [13]. This categorization effort appears to be the first to systematically evaluate a broad set of public data platforms using an explicit feature checklist. The findings offer insights into current practices, informing subsequent design decisions in the next phase of the project as described in Chapter 4. Likewise, these findings have the potential to inform the design decisions of developers and designers of data repositories and portals.

Chapter 4 describes the iterative design of low-fidelity prototypes and their evaluation through semi-structured interviews with data producers. Conducted sessions tested design assumptions, identified usability concerns, and generated actionable feedback that later shaped platform design choices.

Building on these earlier stages, Chapter 5 presents the design and implementation of ViDa. The platform includes a dataset profiler, automated file processing workflows, and a user interface designed to make dataset context more accessible. Implementation details, architectural decisions, and optimization measures are documented to support future development.

Chapter 6 reports on user testing using a mixed-methods evaluation that combined participant feedback with task-based indicators. The results suggest the platform is highly usable under the study conditions. The final chapter 7 discusses limitations of the conducted studies, lessons learned, and directions for future work.

2. Research Background

As data volumes and data-driven methods in science continue to expand rapidly, a new scientific model, the *fourth paradigm* coined by Jim Gray [14], has emerged. This most recent approach, sometimes referred to as data-intensive scientific discovery, is characterized by the reliance on massive datasets captured from instruments, simulations, or experiments, combined with computational methods to derive knowledge and insights [14]. This shift has led to the emergence of new disciplines such as astroinformatics, computational biology, and digital humanities [15]. Deliberate data curation, defined by Johnston et al. [16] as the *"various actions taken to ensure that data are fit for purpose and available for discovery and reuse"*, is a foundational component of both long-established and newly emerging disciplines that enables data reuse [2, 14, 17].

A helpful point of departure lies in historical precedent. Consider Galileo's *Sidereus Nuncius*, which demonstrates the value of meticulous documentation. When recording his observations of Jupiter's moons, he attended to more than just numbers. He captured timing, meteorological conditions, and instrument specifications alongside his measurements. His records included the methods he employed and his analytical reasoning. Goodman and colleagues [17] highlight this precedent precisely because contemporary science, they argue, often neglects such thoroughness. A tension arises where data can exist without being easily usable or discoverable. The mere availability of a dataset does not guarantee its usefulness to others [7]. Datasets that are well-curated and structured for reuse may offer considerable benefits for addressing multi-disciplinary problems, such as climate change or poverty [18, 19, 20]. However, despite numerous reported success stories (e.g. [3, 6]), in practice the reality frequently appears to fall short of this promise [17, 21]. Policy frameworks, research communities and governing bodies of public research institutions advocate increasingly for data reuse [22, 17, 15, 23]. At the same time, the sheer quantity of available data introduces its own challenges [24, 14]. Identifying relevant datasets within one's scientific domain is itself labour-intensive [17, 25, 26]. Researchers have produced extensive recommendations and best-practice guidelines aimed at improving reuse (e.g. [10, 27, 28, 13]). However, their implementation is inconsistent, and there has been little comprehensive, systematic evaluation [7]. This thesis seeks in part to narrow this gap.

2.1. Conceptualizing Data

Numerous characterizations have been introduced in the literature to describe the heterogeneity of data. A frequently referenced definition of data by the National Research Council [29] describes data as encompassing everything from factual information to

alphanumeric or other symbolic representations that can characterize an object, state, idea, condition, event, or similar entities. Among the various attempts, Buckland [30] provides one of the most terse definitions, describing data as an *"alleged evidence"*. Consistent with this definition, which frames data as a snapshot in time, Borgman [15] identifies a range of perspectives on what counts as data, including samples, software, field notes, instrument calibrations, archival records, codebooks, and others. She emphasizes that data may exist as information objects in both physical and digital forms, and that these forms seldom appear in isolation.

Data has been described as anything from geospatial coordinates, numerical values and measurements, literature corpora, images or physical samples that may provide evidence of a phenomenon [13]. Such data may aid the analysis of a particular question. In other framings, particularly within research contexts, Borgman[31] characterizes data as *"representations of observations, objects, or other entities that are used as evidence for the purposes of research or scholarship"*.

2.1.1. Categorizations of Data

Different scientific disciplines often rely on their own specialized terminology to interpret and organize data [32]. The 2005 report of the National Science Board [33] classifies data into four main categories according to their source. The first type, *observational* data, may consist of diverse records of physical phenomena, such as weather and atmospheric measurements linked to specific geospatial locations and temporal attributes. The second category consists of *computational* data, produced by running models and simulations. The third category, *experimental* data, is described as measured results obtained from studies conducted under controlled conditions. The final category may consist of records produced by various institutions and groups, such as government bodies, businesses, social organizations, and documentation of historical events [33]. While this framework may be useful, Borgman [15] argues that such categorization often masks the diversity of data types.

Gregory and Koesten [13] offer a more general outlook. According to their view, data can be understood through two fundamental lenses: either as an *object* or as a *process*. In the former case, data are understood as queryable entities that can be *"stored, found, accessed, retrieved, consumed, maintained, archived, sent somewhere, looked at or sold"* [13]. In the latter case, data are considered in relation to specific scientific practices, procedures, and technologies. They are never treated as standalone entities but are inherently linked to a particular context [34]. Along similar lines, emphasizing that data should not be regarded as isolated entities, Gregory and Koesten [13] argue for adoption of the latter data-as-process perspective. They assert that discovering and making sense of data in the absence of context is often challenging, if not impossible.

2.1.2. Datasets

To facilitate processing, distribution, and publication, data are typically organized into datasets, understood here as themed collections of data assembled for specific

purposes [35, 36]. Datasets can be divided into primary and derived categories, where primary datasets consist of data obtained directly through experiments, whereas derived datasets comprise processed outputs, such as meta-analyses and other analyzed results [4]. Beyond these intuitive definitions, Renear et al. propose a formal characterization grounded in four defining properties: logical grouping, shared content, internal relatedness, and purpose [37]. How data are structured, however, varies considerably across contexts [4]. We can observe spreadsheets (tabular formats such as CSV), graphs, fields, geometric representations, or collections of related artifacts depending on disciplinary needs and application requirements [38, pp. 21-30].

Structured data tend to be created with particular purposes in mind. The arrangement and formatting of such data often reflects adherence to more formalized conventions [39, 40, 4], suggesting an underlying relationship between organizational choices and intended use. This coupling also raises questions about how data originally structured for one purpose might be repurposed or reinterpreted in new contexts.

2.1.3. Metadata

Colloquially, metadata describes information about the data [17]. Yet, determining what actually constitutes metadata as opposed to data itself may often be elusive [38, p. 23]. More formally, metadata might be characterized as structured contextual documentation that supports identification, representation, interoperability, and broader management and use of data [41, 42]. Extending this definition, Loeffler et al. assert that metadata are created for purposes of search, classification, or knowledge derivation [36]. They highlight that metadata answer the *W-questions*: *"What has been measured by Whom, When, Where and Why?"* [36].

Gregory and Koesten [13] frame metadata creation as a sociotechnical process. This perspective recognizes that both human and machine processes shape metadata. Further, they proceed to identify three properties that appear to facilitate data discovery, sense-making, and data interaction. Interoperability of the data coupled with open classification represents the first dimension. Rich, linked documentation forms another. The last dimension involves ongoing creation and curation practices [13].

A useful framework for metadata categorization is offered by Gheel and Anderson [43], who propose distinguishing between *intrinsic* and *extrinsic* forms of metadata. Critically, these categories are defined not only by their inherent nature but also by how they originate. Some are manually assigned while others arise through automated processes. Intrinsic metadata, according to their classification, comprises properties that remain specific to the data object itself. Title, author, category and keywords are examples of manually determined metadata. The automated metadata may include associated URL, file size, or creation date. Properties of extrinsic metadata lie outside the data object rather than being embedded within it. For example, citations, comments, and highlights represent manually determined extrinsic metadata, whereas automatically captured instances might be the number of accesses or the timestamp of the last access. Similarly, documentation of visualization content or revision counts also falls into this category [43].

Metadata serving both machine and human needs necessitates standardization and shared vocabulary [9]. General metadata standards, while theoretically sufficient for locating datasets, may fall short when practitioners need deeper understanding of the data [13]. Just like in the case of data categorization, a critique from Bowker and Star [44] reminds us that classification systems, by their nature, simplify complexity and are therefore inherently reductionist. When applied to metadata standards, this simplification risks masking the disorder and inconsistency that characterize real-world data. Discovery systems typically rely on assigned keywords rather than on the variables contained within datasets themselves. If a researcher uses terminology absent from the metadata record, the dataset may remain undiscovered, regardless of whether it contains relevant information [13].

These challenges suggest that creating a single universal metadata standard is neither realistic nor desirable. Instead , data discovery systems would benefit from accommodating multiple metadata approaches, each suited to particular disciplinary or functional needs. *Schema.org*[1], initially designed for broader structured data description, has gained considerable traction for dataset metadata through its flexibility. Its integration with complementary standards enabled Google's dataset search engine[2], suggesting that domain-specific vocabularies can achieve broader recognition when embedded in established discovery platforms. Alternative approaches exist across different communities. The RDF vocabulary *Data Catalog Vocabulary (DCAT)*[3], *Dublin Core*[4] and its simpler variant *Simple Dublin Core*[5] offer a domain-agnostic alternative to schema.org [45, 46], while discipline-specific standards like *Darwin Core*[6] and *Data Documentation Initiative (DDI)*[7] address particular research contexts [46].

2.1.4. Context

Understanding and reusing data adequately requires an awareness of its context, broadly defined as the collection of information surrounding how and where data originate [47, 48]. Scholars (e.g. [49, 50, 51, 47]) have conceptualized the notion of context in different ways. Using the term situation interchangeably with context, Dervin [49] identifies three perspectives. She first treats context as a discrete set of environmental or situational settings that frame an event from the outside. Her second formulation shifts the meaning to context comprising identifiable elements that influence interactions or events. In her last view, she portrays context as a surrounding backdrop and a necessary condition that enables comprehension of human behavior [49].

Not all scholars accept this framing. Dourish [50] critiques the notion of context as setting, proposing instead that it emerges through relationships among actors, objects and

[1] `https://schema.org/dataset`
[2] `https://datasetsearch.research.google.com/`
[3] `https://www.w3.org/TR/vocab-dcat-3/`
[4] `https://www.dublincore.org/`
[5] `https://www.dublincore.org/resources/glossary/simple_dublin_core/`
[6] Metadata Standard for Biodiversity `https://dwc.tdwg.org/`
[7] Metadata Standard for data in the social, behavioral, economic, and health sciences `https://ddialliance.org/`

the activities they engage in. Building on this argument, Agarwal [51] suggests context as a combination of various elements such as environment, task, the social relationship between actor and source, and temporal factors. These elements seem to operate on different levels. They differ in their magnitude, how they change over time, and how they interact and combine. Actors may perceive and weigh these elements differently than outside observers, who tend to categorize them through in-group and out-group lenses (us vs. them). While in-group observers identify with the actor (e.g. by sharing profession, culture or experiences), perceiving the context more similarly to the actor, out-group observers do not share those attributes and their view may be influenced by stereotypes, assumptions or external biases [51].

In practical terms, describing the context of the data can take different forms. On one hand, contextual information may remain implicit, e.g. embedded in the name of the managing institution or hosting repository itself, along with its reputation in the field. Although these markers may signal something about the data's earlier applications, this connection may not always be obvious [13]. On the other hand, information can be made explicit through structured metadata governed by particular schemata [46]. Yet metadata does not exist solely in specialized formats. It frequently appears in other shapes like codebooks, publications or research annotations [13]. In this respect, the boundary between implicit and explicit information is fluid. This conceptualization aligns with findings from Faniel and colleagues [47] listing three broad contextual categories describing datasets: Data production information, repository information and data reuse information. The first category should provide information about data collection methods, analysis, resulting artifacts, data producers themselves, missing values and objective of the research. In the second category, the implicit contextual information as introduced earlier, includes the reputation and history. The explicit markers are the declared provenance and curation practices of the repository. The last category demands elaboration on prior use, provides guidance for potential reusers and the terms under which the dataset can be used [47].

2.1.5. Data Provenance

Although often challenging to record, the context in which data is produced is the crucial element that enables its reuse [47]. Therefore, a systematic approach to documenting this information is necessary. This concept is referred to as *data provenance* [17]. The W3C Provenance Group defines provenance, in the context of information, as the collection of processes, organizations, individuals, documents, and data that participate in, shape, or deliver a specific piece of information [52]. Goodman and colleagues [17] argue that there is a strong relationship between the quality of provenance information and the extent to which data are reused. Their position is that when data, metadata, and details about the underlying processes are made available, the likelihood of successful data reuse increases. In scholarly publications, researchers typically describe their procedures for data collection, transformation, and analysis in a section labeled methods, analysis, or workflow, with the terminology varying across disciplines [17].

2.2. Data Sharing

Borgman [15] characterizes data sharing as the practice of making research data accessible for others to use. Sharing of data can be carried out via a variety of formal and informal channels. Examples include exchanging data upon private request (e-mail), depositing it in public repositories (both offline and online), making it available in a personal storage, on a website, or providing it as supplementary files to a journal article [28, 23, 53].

In 1997, the US-based National Research Council [54] defended data sharing with the following rationale:

> *"The value of data lies in their use. Full and open access to scientific data should be adopted as the international norm for the exchange of scientific data derived from publicly funded research"*

Above all, sharing data is a defining feature of Open Science, an effort to make scientific research, data, methods, and other outputs openly accessible, transparent, reproducible, and reusable for the benefit of scientists and society [53, 55]. Borgman [15] further expands on the mentioned rationales, identifying four dimensions in favor of data sharing: (1) reproducibility and verification of research, (2) making publicly funded research results available to the public, (3) allowing others to ask new questions, and (4) advancing the state of research. As she concludes, the second and third rationales are largely driven by public interest while the first and partially the last are primarily research-driven. At the same time, she stresses that reproducibility is the most problematic rationale. Although viewed as the *"gold standard"* of research, reproducibility may be exceptionally challenging in certain domains. In line with Sielemann and colleagues [4], she lists fields such as genomics[8], transcriptomics[9] or proteomics[10], where slight deviations from used procedures, data processing or instrument calibration may render vastly different results [15].

2.2.1. Barriers to Data Sharing

There is a spectrum of reasons that explain why researchers may avoid sharing their data. In research, the effort and time devoted to metadata curation are resources that are no longer available for performing the research itself. Although academic research is inherently collaborative, scholars are in constant competition for grants, positions, public platforms, and students [15]. In this respect, despite significantly reducing potential for reuse [56], these competing interests outweigh the benefits and costly data curation can be difficult to justify [23]. Among the various reasons that lead researchers to withhold their data is the frequently cited challenge of imagining who the potential users or beneficiaries of their data might be [57]. Furthermore, data producers may either lack the awareness or necessary expertise [58, 23]. Also, the data may lack the technical infrastructure necessary to be transferable [57]. Lastly, Borgman [15] mentions potential ethical and

[8] Study of an organism's entire genome, including structure, function, and interactions
[9] Study of the complete set of RNA transcripts under specific conditions
[10] Study of the proteome, all proteins in cells of organisms

epistemological constraints such as uncertainty, competing interests, intellectual property and licensing.

2.3. Data Needs

Expanding on Taylor's information needs [59], data needs are the implicit or explicit requirements for data that are required or desired to serve a particular purpose [13, 47]. Gregory and Koesten [13] identify three characteristics associated with data needs, describing them as (1) *specific*, (2) *multiple*, and (3) *diverse*. The first describes how individual researchers and data professionals usually have very specific data requirements that enable them to carry out narrowly defined tasks within their particular domain. Second, the complexity of these tasks necessitates the use of multiple data types, since each type fulfills a distinct role. Finally, relying on several data types implies a diversity of data needs [13, 60, 61].

According to Krämer and colleagues [62], data requirements are dynamic and evolve together with search behaviors. Additionally, data requirements are shaped by a tendency to evolve with broader changes in how communities handle and use data [62]. This view is reinforced by Xiao et al. [63] and Friedrich [64], who argue that individuals' data needs are strongly connected to and shaped by the contextual information surrounding the data. Individual data needs are driven by gradually evolving circumstances, including one's role, professional standards, intended uses, and what data are actually available [63, 64]. Consequently, data needs are best identified and conceptualized within their specific context [47, 63, 64].

Although the data needs of different scientific disciplines are often treated as distinct, a comprehensive study by Gregory et al. [65] shows that most researchers not only identify with multiple scientific domains, but also stress the importance of accessing data from outside their own field. Contrary to the assumptions embedded in many open data policies and guidelines, data-seeking individuals should not be assumed to be exclusively affiliates of the underlying domains. This disciplinary overlap may lead to complex data requirements shaped not only by domain expertise but also by how the data are used [65].

2.4. Data Reuse

The concept of data reuse refers to the utilization of secondary data, meaning the repeated use of data originally generated by another party [66, 67]. As suggested by Borgman's [15] previously mentioned arguments for data sharing, there are numerous reasons to promote and support the reuse of data. Historical data can reveal valuable insights and offer answers and solutions to both past and current problems [14] (as evidenced in [3, 6, 4]). It may open the door to making more effective business decisions [68], optimize business operations and improve consumer behavior [22]. Finally, it can also foster civic participation by ensuring transparency to the public [69, 53]. Walshe and colleagues have suggested [18] that reusing cross-sectional data could even offer a means to tackle a range of global challenges.

Although the reuse of data is increasingly promoted [22, 53, 56, 4], a variety of obstacles still hinder its broad implementation. Reasons largely reflect barriers to data sharing, as outlined earlier in Subsection 2.2.1. In addition, from the data reuser perspective, there may often be warranted concern about the quality and reliability of data [4]. There is broad agreement that, without contextual information, data cannot be effectively reused, because the absence of such context obstructs the interpretation of unfamiliar data [31, 70, 71, 56, 47, 20]. Furthermore, Koesten and Simperl [7] point out that even adherence to established guidelines and best practices in publishing data does not automatically ensure that others can use it easily. Additionally, the evidence is limited that compliance with such guidelines, which includes technical standards and community support, will lead to increased data engagement and reuse [7].

2.4.1. Data Quality

An important attribute of data predicting its reuse is data quality [47]. Wand and Wang's 1996 paper established a foundational framework for data quality organized into four categories: intrinsic (accuracy, objectivity), contextual (relevancy, completeness), representational (clarity, consistency) and accessibility [72]. Batini et al. [73] expanded this with data quality methodologies, listing over 40 dimensions grouped similarly, emphasizing reuse fitness. A 2023 review by Wang and colleagues synthesizes these into core aspects like accuracy, completeness, timeliness, and relevance and dependency on a task [74]. This is in line with Koesten and Simperl [7], describing data quality as a multidimensional concept. According to them, perceived data quality depends on the task which data helps to achieve [7]. Further, the relative importance of the discussed quality dimensions varies across domains. For instance, genomics research may prioritize accuracy and standardized metadata, while social science datasets may emphasize transparency about sampling methodology and potential biases [73]. This reinforces the idea that data quality is fundamentally task and context-dependent.

Standing for four high-level domain-independent dimensions of data, i.e. *Findable*, *Accessible*, *Interoperable*, and *Reusable*, the FAIR principles translate these quality dimensions into practical criteria [9]. The first dimension focuses on assigning persistent identifiers and comprehensive metadata to make the resources indexable and easily searchable. Second, the data must be available through standardized communication protocols. Metadata should remain accessible even when the underlying data are no longer accessible. To comply with the interoperable dimension, data should use shared, formal languages and vocabularies so that different systems can reliably combine and interpret them. Lastly, data should be made reusable by means of clear provenance, licensing and community standards [9, 75, 46].

Although the FAIR principles became dominant, critics argue they fail to guarantee data quality, provenance, ethical considerations and domain-specific usability, often prioritizing machine-readability over intrinsic accuracy and completeness [75, 76]. Furthermore, implementation inconsistencies and accessibility barriers, such as geo-blocking, may undermine reusability despite FAIR compliance [77]. Complementary approaches, such as the Dataset Nutrition Label, assess intrinsic dataset qualities like composition, provenance,

limitations, and suggested uses [10]. Inspired by food nutrition labels, the Data Nutrition Project[11] aims to create standardized data quality indicators and list characteristics of the dataset that help researchers make an informed decision when selecting datasets.

2.5. Data Discovery

Building on Wilson's [78] information science proposals, Gregory and Koesten [13] extend these ideas to data, characterizing *data discovery* as an umbrella term for processes that address a data need. This concept goes beyond purposeful goal-driven data search and data seeking, and extends to serendipitous search. A subset of data discovery, *data search*, is described as the process of using search systems to locate, rank and obtain datasets [35]. The broader concept of *data seeking* can be described as approaches to locating data aimed at fulfilling a particular goal, which may include using resources other than search systems [78].

Data-seeking strategies depend on particular data types and users' needs, drawing on information-seeking models from information behavior and retrieval domains [13, 79]. Gregory's integrated framework for data discovery [80] emphasizes context and community-specific data-handling practices. Building on this, Koesten and colleagues [81] propose a framework with four semi-sequential components: (1) defining whether the task is goal-oriented or process-oriented; (2) locating data through search engines, portals, repositories, or networks; (3) evaluating data by *relevance*, *usability*, and *quality*; and (4) examining datasets (e.g., errors, documentation, summary statistics).

Research identifies multiple data discovery pathways (see [53, 45, 4, 65, 82]). Gregory and colleagues [65] found that researchers locate data primarily through academic literature, followed by search engines and domain-specific repositories. These pathways include search engines with keyword queries, data portals and repositories, academic literature and citations, and social connections that are particularly important for early-career researchers seeking specialized data otherwise inaccessible through standard search [13, 65]. Although citations from literature to data are a common strategy [65], data citations remain relatively uncommon [17] and differ from bibliographic citations [58]. Moreover, they often lack detail about how data were employed [83]. This may hinder researchers seeking data for specific purposes [65]. Although initiatives such as DataCite[12] and Make Data Count[13] are working to formalize data citation, primarily by assigning digital object identifiers (DOIs) or other persistent identifiers to datasets, there is currently no consensus on a standardized format for citing data [13]. Success across pathways depends on metadata quality and accessibility [13], though metadata are often limited or absent altogether [21].

[11] `https://datanutrition.org/`
[12] `https://datacite.org/`
[13] `https://makedatacount.org/`

2.6. Sensemaking

Zhang and Soergel [84] frame sensemaking as a cognitively demanding process in which a sequence of information-processing units receives data as input and generates changes in concepts as output. Similarly, in their study examining the behaviours that occur while individuals attempt to understand data, Koesten and colleagues [61] describe *data-centric sensemaking* as the process of interpreting and making sense of discovered data in relation to its context. Along similar lines, they point out that a closely related concept is *Exploratory Data Analysis* (EDA), which Tukey [85] characterized as an iterative process that involves interacting with data to gain an understanding of the information it contains.

Within the field of human-computer interaction, a large body of work focuses on how individuals make sense of quantitative data [86, 87]. This process can be supported through, for example, visual exploration environments (e.g. [88, 87]) via the use of exploratory data strategies, in which a predefined series of procedures (e.g. calculation of basic statistical indicators) is systematically applied to examine new datasets until a coherent narrative is formed [89, 90]. Summarizing the sensemaking process as requiring information *from* and *about* data, Gregory and Koesten [13] stress that sensemaking is not a linear sequence but instead involves iterative cycles of evaluation, selection, and detailed exploration. They argue that, to interpret data, people initially assess them according to the criteria (as discussed in Subsection 2.4.1) of relevance, usability and data quality [13]. These findings are consistent with the criteria that researchers apply when selecting data, as outlined by Borgman [15]: *usefulness*, *trustworthiness*, and *value*. For data to be perceived as reliable and trustworthy, explicit information about how the data were collected and processed must be communicated. In addition, the credibility of the source (both individuals and the institution behind them) as well as a low rate of errors appear to also be among the key factors [65, 4, 74].

2.6.1. Sensemaking Activities

Contrary to Zhang and Soergel [84], the results reported by Koesten and Gregory [61] indicate that engaging with conflicts present in datasets can actually support data exploration and even function as an *"entry point to sensemaking"*. They argue that conflicts and errors are an inherent part of any dataset, and that participants managed to address these issues by applying a range of analytical methods [61]. Furthermore, as in collaborative data discovery, engaging with peers appears to be an effective strategy for understanding and making sense of the data [13].

In their overview of sense-making activities, Koesten and Gregory [61] distinguish three main categories of actions performed by data practitioners: *inspecting* the data, *engaging* with its content, and *placing* it within various contexts. During the inspection phase, people aim to form a broad overview of the data's structure, size, contents, and thematic focus. This may involve quickly scanning and scrolling through the dataset, identifying the main variables, and noting any missing values. Interacting with the content involves the development of a deeper understanding of the data. At this stage,

users closely examine the data, identifying suspicious or erroneous values that contradict expert expectations and analyzing how variables are constructed in context. To assess the accuracy and reliability of unfamiliar datasets, they look for outliers, formatting anomalies, and deviations from standard reporting practices. By placing data within familiar contexts, users refine their understanding by formulating questions about the entities data represents. Next, they test how the data behave under conditions different from the original environment, considering study design, disciplinary conventions, and temporal as well as geographic factors to judge representativeness, level of detail, and the data's original purpose [61].

3. Survey of Dataset Repositories and Portals

In this chapter, I report on a survey of dataset repositories and portals (DRP), with a focus on their features. The goal of this categorization effort was to develop a preliminary understanding of existing platforms rather than an exhaustive assessment of the current landscape. I approached this by comparing repository features with best practices described in the literature. As far as I could determine at the time of my investigation, the literature appeared to lack a comparable analysis that categorizes a broad set of public data repositories and portals using an explicit feature checklist. Given the chosen sample, the survey results should offer a useful signal into user's expectations and best practices, guiding design choices in the later stages of this work.

3.1. Methods

To conduct the survey, I drew on best practices and recommendations summarized by Gregory and Koesten [13]. After consulting with my supervisors, I surveyed a sample of 14 DRPs that appear widely used in their communities and represent several domains. Although definitions may offer subtle differences, for the purposes of this study I treat Dataset Search Engines as Dataset Portals. In total, I grouped 47 features into eight thematic clusters that capture recurring functional themes. Surveyed DRPs are summarized in table 3.1.

3.1.1. Data Collection

Data collection took place between November 2023 and March 2024, which I treat as the primary study window. Moreover, I revisited and updated the collected data in October 2025 to account for the time gap, which may have introduced changes. This time gap was caused by later phases of the underlying project, primarily from lengthy development, deployment, and testing of the ViDa platform, as described in the Implementation chapter 5. The data collection approach corresponded to purposive sampling of information-rich cases, with sample size determined by feasibility and expected variation.

The data were gathered by manual review of the Search Engine Results Page (SERP) of each DRP and the detail page for each result. For this purpose, in each instance between five and ten datasets were considered. Furthermore, I selected datasets that appeared aligned with the platform's topical focus to capture features that might only emerge in certain cases. For a health-oriented DRP, for instance, I queried terms such

Table 3.1.: Data portals, repositories, and search engines

Name	Type	Domain	URL
London Data Store	Repository	Government	`data.london.gov.uk/`
GESIS	Repository	Social Science	`search.gesis.org/`
Dataset Search	Search Engine	General	`datasetsearch.research.google.com/`
ICPSR	Repository	Social Science	`www.icpsr.umich.edu/sites/icpsr/find-data`
Kaggle	Data Portal	General	`www.kaggle.com/datasets`
GenBank	Repository	Genetics	`www.ncbi.nlm.nih.gov/genbank/`
Inspire	Data Portal	High-Energy Physics	`inspirehep.net/`
Open Science Framework	Repository	General	`osf.io/`
European Data Portal	Data Portal	General	`data.europa.eu/en`
Figshare	Repository	General	`figshare.com/`
Harvard Dataverse	Repository	General	`dataverse.harvard.edu/`
Zenodo	Repository	General	`zenodo.org/`
Statista	Repository	Economics	`www.statista.com/`
Auctus Dataset Search	Data Portal	General	`auctus.vida-nyu.org/`

as `virus` and then inspected the returned records. Moreover, I also consulted official documentation, FAQs, and release notes. When on-site search was limited or absent, I used Google to search within the documentation via site-specific queries combined with feature-related keywords (for example, `site:domain "metadata schema"` or `"API pagination"`). This triangulation appears to improve coverage, though documentation may lag behind deployed functionality or use different terminology. When a given feature appeared to satisfy the qualifying criteria described in 3.1.2, I recorded both the extent (for partial implementations) and the type of implementation. When relevant, I also included a URL, typically to the API documentation, unusual implementation, or metadata curation guides. Overall, the strategy is likely to surface platform-specific functionality, yet it may introduce selection bias and underrepresent edge cases. The findings should be read as illustrative of common patterns rather than an exhaustive inventory of all available features.

3.1.2. Thematic Clusters

Thematic clusters summarized in table 3.2 approach DRPs from multiple perspectives. The following section outlines the ideal criteria used to evaluate features in each cluster.

The first broad category, *General UX*, examines how DRPs handle the established order of pages when searching for datasets. The expected order typically begins with (i) a search page where users enter keywords, followed by (ii) a results page (a.k.a. Search Engine Results Page, SERP), and ends with (iii) a landing or data preview page. Next, a data curation guide or methodology should be made available to properly manage data.

Table 3.2.: Clusters and their feature counts

ID	Cluster	# Features
C1	General UX	5
C2	Social	6
C3	Faceted Search	5
C4	About Data and Provenance	10
C5	Sensemaking	7
C6	Quality Indicators	6
C7	Task Support	4
C8	SERP: beyond 10 blue-link paradigm	4

An important aspect of this cluster is the user interface's ability to orient users. It should indicate where they are, especially when they arrive at a detail page from outside the established order discussed previously. In terms of machine-to-machine communication, it is a widely used practice that systems provide the ability to interface with external services. Accordingly, evaluated DRP should expose API for this use-case. Final criteria in the category, though unusual, is a voice-activated bot that may enhance user experience in hands-free or assistive scenarios.

Next category *Social* focuses on features that connect DRPs with their user communities. Social network engagement refers to built-in widgets that let visitors share datasets or follow the hosting organization's accounts. Dataset community rating provides a quantitative score on a predefined scale, while dataset community reviews capture qualitative feedback in text, such as comments attached to a dataset page. Community-driven documentation denotes a wiki-style space where members can edit procedures, rules, and recommendations so shared practices stay visible and current. Chat considers real time messaging either among users or with bots. Finally, collaborative tool integration is defined as official support for external platforms like Slack, Mattermost, Discord, Microsoft Teams or other external collaborative messaging software.

The category *Faceted Search* captures how DRPs let users narrow results along meaningful dimensions rather than rely on a single keyword query. Topic categorization should assist in a quick categorization of datasets by domain, so that large query results become more manageable. Geography provides another strong option where users can focus on geolocation is relevant to them, whether by region, administrative unit, or coordinates. Two lenses for temporal selection are considered. For temporal filtering, a narrowing tool for selection of a time window is considered. In contrast, temporal zooming should let users pan and scale the timeline to inspect stretches of time without losing context. On the contrary, data exclusion should let users filter results. In practice, faceting turns search into a stepwise conversation with the platform, with each choice trimming ambiguity and bringing the right datasets into view.

The category, *About Data and Provenance*, consolidates core descriptors that explain what a dataset is, where it came from, and how it has changed over time. Format should

inform the user about the file type such as CSV, JSON, or Excel. A metadata schema specifies the standard or framework used to describe the dataset, including fields for origin, collection methods, quality notes, and intended usage, which helps interpretation and interoperability. File size of the dataset should be reported (e.g. MB), to indicate the sample size. Further, license information is assumed to link or to explicitly describe the terms under which users may use the data. A persistent identifier such as DOI should be offered, as a stable, citable reference to the dataset. Further, an information about the last update of records or dataset revision coupled with an update frequency, should be indicated. Original purpose must offer a context, explaining why the data were created or collected in the first place. Along similar lines, information on how the data were collected or derived including techniques, instruments, sampling are expected. Data processing steps should describe cleaning, transformation, aggregation and other derivation or data-manipulating procedures. Finally, versioning information should track releases with identifiers, dates, and potentially change logs, as a concise history of modifications.

The category, *Sensemaking*, brings together features that help users to understand and reason about the dataset. Key variables should highlight the most important fields, pointing to the columns or abstracted dimensions that define the dataset. Textual summaries of the dataset should outline information either about the dataset or individual data columns as short descriptions. Statistical summaries should then provide a quantitative perspective in form of counts, means, medians, or measures of spread. Visual summaries are also expected to provide visual insights via charts, graph or maps at a glance. For finer detail, column-level summaries should then report information for each variable separately. Some DRPs implement such inspection using a spreadsheet view. However, since datasets may contain large number features translated to columns, horizontal scrolling should ensure navigation across all of them.

The category *Quality Indicators* evaluates features that make the state of a dataset visible and useful during discovery and evaluation. Completeness should describes how well the dataset covers its intended scope and whether it suffers from missing values, empty fields (null or undefined). At the same time, empty headers should be flagged as well as columns without meaningful names or descriptions. A description of data quality issues and limitations is expected to record problems and caveats known to the authors. Feature that tracks where the data was used before should lists prior applications or citations. To finish the category, in order to qualify for toggleable quality indicators feature, DRP is expected to display optionally displayed evaluations (warnings or quality control checks) assessing reliability of the dataset.

Moving on to *Task Support*, the features in this category should help users carry out analytical tasks directly within a DRP. For side-by-side comparison, expectation of the feature assumes two or more datasets that can be viewed next to each other. The goal is to support visual comparison of similarities, differences, or trends. Moreover, combining of data should provide mechanisms to merge or append datasets from different sources so that a single, integrated table can be formed. For different datasets highlighting of common attributes should automatically identify shared fields across datasets. Lastly, a

task-based filter is expected to offer narrowing of results based on user's visualization task or use-case.

The last category, *SERP: beyond 10 blue-link paradigm*, addresses what the results page should communicate before any dataset is opened. First it assumes a numerical indicator of matching results for a query. Further, it covers similarities between data sources, highlighting overlaps such as shared variables or common attributes. Providing an early insight, the display of key variables should display main fields for each dataset. Finally, comparisons during search refer to tools that lets users juxtapose datasets or run comparison checks during search. Combination of those features enable SERP of a given repository to become a preliminary analysis tool rather than a simple list of links.

3.1.3. Evaluation Metrics

Each feature was evaluated with yes/Did Not Find (DNF) binary presence (partial yes counted as yes). For each cluster a relative feature coverage (in percent) was calculated. Categories are treated with the same weight, and therefore summed coverages divided by the number of clusters yields the Feature Coverage Index (FCI):

$$\text{FCI} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{p_i}{t_i} \times 100\% \right),$$

where:

- $N = 8$ is the number of clusters,
- p_i is positive features (yes/partial yes) in cluster i,
- t_i is total features in cluster i with $t_i > 0 \ \forall i$.

$\text{FCI} \in [0; 100]$ averages per-cluster coverage ratios to measure aggregated thematic feature support for a given DRP.

3.2. Results

Before examining individual DRPs let us first focus on the aggregate results. Observed Feature Coverage Index across analyzed repositories was low (M=45.087, STD=9.919), suggesting considerable space for improvement. Inspecting the individual categories, we can observe that category C4: About Data and Provenance had the highest average coverage (M=90.71%, STD=13.28%), followed by C1: UX Generally (M=70%, STD=13.01%) and C5: Sensemaking (M=62.24%, STD=24.17%). Conversely, categories C2: Social (M=22.62%, STD=16.80%) followed by C6: Quality Indicators (M=29.76%, STD=23.73%) and C3: Faceted Search (M=30%, STD=21.84%) have the greatest potential for improvement.

As summarized in the table 3.3, the highest FCI was achieved by the data portal Kaggle with 66.458 followed by Open Science Framework (56.309) and Harvard Dataverse

(55.625). At the opposite end, the lowest index had the Google Dataset Search and Gesis both achieving FCI=33.185.

Table 3.3.: FCI with per-cluster coverages (in %)

Name	FCI	C1	C2	C3	C4	C5	C6	C7	C8
London Data Store	44.613	60.00	16.67	60.00	100.00	28.57	16.67	25.00	75.00
Gesis	33.185	60.00	16.67	20.00	70.00	57.14	16.67	0.00	25.00
Dataset Search	33.185	40.00	16.67	20.00	90.00	57.14	16.67	0.00	25.00
ICPSR	43.185	60.00	16.67	20.00	100.00	57.14	16.67	50.00	75.00
Kaggle	66.458	80.00	50.00	60.00	100.00	100.00	66.67	50.00	75.00
GenBank	40.774	80.00	16.67	20.00	100.00	42.86	16.67	25.00	50.00
Inspire	37.738	80.00	16.67	20.00	90.00	28.57	16.67	50.00	50.00
Open Science Framework	56.310	80.00	66.67	80.00	100.00	57.14	16.67	50.00	50.00
European Data Portal	53.304	80.00	16.67	20.00	80.00	71.43	83.33	25.00	75.00
Figshare	39.435	80.00	16.67	20.00	100.00	57.14	16.67	50.00	25.00
Harvard Dataverse	55.625	80.00	33.33	40.00	100.00	100.00	16.67	50.00	75.00
Zenodo	37.649	80.00	16.67	20.00	100.00	42.86	16.67	25.00	25.00
Statista	38.929	60.00	16.67	20.00	60.00	71.43	33.33	50.00	50.00
Auctus Dataset Search	50.833	60.00	0.00	0.00	80.00	100.00	66.67	75.00	100.00
MEAN	45.087	70.00	22.62	30.00	90.71	62.24	29.76	37.50	55.36
STD	9.920	13.01	16.80	21.84	13.28	24.17	23.73	21.37	24.37

3.2.1. Universal Characteristics Across Platforms

Several characteristics appeared consistently across all analyzed platforms. All platforms followed an established order of pages (search, SERP, detail). Notably, Google Dataset Search and Auctus Dataset Search slightly diverge from this pattern by integrating the detail page directly into the SERP. SERPs consistently display the number of available datasets for a given query and the number of results under each filter option. All platforms offer filtration through dynamically generated lists and controlled vocabularies, as well as faceted browsing. Filtering by dataset variable types and common attributes, however, appears less frequently among the surveyed DRPs. Every analyzed repository included some form of temporal filtering. Temporal zooming, however, appeared less frequently. Only six repositories offered this functionality, either through date ranges or fine-grained temporal intervals. Many repositories allow users to exclude terms by using special query operators such as `-` or `NOT`. Eight repositories featured at least partial geographic filtering. London Data Store and Auctus Dataset Search distinguish themselves through map-based interactions, enabling users to define search areas via bounding boxes. With the exception of Google Dataset Search, all platforms offer orienting cues to help users understand their location when arriving from external sources. Platforms typically achieve this through breadcrumb navigation and informative page headers. Notably, none of the analyzed repositories offered side-by-side dataset comparison functionality, suggesting a gap in current platform design.

3.2.2. About Data and Provenance

In the category about data and provenance, all repositories included information about given dataset formats. These were indicated either as file extensions, explicit program formats (Excel, RData, SPSS), or domain-specific formats such as FASTA or ASN.1. Most platforms also displayed file sizes in MB alongside format information, though this detail was occasionally buried within metadata or hidden from immediate view. Further, dataset licensing information was explicitly stated on the detail page or implicitly in accompanying metadata. An unusual implementation of licensing information had the commercial repository Statista, restricting license information to paying users.

Update frequency was documented across all DRPs, though the level of detail varied. Some platforms capture dataset creation, publication dates, and revision histories, often accompanied by commentary. Others, however, provide only a timestamp of the most recent update. Platforms communicated the original intent and contextual information in various ways, most commonly through abstracts or accompanying metadata. Abstracts served as the primary summarizing mechanism for datasets. ICPSR and Kaggle went further by providing lists of related publications. Dataset collection and derivation steps were often detailed within these abstracts. Kaggle and Statista, however, dedicate a specific section (the Provenance panel) to this information on the dataset detail page.

In ICPSR, users can filter results by collection method (survey, census etc.) or mode (face-to-face, CATI, CAPI, questionaire). Apart from mentioning processing techniques in abstract, platforms commonly address data processing steps by attaching supporting documents such as codebooks, interview guides, or processing code scripts. Notably, Kaggle allows users to embed Jupyter Notebooks to document processing steps.

3.2.3. Sensemaking

Ten of the surveyed repositories supported an implementation of a spreadsheet preview with horizontal scrolling. Platforms commonly referred to these as Data Explorers. Predisposition for the explorer is a compatible textual, spreadsheet formats such as CSV or Excel-like formats. Statistical and visual summaries, however, seem to lag behind. While ICPSR implements a basic form within variable search, implementations of Kaggle, Harvard Dataverse, Statista and Auctus Dataset Search are more intricate. Kaggle, for instance, calculates relative statistical indicators for nominal data. The second generation of Harvard's Dataset Explorer goes further and conveniently offers dynamically calculated column-level statistical summaries (mean, standard deviation, min, max and count). For qualitative variables it provides frequency tables and a bar chart. Auctus Dataset Search takes a similar approach and using a profiler it infers data types (Integer, Text, Enum or Null) and generates statistics for columns (unique values, mean, variance) which is then visually supported by histogram. Elegant visualizations are offered by Statista, featuring interactive bar charts, line charts and maps.

Except for London Data Store and Zenodo, all observed DRPs displayed key variables. Many platforms feature those key variables on search page, while some incorporate key variables in abstract. In case of Auctus Data Search and Harvard Dataverse, key variables

are automatically identified. Auctus Data Search goes one step further, and profiles the datasets resulting in variable type recognition (categorical, numerical, spatial).

3.2.4. Quality Indicators

There appears to be a serious gap in assessment of data quality among surveyed repositories. Only Kaggle, European Data Portal and Auctus Dataset Search explicitly inform users about missing values, empty fields and empty headers. In Kaggle, indicators of data quality are optionally shown in Coverage information panel within dataset detail. European Data Portal assesses datasets by evaluating their metadata quality using their Metadata Quality Assessment Methodology (MQA). Statista appears to be the only repository that provides explicit warnings about data quality. Although European Data Portal displays quality flags, users cannot filter or search datasets based on these quality indicators.

3.2.5. Technical Infrastructure and Data Organization

Analyzed repositories adopt different approaches to data organization using metadata schemas. Some mandate a single standard, while others support multiple schemas simultaneously. Many DRPs mandate standardized schemas such as JSON Schema, MARC, DATS 2.2 or Dataset definition from schema.org. Others support additional standardized schemas such as Dublin Core, DCAT, DDI, and DataCite. Some platforms serialize these using RDF formats such as Turtle or N-Triples. Beyond schema requirements, the majority of platforms expect specific conventions described in data curation guide or methodology, designed to allow for reuse.

DRPs implement APIs for different functions, though capabilities differ across platforms. APIs, built typically using REST architecture, offer mostly management of resources, querying results and dataset metadata, and upload or update of datasets. Kaggle and Auctus Dataset Search additionally offer dedicated Python client. Harvard Dataverse and Open Science Framework also offer extensions. Specifically for OSF, connection to Github, Dropbox or Google Drive resources appears to be popular. Harvard Dataverse's integration with the DataCite REST API also enables automatic DOI generation. Where persistent identifiers are used, DOIs represent the standard choice across repositories. The only exception is Gesis which assigns its own stable accession number to datasets. Figshare, Zenodo and Open Science Framework automatically assign DOIs to the published datasets. Harvard Dataverse requires DOIs to be assigned to datasets and extends this by also supporting DOI assignment to accompanying resources such as documentation or code. Severl of those DRPs also link different versions of the same dataset using DOIs to creating a history of changes.

3.2.6. Social Features

With the exception of Auctus Dataset Search all observed platforms offer some implementation of social media engagement, whether through dedicated widgets for sharing

content, or simple links leading to dedicated pages on social networks.

Information about where the dataset was previously used is communicated with different granularity across analyzed DRPs. Although most platforms rely on abstracts and links to related projects, Open Science Framework, European Data Portal and Inspire track DOI and citations of their datasets. Google Dataset Search also links scholarly articles to datasets indexed by Google Scholar. ICPSR manages an extensive bibliography of data-related literature, allowing users to search for publications that have used datasets from their catalog and to discover how datasets have been cited and reused. The Gesis repository offers a distinctive approach through its implementation of InFoLiS[1], a project attempting to find hidden references to datasets in publications.

Returning to the social aspect of analysed repositories, only Kaggle and Open Science Framework supported community rating and reviews. Kaggle calculates usability score for each dataset and encourages user feedback, either via comments or keywords. Similarly, Open Science Framework implements both of those features by enabling users to comment on projects and datasets. It was also the only platform that offers wiki-like crowd-sourced documentation, where collaborators can contribute to documentation. Notably, none of the surveyed repositories integrated collaborative tools such as Slack[2] or Microsoft Teams[3].

3.2.7. Emerging Technologies

Among the repositories examined, Harvard Dataverse appears to be the only one offering DataChat, a multilingual natural language interface for querying datasets. The recent addition of Model Context Protocol (MCP) support for LLM agent integration may, however, signal a shift away from DataChat in the coming months.

[1] https://infolis.github.io/
[2] https://slack.com/
[3] https://teams.live.com/free

4. Low-fidelity Prototypes

In this chapter, I discuss the rationales behind iteratively developed low-fidelity designs intended to approximate the layout and features of the envisioned dataset preview platform. I then discuss insights from four semi-structured interviews conducted with data producers at various career stages in which the prototypes were presented to the participants. Drawing from their perspectives, the objective of the interviews was to gather general insights, to test design assumptions, to identify elements that impede usability, and to collect actionable feedback.

4.1. Designs

Insights from literature coupled with a survey of the data repositories and portals have provided signals about what to include in a dataset preview platform. I incorporated these findings into low-fidelity prototypes, which I created using sketching app Notability[1] on iPad. The digital nature of the prototypes allowed for rapid prototyping through easy duplication and rearrangement of design elements. During the prototyping, it became apparent that the initial scope of the platform may have been overly ambitious. This realization prompted me to narrow the project's scope, focusing instead on a platform showcasing a single dataset i.e. the detail page from the perspective of dataset repositories.

4.1.1. Prototyping Methodology

Guided by the Five Design Sheet (FdS) methodology [91], I began developing low-fidelity prototypes. that incorporated elements from the previous phases. The initial brainstorming phase involved sketching various components of the envisioned user interface in a free-form manner. I organized these components into a tabular catalogue, documenting and justifying the intended function of each element. I drew upon the design language of the widely adopted Google's Material Design System[2] [92]. A key advantage of this design system lies in its consistent cross-platform appearance [93]. Users come across the same visual language, behavior and interaction patterns in web, mobile, and desktop applications. Supported by exhaustive user testing [93], it is among the most widely used design systems, used by default in Android[3] smartphones and Chrome OS[4]. Consequently, users are likely already familiar with its conventions making it an ideal choice for the envisioned platform.

[1] `https://notability.com/`
[2] `https://material.io/`
[3] `https://www.android.com/`
[4] `https://chromeos.google/`

With an established catalogue, I systematically began to position elements on the page to resemble a dataset detail page. This resulted in the first version of the low-fidelity prototype, which underwent six subsequent iterations. Each iteration was informed by feedback from the project supervisory team, gathered over three months starting in April 2023. After conducting semi-structured interviews with potential users (described in Section 4.2), I integrated received feedback into a final design that would later guide the implementation phase. Figure 4.1 summarizes this design process.

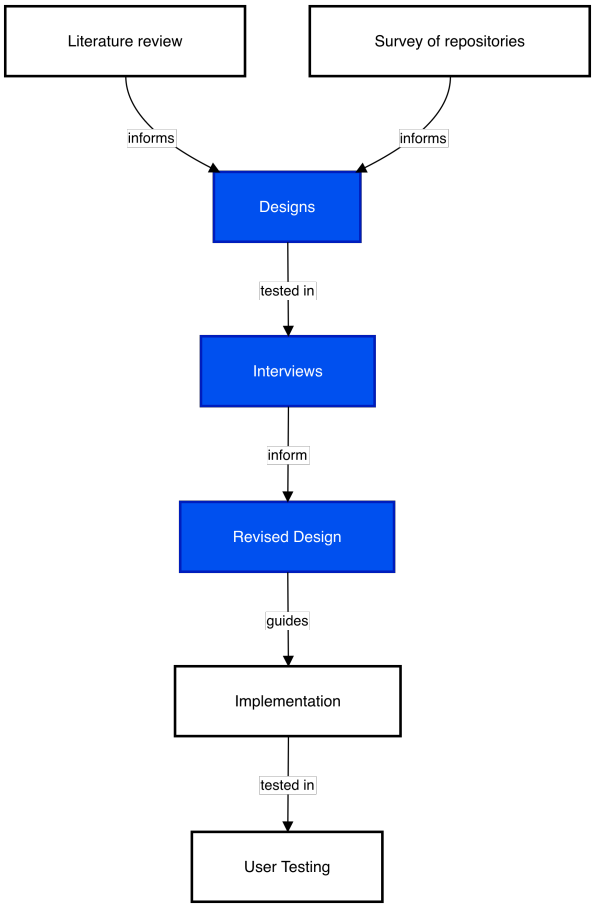

Figure 4.1.: Project methodology with chapter-relevant steps highlighted in blue

4.1.2. Initial Prototype

In the initial version labeled *A V1* shown in Figure 4.2, I laid out the components in a two-column grid layout. Starting from the top, a bread crumb navigation is the first element, intended as an orientation mechanism. Including this component reflects the early conceptual stage when the scope of the platform remained abstract. The element

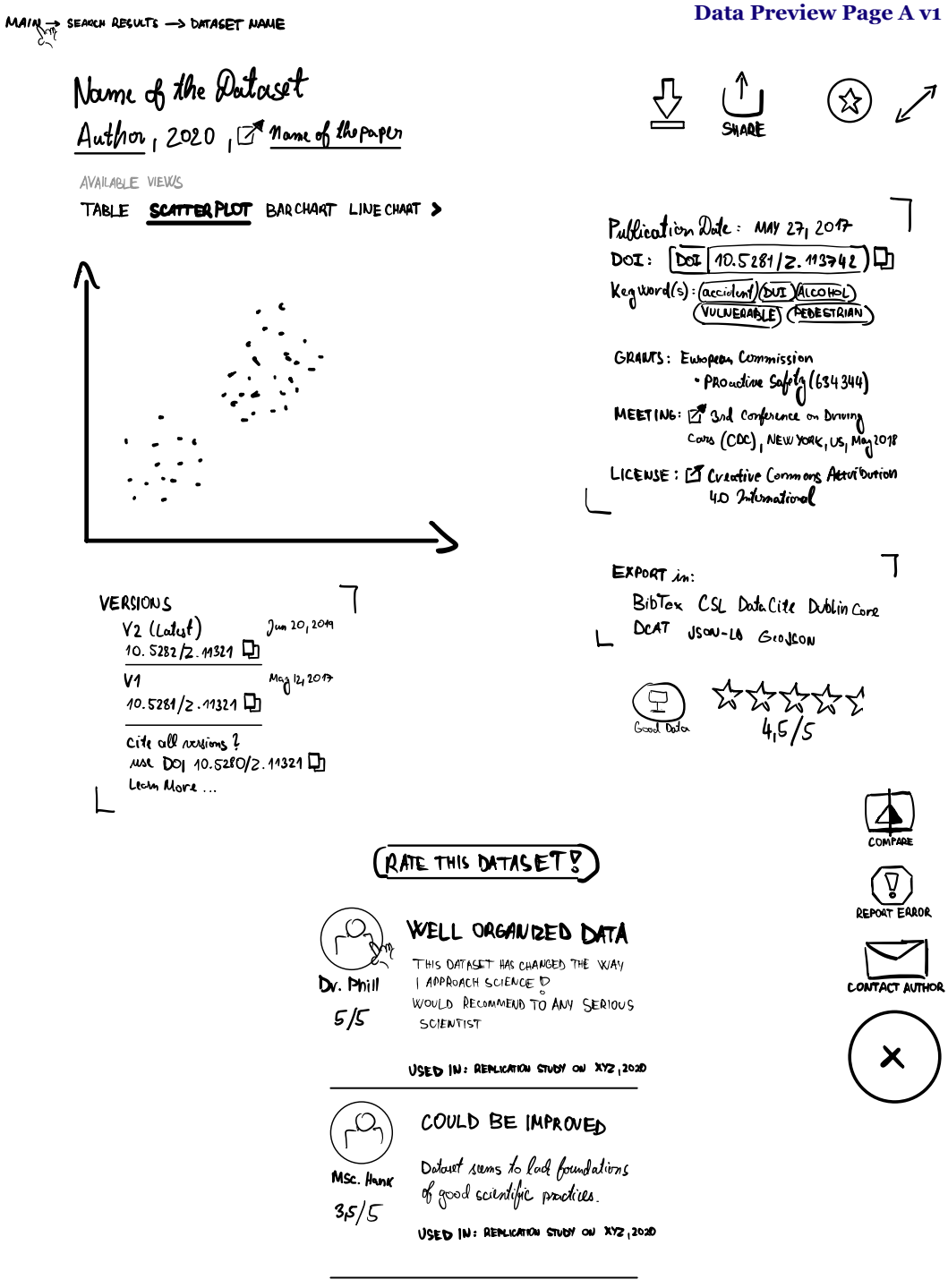

Figure 4.2.: Initial low-fidelity prototype

would be eventually removed as the vision crystallized. On the left, the prototype displays basic introductory information such as the name of the presented dataset, its authors, year of publication and academic paper in which it was first appeared. To the right of the introduction, a horizontal navigation bar offers several actions, including download, share, rate, and expand. The share button in particular was intended to facilitate social engagement. Among other things it was envisioned to enable leaving a rating for the dataset by the visitor similarly to the call to action button described later.

Below the navigation sits a selection of visualization views implemented as tabs, with the scatter plot currently active. Next to it is a panel with highlighted key meta information such as the publication date, persistent identifier (DOI), keywords, conference or meeting where the dataset was presented and license. Below the visualization, a version history lists each release with its persistent identifier and publication date. Presented approach draws inspiration from the Open Science Framework. On the right side of this section, a tile displays various academic citation formats of the dataset for export. Below it, a dataset quality section displays a quality badge and star rating.

The page concludes with testimonials section featuring call-to-action button (rate this dataset). Here, users can leave their reviews, assign a star ratings and describe how they used the dataset. Last two features attempt to address the limited social engagement observed in analyzed data repositories by adapting strategies from e-commerce platforms, particularly the use of social proof. Finally, an opened floating action button in the lower right corner reveals vertical secondary navigation menu. This component maintains a fixed position at the right corner of the viewport. This navigation is implemented as a Material Speed Dial component (a button that reveals options when pressed). In its open state it offers options to compare different datasets, to report an error and contact the author of the dataset.

4.1.3. Layout Refinement

Iterations 2 and 3 established the platform's information architecture. In the second iteration *A V2*, I expanded the layout beneath the visualization from two to three columns to accommodate additional content. To mitigate potential information overload, I wrapped all sections in collapsible rollout panels akin to the native HTML element `details`[5]. Establishing the concept of information panels, users can reveal or hide content on demand by clicking on the header or the accompanying arrow indicating its closed/open state. I added two new information panels to this iteration, namely *Warnings by author* and *Similar Datasets*. The information panel *Warnings by author* replaced the earlier data quality badge, allowing dataset creators to address potential caveats and shortcomings of the dataset directly. The *Similar Datasets* panel provides links to related datasets identified by the author. The visualization component also received two subtle enhancements. First, I added visual feedback for data point selection through brushing functionality. The second change is a two-way-arrow icon in the upper right corner. This icon button enables users to focus on the visualization by enlarging

[5] `https://html.spec.whatwg.org/multipage/interactive-elements.html#the-details-element`

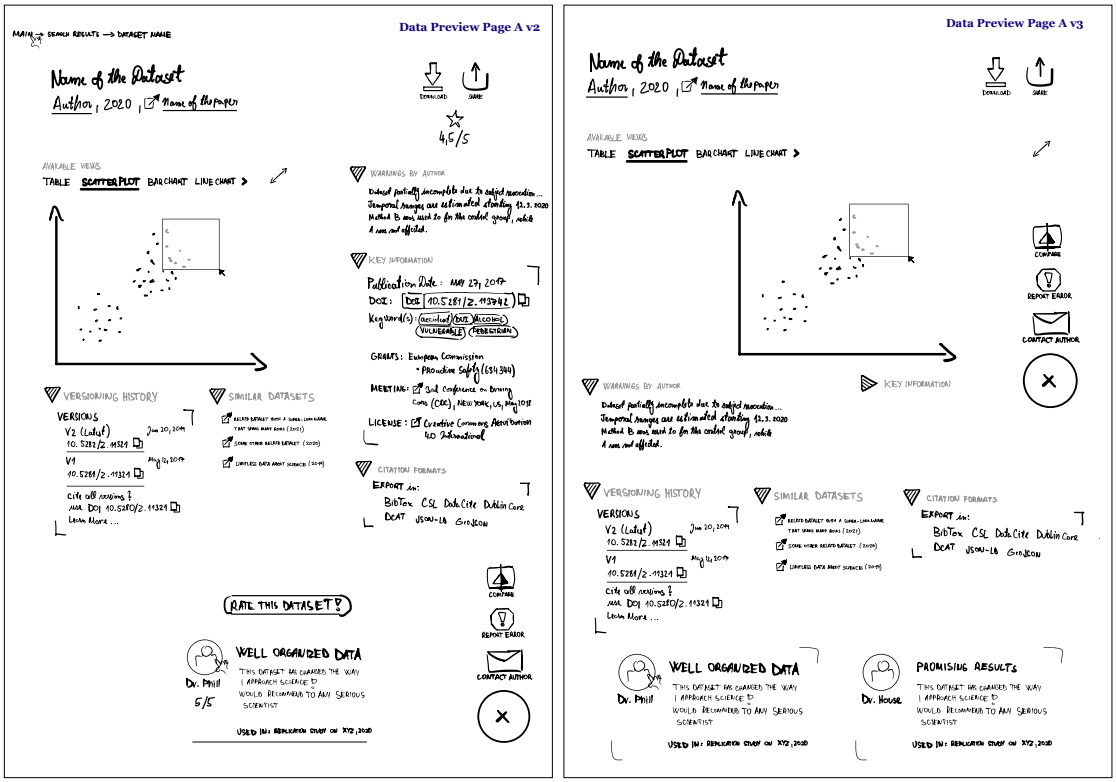

Figure 4.3.: Second and third prototype iterations

the component to fullscreen. The rationale was that a larger display might aid pattern recognition, particularly for dense datasets.

The next iteration *A V3* further emphasized data visualization. I expanded the visualization widget to span the entire width of the page. The reasoning behind this decision was that interactive data visualization may enhance cognitive processing and consequently warrants primary focus. The enlarged visualization forces information panels to be repositioned below, resulting in somewhat irregular layout. The first two information panels (*Warnings by author* and *Key Information*) occupy the first row, while the remaining panels are distributed across three columns. Notably, after careful consideration, I decided to abandon the breadcrumb navigation in this iteration, narrowing the platform's focus to showcasing of a single dataset. I also reconsidered the trustworthiness indicators. The star rating system appeared to undermine professional credibility, evoking consumer platforms rather than scholarly data presentation. Consequently, I removed this component entirely in this iteration.

4.1.4. Enriching Content and Interaction

The fourth and fifth iterations (Figure 4.4) shifted focus toward enriching content and refinement of interactive features. A textual summary (abstract) appeared in all analyzed

repositories, often serving multiple functions beyond its intended purpose, as discussed in the previous chapter. Given this convention, I introduced an abstract section beneath the introductory information in the fourth iteration titled *A V4*. In this phase, I recognized that structured guidelines for creating the abstract would likely be necessary to ensure appropriate use. This iteration also introduced an interactive map view with a discrete color scale, for color coding of the geographic areas. The fifth iteration represented the final design phase before creating a version for participant interviews. To address data quality concerns, I included a simplified version of Data Nutrition Label [10] with a link to the original project. The simplified *At Glance* version of the label, indicates if:

- the dataset contains data about humans and if so whether it was properly anonymized (about humans)
- the dataset has been reviewed for missing or inconsistent data and possibly preprocessed (technical review)
- the dataset has been ethically reviewed, all necessary consents have been obtained and all privacy concerns addressed (ethical review)
- the dataset is updated sufficiently frequently for the intended use-case or there is reliance on outdated data (update frequency)

The visualization section underwent two seemingly minor yet potentially significant updates. To improve immediate chart type recognition, I added icons to all visualization tabs. Rather than naming the underlying tool (type of chart), each icon visually suggests the type of visualization available. Along similar lines, I moved away from conventional type labels (bar chart, scatterplot, line chart etc.). Instead, I named each option according to the analytical task it supports as described by Munzner [38, Ch. 3]. Although the majority of visualization labels directly follow established task typology (e.g. *Discover Groups*, *Compare Trends*) some tasks deviate slightly (namely *Group Distribution* and *Group by Location*). The rationale behind this change from chart names to analytical tasks is to make the visualizations more accessible to lay audiences. Since data visualization should primarily help to obtain an insight, emphasizing what users can accomplish may encourage more purposeful engagement. This framing might reduce the tendency to treat visualizations as merely decorative visual spectacle. Each visualization received a help icon button that clarifies the function and controls of the current tool. Another important update is the addition of axes labels to the scatterplot. Users can control the labels of axes via the Feature Selection control panel placed to the right of the visualization. Axis reassignments will dynamically update positions of all displayed data points. I positioned a summary statistics panel above the feature selection area. When users select data points through the brushing tool, the panel dynamically calculates statistics for that subset. These can be compared with global statistics computed for all visible data points.

In contrast to the previous iteration, I reorganized the information panels below into a regular three-column grid. Following a similar rationale, I added icons to each information panel to aid comprehension. To reflect their perceived importance, I also reordered the

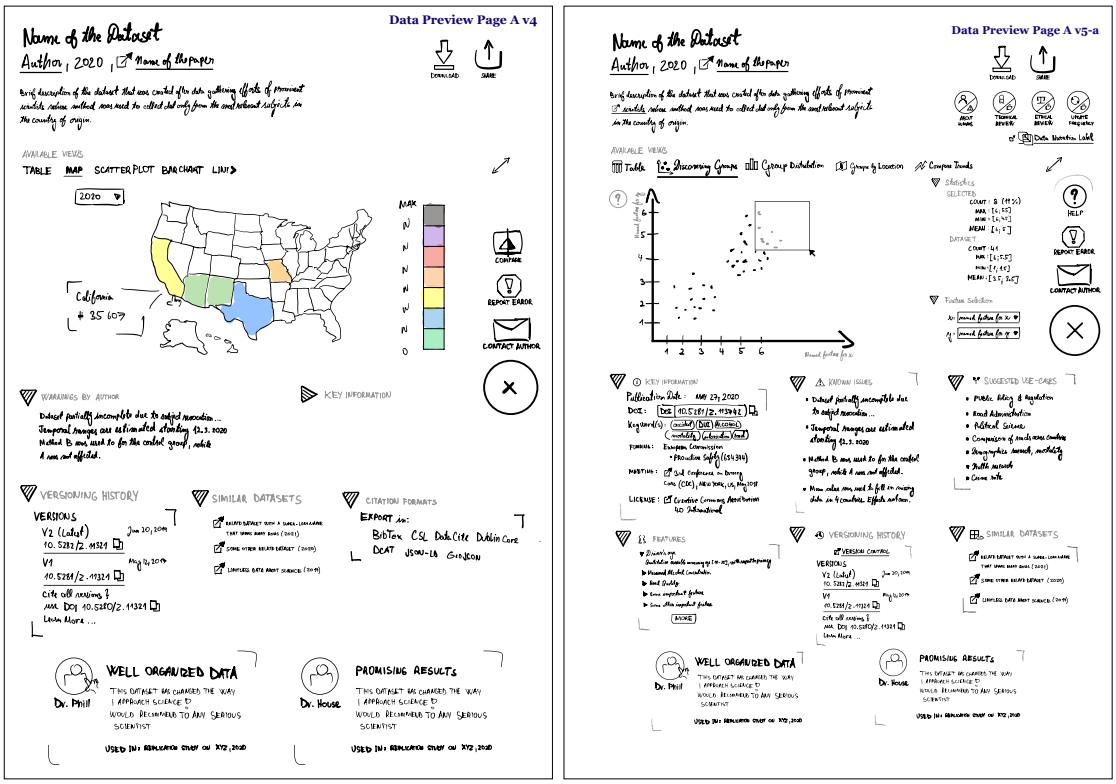

Figure 4.4.: Prototype iterations from version 4 to version 5

panels once more. The first row includes Key information, known issues, and the newly added suggested use-cases.

The last information panel *suggested use-cases* allows dataset authors to propose alternative applications or research contexts for reuse. The second row starts with a new information panel *features*, listing individual dataset features (variables) and their descriptions. The row concludes with *Versioning history* and *Similar dataset* panels. Importantly, in this iteration I reached the decision to remove support for comparison of datasets, found under the speed dial menu in previous iterations. This further narrowed the platform's scope to showcasing a single dataset with all necessary information to encourage reuse. The underlying idea was to enable data producers to automatically create a showcase of their dataset without requiring extensive visualization expertise. At the same time, I aimed to ensure accessibility to as wide audience as possible without sacrificing best data curation practices and standards.

4.1.5. Interview Prototypes

Following the feedback from the last iteration, I produced two design variations (A and B) as shown in Figure 4.5. Each prototype illustrated three states, starting with all roll-out panels closed, then all panels opened and finally mixed configuration. In terms of

changes, version A implemented several information panel refinments. First, I reordered the panels slightly, placing *features* and *Caveats Watchlist* (previously Known Issues) after *Key Information*. Second, I moved testimonials below the information panels also in a rollout panel allowing users to control their visibility.

Version B, by contrast, focused on layout modifications. The goal was to maintain the same content across both versions while exploring alternative spatial arrangements of the components. Starting from the top, I centered the introductory component which lists name, authors, year and publication of the dataset. I then removed the abstract and relocated it into its own dedicated information panel below the visualization. I shifted the data nutrition label component left to the introduction. The menu retained its position but extended its options by those found in the speed dial menu in design A. Scatter plot also differs by positioning of the controls. I centered the visualization and positioned axis label controls directly adjacent to their corresponding axes. Moreover, local statistics for selected features now appear in a dialog near the cursor, while global statistics occupy the left side of the chart. Adding a dedicated information panel *About* as the first tile forced the other elements to shift and reorganize. Although I maintained the three-column grid, the testimonials occupy a single column in this version. This differs from design A, where they span the full footer width.

4.2. Interviews

Interview process is reported using the Consolidated criteria for reporting qualitative research (COREQ) [94]. Where applicable, items from the COREQ checklist are addressed; non-applicable items are omitted. I (HZ, male), the author of this thesis have conducted the interviews as a graduate student working as a DevOps engineer. My technical background and position likely influenced how questions were framed and how responses were interpreted.

4.2.1. Methods

Participants were selected through a mix of purposive and convenience sampling. The two criteria for inclusion were active involvement in dataset production (having produced at least one dataset) and ability to communicate in English. Former criterion was restrictive in practice, limiting the available participant pool. Moreover, given the supplementary role of these interviews within a primarily software-focused project, only four participants were interviewed. Prioritizing in-depth feedback rather than scale, the sample size was considered pragmatic rather than driven by theoretical saturation, though some recurring patterns in responses did emerge across interviews. All four individuals were contacted via email. Three were recommended by the thesis supervisor, and one was a scholar known to me professionally. All four agreed to participate and completed interviews within the established time limit. The sample included one undergraduate student, two PhD students, and one senior researcher, all with experience with dataset production.

The semi-structured interviews were conducted online and recorded using videocon-

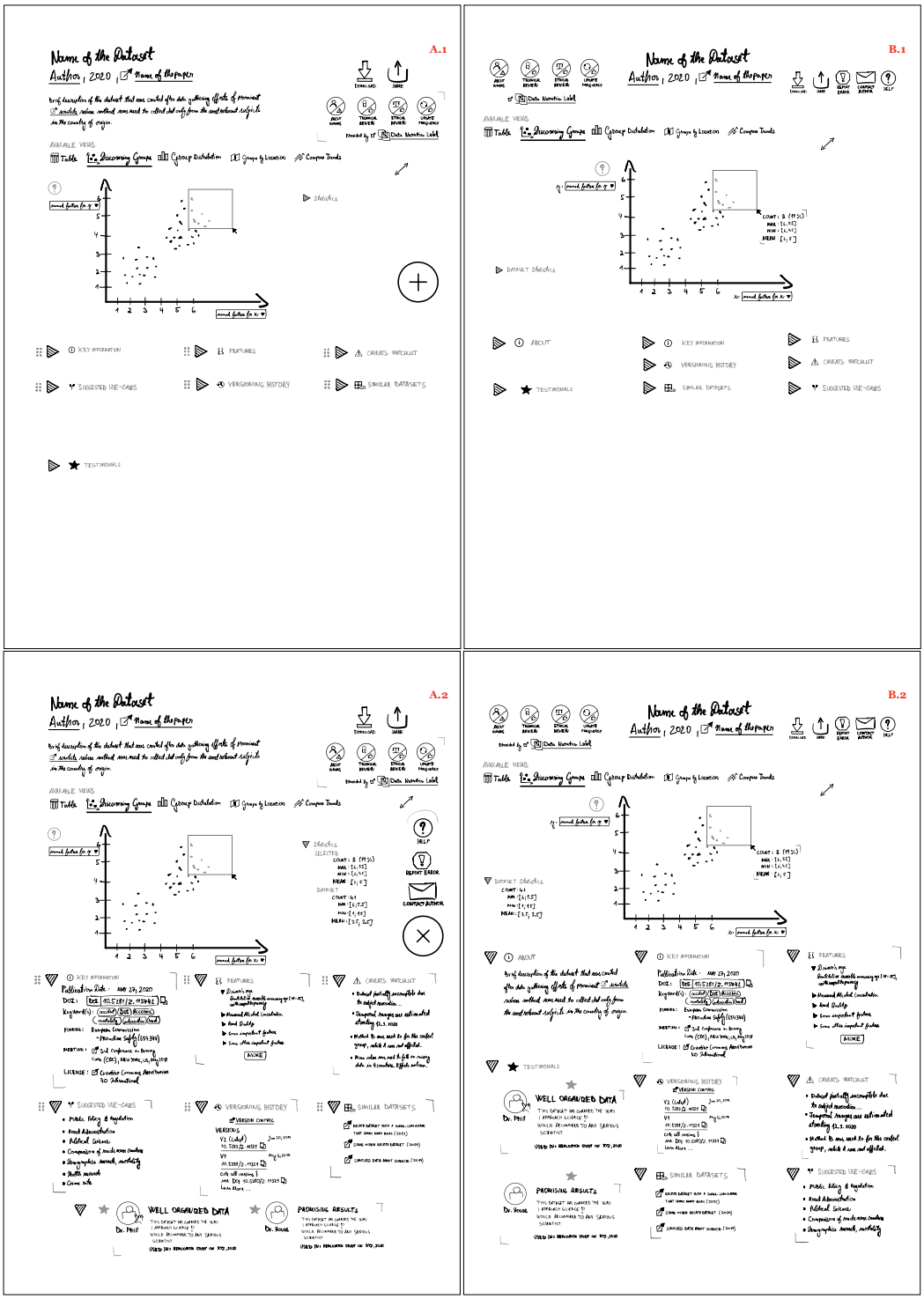

Figure 4.5.: First two pages of the prototype used in the interviews (closed and opened rollout panels)

ferencing platform Zoom. Each interview was limited to 50 minutes. Interview format followed a three-part structure: Introductory Questions, Dataset Questions, and Design Relevancy. Each part contained open-ended questions designed to address specific objectives. Questions were formulated to be non-leading and accessible in an effort to maintain neutrality, though some degree of framing is inherent in any interview design. To prevent bias in the answers, I withheld the project's goal from participants until after interviews were completed. Probing questions helped explore responses more deeply. In the final section (Design Relevancy), participants evaluated how effectively the designs showcased datasets. The low-fidelity designs of an envisioned data showcase application were central to this investigation. Questions were designed to evaluate the mock-ups against several criteria, including clarity, engagement, context, and trustworthiness. The interview guide with color-coded questions is included in the Appendix A.3.

At the start of each session, I reiterated the instructions and participants' rights as outlined in the consent form (see Appendix A.2). I also provided an overview of the study, mentioning that the interview is part of my master's thesis. I then asked participants for verbal consent to record the session before proceeding with introductory questions.

Analysis was approached using a mix of deductive and inductive thematic analysis as described by Braun and Clarke [95]. This allowed themes to emerge from participant responses while considering evaluation criteria from the literature review and repository survey. Interview recordings were transcribed using the transcription feature in Microsoft Word. Participants were not asked to review transcripts or validate findings. Transcripts were then imported into MAXQDA[6] for systematic coding. When transcript content required clarification, I revisited the original video recordings. I conducted all coding independently. While focusing on specific codes, the coding process remained open to unexpected themes. The codes used are summarized in Table 4.1. Discussions ultimately clustered around five areas centered around the prototype: Dataset Credibility Indicators, Summary, Layout, Data Access, and Detailed Information.

4.2.2. Changes Among Interviews

A few minor changes were introduced between the initial and subsequent interviews to improve the respondent experience. The most notable change was an altered order of designs. The first page initially displayed designs with all information panels expanded (*A.2* and *B.2* in Figure 4.5). However, the pilot participant confessed feeling overwhelmed by the amount of information presented during the initial encounter. In the interviews that followed, the first design shown (*A.1* and *B.1*) had all information panels closed. I then presented designs with expanded and semi-expanded roll-out panels on subsequent pages. This decision allowed participants to become familiar with a cleaner interface before introducing additional cognitive load with expanded roll-out panels.

The pilot interview also prompted a minor adjustment to the displayed design. Initially, users could select a feature associated with an axis of a respective visualization, with the selection buttons grouped and aligned to the right of the visualization. Following a

[6] Qualitative data analysis software `https://www.maxqda.com/`

Table 4.1.: Coding scheme for qualitative analysis

Code	Coded Segments	Description
Latent Need	42	Implicit need identified when the interviewee described their research
Potential Users	14	Potential users that can benefit from using the data, envisioned by the data producer
Expectation	48	Feature of an application or design that the interviewee expects
Concerns	20	Represents any concerns raised by participants
Familiar Software	9	Software with which the interviewee is familiar, which might be helpful for creating a common design, making the design more intuitive to use.
Confusion	15	Participants seek clarification or additional information about the design
Suggestion	28	When a participant suggests a missing feature or proposes an improvement to an existing design.
Negative Feedback	13	A participant explicitly gives negative feedback for a design feature
Positive Feedback	49	A participant explicitly expresses praise for a feature of the design

participant's suggestion to place the selection buttons in close proximity to the respective axis, this change was implemented in the subsequent design.

In terms of the interview format, since the nature of the questions remained unchanged, the pilot interview was included in the resulting study. Similar to the adjustments to the design, the order and explicitness of the questions were subject to adjustment. The first participant in the initial interview provided responses that did not diverge from answers from other participants. The remainder of this chapter discusses the results of the interviews.

4.2.3. Dataset Credibility Indicators

Since trust appears to influence the decision to reuse data substantially, users need assurance that the information presented is credible. In this context, participants raised concerns about the accessibility and usability of the platform. One senior researcher cautiously noted that existing solutions primarily target experts. This suggests a need for more accessible alternatives. Another participant offered a contrasting perspective, suggesting that even with improved accessibility, experts might not readily adopt such platforms. From this perspective, senior researchers often obtain data for reuse through personal interactions (conferences, e-mails) with other high-profile researchers rather than through online repositories.

> *"This is, I would say, very much field of expertise in the field that you know, [one knows] what is in the field, right ... We are serious researchers. We*

meet on conferences, you know, so that is how we exchange data."

Along similar lines, participants raised concerns about the discoverability of datasets. The participants emphasized the challenge of finding specific datasets without prior knowledge of their existence. This concern highlights the importance of making datasets discoverable and searchable, both within and beyond target research communities.

"So what is very hard to do as well is if you don't know already what the data set is about to find datasets that are just there."

When discussing credibility, participants stressed that peer review information of the associated papers should be integrated into dataset repositories. Two respondents expected this information to be readily available. Notably, within the application's design, analogous information is already featured next to the label *meeting* under the section *Key Information.* The highlighted meeting name links directly to the archive of the event where the paper was submitted. The design thus prioritizes dataset-specific information over papers that discuss results derived from the dataset. This approach emphasizes the data itself.

Turning now to identifying mechanisms, participants also raised concerns about what the provided Digital Object Identifiers (DOIs) actually reference. Since in the scientific community DOIs tend to be associated with academic literature rather than datasets, assigning DOI to other digital objects may be surprising. One respondent voiced confusion regarding the intended reference of the provided DOI. The solution to this ambiguity was to label the entry with *dataset.*

In line with efforts to enhance usability and facilitate scholarly workflows, participants advocated for the inclusion of features such as export to citation managers. This suggestion, may reflect a latent need to simplify how datasets are cited. One participant highlighted the significance of including citation information within datasets to accommodate users who primarily interact with the data rather than accompanying documentation.

"We also included the citation of the data set how to cite also in the data set, because many data users, they never download the document in material. They only download the data set."

Continuing the exploration of dataset credibility indicators, attention was drawn to various design elements. A notable design feature, derived from commercial practices, is the inclusion of a *Testimonial section.* This feature elicited mixed reactions. All participants expressed skepticism and advocated for it being in the hidden state by default. Cited reasons for the concerns were manifold, including testimonial authenticity, distrust in claims, scientific integrity, potential biasing, and the need for content moderation. In contrast, one participant offered a counterargument in favor of its inclusion. This participant believed that testimonials could improve trustworthiness by showcasing real-world applications and outcomes associated with the dataset.

> *"It also is trustworthy, if I have some kind of testimonials down there where I really see that people have already analyzed the data or worked with the data and actually produced something with this data"*

Another reportedly innovative element that was discussed was the simplified Data Nutrition Label. One participant found value in how the Data Nutrition Label summarizes key information. They emphasized ethical considerations and the importance of documenting both technical review processes and update frequency.

> *"I also like that you included the ethical and technical review and update frequency. These things are really important if you want to reuse data, you need to make sure that the data that you are trying to access were, for example, collected in an ethically accepted manner so that some ethical institution approved the collection of the data."*

Most participants, however, were unfamiliar with the Data Nutrition Label project. This lack of recognition suggests that awareness of such frameworks may still be limited among potential users.

4.2.4. Summary (Abstract)

All participants recognized the importance of the introductory description, which they often referred to as an abstract. This consensus appears to reflect a common search pattern where users typically look for an overview first.

> *"I really like the brief description of the data set on A1. The thing is that it's visible right away so you know it saves time. I don't have to click through to actually find out what's going on which I really appreciate ... It's like when you have an abstract for a paper."*

Returning to the issue of credibility, two participants emphasized the importance of detailed dataset documentation. They perceived a direct link between the amount of detail provided and how trustworthy a dataset appears. One of the respondents articulated it by suggesting that including as much information as possible increases the dataset's credibility.

> *"You can add as many details about the data set as possible, which I would think is something that we would find for example here in the key information that also makes the data set credible."*

Participants also valued including the name and link to the source paper where the dataset was first used. All participants considered this essential since the references help users trace a dataset's origins and find additional context.

Rejecting specialized terms and collocations as potential barriers to comprehension, participants favored plain language over specialized terminology. They further noted that clear, accessible language improves understanding, especially in technical fields where jargon tends to obscure meaning.

> *"I think that's also helpful [that] it has the verbalized description. So, I like this one [design A] better because it has a verbalized description just like in the common language."*

Not all participants agreed on the value of an abstract. While the majority acknowledged its relevance, one participant, disregarded the description, questioning its necessity for the dataset documentation. This unexpected contrary viewpoint suggests that information priorities may differ across users, pointing to the value of customizable interface options.

4.2.5. Layout

Participants universally evaluated the mockup interfaces as intuitive, voicing confidence in their ability to navigate and interact with the presented designs. As the presented design builds on various layout characteristics borrowed from existing online repositories, experienced data (re-)users would likely find the environment familiar and easy to navigate. One respondent remarked on the familiarity of the design layout, by likening it to similar repositories.

> *"To me this kind of dashboards and data download pages are very familiar so, for me there's nothing to figure out. I've seen such presentations."*

While most participants evaluated the compactness of the design as a favorable aspect, there were also contrary opinions. One respondent disliked the compactness, describing it as cluttered and needlessly utilizing all available space. Arguing for more breathing room around elements in the final application, they articulated concerns regarding the excessive density of elements.

> *"I imagine this on the web page right now and not just [on] a sheet of paper because for me, it's like a sheet of paper like you have an A4 sheet of paper and you just gonna use the entire space that you have. I would prefer to have it arranged in a way where maybe the header gets smaller and you have like more space on the sides and then it branches out again and you have the graphics displayed and then maybe the little squares with the informative parts they're arranged, in a more fun way"*

Among the two presented designs, Design A emerged as the preferred choice. Participants found Design A's layout to be more organized and clean, likely translating to easier comprehension. Specifically, the left justification of elements in Design A was favored over the centered style of Design B. One participant explicitly rejected the centered approach.

Although Design A was favored overall, different preferences were observed among participants. One participant expressed a preference for Design B's menu, where the complete set of menu items is visible compared to Design A, where only the most important items are shown. The participant explained that they wanted essential information visible at a glance.

> *"I really like to see these things right away, for example, contact information, because this is something that you will need, if you find some data set that you would like to reuse or you have some questions about the data set, the contact author, uh icon should be right away visible"*

Participants noted the presence of the speed-dial button. However, likely due to the static nature of the presented design, participants questioned whether this element would actually be useful in a functioning application. Reflecting these comments, I removed the speed-dial element from the final design.

Having described visual interface preferences of the participants, let us now turn to discussed abstract aspects of the design. One respondent raised concerns regarding the potential cultural bias inherent in the design choices. They further expressed skepticism about the universality of the design's accessibility. They suggested that the design might be tailored to Western audiences, potentially excluding or alienating users from other cultural backgrounds.

From the aesthetic point of view of one respondent, a comment was made that the grayscale presentation may not accurately reflect the intended visual esthetic of the final application. In this respect, the participant raised doubts about the fidelity of the presented designs and their alignment with the envisioned end product.

Lastly, participant articulated an interest in having a scaled version of the presented concept. Envisioned solution would combine repository of datasets with the presented solution, constituting platform where users could access a variety of datasets along with comprehensive details managed by the application.

> *"I would really love to have such a website, for instance. Where we have all the datasets in there with all this information that is actually there and to just browse through different datasets and look at the datasets more closely"*

4.2.6. Data Presentation

In their feedback, all participants unanimously identified the visualization component as the most important aspect of the presented designs. The interactive selection of specific data points and the general interactivity of visualizations received positive ratings from participants. The ability to view different visualizations for a given dataset was highlighted as a particularly useful feature. Emphasizing the importance of an engaging default visualization, one participant described the visualizations as *"inviting to play around"*.

Another participant suggested incorporating a signature piece that captures attention and portrays an interesting insight.

> *"The dashboard should not be empty, but it really should show you like one signature piece or something so that there is already a graph that looks nice ... If there's just a random graph that can actually turn people off ... I would make sure that this looks beautiful"*

While the general consensus among participants was to prioritize the display of visualizations, one participant advocated for giving precedence to the information displayed below the visualization.

> *"If I was going to start working with this dataset, I think personally I would need these information that are featured below the visualization first before the graph would make sense to me."*

In contrast, another participant expressed concern about the lack of visibility of raw data. This suggests that a more prominent display of the table view or making it a default view option. Additionally, as most elements in the design can be hidden on demand, one participant recommended an option to hide visualizations to be made available.

In the scatter plot visualization, concerns were raised regarding the clarity of calculated statistics. For displaying local statistics, one participant explicitly favoured solution in the Design B, which avoids any ambiguity by displaying calculated values next to the lower right corner of a brushed rectangular area. Along similar lines, global statistics were expected to be displayed by default.

Participants unanimously agreed on the significance of presenting summary statistics when attempting to understand data, along with the importance of a quick reliability indicator. These features play a crucial role in providing users with a comprehensive overview of the dataset's characteristics and aiding in its interpretation and analysis.

Participants unanimously expressed a latent need to display univariate statistics such as measures of central tendency, variability, or frequency distributions. They uniformly identified this as the initial task when working with datasets.

> *"If we have some quantitative data, then I usually calculate the mean, the standard deviation and I look at the distribution of the quantitative data values"*

During a discussion about their research, participants identified the need to visualize changes in data over time, emphasizing the critical role of temporal representations in elucidating trends and patterns.

> *"Maybe just showing mean levels changing over time ... we need to sort of compress it more so we just show the mean trend for a different category. Yes, so that was sort of univariate statistics that we use most of the time."*

A participant disclosed the challenges faced in paper publication, where severe limitations are imposed on visualizations by publishers due to page limitations for each paper. Participants expressed a desire to incorporate additional visualizations to provide a more comprehensive understanding of the data. Reportedly, the minimum constraint often fails to convey the entirety of the story. This constraint presents an opportunity to explore innovative ways to convey insights visually, unconstrained by limited space and format.

While the general consensus among participants was to prioritize the display of visualizations, one participant advocated for giving precedence to the metadata displayed below the visualization.

> *"If I was going to start working with this dataset, I think personally I would need these information that are featured below the visualization first before the graph would make sense to me"*

4.2.7. Data Access

During the interview, participants were asked about the software they use to gather and process data. The rationale behind this topic was to understand the user interface elements participants are familiar with and potentially leverage this knowledge for designing a familiar interface. In that respect, participants reported using spreadsheet editor software such as Microsoft Excel or Google Sheets for initial data inspection and manipulation. Additionally, SPSS[7], Stata [8], and R[9] were mentioned for calculating statistical models and analysis. For visual inspection and discovery, participants mentioned using Stata and Tableau[10]. Qualtrics[11] was reported as the tool of choice for conducting questionnaires.

One participant emphasized a preference for solely utilizing the design to download the dataset, suggesting that the download button should be the most prominent element. As a reaction for clicking the download button, all participants anticipated triggering of a modal dialog presenting options to download the dataset. Alternatively, before presenting the user with format options, licensing window was expected that must be agreed to before proceeding with the download. One participant expressed a preference for downloading the dataset along with all accompanying resources as a single archive.

> *"I think the most important thing for me would be that it downloads everything combined, including the documentation at once ... in my view it should be a single click and you should get all the relevant files."*

Additionally, the participant argued for the inclusion of identifying information and metadata in the common archive with the data. Drawing from their experience, they highlighted that many data reusers do not read accompanying papers or documentation, stressing the importance of including essential information directly in the dataset.

> *"Much of this information also needs to be included in the data set, yeah, because also many data users they read nothing. They don't read the data paper, and they don't read the method report. So also, in the data set is also very important that for example the date of the interview"*

Returning briefly to the subject of expected license dialog, this seemingly trivial operation might be one of the obstacles in reusing data. One participant highlighted the challenges associated with downloading datasets. They enumerated numerous complexities and inconveniences when attempting to obtain datasets. The difficulties encountered

[7] IBM SPSS Statistics `https://www.ibm.com/products/spss-statistics`
[8] `https://www.stata.com/`
[9] programming language R `https://www.r-project.org/about.html`
[10] `https://www.tableau.com/`
[11] `https://www.qualtrics.com/`

have varied, spanning from redirection to external providers, to the necessity of signing up or registering using an academic email, and even requiring the submission of an official email request followed by a multi-week processing period. The respondent noted that navigating the intricacies of dataset download often requires dedicating an entire lecture to this topic.

> *"We spent an entire sessions only on downloading, because one cannot imagine what kind of hurdles [there] are in the way. You know, where are they [the data]? You have to sign up here, request it there, click here and so on. Where are the guidelines [how to download it], where are they?"*

Moving on, prior to the presentation of the design, two participants emphasized the necessity of being easily contactable. This underscores the significance of the feature enabling straightforward communication with the author and potentially guiding interested parties to additional relevant resources. Doubting its practicality, one participant expressed skepticism about the usefulness of the share button. They recounted previous experiences of media engagement, emphasizing the necessity of facilitating easy contact with the author and to redirect interested parties to relevant resources.

> *"We did a lot of media stuff, so a lot of journalists called us all the time because they wanted to know about what the data are showing, we wrote a lot of blog posts to disseminate the research to the to the general audience"*

4.2.8. Detailed Information

Let us turn to the metadata information presented below the visualizations. All participants unanimously agreed that the order of information panels is intuitive. One participant suggested personalizing the order of panels to better suit individual preferences. Furthermore, all participants understood the feature for hiding panels that contain information. They appreciated the ability to hide information that is not immediately useful, allowing them to reveal or conceal information as desired. Furthermore, it was noted that this feature helps users to focus on elements they consider important.

As mentioned in the changes section, participants collectively agreed that the amount of information becomes overwhelming when all information panels are unfolded. This remark led to the aforementioned change in the presentation of designs.

Explicitly favorable mention was made of several panels, including Key Information, Features (description of the features), Caveats Watch List, Similar Datasets, Use-Cases, and Versioning History. In the Key Information panel, alongside funding and dataset license information, the publication date was highlighted as a particularly important component. Further, all participants unanimously agreed that key information should be the first information panel, in contrast to the short description as shown in design B. One participant stressed the importance of having access to information contained within the Key Information section to facilitate comprehension of the contextual nuances inherent in the utilization of foreign datasets.

> *"I find key information very important when dealing with datasets. ... You need this for anything, you need it to understand what you're looking at."*

Continuing discussion about elements included in the Key Information panel, a noteworthy comment was made regarding the nature of keywords in repositories. While one participant expected keywords associated with the dataset to be shown, they also highlighted a known issue in data repositories where keywords intended to better organize and describe datasets tend to have the opposite effect. Specifically, it was communicated that these keywords make the dataset less understandable and discoverable. Arguing for comprehensible language, another participant asserted that the problem with keywords is that they are often artificially manufactured by repository maintainers. Adding to this sentiment, such individuals are hardly ever domain experts for the given datasets.

Concluding the insights about the Key Information panel, the need to incorporate details about funding was validated by two participants. The respondents have mentioned it prior the design presentation. As discussed in the credibility section, respondents acknowledged the pivotal role funding transparency plays in establishing trustworthiness and in guiding the appropriate level of skepticism towards the data.

Moving on, a latent need for a feature that would aid understanding and of dataset features was identified prior to the presentation part of the interview. Additionally, the need to understand data features was also mentioned by participants explicitly, supporting the role of the data description (Features) panel. Aligned with respondents' methodology for comprehending data, one participant characterized their strategy for assessing datasets.

> *"I would have a short look on the data features that are in there. If there are the features there that I actually need, and if I find something interesting, then I would go into the table and really look at the data in the raw worksheet for instance."*

Along the similar lines with artificiality of keywords, one participant described the challenge with assigning names to data features from the perspective of a data producer. Explaining the difficulty of mapping feature names to the collected data, they confessed that many scientific disciplines lack authoritative reference for data variable names.

> *"... I had no proper documentation, so nothing to explain the variable names, the phrasing of survey items itself was not connectable easily with the variable name and so on"*

An idea that resonated among participants was that data producers should be transparent about the methods used to collect and process data, and report any non-standard incidents that might have influenced data collection and processing. This supports the idea that there is a need to address known problems that occurred over the lifetime of the dataset, as evidenced by the creation of a caveats watch list. One participant emphasized the importance of transparency, stating:

> *"I find [it] really, really important because [the] first [caveat] already says that this dataset is partially incomplete, and this is really important to know. I could also look at the raw data and see that there are data points missing, but it's nice to have it explicitly stated."*

One participant expressed difficulty in understanding the term "caveat". Contrary to the consensus around the importance of knowing about problems in the data, the participant concluded that since they didn't understand the term, it is not important. This insight exemplifies a concerning scenario wherein ambiguous language leads the user to reject an otherwise well-founded feature. In an attempt to avoid further misunderstanding, the name of the section was changed to *To consider*.

Commenting on the panel listing potential dataset use-cases, one participant believed that envisioning use-cases other than the original purpose of the dataset was a valuable exercise for data producers. Reportedly, envisioning alternative use-cases encourages data producers to think about uses outside of their domain, promoting reuse by researchers from other disciplines.

Coming back to the question of credibility, one participant pointed out that information panels lack clear information about the number of participants, sample size, duration of data collection, or date of data collection. They suggested including general information regarding the sampling design, fieldwork conductors, interview duration, interview mode, participant compensation, incentives used, study funding, and related details. Additionally, participants missed having an explicit explanation of the method used for data collection, including detailed information such as survey software or programming language used. Participants argued that such information is often omitted, making it difficult to assess the dataset's quality and compilation process.

> *"I would like to know how many participants did they have, or how it was conducted. It's very hard to understand what it [the dataset] does, how it has been compiled, what the quality of it is and so on"*

It is a valid assertion that details regarding the sampling methods employed for data collection should be incorporated into the dashboard. Some existing elements, such as the key information section, abstract, and caveats watchlist, partially address this need. User feedback suggests, however, that more detail would be valuable. In response to this feedback, supplementary information regarding dataset sample will be made available as an optional feature within the Key Information section.

Finally, information panel describing versioning information was a welcome feature among participants. In addition to documenting alterations over time, two participants voiced a sentiment that planned support would be a useful addition. This could serve as a reminder for data reusers to anticipate periodic updates. Drawing upon these insights, I incorporated an optional support indicator into the panel.

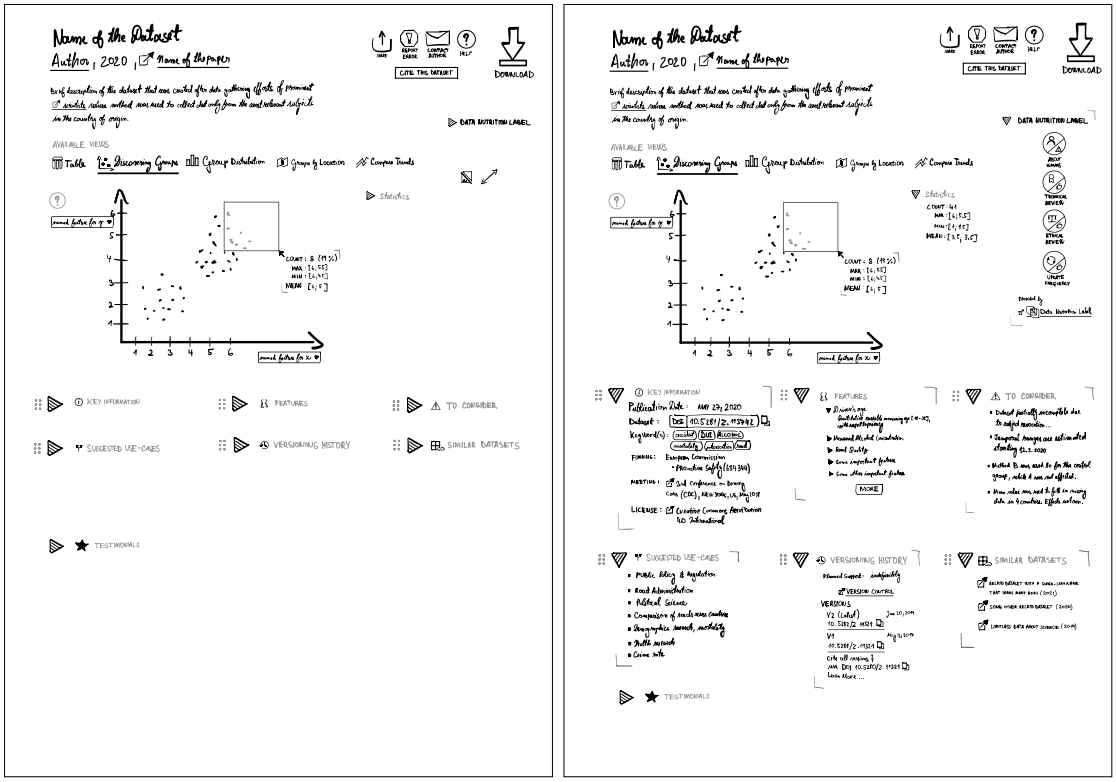

Figure 4.6.: Final low fidelity design informed by the interviews

5. Implementation

This chapter introduces the dataset visualization platform. Building on the findings of the preceding project stages, the chapter examines key architectural and implementation decisions. It begins with a high-level overview of the system architecture, followed by a detailed description of the user interface and front-end functionality. Subsequently, the chapter presents the back-end file processing workflow and the algorithmic design of the dataset profiler. The chapter concludes with a description of the application deployment, including an enumeration of the supported execution modes and their configuration. In addition, it highlights aspects of both code and project-management practices, as well as the optimization measures used. Throughout the exposition, the encountered challenges are systematically identified and their corresponding solutions are discussed.

Seeing the growth of the project to become a unique solution, I have decided that an appropriate name is warranted. In an effort to name the system to reflect its purpose, I also wanted to make it witty, hiding a non-obvious reference or message. After brainstorming various names, I have settled on the name **ViDa** short for **Vi**sual **Da**tasets. The twist is that it also refers to the Czech interjection or expression for astonishment *"No vida!"* or *"Vida Vida!"* roughly translated to English as *"Well, well, well!"*.

5.1. Architecture

Similarly to the low-fidelity design phase, the architecture of the application has evolved over time as technical challenges emerged. The final architecture of the application shown in Figure 5.1 depicts all components of the multi-service application, emphasizing the optional inclusion of components. Depending on the mode of deployment (see Subsection 5.4.1), ViDa runs in two containers, orchestrated by a compose file. The Front-end can optionally communicate over HTTP to the Transformation service, which can further store and retrieve files from Azure Blob Storage. The Frontend application built with NextJS consists of two high-level components, React-based user interface (`Frontend` and `API proxy`) implemented using the built-in NextJS API Server running on the NodeJS server. Front-end sources its configuration files from mounted Volume, which is set up on container startup.

The Transformation-service, built using the Go framework Gin, offers REST-like endpoints for processing incoming data. Optionally, it can store or retrieve the processed data to a connected Azure Blob Storage instance and relay those resources to the front-end. Containerization of applications enables isolation of environments. This separation of concerns enables ViDa to fall back on the data in Persistent Storage (Volume) when transformation-service or Blob Storage become unavailable and vice versa. In the unlikely

scenario that the volume dismounts or otherwise fails, the front-end withstands the fault by using sensible defaults including a model dataset. Part of the implemented fault tolerance mechanism is notification of the user via alerts and error logs.

5.1.1. Architectural Decisions

Led by the KISS[1] design principle, the initial concept considered only the Front-End, which directly communicated with Azure Blob Storage as the only default store for all files. The rationale behind this was that centralized remote storage of files for all showcases eliminates a class of caching and networking traffic limitations. The core assumption was that users don't want to host their dataset showcase themselves. However, it became evident that this solution introduces unnecessary complexity and multiple points of friction, evoking an enterprise extension rather than a core feature. Although Azure cloud environment guarantees high availability, it introduces a single point of failure, requiring an excessive administrative setup and routine rotation of security keys. Such a system also relies on the proprietary solution and API of a single vendor, which may potentially lead to vendor lock-in. Perhaps most importantly, the user may not wish to store their data in Cloud Storage. For that reason, I reconceptualized the architecture of the system to source configuration and other data locally from a Volume by default and let the data producer opt in to using Azure Blob Storage.

Direct communication between the front-end and the external API is also problematic. Since all client-side data are public, this approach risks exposure of API credentials. Moreover, modern browsers block cross-origin requests unless the server explicitly sets appropriate CORS[2] headers. In response to these issues, I implemented the Backend for Frontend Pattern (API Proxy) built on top of client-server model [96]. In this pattern, the client communicates with the intermediary server, which then appropriately relays the call. Furthermore, when a user chooses to store their data in Azure Blob Storage, this call does not go to Azure directly. Rather, it is relayed through the Transformation Service, responsible for dataset processing. This helps maintain the separation of concerns and keeps the responsibilities of individual components within their bounded contexts, as recommended in Domain-Driven Design [97].

One of the largest bottlenecks was the processing of the dataset in the browser. Although there are libraries that allow parsing of common data formats such as CSV on the client, it would likely negatively influence the Quality of Experience (QoE). Depending on their machine specification, users would likely experience delays for larger datasets. To offload the client, the stateless Transformation Service addresses this problem by processing the uploaded data and returning it in an optimized JSON format. It is important to emphasize that the transformation process takes place only once[3] during the generation of the ViDa showcase and the resulting artifacts are stored and served from the front-end container volume afterwards.

[1] Acronym for Keep It Simple Stupid

[2] Cross-Origin Resource Sharing, a browser-level security mechanism that regulates access to resources from foreign domains

[3] And potentially during an update of the dataset

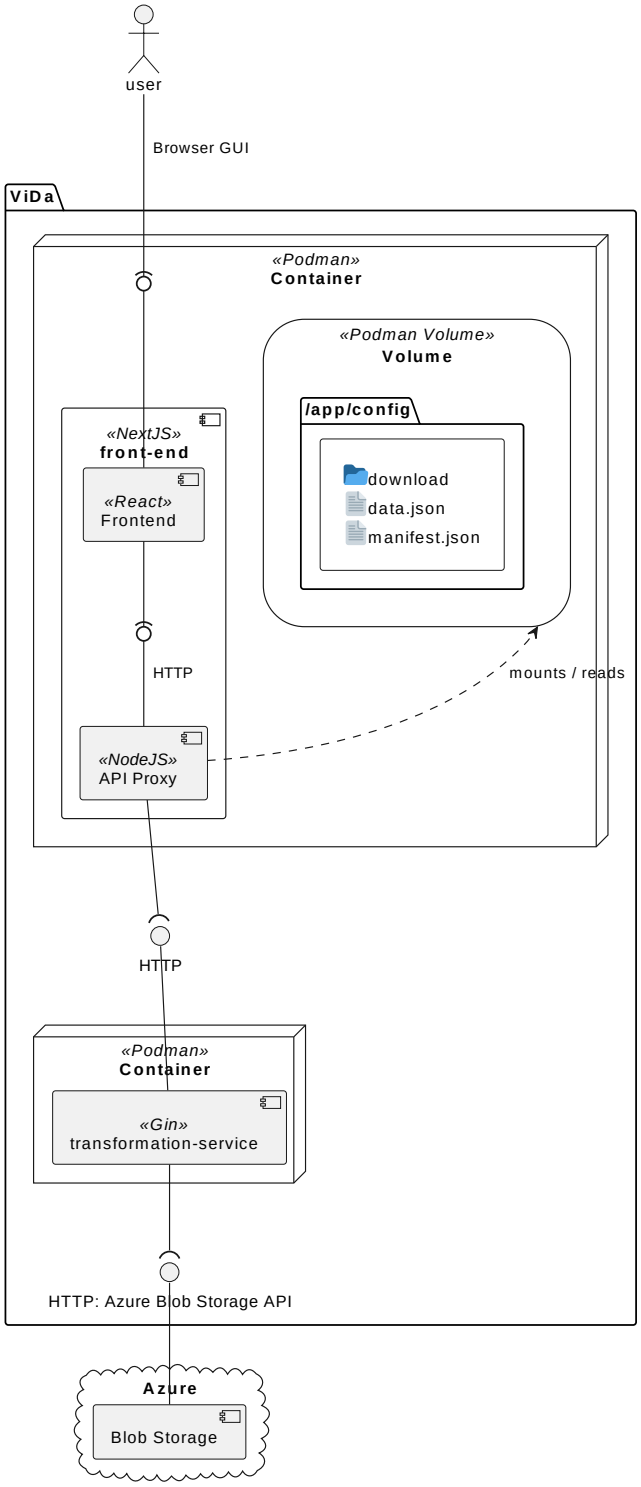

Figure 5.1.: ViDa Architecture

5.2. Front-End

The base of the front-end uses the open-source React[4] Framework, NextJS. I chose this framework by Vercel[5] because at the time of writing, it is the de facto standard choice for full-stack projects built with React. Although developed for composing user interfaces, the main advantage of React is its vast ecosystem covering all concerns of web development. NextJS bundles and extends many popular React extensions and eliminates otherwise tedious aspects of setting up and managing React projects. Key project-relevant features of the framework include zero configuration for TypeScript[6], CSS-in-JS, file-based routing, code-splitting, image optimization and elaborate tooling for development and production build. Furthermore, since React by design supports only client-side rendering, NextJS bundles an optional Node server for server-side concerns. Although both yarn and npm package managers for managing external dependencies are supported, I opted for npm[7].

For the development of the user interface, I used a React implementation of Google's Material Design, Material UI (MUI)[8]. Initially, I implemented the visualizations and statistical summaries exclusively using the D3[9] library. However, this approach soon encountered technical and efficiency challenges, especially inconsistent presentation and behaviour on different viewport sizes and lower performance inherent to vector graphics. Addressing both accessibility and edge-case display concerns proved unnecessarily laborious when extensible optimized solutions were already available. To mitigate the performance bottleneck for visualization of large datasets, I customized and extended charts from a bitmap-based visualization library Chart.js. The library allowed me to focus on parsing incoming data into the format expected by the visualization and not on appearance implementation details.

Forms in React tend to be notoriously difficult to manage at scale. ViDa implements complex multi-step forms for describing dataset metadata. Due to the stateless design of the system, input validation takes place on the client. To validate and manage the state of the form fields, I used the React form library formik.js[10] and described validation schemas using Yup[11]. Finally, to help users navigate the platform and interactively discover available features, I implemented guided tours using Driver.js[12].

5.2.1. Pages

ViDa's front-end is structured around three pages as shown in Figure 5.2:

1. **Main Page**: Dataset Dashboard with visualizations and contextual metadata

[4]Declarative JavaScript library for building User Interfaces `https://react.dev/`
[5]`https://vercel.com/`
[6]popular superset of JavaScript by Microsoft `https://www.typescriptlang.org/`
[7]Node Package Manager `https://www.npmjs.com/`
[8]Open-source component library Material UI: `https://mui.com/`
[9]D3.js: Data Driven Documents `https://d3js.org/`
[10]`https://formik.org/`
[11]Object schema validation library `https://github.com/jquense/yup`
[12]Driver.js `https://driverjs.com/`

2. **Wizard**: Guided walkthrough dialog that helps to describe the dataset and generates a configuration file that powers the Main Page

3. **Publishing**: Static page with instructions on how to generate and publish ViDa showcase.

Users start on the main page where the dataset is presented. They can inspect associated metadata, interact with dataset-specific views and download the dataset together with supplemental files. Users can also share and cite the dataset, report a discovered platform-specific bug or a problem with the current dataset. If the author(s) provide their e-mail address, users can also contact them. Contacting via e-mail is solved by triggering the default e-mail client on the user's device with pre-filled information.

From the main page, users can navigate through the first option in the speed-dial menu (*New Dataset*) or link button (*Create Your Own Dataset Showcase*) from a help dialog to the Wizard page. Located at the path `/wizard`, it offers a guided stepper dialog with forms designed for accurate description of a dataset. Successful completion of the wizard triggers the download of the configuration archive `vida_config.zip`, containing configuration and dataset files. The resulting success screen offers the user to either preview the showcase or learn how to publish the dataset. Previewing the dataset on the main page is only available immediately after completing the wizard until the user closes the browser or presses the *Close Preview* button next to the *Preview* banner in the top left corner.

The last page `/publishing` enumerates technical requirements for deployment of the ViDa showcase. It can be reached from the help dialog via the *Learn How To Publish* link button or via the same button in the success screen of the wizard page. Although ViDa is conceptualized to be a desktop-first platform, all pages and components are responsive, supporting tablet and smartphone screen resolutions. A notable limitation of the visualization components is that they may become small and difficult to operate on smaller screens.

5.2.2. Data Views

The core section on the main page is the data visualization. Displayed views are determined based on the variable types inferred during dataset profiling as described in Subsection 5.3.1. The final selection of implemented data views listed in Table 5.1 reflects the effort to avoid specialized visualizations as described by Franconeri and colleagues [98]. Rather, the goal is to provide common visualizations that are widely understood and can be used by both professionals and lay audiences alike. The second rationale is that the provided visualizations should cover most of commonly occurring dataset shapes [98].

For visual consistency and predictability, I paid special attention to establishing common visualization conventions in the platform. This includes positioning of visualization controls and helper tools, appearance of the elements and consistent use of color [98]. The following list of conventions should also be used as a guide for future implementation:

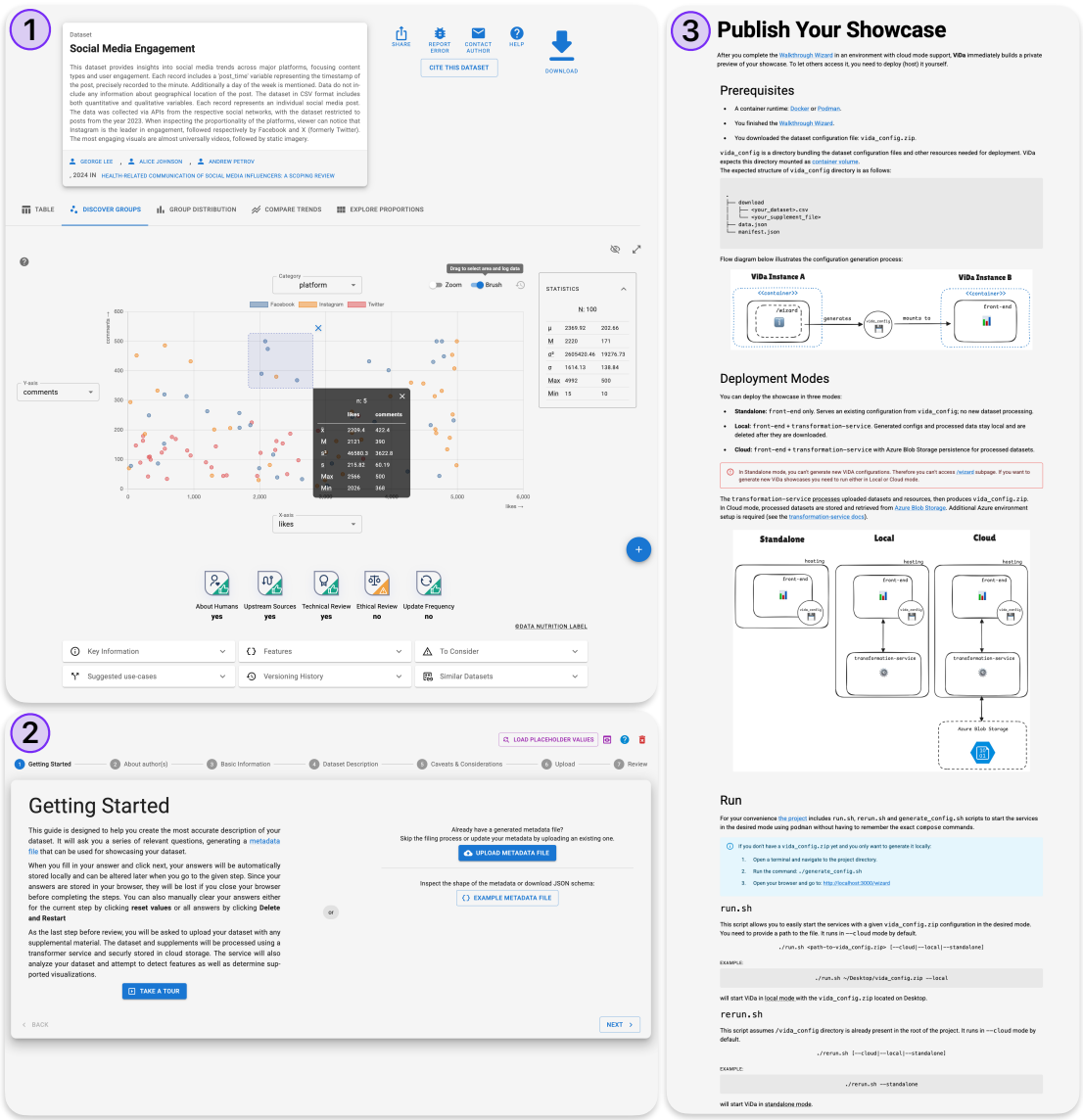

Figure 5.2.: Overview of ViDa Pages, (1) Main Page (2) Wizard page (3) Publishing

- Each visualization has a help icon-button in the upper left corner that shows on hover the name of the data view and what tools are at the user's disposal.
- Visualization can always be hidden using the eye icon-button. This mitigates visual distraction when trying to focus on other elements.
- The neighboring bidirectional arrow icon-button expands the visualization to full-screen in a modal dialog.
- Optional visualization tools in the **toolbar** are located in the top right corner above the visualization canvas. Where relevant, tools are usually implemented as a toggle button or icon button.
- Color-coded variables are shown in a legend displayed at the top of a given visualization. Users can toggle displayed groups by clicking on the data group legend. Groups that are currently hidden are marked with strike-through text.
- For color-coding of logical groups, visualizations use a discrete color scale using d3.schemeTableu10 as demonstrated in Figure 5.3.
- If visualization displays data along axes, categorical variables (column headers) labeled as categories can be changed using a drop-down located above the data view canvas (above the legend).
- When relevant for visualization, axes labels can be controlled using drop-down controls located in close proximity to their corresponding axis (left of the y axis and below the x axis).
- Scales of both axes start with zero.
- Initial display and data updates are animated.

5.2.3. Table

When the user visits the page, the default selected data view in the visualization section is table. This is because it is guaranteed to display the data regardless of the variable

Table 5.1.: Visualization types and their dataset variable requirements

Visualization	**Tab Label**	**Dataset Variable Requirements**
Table	Table	none
Scatter Plot	Discover Groups	2 quantitative variables
Bar Chart	Group Distribution	1 categorical and 1 quantitative 1 quantitative and 1 unique
Line Chart	Compare Trends	1 temporal and 1 quantitative
Tree Map	Explore Proportions	1 categorical and 1 quantitative

Figure 5.3.: Discrete color scale used for color-coding in visualizations

types detected in the dataset. Implemented using MUI X Data Grid[13], the component enables users to interactively browse a dataset in tabulated form. Data records are organized in pages which can be navigated using controls in the bottom right of the table. The number of displayed rows per page is also customizable. Users can search, hide and filter columns using the tools menu located at the top of the table. From the same menu, users can also control the width of the rows. Columns can also be granularly controlled by clicking on the vertical three-dot menu button next to the header of the column, opening a pop-up menu. Users can sort records alphanumerically in ascending or descending order by clicking on the option in the pop-up menu or by clicking on the arrow next to the header indicating the current sorting mode. The density button in the top-left table menu allows users to control the width of the data rows. Finally, modified data can be printed or exported in CSV format.

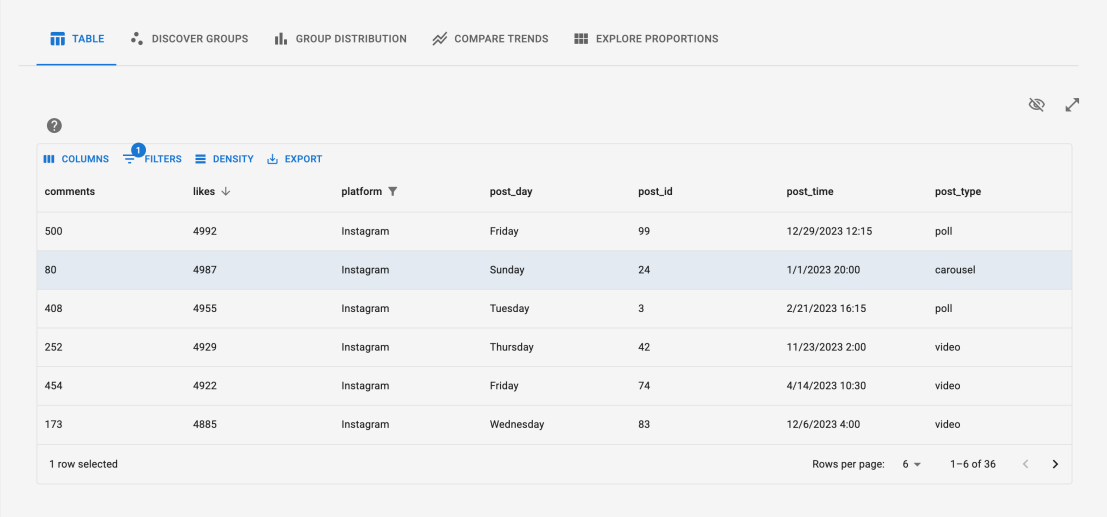

Figure 5.4.: Table view with a selected row, an activated descending sorting of records (column `likes`) and filter for column `platform`

5.2.4. Scatter Plot

Labeled *Discover Groups*, the Scatter Plot was implemented by extending the Scatter Chart component from ChartJS with a zoom plugin and custom brushing component. The view is offered when the dataset contains at least two quantitative (numerical)

[13]Extendable react data table `https://v6.mui.com/x/react-data-grid`

variables. By default, the first two numerical variables are assigned to the scales of the x and y axes, and the positions of the data points are then spatially encoded into the plot. When hovered over, data points enlarge slightly and a tooltip pointing to the selected point is displayed. The tooltip lists the associated color-coded group, name and the x and y axis values, respectively. Data points are color-coded based on their group indicated in the legend above the chart. Above the legend, users can change the category using the dropdown button. Groups are determined based on categorical variables (columns) found in the dataset. Reoccurring values from the column become color-coded groups. Users also have the option to control the assignment of x and y axes, representing the quantitative variables.

Users have two more options for how to examine data. By enabling the *Zoom* toggle button from the toolbar, users can zoom in and out of the data clusters using mouse scroll. Alternatively, on touch screens or track pads, users can use the pinch-to-zoom gesture. Additionally, to move around the zoomed-in data points, users on desktop can pan the canvas by holding `shift` and dragging it. Zoom dynamically recomputes the scales of the axes and repositions data points. The zoomable/pannable areas are indicated by a change of cursor to a magnifying glass or hand cursor. The listed instructions are also displayed in a tooltip pointing to the enabled toggle button. Users can also reset zoom and pan using an icon-button in the aforementioned controls row.

The second toggle button in the toolbar enables the brushing tool. When the brush is enabled, Zoom is automatically disabled (while staying zoomed in), and vice versa, to avoid conflicting interaction. Using the brush tool, users can select a rectangular area on the visualization canvas. When there are data points within the bounds of the brushed selection, summary statistics are computed and tabulated in an associated tooltip. The displayed statistics are computed for both the currently selected x and y axis variables. Statistics include the number of points n, arithmetic mean $\overline{x}$, median M, variance s^2, standard deviation s, maximum Max, and minimum Min. Both the brushed area and local statistics can be dismissed using a cross icon button displayed in the top right corner of the components. Local statistics can be compared with the same global statistics computed for all points in the viewport and tabulated in the rollout panel next to the visualization. To highlight that the statistics are computed for all points, the measures in the same order are labelled N, μ, M, σ^2, σ. To aid accessibility for lay audiences, all statistical symbols display a textual label when hovered over.

Being the most complex and time-consuming visualization to implement, the component warrants further commentary. Implemented using D3, the initial solution used the vector-based `<svg>` element. This implementation is still present in the code base in the file `SVGBrushableScatterPlot.tsx`. Facing inconsistent animation behavior across browsers, sizing issues and degraded performance on the client when using larger datasets, I decided to reimplement the visualization in a raster-based format. Unfortunately, transferring to a different graphics paradigm introduced unexpected complexity. With 1123 lines of code, the solution in `BScatterPlot.tsx` implements the visualization with the `<canvas>` element. The main challenge of this solution was synchronization of the tracked data points both for animation and brush-selection purposes, which is

handled using DOM elements in the vector implementation. Additionally, canvas-based animations require explicit synchronization of animation frames using the standard browser `requestAnimationFrame` method[14], as opposed to optimized CSS styling of DOM elements. Despite being functional, the solution suffers from occasional repainting when interacting with the brush tool, causing flashing of the canvas content. Facing time-consuming debugging sessions with these relatively benign concerns, I ultimately decided to switch to the optimized ChartJS component instead. In this final solution located in `ScatterPlot.tsx`, to avoid needless complexity, I implemented the brushing functionality by superimposing a `<div>` element over the bitmap canvas. This approach advantageously combines both vector and bitmap approaches for performance-critical tasks.

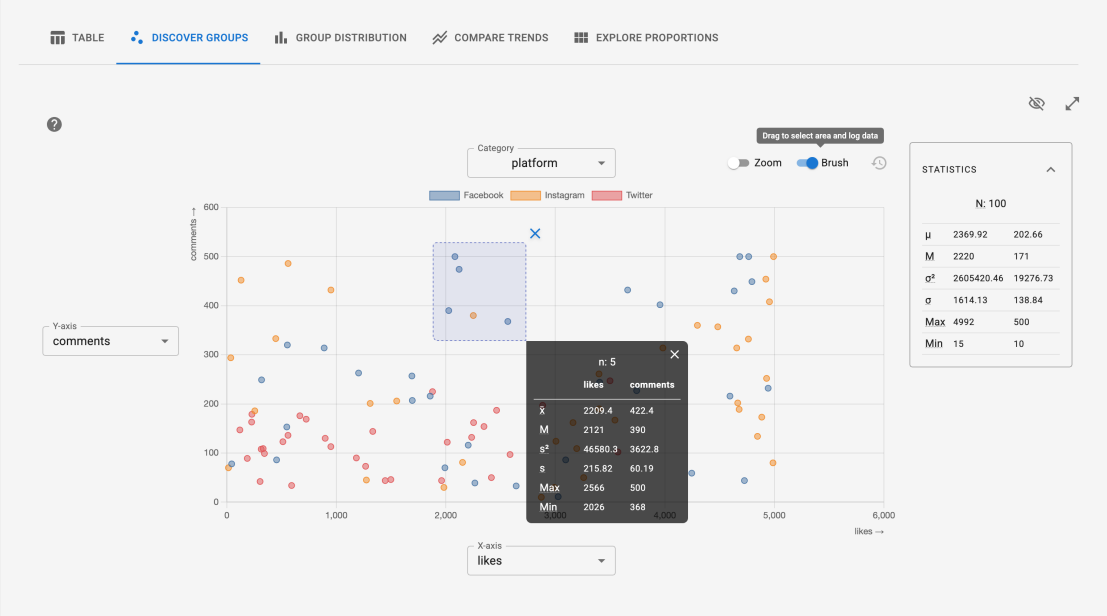

Figure 5.5.: Scatter Plot View with an enabled brush

5.2.5. Bar Chart

Labelled as *Group Distribution*, the bar chart view is a customized Bar Chart component offered by ChartJS. The component operates in two modes, showing either derived or single values. The default derived mode is supported when the dataset contains at least one categorical and one quantitative variable. In derived mode, the bar chart displays derived aggregated values computed from the categorical variables. Derived values encoded by the length of the bars (one bar per category from a categorical variable) are juxtaposed in the view for each quantitative variable. Users can control the currently

[14] `https://developer.mozilla.org/en-US/docs/Web/API/Window/requestAnimationFrame`

displayed values by selecting a value from the dropdown to the left of the bar chart. As indicated by the label of the dropdown, when a value is selected, it changes the y-axis label, its scale and recomputes the dimensions of the bars. Dropdown options for computing derived values are *total*, *mean*, *median*, *mode*, *variance* and *standard deviation*. The component also supports changing of categories as well as hiding individual color-coded category groups. For the single mode to be supported, the dataset has to contain at least one unique variable to which at least one quantitative variable can be mapped. Users can enable Single mode using the Single Value toggle button from the toolbar. In the single mode, dropdown controls for both x and y axes are supported. The dropdown control for the x axis lists unique values, while the control for the y axis lists quantitative variables. Additionally, bars can be sorted in ascending order using the Sort toggle button in the toolbar. In both modes, when a bar is hovered over, a tooltip is displayed with the name of its quantitative variable, color-coded category and the numerical value. Using the first toggle button in the toolbar, *Horizontal*, users can switch between horizontal and vertical bar layout.

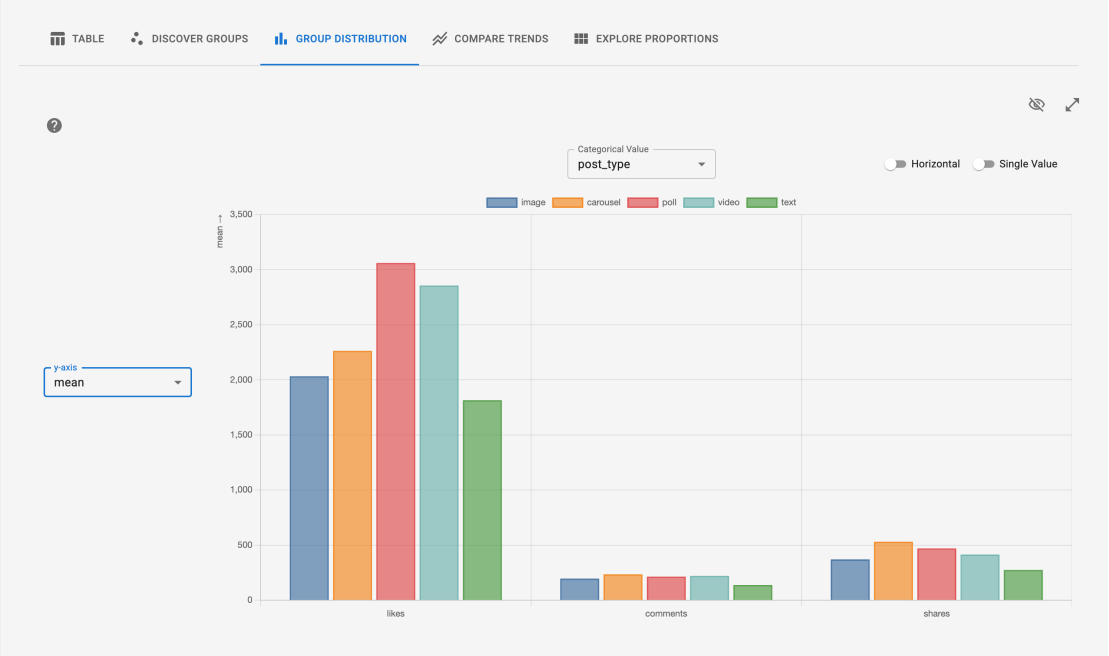

Figure 5.6.: Bar Chart in default mode, depicting derived mean values for individual categories

5.2.6. Line Chart

Named *Compare Trends* in the Tab views, the line chart view is a customized ChartJS Line Chart component. The component requires the dataset to contain at least one quantitative and one temporal variable. Since categorical variables are not relevant, a

slight variation in comparison to the previous charts is that color-coded lines represent quantitative variables. These are accordingly displayed in the legend and can be hidden in the same fashion as in the previous components. In the view, lines connect discrete points that represent the intersection of a temporal variable on the x axis and a numerical variable on the y axis. Similarly to the scatter plot, when the user hovers with a cursor over a point, it increases in circumference and a tooltip is displayed pointing to it. This tooltip displays the current value of the temporal variable, the associated color and value of the quantitative variable and the numerical value. Users have the option to adjust the displayed time interval using the range slider below the view. This can be accomplished by dragging the handles. When hovered over with the cursor, the slider handle displays the current temporal boundary displayed in the view. Dragging one of the handles triggers an update of the x axis and reanimates the line shapes to match the time frame.

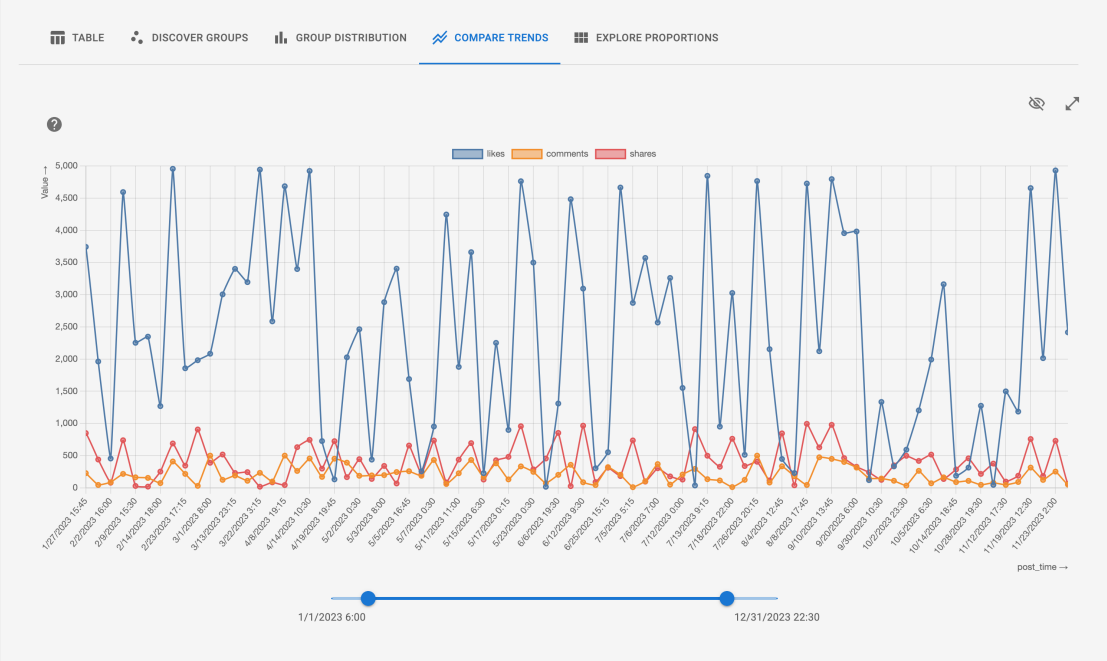

Figure 5.7.: Line Chart with an adjusted temporal slider

5.2.7. Tree Map

The last available visualization, Tree Map, has been built from the ground up using D3. This view requires at least one categorical and one quantitative variable. The visualization spatially encodes the distribution of a numerical variable among categories of a categorical variable. The relationship is depicted using the size of rectangles representing the relative proportion of a category. Each rectangle displays its category label in the upper-left corner. In the lower left corner of the rectangle, the absolute number followed by the

relative percentage in parentheses are displayed. When a rectangle is hovered over, a tooltip appears next to the target rectangle, listing the category, value and quantitative variable. When a subcategory rectangle is selected, the category is written in the format `parent category / subcategory`.

Since the visualization does not use axes, the placement of dropdown controls is different from the established conventions. All controls of the Treemap are located to the left of the view. By default, the visualization uses the first available quantitative variable and the first categorical variable. Starting from the top, the summed up total of the currently selected quantitative variable is displayed. This value represents the reference value distributed among categories. Beneath it, a dropdown control indicates the currently selected category. Moving further below, when there are at least two categorical variables, users can select an optional sub-category in the adjacent dropdown. This triggers subdivision of the rectangles according to the selected subcategory. Since rectangles already have a label, the legend listing color-coding of the parent categories is displayed only when a sub-category is selected. The label of the subdivided rectangles shows the sub-categories. In this case, it does not make sense to make the legend toggleable like in the previous visualizations, so it is not supported. Hiding a category would either cause a blank space in the visualization or the recomputed rectangles would no longer reflect the real distribution of values. Finally, the last dropdown in this side-menu controls the qualitative variable.

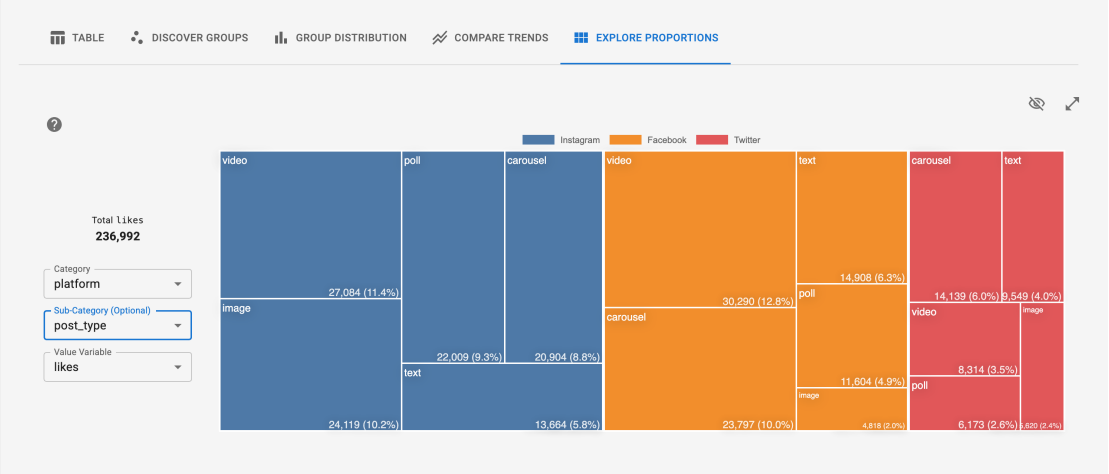

Figure 5.8.: Treemap with nested sub-categories

5.2.8. Dataset Attribution

To credit datasets, users can use multiple mechanisms to reference or properly cite the dataset authorship. Emphasizing the social aspect of the platform, via the share dialog, users have multiple options for how to share the dataset. They can copy the link to the website to their clipboard, they can copy the DOI, or they can share the

dataset via built-in mechanisms of supported social-media platforms. To use the social media buttons, users have to be logged in to their account on their current client. Via the dedicated button for the citing dialog, users can cite the dataset using one of the supported formats. Supported citation formats are APA Style (7th edition), Harvard Style, Vancouver and Dataverse Standard. Although the initial intention for this function was to use a popular library like citation.js[15], to save on the final size of the shipped code bundle, the component was built manually as a standalone React component `Cite.tsx`.

A common feature of both dialogs is the characteristic appearance of DOIs. Also appearing in Key Information and Versioning History, DOIs are embedded within a static badge set up using the public API of shields.io[16]. The service offers a variety of badges and is commonly used by services and online documentation to track various statuses. The advantage of the service beyond its standardized look is that it can be called without requiring authentication. This allows easy embedding in the code without the need to include it as a dependency, reducing the amount of shipped code.

5.2.9. Dataset Description Wizard

The motivation behind the wizard page was manifold. From a purely technical perspective, the wizard is designed to avoid friction and enable users without technical expertise to describe metadata used to source a ViDa showcase via forms. This also includes updating an existing ViDa showcase. Furthermore, the wizard can also serve as an exploratory tool. Data producers don't necessarily need to understand the shape and variable types in their dataset. Built-in dataset analysis will identify variables and their types and suggest supported visualizations. Importantly, this guided collection process should elucidate the meaning of the expected values, nudging the data producer to adopt the data description perspective inherent to the FAIR principles [9].

After several iterations, I consolidated the creation process of the ViDa showcase into seven steps as described in Table 5.2. Steps include evaluation of the guided description, dataset quality indicators, profiling of the dataset and review of components as they would be displayed in the final showcase. After finishing all steps, the wizard concludes with a download of the `vida_config.zip` archive, used as a source for the platform.

As explained in the provided interactive tour reachable from both the initial step and the help dialog, the wizard page layout is comprised of three sections: action menu, step indicator and step content. Starting from the top, the *action menu* offers options to inspect the steps ahead of time, help and global reset of filled-out values. All of the buttons trigger a modal dialog explaining what a given feature does. The first button (*Load Placeholder Values*) allows users to pre-fill all steps with placeholder values, while the neighbouring icon-button activates *overlord mode* in which no values need to be filled in order to proceed to the next step. Below the action menu, sequentially ordered steps are depicted, indicating the progression of the wizard.

The last section is a *step container* rendering content for the current step with common

[15] `https://citation.js.org/`

[16] API service for creating consistent badges `https://shields.io/`

Table 5.2.: Steps in the wizard page

Step	Goal
1. Getting Started	How to use, load existing manifest to update a showcase, example JSON structure, JSON Schema
2. About Author(s)	Author contact information
3. Basic Information	Foundational information (name, funding, license etc.), links to social media, DOI, similar datasets, versioning information
4. Dataset Description	keywords, guided abstract formulation, alternative use-cases
5. Caveats & Considerations	Quality assessment using simplified Data Nutrition Label
6. Upload	profiling of the dataset, description of detected and manually added features
7. Review	Preview of all elements as displayed in the ViDa showcase

control buttons located at the bottom. Users can go back to previous steps to check or correct their answers. Filled-out answers are stored in the browser using the built-in session storage. Data are always stored locally and are only stored remotely in the last step when the platform is running in the cloud mode as described in Subsection 5.4.1. In every step, users can also save a partially filled-out step by clicking a dedicated *save values* button below the form of the current step. Alternatively, users can remove values for the current step via the neighboring button *reset values*.

5.2.10. Dataset Description and Schema

Facilitating interoperability of the ViDa platform, in the initial step, users can inspect, copy and download the example JSON configuration and the JSON Schema used for the validation. Beyond the Getting Started Page, the option to inspect the files in the built-in editor is also available from the help icon button in the Action Menu. I have selected JSON configuration for its flexibility and ubiquitous support, especially in the browser environment. Rather than using a standardized solution, the used schema was incrementally extended to meet the data needs of the showcase as they arose. It is automatically generated based on the TypeScript metadata types defined in `metadataTypes.ts` using an npm script `npm generate-schema` defined in `package.json`.

Following the initial introduction step, the second step *About Author(s)* asks about the author's contact details. The required information is only their name as the author wishes to be displayed in the final showcase and their e-mail. Optionally, the author can add their personal website. There can be an arbitrary number of authors, but at least one.

In the next step *Basic Information*, the wizard prompts the user to input foundational information about the dataset. These include information about both the dataset and its contextual metadata. Although only the dataset name, date of publishing, funding and license are required fields, the page encourages users to enrich this dataset with further information. Among this enriching information are the DOI assigned to the dataset, the paper in which the dataset was first used and the meeting where the dataset was

presented. Associated URLs can be assigned to the paper, funding and meeting fields to promote the trustworthiness of the metadata.

The user is further asked for optional URLs for the associated website and social media pages. In the last two sections of the step, the data producer can link similar datasets and versioning information. Both sections can be disabled using their respective toggle buttons (*No similar datasets*, *No versioning*). In the versioning section, the data producer is prompted for a URL to a Version Control platform page such as GitHub, GitLab, BitBucket and others. Each version can be assigned a unique DOI. If *No versioning* is enabled, the current version becomes v1.0 and the current date is selected. In the fourth step *Dataset Description*, the users are prompted to input keywords, formulate a dataset summary and suggest alternative use-cases or disciplines which could reuse the dataset. Notably, in the abstract section, the questions guide the user to formulate a comprehensive dataset summary using the outline proposed by Koesten and colleagues [27]. Input fields implement the questions in several layers. The input label summarizes the goal of the question in a few words, the helper text asks the question itself and an appended icon-button with a question mark displays further clarifying information including expected values. Ultimately, the user can opt out of this guided description by enabling the toggle button below, titled *I want to use my own description instead.* This provides the user with a single text area to formulate a custom summary.

5.2.11. Caveats and Considerations

In the fifth step of the wizard page, the dataset is evaluated against several quality dimensions and the user is encouraged to list all potential dataset issues. The dimensions are an implementation of a simplified Data Nutrition Label, generating "at-a-glance dataset badges" as introduced in the previous Chapter 4.1.4. The awarded badge can be one of the following types:

- **Ok**: green corner with a thumbs-up icon, generated when all concerns are addressed
- **Warning**: orange corner with an icon of a triangle with an exclamation mark inside, generated when at least one concern has not been addressed
- **Unknown**: same as Warning with a *Not Known* label, generated when the option *not sure* is selected by the user

Evaluation is implemented as a dynamic questionnaire with radio buttons for each question as summarized in Table 5.3. Based on the selected answer, a follow-up field may be displayed. In each evaluated dimension, the answers are either yes, no or unsure. Based on the nature of the question, follow-up questions have only binary selection of radio buttons yes/no. If the response warrants further commentary, additional text input appears for further clarification. Such answers are then displayed in the *To consider* information panel along with other manually added caveats, as illustrated by Figure 5.9. The badge evaluation algorithm is implemented in `Transformer.ts` and the evaluation takes place locally in the browser. The flow of the algorithm is depicted in Figure 5.10.

Table 5.3.: Simplified Data Nutrition Label assessment dimensions

Dimension	Follow Up	Description
About Humans		This dataset contains information about humans
	Sensitive Personal Information	The dataset includes sensitive personal information (e.g., health data, financial data)
	Anonymized	The data has been anonymized/de-identified
Ethical Review		This dataset has been reviewed for ethical consideration
	Consent Obtained	Consent has been obtained for the use of human data.
	Privacy Concerns Addressed	Privacy and ethical concerns were addressed in data collection.
Technical Review		This dataset has been reviewed for technical issues (missing or inconsistent data).
	Unresolved Values	Missing values have been handled appropriately
	Preprocessing	The dataset has undergone preprocessing for quality assurance
Upstream Sources		This dataset contains information from other sources
	Data Sources Verified	Data sources have been verified for quality and reliability
	Diverse Sources	The dataset originates from a diverse range of sources
Update Frequency		This dataset is updated frequently
	Sufficient Update Frequency	The update frequency is sufficient for the intended use case
	Reliance on Outdated Data	The dataset relies on outdated or stale information

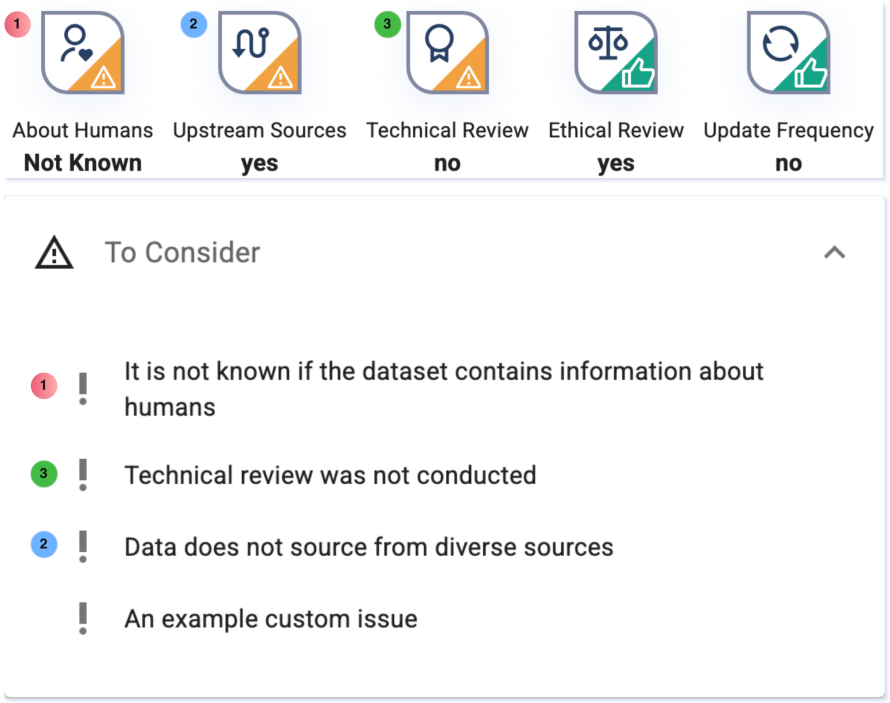

Figure 5.9.: Example of possible badge combinations mapped to justifications in the information panel, (1) Warning issued when the dimension is not known (2) Positive answer does not automatically warrant Ok badge (3) Missing technical review is evaluated negatively

5.2.12. Dataset Upload and Annotation

The next-to-last step in the walkthrough is the data upload. The step informs the user with an alert banner that the expected file format needs to be in CSV (`text/csv`) format, using the comma symbol (`,`) as a separator. It further informs the user that they can help the profiler by annotating dataset headers (variables) with one of the data type tags:

- `<quantitative>`(arithmetic can be performed)
- `<categorical>` (repeating values i.e. categories)
- `<temporal>` (having a time component in a common date format)

By clicking on the *Show Example* button, the user can display an example dataset with both annotated and unannotated headers. Below the example, I also included a few hyperlinks for the user's convenience leading to tutorials on how to transform an existing dataset to CSV using Excel, Libre Office Calc and Stata.

In the first step, the user is asked to select a dataset using a file picker. Picking a dataset displays a preview table with the file name, size and type. At this point, the dataset has not been uploaded yet and can be removed using a trash-icon button.

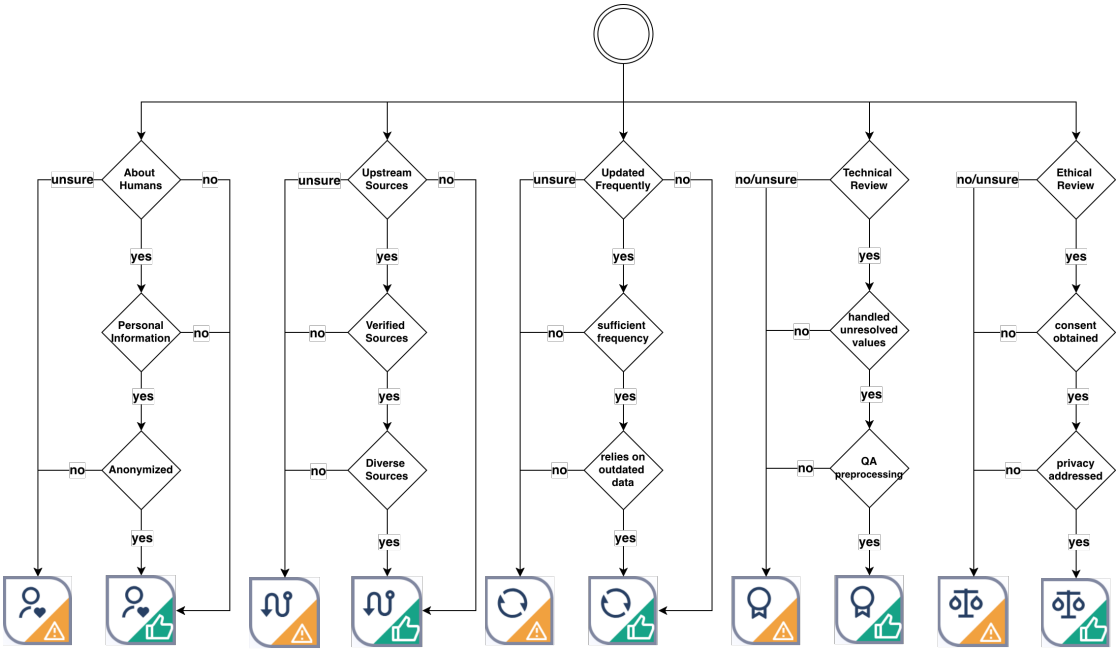

Figure 5.10.: Badge generating algorithm for simplified Data Nutrition Label

When the user clicks the Upload button, the dataset is sent to the transformation-service for profiling, returning a list of detected features and other metadata stored temporarily in the session storage of the browser. If the profiling is successful, a combination of `Feature Name : Feature Description` input pairs for all detected features are enumerated in the *Detected Features* section of the current step. Users are encouraged to give a brief description of the features and even manually add some if some do not display in the detected features. Feature input pairs can also be removed using the trash icon button next to them. This will not remove a feature from the dataset, but merely omits the description in the feature description information panel.

When the dataset is successfully uploaded, users can optionally upload additional supplemental materials. The upload process follows the same steps as in the case of uploading the dataset. The main difference is that the user can pick multiple files. These can include various document types, scripts, images or even other datasets. Ultimately, users on the ViDa home page can choose between downloading the dataset or downloading an archive containing all files uploaded by the dataset author.

5.2.13. Review and Finish

In the last step of the Wizard page, the user can preview all components including visualizations filled with the metadata as collected and stored from the previous step. Components are displayed within roll-out panels giving the user the option to reveal and hide the content. Thanks to the reusable self-contained component approach inherent

to React, the components display exactly as they will be shown on the final showcase page. Below the previews, users have the option to download the JSON metadata used to hydrate the components. This can be reused to prefill the values of the wizard in the future in case the user decides not to finish the process. This can be accomplished by loading this metadata file at the initial *Getting Started* step via the file picker button *Upload Metadata File*.

The user may notice in the last step that the text of the *Next* button is changed to *Finish & Publish*. Clicking the button will redirect the user to the success page, triggering the download of the `vida_config.zip` configuration archive used to power the showcase. On the success page, the user can either go back to the home page[17] or to the page documenting how to run and publish the ViDa showcase.

5.2.14. Interactive Tours

A common feature available on the main page and the wizard page is the help dialog. When opened by clicking on the help icon-button with a question mark symbol, the user has the option to take an animated tour of the platform. Tours are tailored to present features of the current page. In such an interactive tutorial, the library driver.js highlights individual elements by targeting a corresponding DOM element, overlaying the rest of the page with a semi-transparent dark backdrop. This removes distractions, focusing the user's attention on the predetermined element on the page. A pop-over dialog pointing to the highlighted element then provides contextual information. Although it can contain arbitrary HTML content, in the ViDa platform it uses textual descriptions explaining how to use a given feature. The available *next* and *previous* control buttons allow the user to navigate the tour with the progress being displayed in a pagination-like fashion. The tour can be exited by clicking outside of the highlighted element, by clicking on the cross button from the description dialog or by finishing the tour.

5.3. Transformation Service

The main requirement for the transformation service was to effectively handle simultaneous processing of potentially large CSV files while consuming relatively few resources with minimal response time. At the same time, I wanted to avoid overly robust language syntax and complicated development setup. While Python[18] tends to be a common choice among data scientists for its simple syntax and large ecosystem of data processing tools, running in an interpreted language did not meet my performance requirements. I have decided to implement the service in the programming language Go, which routinely outperforms Python in processing speed and memory consumption. Supported by Google, Go achieves its runtime performance by compiling its source code into the binary executable of the target platform. It also has the benefit of lightweight syntax, while demanding verbose error handling. Furthermore, Go's rich standard library allowed me to reduce

[17] Where, depending on the mode, the user can preview the dataset page
[18] Python programming language `https://www.python.org/`

external dependencies to an absolute minimum. The largest dependency is the Gin[19] web framework, implementing REST-like endpoints. An overview of the features available via the endpoints is provided in Table 5.4.

Table 5.4.: Transformation-service API endpoints and their descriptions

Method	Path	Description
GET	`/api/v1/isalive`	Telemetry endpoint, returning availability status in JSON
POST	`/api/v1/dataset`	Uploads, profiles, and stores original and transformed dataset, returning `manifest.json` containing dataset `id`
POST	`/api/v1/supplements/:id`	Uploads and assigns supplements to an existing dataset identified by `id`, returning updated `manifest.json`
PUT	`/api/v1/metadata/:id`	Updates `manifest.json` for a dataset identified by `id`
GET	`/api/v1/dataset/:id`	Retrieves original CSV dataset identified by `id`, stored as `dataset_<hash>.csv`
GET	`/api/v1/transformed/:id`	Retrieves transformed dataset `data.json` identified by `id`, optimized for visualizations
GET	`/api/v1/download/:id`	Retrieves dataset and associated supplements identified by `id` and compressed resources are returned as a zip archive `dataset_bundle.zip`
GET	`/api/v1/folder/:id`	Retrieves all files associated with a dataset identified by `id`, compresses them and returns a zip archive `vida_config.zip`

The transformation service has three high-level use-cases. It (1) processes incoming files, for storing and retrieving them it (2) profiles datasets, and it (3) implements the Azure Blob Storage client. Since the service is stateless, it records all changes to the `manifest` section of the `metadata.json` stored on the client as demonstrated in Figure 5.11. For each obtained file, the service inspects the resource, determines the format, calculates a checksum hash together with a UUID[20] and returns parsed file metadata. Fields also include `upload_date` and `size_in_bytes` records. Uploaded files can either be a dataset or one or more supplements. To prevent naming collisions, the service renames files using a common naming scheme to `dataset_<checksum>.csv` and `supplement_<checksum>.<filetype>` respectively. Specifically, when handling the upload of a dataset, the service generates a `resource_group_id` which organizes all files under a common folder and serves as the identifier for HTTP calls (`id`). The dataset processing flow also adds profiling and conversion steps, resulting in an additional field `row_count`.

[19] Gin web framework for Go `https://gin-gonic.com/`
[20] Universally unique identifier

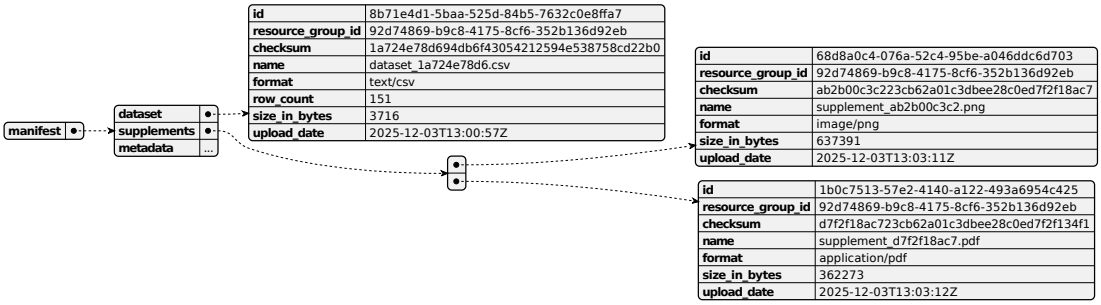

Figure 5.11.: Example JSON structure for the Iris Flower Dataset [1] with PNG image and PDF document supplements

5.3.1. Dataset Profiling

A distinguishing feature of the ViDa platform is its transformation service with the ability to detect variables and heuristically infer their types. The service expects a CSV file with at least two rows, i.e. header and one data row. The profiler selects each header as a variable and analyzes values contained in the corresponding column. During implementation, I did not anticipate ingestion of incomplete datasets. Although datasets with missing values are common, I designed the front-end initially around an idealized mock dataset. Thanks to the employed test-driven approach during development of the service, I encountered this issue early on. For the front-end, it meant simply ignoring missing values, while in the service, the need to track those empty values arose. I first implemented the value inference using a naive approach by sampling the first 100 records (or maximum for smaller datasets) of each column (i.e. variable). Although effective, this approach did not work for datasets that had missing values for the first 100 records or the same value. I then attempted to use a stratified sampling approach where I inspected different locations of the dataset column, e.g. at the start, the middle and towards the end. However, the approach was still too limited and vulnerable to sampling bias. Facing the limitation of the naive approach, I implemented a combination of statistical and HashMap-based approach. This approach analyzes the whole column, eliminating any sampling bias. It divides the profiling into two steps, sampling and subsequent inference. For each column, a `ColumnStats` object is initialized, tracking encountered unique values and their frequency (`UniqueValues`), count of valid non-null/non-empty values (`NonEmptyCount`), numeric values `NumericCount`, and 20 `TemporalSamples` for date/time detection. Based on this aggregate object, the service decides the variable type of the column. The inference algorithm depicted in Figure 5.12 distinguishes between five variable types:

- `Unique`: all non-empty values in the column are unique
- `Quantitative`: all non-empty values in the column can be parsed as 64-bit float
- `Categorical`: all non-empty values in the column contain at least two repeating

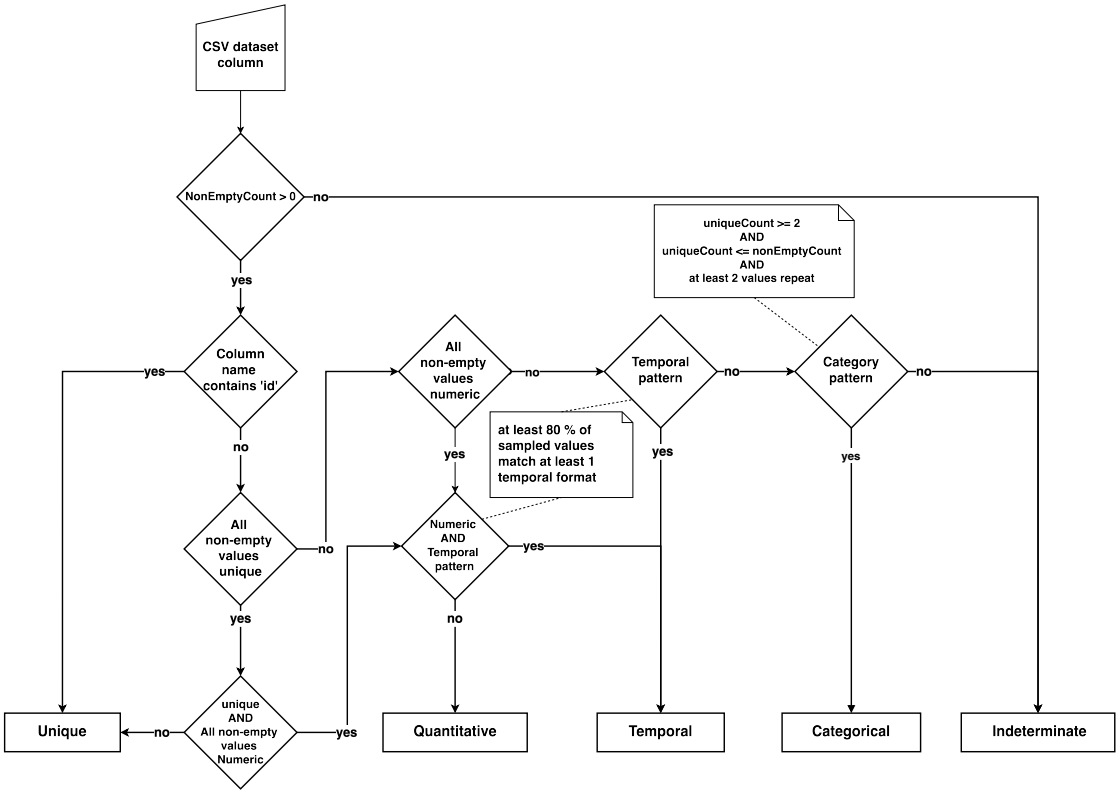

Figure 5.12.: Dataset variable type inference algorithm

non-unique values (categories) and are not of type quantitative.

- `Temporal`: at least 80% of tested non-empty values in the column conform to one of the 47 tested date/time formats
- `Indeterminate`: fallback for when all heuristics fail

As mentioned previously in Subsection 5.2.12, data producers can assist the profiler by annotating header columns with a tag. When the profiler encounters a tag, it skips analysis for the tagged column and categorizes the variable under its type. Based on the dataset's available variable types, the service determines appropriate visualizations as outlined in Table 5.1 in Subsection 5.2.2. Results of this analysis are subsequently stored under the `metadata` key nested within `manifest` of the `metadata.json` (see example in Figure 5.13).

5.3.2. Storage

As the requirements for storing the ViDa resources extended from a remote to a local-based approach, the underlying directory structure has been adapted, resulting in

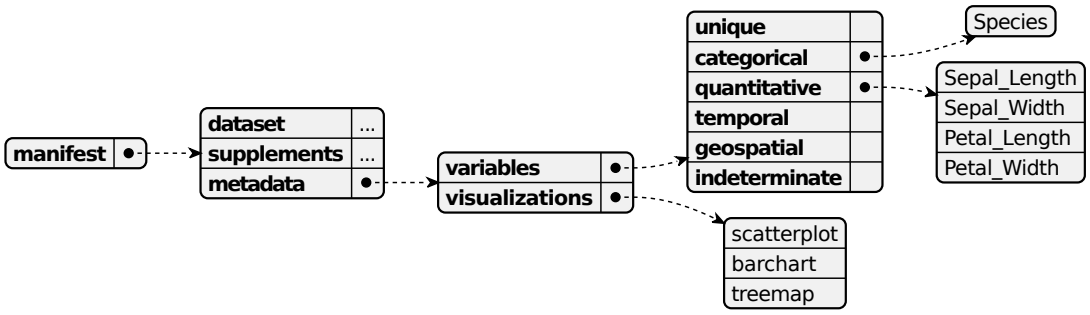

Figure 5.13.: Example JSON structure of inferred variable types and corresponding visualizations for the Iris Flower Dataset [1]

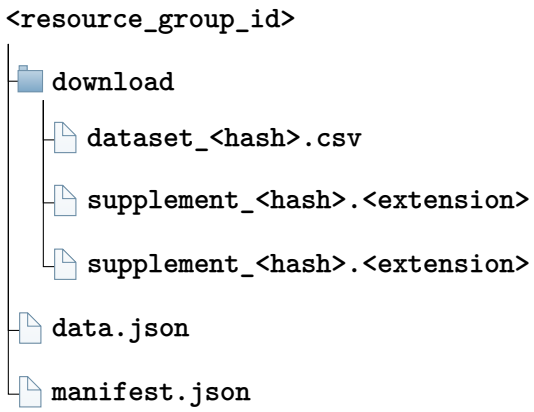

Figure 5.14.: Virtual folder structure of a dataset adopted by `vida_config`.

`vida_config.zip`. The storage is queried on the ViDa page load and during completion of the wizard walkthrough. Since Azure Blob Storage was supposed to be an exclusive storage solution for all resources, the transformation service implements a client with Azure Blob API connectors. In this sense, the transformation service also serves as an intermediary between the front-end and the cloud storage. It defines the folder structure in the target cloud instance. Blob Storage only supports virtual folders, i.e. all resources are located under one common bucket (container) and the names of resources are prefixed with the virtual path interpreted by the storage as a folder. To categorize the resources for a ViDa showcase, I elected to use the prefix `resource_group_id` to signal a common folder. The resulting virtual folder structure of a dataset resembles a tree depicted in Figure 5.14.

One of the major disadvantages when using the Azure infrastructure is challenging Azure portal navigation and service configuration. A user with an Azure account must create a storage account resource. In the resource, the user can create a container, which will be the target location of the stored resources that can be inspected in the Azure

portal. To enable remote access to the container, the user needs to further register an App and authorize it for access and contribution to the created Blob Storage Container. Azure then generates authentication credentials for access, namely `Directory (tenant) ID`, `Application (client) ID` and `Application Secret`. Using the multilingual SDK[21] provided by Azure, users can authenticate against the generated credentials and control the authorized storage container. Importantly, generated credentials have expiration dates (half a year by default) and need to be routinely rotated. The outlined steps for setup and configuration instructions for Azure Blob Storage are documented in detail in the source-code repository in `/transformation-service/README.md`.

With the architectural decision to offer Azure Blob Storage as an optional dependency, there was a need to address storing of transient files during wizard completion. In this mode, resources are stored temporarily in the local file system of the server image. To avoid reimplementation of the existing solution, the local client implements an Adapter Pattern [99], mimicking the Azure client by accepting the same parameters. It produces the same resource structure as the Azure Blob Storage. To avoid pollution of the container's storage, the local client implements a clean-up function, executed when the user downloads `vida_showcase.zip`.

5.4. Deployment

As mentioned in Subsection 5.2.1, on page `/publishing` users can access detailed documentation on how to deploy their ViDa showcase. Further details are available in the source-code repository `/README.md` file. Prerequisites for publishing are a container runtime (such as those provided by Docker or Podman[22]) and `vida_config.zip` obtained by finishing the stepper dialog on the `/wizard` page. Since Docker on macOS requires either Docker Desktop[23] or Colima[24] virtualization, I opted for Podman[25] and set up all helper scripts using it. Coupled with Podman's daemonless architecture, containers tend to be faster to build and to start. The structure of commands is almost identical to that of Docker.

A significant challenge encountered during the build of service images was the different processor architectures. While my local machine features an arm processor, servers generally work with amd architecture. Naturally, published arm images fail to run when deployed to the server. For multi-architectural (multiarch) builds, I implemented automation shell scripts for both front-end and transformation-service. Scripts build images for both architectures, generating a manifest that differentiates between them. It then proceeds to push the images with the manifest to a remote repository. Despite my best efforts, local builds of images targeted to a foreign architecture are both

[21] `https://azure.github.io/azure-sdk/releases/latest/go.html`
[22] open-source solution for managing containers by Red Hat, Podman `https://podman.io/`
[23] `https://www.docker.com/products/docker-desktop/`
[24] `https://github.com/abiosoft/colima`
[25] I alternated between developing on a personal Mac machine and a company-issued Mac. Using Docker Desktop on a laptop owned by a commercial institution requires a commercial license

time-consuming and error-prone. Since the local client needs to leverage an additional virtualization layer, builds for foreign architecture tend to freeze or non-deterministically fail. Facing this limitation, I implemented automated remote image builds for both arm64 and amd64 processor architectures. When attempting to run containers, before triggering a local build, the container runtime attempts to pull the remote image, automatically selecting the appropriate architecture.

Although containers for both services can be run manually, in production, the services are intended to be orchestrated using a YAML `docker-compose` file. The file declaratively describes attributes of services and the `compose` engine builds/runs images, sourcing the build instructions from their respective `Dockerfile` configurations. It also automatically injects required variables via matching environment variables present in the execution environment or ingests variables defined in the `.env` file found in the current folder. Among the attributes, compose files solve network concerns and handle mounting of volumes to the image. This ensures that the `vida_config` can be sourced and used by the front-end. There are multiple modes with different use-cases in which ViDa can run. For each mode, I created a specific `docker-compose` file. I named these files using the schema `docker-compose.<mode>.yaml`. To ease the building and running of the platform, I added convenience scripts `run.sh`, `rerun.sh` and `build.sh`, leveraging the `compose` commands.

5.4.1. Modes

With the architectural decision to offer Azure Blob Storage as an optional dependency and to attempt to make ViDa more flexible, the platform required different modes of operation. To accommodate different deployment settings, I modified the ViDa source code to offer three modes:

- **Standalone:** Serves an existing configuration from `vida_config`, using only the front-end. No new showcases can be created[26].
- **Local:** Serves an existing configuration from `vida_config`, both `front-end` and `transformation-service` are used. Generated configs and processed data stay local and are deleted after they are downloaded.
- **Cloud:** An existing configuration is sourced from Azure Blob Storage with a fallback to local `vida_config`. Both `front-end` and `transformation-service` are used with Azure Blob Storage persistence for processed datasets.

Each mode is associated with its own compose file. Providing a specific flag to one of the convenience scripts selects the corresponding compose file. The ViDa modes, along with their compose files and script flags, are summarized in Table 5.5.

Compose files assume the `vida_config` directory to be available at the current working directory. Using the convenience script, users can define the location of the configuration

[26] Attempting to access the `/wizard` page results in a redirect to the home page

Table 5.5.: Deployment modes and configuration files

Mode	docker-compose file	convenience scripts flags
Standalone	`docker-compose.standalone.yaml`	`-standalone`
Local	`docker-compose.local.yaml`	`-local`
Cloud	`docker-compose.cloud.yaml`	`-cloud`

archive. An example of a user running the ViDa showcase in local mode using a config located in the Desktop directory would therefore have the following shape:

```
./run.sh ~/Desktop/vida_config.zip --local
```

The script checks for installation of Podman, then proceeds to extract the configuration archive to the `./vida_config` directory and calls the compose command:

```
podman compose -f docker-compose.local.yaml up
```

When a user does not have an option to generate the configuration archive online, they can run a minimal version of the platform using the `./generate_config.sh` script and in the browser navigate to the `/wizard` page.

5.4.2. Remote Configuration

In addition to the presented modes, I extended all existing compose files with remote versions following the naming pattern `docker-compose.<mode>.remote.yaml`. These remote versions were a reaction to the deployment environment constraint, where I did not have access to the underlying server file system. To solve this issue, I configured the front-end `Dockerfile` to use a custom script via the `ENTRYPOINT` instruction, known as the *entrypoint wrapper script* [100]. The script checks for `REMOTE_CONFIG_URL`, downloading the `vida_config.zip` archive from the provided remote address. It then unzips the archive and stores the `vida_config` directory in the filesystem of the running container, thereby making the resources available for the Volume mount.

5.4.3. Performance

In both services, I implemented a number of environment-specific optimization techniques, reducing size and runtime demands. To keep the image memory footprints low for both services, I used multi-stage container builds [100]. This pattern divides build steps defined in the `Dockerfile` into multiple stages. In the initial stages, compiling and packaging of the application requires bundling of the complete toolchain and build dependencies. In later stages, only the produced artifact from the previous step is copied to a lightweight runtime image. The building environment from the previous stages can be dropped and the resulting image has a fraction of the build stage size. Another size-reducing measure was reducing the number of dependencies. I established the policy that unless including

an external package is warranted, standardized APIs should be preferred. For instance, on the front-end this meant dropping the Axios[27] library in favor of the standard Fetch API[28]. Specifically in the React ecosystem, the Redux[29] library is a common choice for management of global state. For this use-case, I replaced Redux with React's built-in `createContext` and `useContext` hooks. In the transformation-service, the established dependency policy coupled with multi-stage builds resulted in an executable of 14.4 MB with the size of the final image 22.2 MB.

For low-level optimizations, the front-end benefits from memoization using the `useMemo`[30] React hook. This caching technique is especially impactful for visualizations of large datasets that may otherwise trigger expensive recalculation and re-rendering by an otherwise unrelated event. Furthermore, the components in the front-end benefit from a combination of lazy loading and code splitting provided by NextJS. Resources are loaded on demand as opposed to all at once. For example, dataset views are not all loaded on page load. Rather, appropriate code chunks are dynamically requested when the user clicks on the visualization. This minimizes the amount of code shipped to the browser to the necessary minimum. In the transformation-service, beyond the performance benefits of the binary executable, I leverage Go routines, Go's well-architected concurrency mechanism. Parallel processing enables serving multiple users simultaneously and faster file processing.

5.4.4. Code Management

For the version control of the source code, I used Git. The code is maintained in a monolithic repository, organizing individual services in their own directories. The code is also stored remotely in a repository[31] on the popular platform GitHub[32]. Since I was the sole developer, I opted for a simple two-branch system, `main` and `test` with an occasional feature branch for experiments. In this workflow, I developed the ViDa platform in the `test` branch. When I deemed the changes to be stable, I merged them to the `main` branch. Changes pushed to the `test` branch would not trigger any follow-up processes. However, following a pull request and a subsequent merge of changes to the `main` branch, I leveraged CI/CD[33] features offered by the GitHub platform. During the initial stages, a merge to the `main` branch would trigger an automatic deployment of the front-end to the hosting platform Netlify[34]. Since the hosting platform offers a generous free tier, I used a subdomain of my personal website for online publication. This allowed me to continuously inspect and share developing changes with the project stakeholders.

Netlify does not support hosting of containers. Therefore, in the later stages of the

[27] `https://axios-http.com/`
[28] `https://developer.mozilla.org/en-US/docs/Web/API/Fetch_API`
[29] Global State Management Library Redux `https://redux.js.org/`
[30] `https://react.dev/reference/react/useMemo`
[31] Source-code repository of the ViDa project `https://github.com/Hajneken/data-reuse-showcase`
[32] `https://github.com/`
[33] Acronym for Continuous Integration and Continuous Development
[34] Cloud hosting platform for static sites `https://www.netlify.com/`

project, I configured the merge event to leverage GitHub Actions[35]. The merging now results in a GitHub workflow which builds and stores images for the front-end and transformation-service in the GHCR[36] image registry. Published images are also linked to the repository in the packages section.

5.4.5. Project Management

Choosing GitHub had several advantages beyond code management and deployment functions. From a project management perspective, the platform allows the user to create a project linked to a code repository. In a project, a user can define a Kanban-style board to track tasks as they move between user-defined stages. Each task tile added to the board represents an issue which can be referenced in the Git branch by using an issue identifier.

For the purpose of my project, I created five columns in the project board. The columns are *Backlog*, *Todo*, *In Progress*, *Review* and *Done*, respectively. In the first stage, tasks would first be roughly defined in the backlog and then further refined in *Todo*. When a task was actively worked on, I relocated it to the *In Progress* column. Implemented items ready for testing and feedback were then shifted to the *Review* column. When the feature withstood the scrutiny and was deemed stable, it was moved to the final *Done* stage. This approach allowed me to track the progression of the project and provide precise reports on the status of implemented features. Importantly, it was a valuable psychological guardrail, documenting tangible results when implementation progress was not immediately obvious.

[35] `https://github.com/features/actions`
[36] GitHub Container Registry using `ghcr.io` namespace

6. User Testing

This chapter reports on user testing of the ViDa platform. In designing the study, I prioritized pain points identified in the initial project phases and combined reports of participants with task-based indicators. Findings from the mixed-methods evaluation suggest the platform appears highly usable for the participants under the study conditions. That judgment is cautious as it is tied to the employed scenarios and may shift as the platform evolves or is tried by other user groups.

6.1. Methods

To test the usability of the ViDa platform, user testing mixed with a semi-structured interview was performed. The process is reported using the Consolidated criteria for reporting qualitative research (COREQ) [94] where applicable, and may not cover every procedural detail of this mixed-methods design. User testing was conducted by the author of this work (HZ, male), at the time a graduate student, working as a DevOps engineer. HZ's position as a graduate student and DevOps engineer may have shaped question framing and interpretation. In the study, HZ applied a mixed-methods usability evaluation. Specifically, the design used quantitative task success metrics and qualitative user feedback in semi-structured interviews.

6.1.1. Data Collection

Prior to their participation, participants were asked to review and sign an informed consent sheet, sent via e-mail. The consent form listed instructions for the upcoming user testing session and necessary information about the study. Furthermore, it included important information about participants' rights, including data retention periods, the right to withdraw without any reason, an emphasis on the anonymization of the collected data, how the data will be used, and points of contact. After each session, participants were encouraged to articulate additional thoughts in writing. This post-session reflection seemed valuable for capturing insights that might not surface during the live interview. In my experience, some participants prefer to reflect after the session, which is why written comments were invited.

User testing was conducted online via Zoom in recorded 50-minute sessions held in English. Only the interviewer and the participant were present on the call. Participants were asked to join the call on a desktop or laptop with a modern browser installed. It was piloted with a participant from prior design testing. This person was excluded from the participant sample intentionally as the goal was to test the format. Furthermore,

prior familiarity could influence task performance, which is why the pilot participant was excluded from analysis. Recordings were stored on the private Zoom instance storage at the University of Vienna for a period necessary to download the records of the sessions and deleted right after. Audiovisual material was then transcribed using the transcription feature of Microsoft Word, licensed by the University of Vienna. Transcripts were then corrected for processing errors. Resulting files were not further shared with anyone and were used during analysis to clarify ambiguous parts of the field notes.

During the session HZ reiterated instructions and participants' rights outlined in the consent form and provided an overview of the study and its objectives. It was also mentioned that the user testing was part of the master's thesis. Next, HZ asked participants for verbal consent to record the session. Following the oral consent, HZ proceeded with introductory questions designed to establish demographic qualities of the participant. Afterwards HZ sent a URL[1] address in the chat of the Zoom call. The participant was asked to open the sent URL in a modern web browser and share their screen. The URL led to the deployed version of the ViDa platform presenting a fictitious dataset, modelling user behaviour on various social media platforms. The session proceeded with questions aimed at initial impressions of the platform followed by a mix of tasks independent of the platform and tasks specific to the presented dataset. The session was concluded using closing questions designed to probe for problems, familiarity with other tools and usefulness of the platform.

The interview guide with the sequence of color-coded questions and tasks, their objectives, and success criteria is provided in Appendix A.6. In each session, a duplicate of this template was used to record field notes. Table 6.1 summarizes the phases and question types, along with the shorthand used in question identifiers.

Table 6.1.: User Testing session outline

Short	**Description**
Before	send consent form
INTRO	Introduction, demographics, rapport
Q1	Question about initial impressions after first opening the platform
TG	General Tasks, independent of the dataset, one point per question when success criteria is fulfilled
TS	Specialized Tasks that are specific for the given dataset to test the user interaction with the platform, one point per question when success criteria is fulfilled
Q2	Closing Comments
After	send message with thanks and possibility for reflections

[1]ViDa instance used for the user testing has been deployed in local mode at the university hosting environement managed by Portainer `https://vida.vda.univie.ac.at/`

6.1.2. Usability Metrics

Usability of the platform has been evaluated through two different lenses. In the first case, participants were evaluated based on completion of the task. In the second case, participants reported their subjective rating of task difficulty. For each successfully carried out task, a participant could obtain up to one point (17 in total) depending on fulfillment of the respective success criteria. In cases of partial fulfillment, participants received fractional points (0.5, 0.75). While fractional scoring appears to introduce more precision than may be warranted by the task definitions, it offers a practical way to capture partial completion.

Points were distributed among two categories, General (TG, max six points) and Dataset-specific tasks (TS, max 11 points). An additional third category VIS (a subset of TS tasks) was introduced to evaluate the operation of the visualizations (7 points). In this category, tasks were designed with an emphasis on forcing users to use all the available tools for given visualization.

All four metrics were combined to produce final **Usability Score** US, defined as the weighted average of the four average success scores T, TG, TS, VIS

$$\text{US} = \frac{w_T T + w_{TG} TG + w_{TS} TS + w_{VIS} VIS}{w_T + w_{TG} + w_{TS} + w_{\text{VIS}}}$$

where:

- T = Average Total Success Score (%)
- TG = Average General Task Success Score (%)
- TS = Average Specific Task Success Score (%)
- VIS = Average Visualization Success Score (%)

and $w_T, w_{TG}, w_{TS}, w_{VIS} > 0$ are the respective weights representing maximal score for each success score, with $\text{US} \in [0; 100]$

Finally, to report the perceived difficulty of each task, participants were asked to choose on a simple discrete scale (easy, medium, difficult). Answers were counted for each question and assigned to the respective frequency category. The most frequent answer was then calculated for each task as the mode of the respective answers. The resulting value was then compared to the mean success score for the given question.

6.1.3. Sample

Convenience sampling was used to recruit initial participants from the author's professional network. Snowball sampling was then employed as participants referred colleagues who met the inclusion criteria. Inclusion criteria included individuals who had at least started undergraduate education and who could effectively communicate in English. An emphasis was put on selecting different professions including academic researchers and professionals familiar and competent with visualizations (data scientists, actuarial scientists, IT

professionals) as well as professionals without visualization training (marketing, product management, law practitioners). This was to ensure that not only people with visualization competency but also a lay audience could operate the platform.

The majority of participants were approached directly face-to-face, followed by an emailed consent form to record an upcoming session, with a description of the study expectations and participants' rights. Three participants were approached via e-mail only. Recruited participants were familiar only with the field of expertise of the interviewer and the general task they would be expected to do during the session. All participants successfully completed their sessions in a timely manner. Participants were deliberately not told details about the study in advance and were asked not to further communicate any details about the process to others until the study was concluded.

Final participant sample summarized in the table 6.2 included 10 professionals from various disciplines, four males and six females. Participants were 26-43 years old ($\bar{x}_{age} = 31.2$, $s = 6.29$). Data collection took place between October 1 and October 10, 2025.

Table 6.2.: User Testing Participant demographics

Participant ID	**sex**	**Age**	**Education**	**Proffesion**
P#1:10-01-2025	female	26	graduate	Academia
P#2:10-02-2025	female	26	undergraduate	IT
P#3:10-03-2025	male	27	graduate	IT
P#4:10-06-2025	female	31	graduate	Marketing
P#5:10-06-2025	female	43	undergraduate	Project Management
P#6:10-08-2025	female	41	graduate	Project Management
P#7:10-08-2025	male	33	undergraduate	Law
P#8:10-11-2025	female	27	undergraduate	Finance
P#9:10-12-2025	male	26	graduate	Finance
P#10:10-13-2025	male	32	graduate	IT

6.2. Results

For the open-ended part of the sessions, multiple participants described the interface as *"clean"*, and *"intuitive"*, having *"not too many information"*, being *"not overwhelming"* and *"easy to use"*, suggesting that the problematic cognitive load from the design phase has been appropriatly addressed.

Majority of participants have reported experience with an array of different data processing and visualization platforms tailored for their given profession. All participants have mentioned excel and similar spreadsheet data processesors. Among IT professionals, platform such as Grafana, various python libraries, powerbi were a typical responses. One IT professional has further mentioned Tableu and Matlab as his tools of choice for data visualization. Participant from academia frequently, mentioned r language. In terms of the visual evaluations, participants have acknowledged a modern appearance of

the platform and self-explanatory functions of the menu buttons. On the contrary, one participant has reported that they believe the appearance of the platform to be outdated. This divergence appears to reflect different aesthetic expectations and may suggest that visual preferences vary with prior exposure to analytical tools.

Interestingly, in the closing comments one participant mentioned that they missed a search field. However, upon further probing the participant could not pinpoint what function the search would fulfill. This may point to a general expectation that data platforms include search, rather than a clear need within the specific workflows tested.

Overall participants have achieved high Total Score ($\bar{x} = 15.575$, $s = 1.3335$). As expected, participants with the highest scores were individuals with previous experience working with datasets and visualizations. Two of those participants achieved the maximum score. The data suggest the opposite trend as well, where the lowest scores were obtained by participants without previous experience working with datasets. While this pattern is plausible, it is also possible that familiarity with testing situations or general digital literacy contributed to the gap.

Calculated high Usability Score **92.13%** supports participants' closing comments that the platform was intuitive to use. Participants achieved higher success scores with dataset-specific tasks (TS) than in general tasks (TG). This pattern may indicate that the data views and their affordances were clearer than some of the cross-cutting platform operations.

Although still high, TG tasks with the lowest scores were TG-22 (82.5%) and TG-07 (80%), designed for the users to create their own dataset showcase and to download all available files (dataset and supplemental files) at once respectively. One explanation for lower scores in the former case might be that participants were not primed prior to their session about the nature of the project, making it unexpected that there was an option for generating their own showcase. The tasks required participants to explore the interface and consult the help section. Ultimately, time pressure may have contributed to fewer exploratory endeavors, resulting in some participants giving up on the task. In the latter case, the order of the questions may have influenced the score as every participant closed the Download Modal Dialog after completing the preceding task, designed to test if users could download the dataset only. This preceding question, in contrast, had a maximal success score. Suspicion that the task was too easy might have paradoxically caused participants to search elsewhere rather than in the place where they had just been. Out of the four categories included in the calculation of the Usability Score, the highest success rate was in VIS at 94.64%. Category result aligns with *VIS* tasks being subset of the TS tasks, further suggesting that interactive views designed to visualize and explore presented data are intelligible.

6.2.1. Percieved Difficulty of Tasks

The most frequent answer reported by the participants was *"easy"*. This could point to an intuitive design as reported at the beginning, though response bias or task familiarity may also play a role. The easiest tasks were reported to be TS-11 and TS-19, receiving a unanimous *"easy"* rating. Supported by the high success score in the VIS category,

Table 6.3.: Scores summary

	Success	$\bar{x}$	s	w
Total Score	91.62%	15.575	1.333593725	17
TG	87.92%	5.275	0.9010025281	6
TS	93.64%	10.3	0.5502524673	11
Vis	94.64%	6.625	0.4602233757	7
Usability Score	**92.13%**			

these tasks dealt with visualization views, specifically the table and the line chart. It is important to note, however, that familiarity with tables and line charts may have influenced responses, independent of interface quality.

Self-reported difficulty largely aligned with observed performance (Success Score), although a few items with medium mode ratings (for example, TS-08 and the previously discussed TG-22) still showed relatively high success, which may suggest that participants experienced brief uncertainty that they could resolve during the task. Specifically, in the latter case, the task was to identify the author(s) of the dataset and their institution. Though the authors were found almost immediately, the institution search took the participants considerably longer. In one case, participant has misattributed the dataset the author of the platform by searching in the footer. This hints at the belief verbalized by some participants that seperation between the platform and the presented dataset is not immediately obvious. However, in the observations, several participants hesitated briefly before proceeding, then resolved the step without assistance. This pattern appears consistent with an interface that can be learned during use, even if initial attempts are not always immediate.

In the results, we can observe a slight disconnect between perceived difficulty and task success for the aforementioned TG-07 (downloading) and TG-14. The latter task asked participants to find proper attribution in a scenario where they had decided to use the dataset. Although the correct approach was to use the *cite this dataset* dialogue, some users came up with alternative valid approaches, such as copying the URL of the page or referencing the article in the journal in which the dataset first appeared. Though valid, the alternative attribution approaches were neither thorough nor formally compliant in the scientific sense and were awarded partial points. The partial-credit outcomes might indicate that users prioritize practical attribution outside academic contexts, a priority that does not always align with formal citation standards. Along similar lines, limited awareness of proper attribution and citation standards outside academia might ultimately also contribute to this disconnect.

Relatively lower success rates in VIS tasks TS-16 and TS-18 appear to tell a story of learnability that emerges with exposure. Both tasks have suffered from partial point awarding. In TS-16, participants were asked to locate a specific data point in a scatter plot. Although every participant eventually found it, the target was intentionally difficult to spot amid surrounding points. To eliminate guesswork, the intended strategy was

to filter irrelevant subgroups by clicking their labels. The feature may not have been immediately discoverable. Once participants became aware of this option, they transferred the strategy to other visualizations. This pattern suggests that earlier exposure to any filtering-enabled visualization might have produced a similarly high success score on this task. In TS-18 participants were required to report the highest value in a bar chart. Without sorting feature enabled, the task was more laborious, though still achievable. When further probed whether the participants could have made their search easier, majority have identified the feature to be helpful and sensible. Once again, time pressure may have discouraged exploration and reduced the likelihood of consulting the help section when in doubt. Finally, in task TS-19, a recurring comment surfaced among four participants. The task asked users to restrict the time window with a slider. Even though it was rated as easy and reached a perfect success rate, the participants who raised the issue said they would rather use a date picker, which they described as a more elegant and more precise tool. This may suggest that the perceived fit of the chosen control (slider) does not always track measured performance.

Table 6.4.: Task perceived difficulty and success

Task	**easy**	**medium**	**difficult**	**Mod**	**Success**
TG-06	9	1	0	easy	100.00%
TG-07	8	0	2	easy	80.00%
TS-08	2	6	1	medium	85.00%
TG-09	7	3	0	easy	95.00%
TS-10	9	1	0	easy	100.00%
TS-11	10	0	0	easy	100.00%
TG-12	9	1	0	easy	90.00%
TS-13	7	1	1	easy	85.00%
TG-14	8	1	0	easy	80.00%
TS-15	6	3	1	easy	97.50%
TS-16	7	3	0	easy	82.50%
TS-17	9	1	0	easy	97.50%
TS-18	9	1	0	easy	82.50%
TS-19	10	0	0	easy	100.00%
TS-20	9	1	0	easy	100.00%
TS-21	9	1	0	easy	100.00%
TG-22	4	5	0	medium	82.50%

7. Discussion

It should be reiterated that the theoretical section of this work draws extensively on the research conducted by the thesis co-supervisors [7, 13, 65, 80, 101, 20, 61, 27]. Along similar lines, an explicit feature checklist used for the survey of the data repositories was adapted from their work [13].

Findings from the survey show that investigated platforms offer diverse but unevenly distributed set of functionalities that support dataset reuse. Not all evaluated features were relevant or applicable to all repositories. At the same time, higher scores should not necessarily be viewed as a measure of an innovative platform. Still, the observed average Feature Coverage Index score of 45.087 provides a signal that improvement potential remains. While basic discovery and access features appear universal, sophisticated features supporting deeper understanding and confident reuse remain concentrated in a minority of platforms. Kaggle (66.458), Open Science Framework (56.309) and Harvard Dataverse (55.625) achieved the highest FCI scores. On the other end of the spectrum, Google Dataset Search and Gesis both achieved the lowest score of 33.185. While most repositories implement spreadsheet-style data explorers, automatic statistical summaries and visualization remain limited. Auctus Dataset Search and Harvard Dataverse stand out by automatically identifying key variables and featuring visualization from derived data. The results also indicate that data quality assessing remains underdeveloped. Platforms treat such indicators as either optional or not taking such indicators into account at all. Only three platforms (Kaggle, European Data Portal, Auctus Dataset Search) explicitly communicate data quality issues such as missing values or empty fields. Statista appears to be the sole repository that issues explicit warnings akin to the *To Consider* panel implemented by ViDa. Though platforms appear to have matured in the context of foundational discovery features, future development should prioritize explicit quality indicators and comparative analytics.

Informed by the interfaces used in practice, I combined surveyed elements into low-fidelity prototypes, the evolution of which I describe in Chapter 4. The main effort was to develop a familiar interface resembling popular data processing tools. I further supported this familiarity by incorporating Material Design elements. In the follow-up interviews, participants did not favor any design in particular. Rather, they labeled elements they preferred from both designs. Recurring choices and well-justified preferences were implemented into a merged final design. Participants rejected both the comment section and ratings features. What they did value was a prominent visualization alongside contextual information, which they saw as crucial for establishing credibility. Their unique perspectives and backgrounds also provided insights into desired features of the envisioned platform. One unexpected instance was the demand to significantly enlarge the download button and simplify the download process as much as possible. Another

important factor was the adjustment and simplification of the feature labels. Sessions revealed that misunderstanding of a word such as *"caveat"* led to users losing interest in the concerned feature of the product.

Implementation of ViDa based on the designs was a long iterative process that evolved into a platform exceeding the original vision. Interestingly, the idea to include a guided metadata curation was conceptualized in the later stages as a secondary concern. However, in combination with the profiling algorithm and simplified data nutrition label, it ultimately became a distinguishing feature of the platform. In terms of the architecture, the component-based front-end and decoupled backend also have longer-term implications. Visualization views can be easily extended without extensive redesign, and the backend can be repurposed for alternative interfaces. Established dataset visualization ecosystem specifically ecourages extension. The implemented but not used map chart, supported by geospatial metadata fields in the manifest, illustrates how new visualizations can be readily integrated. Flexible implementation of the profiler algorithm in the transformation-service also provide opportunities to infer additional specialized data types, which would in turn support new visualization options. Containerized deployment is further simplified through convenience scripts and comprehensive documentation, reducing barriers for potential adopters.

To assess whether implemented design choices translated into usability, I structured the user testing to capture both impressions and task performance. Participants reported on their experience while completing a series of tasks that fell into two categories: general tasks applicable across any dataset, and those specific to the particular data presented. This dual approach allowed me to distinguish between platform-level usability and context-dependent challenges. I also tracked visualization interactions separately, recognizing that the effectiveness of individual charts and graphs warrants independent assessment. The results appear encouraging. A Usability Score of 92.13% as defined in Subsection 6.1.2 aligned with how participants rated individual tasks. Qualitative feedback reinforced this pattern. Participant comments indicated satisfaction with the user interface, though some also voiced specific friction points. Overall, these findings suggest the platform is likely to be intuitive and functional under the tested conditions.

7.1. Declaration of Generative AI and AI-assisted Technologies During the Writing Process

During the preparation of this thesis, I used ChatGPT 4o and Writefull[1] models via plugins provided by Overleaf[2] as auxiliary tools. I have used the models to check grammar and improve the readability of the text. The models did not replace my independent academic work. Furthermore, I used translation service DeepL[3] to translate the Abstract from English to German. I used all disclosed tools in accordance with the guidelines of the University of Vienna [102] and the accompanying principles of good academic

[1] `https://www.writefull.com/`
[2] LaTeX editor environment `https://www.overleaf.com/`
[3] `https://www.deepl.com/en`

practice. I critically reviewed and edited the content and take full responsibility for the content of the thesis.

7.2. Limitations

While ViDa may occupy a relatively novel position among open-source visualization tools, numerous opportunities for refinement remain. Similarly, it is important to recognize the limitations of the underlying research. The reliance of the DRP survey on established repositories and portals may have introduced sampling biases. The qualitative interviews involved only a small sample of participants, the coding approach deviated from established qualitative methods and user testing has only explored usability from the reuser's perspective.

7.2.1. Survey Sampling

Although the findings of the survey of DRPs should not be read as generalizable, the sample selection process introduces potential biases that warrant acknowledgment. First, the reliance on recommendations of the supervisory team and literature citations, while pragmatic, likely privileged repositories with greater visibility and established track records. This preference for well-resourced repositories may have underrepresented or excluded smaller, emerging initiatives that operate under different constraints. The underrepresentation of commercial platforms (only one out of fourteen) means that findings may not transfer to the private sector. Domain representation is somewhat limited as well. Most repositories in the survey serve general purposes rather than being discipline-specific. The dominance of tabular datasets in the surveyed repositories suggests that repositories handling other data types may have been underrepresented. Specialist fields such as astrophysics remain largely absent from the analysis. Taken together, these patterns suggest our findings reflect best practices primarily suited to established, general-purpose, and well-resourced repositories, rather than offering comprehensive insight into the broader landscape of dataset repositories and portals.

7.2.2. Interviews

Several limitations related to the qualitative interviews should be noted. The small sample size ($n = 4$), while appropriate for exploratory design evaluation, limits the transferability of the findings. Although ongoing supervisory consultations still offered some analytical feedback, relying on a single coder, along with not conducting member checking, limits the possibility of study's validation. Furthermore, the restrictive sampling criterion (data producers) and recruitment through professional networks may have introduced selection bias. Moreover, participants were not asked to review transcripts or validate findings (member checking), which constitutes a limitation regarding the validation by participants. Finally, as interviews performed a supplementary design-validation role within a primarily software-focused thesis, the scope for in-depth qualitative investigation was inevitably limited.

7.2.3. Findability

According to the FAIR principles, the primary requirement for data reuse is that datasets must be easily findable [9]. Online discoverability is fundamentally connected to how search engines index content and present it as organic search results to users. Although NextJS provides several SEO[4] features, ViDa loads metadata dynamically from configuration files at runtime, which complicates crawler interpretation. Search engine crawlers may struggle to load or correctly interpret such dynamically sourced metadata. Although the SEO score in Chrome Lighthouse[5] is 100 %, citing properly implemented technical prerequisites, ViDa currently does not provide machine-readable context for the dataset. This shortcoming could be alleviated in the future by parsing structured metadata[6] ahead of time, injecting parsed metadata into the `meta` element in the `header` of the shipped HTML code. Open Graph tags would complement this approach, enabling customized previews when URLs are shared across social platforms.

7.2.4. Dashboard Features

In the early stages, I planned to implement a choropleth map visualization as one of the core visualization types, since it is easily understood by a broad audience. Although a prepared component implemented using D3 is available in the code base under `MapChart.tsx` and the `geospatial` property is part of the introduced JSON specification of profiled data (Figure 5.13), practical limitations ultimately made implementation untenable. First, I did not find implementation of a reliable heuristic for automatic inference of geospatial data. Moreover, spatial data inherently depend on geographical boundaries, which vary in scope and shape. Supporting diverse geographical scales would demand multiple map implementations, each calibrated to different levels of political and administrative division. This ambiguity quickly expands the number of use cases that need to be assessed or alternatively limits the number of datasets that can be visualized. Requiring users to match their data to predefined geographic boundaries would contradict the platform's core principle that no technical or metadata curation is required. Given these constraints, the required effort outweighed the benefits, and I ultimately prioritized other aspects of the platform.

During the testing sessions, users revealed a consistent frustration with the line chart interaction. While the current slider allows continuous range adjustment, participants overwhelmingly favored discrete date selection instead. Rather than replacing the slider, a hybrid approach could be implemented. A date picker would explicitly select the temporal value, while the slider would describe the context by indicating the current temporal resolution.

Datasets are rarely provided alone, isolated of other accompanying resources. PDFs,

[4] Search Engine Optimization

[5] Audit tool built into Chrome Developer Tools `https://developer.chrome.com/docs/lighthouse`

[6] Standardized code (typically JSON-LD) embedded in HTML to provide search engines with explicit context about page content `https://developers.google.com/search/docs/appearance/structured-data/intro-structured-data`

images, and other supplemental materials often accompany them. Several data repositories offer built-in previews for these materials, supporting both raster and vector formats. ViDa currently prioritizes dataset visualization over supplemental materials, leaving such resources outside its scope. Rather, it focuses on visualization of the dataset itself. Integrating a previewer for images and documents might reduce friction for reusers, allowing them to assess supplemental materials before deciding to download.

In terms of downloadable files, when a user wants to download resources, they must consent to the agreement first:

> *"I agree to use the dataset in accordance with the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0) license. I acknowledge that I must provide appropriate credit, link to the license, and indicate if changes were made."*

This agreement is currently limited to the linked CC BY 4.0 license[7]. This constraint stems from the metadata workflow, which accepts freeform license information rather than enforcing a predefined selection. Implementing a standardized license registry would enable the system to enforce license-specific restrictions at the point of download. In case of more restrictive licenses, the download could be restricted or disabled entirely. Importantly, such restrictions would need to need to extend beyond the main download interface. The Table view currently allows users to export filtered or customized data, which would also require license enforcement.

A limitation of the metadata section in the implemented dashboard is that the dataset versioning information panel does not contain release notes. Currently, the panel lists the version number, publication date, and DOI link. Adding optional release notes would allow data producers to document what changed in each iteration, helping reusers understand the evolution of a dataset.

7.2.5. Performance and Memory Testing

Although I tested ViDa using various workloads, the platform would benefit from systematic stress testing. The performance was assessed by monitoring CPU and memory consumption of containers on a local machine and on the server. In both cases, only a single user scenario was tested, processing and downloading datasets of various sizes (33 kB to 90 MB). Accordingly, the recommended scenario would be to simulate multiple users attempting to generate a showcase. Observing the system's behavior under load may reveal unexpected bottlenecks.

In terms of memory consumption, I recognize that server-side caching of the wizard page may pose a memory problem. In Local mode, the service stores processed resources in the temporary filesystem of the image, mimicking the directory structure that would otherwise be created in Azure Blob Storage. At the moment, these resources are removed when `vida_config.zip` is successfully downloaded. This may potentially lead to the image running out of storage space, as incomplete sessions accumulate. To alleviate

[7] `https://creativecommons.org/licenses/by/4.0/`

this problem, a potential solution would be to modify the transformation-service by establishing a time limit for wizard completion. While session timeouts would prevent storage buildup, they don't fully address the underlying issue. Local deployments may lack the capacity for sustained multi-user activity. In such scenarios, using cloud mode is recommended, as Azure Blob Storage offloads storage from the server.

7.2.6. User Testing

Although the platform targets both data producers and reusers, testing was focused exclusively on the latter group. While rooted in literature, the guided data curation in the wizard page remains untested. Consequently, future research should explore the perceived usefulness of guided curation. It is unclear whether users would favor the stepper form over their customized generation workflows. A comparative study examining whether the guided approach in ViDa yields more complete or consistent metadata than self-directed curation could be informative. Additionally, discipline-specific testing might also reveal domain metadata needs that the current guide overlooks.

To maintain comparable conditions, all participants encountered the same dataset and tasks. Although high average success scores suggest the platform is usable under the study conditions, the results may not reflect real-world usage. There is a risk that the tasks may have been too easy, potentially overstating actual usability. Future testing could vary the datasets and task difficulty intentionally, allowing to observe whether performance differences emerge or whether average scores remain consistent.

The sample featured professionals from various backgrounds as listed in Table 6.2. However, only one participant could be categorized as lay audience, having minimal visualization experience, as they did not require it professionally. This skew towards visualization-literate participants may mask usability barriers for less experienced users. Notably, this participant achieved the lowest score. It would be worth investigating whether similar individuals with comparable experience would encounter similar difficulties.

Lastly, although participants occasionally voiced sentiments of perceived platform usefulness, user testing examined this aspect only marginally. The task-based approach measured task completion, not satisfaction or perceived value. To move beyond task-based metrics, a longitudinal deployment tracking usage through embedded analytics and periodic user feedback would provide more informative data about perceived usefulness.

7.3. Lessons Learned

As described in Chapter 4.2, following the interviews of the low-fidelity prototypes I analyzed resulting transcriptions through coding. To assist with this process, I attempted to use MAXQDA's AI-assisted coding feature. Unfortunately, despite the recent leaps with large language models, AI tools in the used coding software provided limited, often hallucinated results. In hindsight, since the context of those transcriptions is large, the likely culprit of the unsatisfactory results is *"context rot"*, degrading performance as

the input grows [103]. Facing reduced efficiency during the coding stage, I ultimately abandoned LLM-assisted coding results and completed the analysis manually.

During deployment, unexpected constraints and platform-specific frictions have driven innovative solutions. Constraints in the deployment environment has forced me to eventually discover the entrypoint script pattern [100], illuminating how to leverage container for initialization jobs. Similarily, non-ideal progression of development of the visualization discussed in Subsection 5.2.4 triggered a search for existing well-tested solution. In this respect, not having to reeinvent the wheel by not addressing solved problems may accelerate the development velocity.

During development, the temptation toward premature optimization led to the inclusion of presumptive features [104]. The problem was that I have started optimizing for a scaled up solution without proper analysis of the trade offs. These tradeoffs became apparent to me only after the development has began. Although guided by the KISS philosophy when doing the architecture I have completely neglected on another similar concept. Coming from agile sofware development, violation of YAGNI[8] principle by building presumptive features introduces delay costs [105]. Concretely, we talk about Azure Blob Storage. Though time spent on architecting the solution gave roots to the eventual structure of the configuration archive, hours of integration effort could have been spent on local solution which in the long run seem to have better chances at reuse than the other. The lesson here is that good intention can exponentially increase the complexity of the software. Ultimately, it felt like going backwards, pruning functionality to arrive at the standalone mode, resembling a static website. At the same time I have to acknowledge that it it difficult to judge if I would have come to the same structure should this process never took place.

7.4. Future Work

Beyond the solutions to the shortcomings described in the previous sections, there are multiple promising directions in which the platform can evolve. As discussed in subsection 7.2.4, implementation of choropleth maps represent a natural extension. Implementation should primarily focus on macro scale (world and continents), with clear guidance for mapping data to geographical regions. Additionally, I envision the addition of a feature that would allow users to upload custom map shapes in GeoJSON. This would allow visualizations targeted to their specific geographic regions.

In terms of metadata, the checksum property (hash value) in the manifest describing uploaded files (see Figure 5.11) is currently not utilized. General practice involves using cryptographic hash algorithms to evaluate file integrity. Cryptographic verification would be a reasonable step towards strengthening of the platform security. Alternatively, the platform could present hash values for individual items in the download section, leaving it to users to verify integrity themselves.

Data quality indicators present a trade-off. The platform currently implements a simplified Data Nutrition Label rather than the full framework. Although this approach

[8]Acronym for *"You Aren't Gonna Need It"*

is more time efficient when describing the dataset, it risks missing nuanced quality issues. For example, in the simplified version, unknown properties are flagged as warnings. Adding an optional full analysis would preserve simplicity and enable complete assessment for interested users.

Recognizing that not everyone wants to store data in Azure Blob Storage, the project would benefit from options to store configuration files elsewhere. Introducing an additional mode that provides integration with various storage solutions may boost adoption. However a more fundamental concern may inhibit platform's appeal. ViDa currently requires an underlying hosting environment to support container runtime. While containerization eliminates certain deployment challenges, this layer may complicate integration with traditional web infrastructure and repository ecosystems. The `standalone` mode may offer a pathway forward. Extending it to static HTML would eliminate the container requirement entirely. Although such an approach would reduce flexibility in dataset updates, it would radically simplify deployment. Resources contained in `vida_config` could be made available through the `public` directory. Deployed as a static website would offer multiple benefits, including no back-end attack surface, easier institutional hosting, and potential integration as a repository plugin. The main limitation of this approach would be the inability to provide the metadata curation interface tied to the back-end logic.

7.5. Conclusion

Guided by theories of data reuse and a unique survey of 14 dataset repositories and portals, this work translated conceptual and empirical insights into a series of low-fidelity interface prototypes. Although it was not possible to apply the criteria uniformly to every repository, the analysis showed that surveyed instances leave considerable room for improvement, particularly in data quality indicators. Through multiple iterations and subsequent semi-structured interviews with data producers, I refined the designs and incorporated targeted changes. Collectively, these steps culminated in ViDa, a dataset visualization platform that represents the primary contribution of the thesis.

Although the landscape of data visualization tools continues to evolve, to my knowledge, no comparable open-source solution currently exists. Rather than pursuing advanced technical features, ViDa prioritizes ease of use for non-technical users. The platform eases documentation through guided metadata workflows and automated dataset profiling. It then generates dashboards with interactive visualizations without requiring users to possess technical skills or knowledge of visualization libraries. These dashboards are further augmented with author-provided metadata, resulting in a structured overview of each dataset. In practice, this approach should mitigate prevalent obstacles to data exploration and reuse.

Conducted user testing suggests that the interface meets core usability objectives. Participants found it intuitive and successfully completed visualization tasks. However, these encouraging results warrant careful interpretation. Tests involving lay audiences were largely underrepresented. Moreover, the extent to which guided workflows truly

reduce documentation effort, and whether the visualizations meaningfully support reuse decisions, remains to be demonstrated. Future work is likely to benefit from longitudinal deployments, integration with existing repositories, and closer examination of how quality indicators are understood. This thesis proposes an approach to translating theoretical understanding of data reuse into practical tooling. I recognize that broader adoption and critical feedback will ultimately determine its value.

Bibliography

[1] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[2] B. Hanson, A. Sugden, and B. Alberts, "Making Data Maximally Available," *Science*, vol. 331, no. 6018, pp. 649–649, Feb. 2011. [Online]. Available: https://www.science.org/doi/10.1126/science.1203354

[3] P. Aylin, A. Yunus, A. Bottle, A. Majeed, and D. Bell, "Weekend mortality for emergency admissions. A large, multicentre study," *Quality & Safety in Health Care*, vol. 19, no. 3, pp. 213–217, Jun. 2010.

[4] K. Sielemann, A. Hafner, and B. Pucker, "The reuse of public datasets in the life sciences: potential risks and rewards," *PeerJ*, vol. 8, p. e9954, Sep. 2020. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7518187/

[5] M. Saito, R. McGready, H. Tinto, T. Rouamba, D. Mosha, S. Rulisa, S. Kariuki, M. Desai, C. Manyando, E. M. Njunju, E. Sevene, A. Vala, O. Augusto, C. Clerk, E. Were, S. Mrema, W. Kisinza, J. Byamugisha, M. Kagawa, J. Singlovic, M. Yore, A. M. v. Eijk, U. Mehta, A. Stergachis, J. Hill, K. Stepniewska, M. Gomes, P. J. Guérin, F. Nosten, F. O. t. Kuile, and S. Dellicour, "Pregnancy outcomes after first-trimester treatment with artemisinin derivatives versus non-artemisinin antimalarials: a systematic review and individual patient data meta-analysis," *The Lancet*, vol. 401, no. 10371, pp. 118–130, Jan. 2023. [Online]. Available: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(22)01881-5/fulltext

[6] N. Waithira, M. Mukaka, E. Kestelyn, K. Chotthanawathit, D. N. Thi Phuong, H. N. Thanh, A. Osterrieder, T. Lang, and P. Y. Cheah, "Data sharing and reuse in clinical research: Are we there yet? A cross-sectional study on progress, challenges and opportunities in LMICs," *PLOS Global Public Health*, vol. 4, no. 11, p. e0003392, Nov. 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11578489/

[7] L. Koesten and E. Simperl, "UX of data: making data available doesn't make it usable," *Interactions*, vol. 28, no. 2, pp. 97–99, Mar. 2021. [Online]. Available: https://doi.org/10.1145/3448888

[8] L. Peer, A. Green, and E. Stephenson, "Committing to Data Quality Review," *International Journal of Digital Curation*, vol. 9, no. 1, pp. 263–291, Jun. 2014. [Online]. Available: https://ijdc.net/index.php/ijdc/article/view/9.1.263

[9] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, p. 160018, Mar. 2016, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/sdata201618

[10] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski, "The dataset nutrition label: A framework to drive higher data quality standards," 2018. [Online]. Available: https://arxiv.org/abs/1805.03677

[11] R. Baskerville, "What design science is not," *European Journal of Information Systems*, vol. 17, no. 5, pp. 441–443, 2008, _eprint: https://doi.org/10.1057/ejis.2008.45. [Online]. Available: https://doi.org/10.1057/ejis.2008.45

[12] P. Johannesson and E. Perjons, *An Introduction to Design Science*, ser. Computer Science. Springer International Publishing, Sep. 2021, google-Books-ID: nRpEEAAAQBAJ. [Online]. Available: https://books.google.at/books?id=nRpEEAAAQBAJ

[13] K. Gregory and L. Koesten, *Human-Centered Data Discovery*, ser. Synthesis Lectures on Information Concepts, Retrieval, and Services. Cham: Springer International Publishing, 2022. [Online]. Available: https://link.springer.com/10.1007/978-3-031-18223-5

[14] A. J. G. Hey, S. Tansley, and K. M. Tolle, *The Fourth Paradigm: Data-intensive Scientific Discovery*. Microsoft Research, 2009, google-Books-ID: oGs_AQAAIAAJ.

[15] C. L. Borgman, "The conundrum of sharing research data," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 6, pp. 1059–1078, 2012. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22634

[16] L. R. Johnston, R. Curty, S. M. Braxton, J. Carlson, H. Hadley, S. Lafferty-Hess, H. Luong, J. L. Petters, and W. A. Kozlowski, "Understanding the value of curation: A survey of US data repository curation practices and perceptions," *PLOS ONE*, vol. 19, no. 6, pp. 1–23, Jun. 2024. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0301171

[17] A. Goodman, A. Pepe, A. W. Blocker, C. L. Borgman, K. Cranmer, M. Crosas, R. D. Stefano, Y. Gil, P. Groth, M. Hedstrom, D. W. Hogg, V. Kashyap, A. Mahabal, A. Siemiginowska, and A. Slavkovic, "Ten Simple Rules for the Care and Feeding of Scientific Data," *PLOS Computational Biology*, vol. 10, no. 4, p. e1003542, Apr. 2014, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003542

[18] R. Walshe, "Introduction to the Special Issue on: Big Data/AI Standardization in the Journal of ICT Standardization," *Journal of ICT Standardization*, pp. 1–1, Apr. 2020. [Online]. Available: https://journals.riverpublishers.com/index.php/JICTS

[19] A. Hey, S. Tansley, and K. Tolle, "Jim Gray on eScience: a transformed scientific method," *The Fourth Paradigm*, 2009. [Online]. Available: https://www.semanticscholar.org/paper/Jim-Gray-on-eScience%3A-a-transformed-scientific-Hey-Tansley/80fa526c43dbb2fcbf725dcfd579cb214788624e

[20] L. Koesten, P. Vougiouklis, E. Simperl, and P. Groth, "Dataset Reuse: Toward Translating Principles to Practice," *Patterns*, vol. 1, no. 8, p. 100136, Nov. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666389920301847

[21] D. G. Roche, L. E. B. Kruuk, R. Lanfear, and S. A. Binning, "Public Data Archiving in Ecology and Evolution: How Well Are We Doing?" *PLOS Biology*, vol. 13, no. 11, p. e1002295, Nov. 2015. [Online]. Available: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002295

[22] D.-G. for Research and Innovation (European Commission), *Turning FAIR into reality: final report and action plan from the European Commission expert group on FAIR data.* LU: Publications Office of the European Union, 2018. [Online]. Available: https://data.europa.eu/doi/10.2777/1524

[23] J. L. Rosenbloom, "Sharing research data: researcher behaviour and attitudes," *Science and Public Policy*, p. scaf041, Aug. 2025. [Online]. Available: https://doi.org/10.1093/scipol/scaf041

[24] M. J. Eppler and J. Mengis, "The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines," *The Information Society*, vol. 20, no. 5, pp. 325–344, Nov. 2004. [Online]. Available: https://doi.org/10.1080/01972240490507974

[25] M. Muller, I. Lange, D. Wang, D. Piorkowski, J. Tsay, Q. V. Liao, C. Dugan, and T. Erickson, "How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–15. [Online]. Available: https://doi.org/10.1145/3290605.3300356

[26] D. Kern and B. Mathiak, "Are There Any Differences in Data Set Retrieval Compared to Well-Known Literature Retrieval?" in *Research and Advanced Technology for Digital Libraries*, ser. Lecture Notes in Computer Science, S. Kapidakis, C. Mazurek, and M. Werla, Eds. Cham: Springer International Publishing, 2015, pp. 197–208.

[27] L. Koesten, E. Simperl, T. Blount, E. Kacprzak, and J. Tennison, "Everything you always wanted to know about a dataset: Studies in data summarisation," *International Journal of Human-Computer Studies*, vol. 135, p. 102367, Mar. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581918306153

[28] B. Fecher, S. Friesike, and M. Hebing, "What Drives Academic Data Sharing?" *PLoS ONE*, vol. 10, no. 2, p. e0118053, Feb. 2015. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC4340811/

[29] N. R. Council, *Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources*. Washington, D.C.: National Academies Press, Apr. 1995. [Online]. Available: http://www.nap.edu/catalog/4871

[30] M. K. Buckland, "Information as thing," *Journal of the American Society for Information Science*, vol. 42, no. 5, pp. 351–360, 1991. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199106%2942%3A5%3C351%3A%3AAID-ASI5%3E3.0.CO%3B2-3

[31] C. L. Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World*. The MIT Press, 2015. [Online]. Available: https://www.jstor.org/stable/j.ctt17kk8n8

[32] A. Martínez and S. Mammola, "Specialized terminology reduces the number of citations of scientific papers," *Proceedings of the Royal Society B: Biological Sciences*, vol. 288, no. 1948, p. 20202581, Apr. 2021. [Online]. Available: https://doi.org/10.1098/rspb.2020.2581

[33] N. S. Board, "Long-lived digital data dollections: Enabling research and education in the 21st century (No. NSB0540)," 2005. [Online]. Available: https://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

[34] S. Leonelli, "What Counts as Scientific Data? A Relational Framework," *Philosophy of Science*, vol. 82, no. 5, pp. 810–821, Dec. 2015, publisher: Cambridge University Press. [Online]. Available: https://www.cambridge.org/core/journals/philosophy-of-science/article/abs/what-counts-as-scientific-data-a-relational-framework/33A05C7F71958D1FEC37AE89C2866A72

[35] A. Chapman, E. Simperl, L. Koesten, G. Konstantinidis, L.-D. Ibáñez, E. Kacprzak, and P. Groth, "Dataset search: a survey," *The VLDB*

Journal, vol. 29, no. 1, pp. 251–272, Jan. 2020. [Online]. Available: https://doi.org/10.1007/s00778-019-00564-x

[36] F. Löffler, V. Wesp, B. König-Ries, and F. Klan, "Dataset search in biodiversity research: Do metadata in data repositories reflect scholarly information needs?" *PLOS ONE*, vol. 16, no. 3, p. e0246099, Mar. 2021, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0246099

[37] A. H. Renear, S. Sacchi, and K. M. Wickett, "Definitions of dataset in the scientific and technical literature," *Proceedings of the American Society for Information Science and Technology*, vol. 47, no. 1, pp. 1–4, 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/meet.14504701240

[38] T. Munzner, *Visualization Analysis and Design*. New York: A K Peters/CRC Press, Oct. 2014.

[39] R. M. Losee, "Browsing mixed structured and unstructured data," *Information Processing & Management*, vol. 42, no. 2, pp. 440–452, Mar. 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457305000130

[40] L. Koesten, "A User Centred Perspective on Structured Data Discovery," in *Companion Proceedings of the The Web Conference 2018*, ser. WWW '18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, Apr. 2018, pp. 849–853. [Online]. Available: https://dl.acm.org/doi/10.1145/3184558.3186574

[41] J. Pomerantz, *Metadata*. The MIT Press, Nov. 2015. [Online]. Available: https://direct.mit.edu/books/book/3111/Metadata

[42] J. Williams, "Introduction to Metadata," *Collections*, vol. 13, no. 1, pp. 47–55, Mar. 2017, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.1177/155019061701300104

[43] J. Gheel and T. Anderson, "Data and metadata for finding and reminding," in *1999 IEEE International Conference on Information Visualization (Cat. No. PR00210)*, Jul. 1999, pp. 446–451, iSSN: 1093-9547.

[44] G. C. Bowker and S. L. Star, *Sorting Things Out: Classification and Its Consequences*. The MIT Press, Aug. 2000. [Online]. Available: https://direct.mit.edu/books/book/4738/Sorting-Things-OutClassification-and-Its

[45] Luis-Daniel Ibáñez, Emilia Kacprzak, Laura Koesten, and Elena Simperl, "Analytical Report 18: Characterising Dataset Search on the European Data Portal | data.europa.eu," *European Data Portal*, Sep. 2020. [Online]. Available: https://data.europa.eu/en/publications/datastories/analytical-report-18-characterising-dataset-search-european-data-portal

[46] National Institutes of Health, Office of Data Science Strategy, Data Management Center of Excellence, "Workbook of Metadata Fundamentals: The What, Why, and How of Metadata," 2024. [Online]. Available: https://datascience.nih.gov/sites/default/files/2025-06/workbook_of_metadata_fundamentals_v1.1_10-25-2024_508.pdf

[47] I. M. Faniel, R. D. Frank, and E. Yakel, "Context from the data reuser's point of view," *Journal of Documentation*, vol. 75, no. 6, pp. 1274–1297, Sep. 2019. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/JD-08-2018-0133/full/html

[48] K. M. Gregory, H. Cousijn, P. Groth, A. Scharnhorst, and S. Wyatt, "Understanding data search as a socio-technical practice," *Journal of Information Science*, vol. 46, no. 4, pp. 459–475, Aug. 2020. [Online]. Available: https://doi.org/10.1177/0165551519837182

[49] B. Dervin, "Given a context by any other name: Methodological tools for taming the unruly beast," in *Information seeking in context*. Taylor Graham, 1997, pp. 13 – 38.

[50] P. Dourish, "What we talk about when we talk about context," *Personal and Ubiquitous Computing*, vol. 8, no. 1, pp. 19–30, Feb. 2004. [Online]. Available: https://doi.org/10.1007/s00779-003-0253-8

[51] N. K. Agarwal, *Exploring Context in Information Behavior: Seeker, Situation, Surroundings, and Shared Identities*, ser. Synthesis Lectures on Information Concepts, Retrieval, and Services. Cham: Springer International Publishing, 2018. [Online]. Available: https://link.springer.com/10.1007/978-3-031-02313-2

[52] "PROV-Overview: An Overview of the PROV Family of Documents." [Online]. Available: https://www.w3.org/TR/prov-overview/

[53] C. Tenopir, N. M. Rice, S. Allard, L. Baird, J. Borycz, L. Christian, B. Grant, R. Olendorf, and R. J. Sandusky, "Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide," *PLOS ONE*, vol. 15, no. 3, p. e0229003, Mar. 2020. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0229003

[54] N. R. Council, *Bits of Power: Issues in Global Access to Scientific Data*. Washington, D.C.: National Academies Press, Aug. 1997. [Online]. Available: http://www.nap.edu/catalog/5504

[55] UNESCO and Canadian Commission for UNESCO, *An introduction to the UNESCO Recommendation on Open Science*. UNESCO, Nov. 2022. [Online]. Available: http://dx.doi.org/10.54677/XOIR1696

[56] I. Faniel and E. Yakel, "Practices Do Not Make Perfect: Disciplinary Data Sharing and Reuse Practices and Their Implications for Repository Data Curation," in *Curating Research Data*. Association of College and Research Libraries, a division of the American Library Association, Jan. 2017, pp. 103–126.

[57] M. Mayernik, "Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators," Rochester, NY, Jun. 2011. [Online]. Available: https://papers.ssrn.com/abstract=2042653

[58] C. L. Borgman, "Data Citation as a Bibliometric Oxymoron," in *Theories of Informetrics and Scholarly Communication*, C. R. Sugimoto, Ed. Berlin, Boston: De Gruyter Saur, 2016, pp. 93–116. [Online]. Available: https://doi.org/10.1515/9783110308464-008

[59] R. S. Taylor, "Question-Negotiation and Information Seeking in Libraries," *College & Research Libraries*, vol. 29, no. 3, pp. 178 – 194, 1968.

[60] I. V. Pasquetto, C. L. Borgman, and M. F. Wofford, "Uses and Reuses of Scientific Data: The Data Creators' Advantage," *Harvard Data Science Review*, vol. 1, no. 2, Nov. 2019. [Online]. Available: https://hdsr.mitpress.mit.edu/pub/jduhd7og/release/10

[61] L. Koesten, K. Gregory, P. Groth, and E. Simperl, "Talking datasets – Understanding data sensemaking behaviours," *International Journal of Human-Computer Studies*, vol. 146, p. 102562, Feb. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581920301646

[62] T. Krämer, A. Papenmeier, Z. Carevic, D. Kern, and B. Mathiak, "Data-Seeking Behaviour in the Social Sciences," *International Journal on Digital Libraries*, vol. 22, no. 2, pp. 175–195, Jun. 2021. [Online]. Available: https://doi.org/10.1007/s00799-021-00303-0

[63] F. Xiao, R. Ma, and D. He, "Task-based human-structured research data interaction: A discipline independent examination," *Proceedings of the Association for Information Science and Technology*, vol. 57, no. 1, p. e308, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/pra2.308

[64] T. Friedrich, "Looking for data: nformation seeking behaviour of survey data users," doctoralThesis, Humboldt-Universität zu Berlin, Nov. 2020, accepted: 2020-11-30T12:16:51Z. [Online]. Available: https://edoc.hu-berlin.de/handle/18452/22813

[65] K. Gregory, P. Groth, A. Scharnhorst, and S. Wyatt, "Lost or Found? Discovering Data Needed for Research," *Harvard Data Science Review*, Apr. 2020. [Online]. Available: https://hdsr.mitpress.mit.edu/pub/gw3r97ht/release/6

[66] K. M. Fear, "_USMeasuring and Anticipating the Impact of Data Reuse." Thesis, University of Michigan, 2013, accepted: 2014-01-16T20:41:59Z. [Online]. Available: http://deepblue.lib.umich.edu/handle/2027.42/102481

[67] S. v. d. Sandt, S. Dallmeier-Tiessen, A. Lavasa, and V. Petras, "The Definition of Reuse," *Data Science Journal*, vol. 18, no. 1, p. 22, Jun. 2019, number: 1 Publisher: Ubiquity Press. [Online]. Available: http://datascience.codata.org/articles/10.5334/dsj-2019-022/

[68] Gartner, "Data and Analytics: Everything You Need to Know." [Online]. Available: https://www.gartner.com/en/topics/data-and-analytics

[69] P. McDermott, "Building open government," *Government Information Quarterly*, vol. 27, no. 4, pp. 401–413, Oct. 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0740624X10000663

[70] J. P. Birnholtz and M. J. Bietz, "Data at work: supporting sharing in science and engineering," in *Proceedings of the 2003 ACM International Conference on Supporting Group Work*, ser. GROUP '03. New York, NY, USA: Association for Computing Machinery, Nov. 2003, pp. 339–348. [Online]. Available: https://doi.org/10.1145/958160.958215

[71] C. A. C. Lee, "A framework for contextual information in digital collections," *Journal of Documentation*, vol. 67, no. 1, pp. 95–143, Jan. 2011, publisher: Emerald Group Publishing Limited. [Online]. Available: https://doi.org/10.1108/00220411111105470

[72] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," *Commun. ACM*, vol. 39, no. 11, pp. 86–95, Nov. 1996. [Online]. Available: https://dl.acm.org/doi/10.1145/240455.240479

[73] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 16:1–16:52, Jul. 2009. [Online]. Available: https://dl.acm.org/doi/10.1145/1541880.1541883

[74] J. Wang, Y. Liu, P. Li, Z. Lin, S. Sindakis, and S. Aggarwal, "Overview of Data Quality: Examining the Dimensions, Antecedents, and Impacts of Data Quality," *Journal of the Knowledge Economy*, pp. 1–20, Feb. 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9912223/

[75] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L. O. B. da Silva Santos, and M. D. Wilkinson, "Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud," *Information Services and Use*, vol. 37, no. 1, pp. 49–56, Feb. 2017. [Online]. Available: https://doi.org/10.3233/ISU-170824

[76] M. Boeckhout, G. A. Zielhuis, and A. L. Bredenoord, "The FAIR guiding principles for data stewardship: fair enough?" *European Journal of Human Genetics*, vol. 26, no. 7, pp. 931–936, Jul. 2018. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6018669/

[77] H. Shanahan and L. Bezuidenhout, "Rethinking the A in FAIR Data: Issues of Data Access and Accessibility in Research," *Frontiers in Research Metrics and Analytics*, vol. 7, p. 912456, Jul. 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9363602/

[78] T. Wilson, "Models in information behaviour research," *Journal of Documentation*, vol. 55, no. 3, pp. 249–270, Jan. 1999, publisher: MCB UP Ltd. [Online]. Available: https://doi.org/10.1108/EUM0000000007145

[79] Donald O. Case and Lisa M. Given, *Looking for Information : A Survey of Research on Information Seeking, Needs, and Behavior*, ser. Studies in Information. Bingley, UK: Emerald Group Publishing Limited, 2016, vol. Fourth edition. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1427926&site=ehost-live

[80] K. Gregory, "Findable and reusable?: Data discovery practices in research," Doctoral Thesis, Maastricht University, Maastricht, 2021, iSBN: 9789464231304.

[81] L. M. Koesten, E. Kacprzak, J. F. A. Tennison, and E. Simperl, "The Trials and Tribulations of Working with Structured Data: -a Study on Information Seeking Behaviour," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: Association for Computing Machinery, May 2017, pp. 1277–1289. [Online]. Available: https://dl.acm.org/doi/10.1145/3025453.3025838

[82] E. Kacprzak, L. M. Koesten, L.-D. Ibáñez, E. Simperl, and J. Tennison, "A Query Log Analysis of Dataset Search," in *Web Engineering*, ser. Lecture Notes in Computer Science, J. Cabot, R. De Virgilio, and R. Torlone, Eds. Cham: Springer International Publishing, 2017, pp. 429–436.

[83] G. Silvello, "Theory and practice of data citation," *Journal of the Association for Information Science and Technology*, vol. 69, no. 1, pp. 6–20, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23917

[84] P. Zhang and D. Soergel, "Cognitive mechanisms in sensemaking: A qualitative user study," *Journal of the Association for Information Science and Technology*, vol. 71, no. 2, pp. 158–171, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24221

[85] J. W. Tukey, *Exploratory Data Analysis.* Addison-Wesley Publishing Company, 1977.

[86] G. W. Furnas and D. M. Russell, "Making sense of sensemaking," in *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '05. New York, NY, USA: Association for Computing Machinery, Apr. 2005, pp. 2115–2116. [Online]. Available: https://dl.acm.org/doi/10.1145/1056808.1057113

[87] Y.-a. Kang and J. Stasko, "Examining the Use of a Visual Analytics System for Sensemaking Tasks: Case Studies with Domain Experts," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2869–2878, Dec. 2012. [Online]. Available: https://ieeexplore.ieee.org/document/6327293

[88] G. Marchionini, S. Haas, J. Zhang, and J. Elsas, "Accessing government statistical information," *Computer*, vol. 38, no. 12, pp. 52–61, Dec. 2005, conference Name: Computer.

[89] J. Baker, D. Jones, and J. Burkman, "Using visual representations of data to enhance sensemaking in data exploration tasks," *Journal of the Association for Information Systems*, vol. 10, no. 7, p. 2, 2009.

[90] G. Marchionini, "Exploratory search: from finding to understanding," *Commun. ACM*, vol. 49, no. 4, pp. 41–46, Apr. 2006. [Online]. Available: https://dl.acm.org/doi/10.1145/1121949.1121979

[91] J. C. Roberts, C. Headleand, and P. D. Ritsos, "Sketching designs using the five design-sheet methodology," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 419–428, Jan 2016.

[92] J. Yew, G. Convertino, A. Hamilton, and E. Churchill, "Design systems: A community case study," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–8. [Online]. Available: https://doi.org/10.1145/3334480.3375204

[93] F. Bentley, L. Schmidt, M. Gilbert, J. Feldman, and M. Pielot, "Expressive Design: Google's UX Research." [Online]. Available: https://design.google/library/expressive-material-design-google-research

[94] A. Tong, P. Sainsbury, and J. Craig, "Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups," *International Journal for Quality in Health Care*, vol. 19, no. 6, pp. 349–357, Sep. 2007, _eprint: https://academic.oup.com/intqhc/article-pdf/19/6/349/5070113/mzm042.pdf. [Online]. Available: https://doi.org/10.1093/intqhc/mzm042

[95] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, pp. 77–101, Jan. 2006.

[96] P. Avgeriou and U. Zdun, "Architectural Patterns Revisited - A Pattern," in *Proceedings of the 10th European Conference on Pattern Languages of Programs, 2005*, ser. 10th European Conference on Pattern Languages of Programs (EuroPlop 2005), Irsee, A. Longshaw and U. Zdun, Eds. UVK Verlagsgesellschaft, 2005, pp. 1–39.

[97] E. Evans, *Domain-driven Design: Tackling Complexity in the Heart of Software*. Addison-Wesley Professional, 2004, google-Books-ID: xColAAPGubgC.

[98] S. L. Franconeri, L. M. Padilla, P. Shah, J. M. Zacks, and J. Hullman, "The Science of Visual Data Communication: What Works," *Psychological Science in the Public Interest*, vol. 22, no. 3, pp. 110–161, Dec. 2021. [Online]. Available: https://doi.org/10.1177/15291006211051956

[99] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*. Pearson Education, Oct. 1994, google-Books-ID: 6oHuKQe3TjQC.

[100] N. Poulton, *Docker Deep Dive: Zero to Docker in a Single Book*. Packt Publishing Ltd, Jul. 2023, google-Books-ID: tJnMEAAAQBAJ.

[101] K. Gregory, P. Groth, H. Cousijn, A. Scharnhorst, and S. Wyatt, "Searching Data: A Review of Observational Data Retrieval Practices in Selected Disciplines," *Journal of the Association for Information Science and Technology*, vol. 70, no. 5, pp. 419–432, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24165

[102] University of Vienna, "AI in Academic Writing," University of Vienna. [Online]. Available: https://studieren.univie.ac.at/en/using-ai-in-your-studies/academic-work-and-writing-with-ai/ai-in-academic-writing/

[103] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the Middle: How Language Models Use Long Contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, Feb. 2024, _eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00638/2336043/tacl_a_00638.pdf. [Online]. Available: https://doi.org/10.1162/tacl_a_00638

[104] R. Kohavi, T. Crook, R. Longbotham, B. Frasca, R. Henne, J. L. Ferres, T. Melamed, and J. M. L. Ferres, "Online Experimentation at Microsoft," Jul. 2009. [Online]. Available: https://www.microsoft.com/en-us/research/publication/online-experimentation-at-microsoft/

[105] M. Fowler, "Yagni," May 2015. [Online]. Available: https://martinfowler.com/bliki/Yagni.html

A. Appendix

A.1. Survey of Data Repositories and Portals

DNF = Did Not Find

		#	1	2	3	4	5	6	7	8	9	10	11
		Cluster	UX Generally					Social					
Name	Url	Feature	(i) search (ii) SERP (iii) Detail	Data curation guide / methodology	(landed on detail) where am I	Support for external systems (API)	Voice assistants	Social Network Engagement	Dataset Community Rating	Dataset Reviews	Community-driven documentation	Chat	Colli
Gesis	https://search.gesis.org/		☑ no data previews	☑ methodology guide	☑ does not link to external repositories	☐ DNF	☐ DNF	☑ simple Twitter, Facebook and LinkedIn Share Buttons in footer on the main page	☐ DNF	☐ DNF	☐ documentation by institute	☐ DNF	☐
London Data Store	https://data.london.gov.uk		☑ data previews only in some results	☐ DNF	☑ some results link to external websites	☑ DataPress API (REST) supports GET POST PATCH	☐ DNF	☑ Latest Twitter feed on the homepage, @LDN_Data	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☐
Dataset Search	https://datasetsearch.res		☑ SERP integrates with Detail (split screen, results left, detail right)	☑ simple guide for providers; reference to DataSet developer guide	☐ only links to external repositories/websites	☐ DNF	☐ DNF	☑ social media share dialog for each dataset	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☐
ICPSR	https://www.icpsr.umich.e		☑ standard	☑ Guide to Social Science Data Preparation and Archiving; characteristics of qualitative data	☑ provides breadcrum navigation, some results link to external websites	☐ DNF	☐ DNF	☑ simple Facebook, Instagram, Twitter, LinkedIn and YouTube links leading to institute's respective social pages; social media share dialog for each dataset	☐ DNF	☐ DNF	☐ documentation by institute	☐ DNF	☐
Kaggle	https://www.kaggle.com/c		☑ standard	☑ Community driven tutorials, getting started documentation	☑ provides breadcrum navigation, some results link to external websites	☑ python CLI tool available on GitHub	☐ DNF	☑ social media share dialog for each dataset	☑ Usability score calculated by Completeness, Credibility and Score; Textual feedback from users under each dataset, keyword-based feedback	☑ Textual feedback from users under each dataset, keyword-based feedback	☐ documentation by institute	☐ DNF	☐
GenBank	https://www.ncbi.nlm.nih.g		☑ standard	☑ submission/format guidelines and procedure	☑ breadcrumbs navigation, page headers	☑ FTP servers, NCBI e-utilities, Entrez	☐ DNF	☑ simple Facebook, Instagram, Twitter, LinkedIn, Github links leading to institute's respective social pages	☐ DNF	☐ DNF	☐ documentation by institute	☐ DNF	☐
Inspire	https://inspirehep.net/		☑ standard, data specific search in BETA	☑ Technical Guidance for the implementation of INSPIRE dataset and service metadata based on ISO/TS 19139:2007	☑ page headers, links to external repository Hepdata	☑ INSPIRE REST API metadata/standard docs, arXiv, ORCID	☐ DNF	☑ social media sharing options in detail page	☐ DNF	☐ DNF	☐ system documentation only	☐ DNF	☐
Open Science Framework	https://osf.io/		☑ standard	☑ detailed guides for Organizing Data	☑ project dashboards, breadcrumbs, OSF header	☑ guide for how to integrate RESTFul OpenAPI, integration with multiple external tools via Add-Ons (GitHub, Dropbox, Google Drive), OAuth2 authentication	☐ DNF	☑ social media sharing options for Open Science social accounts, visible on the main site	☑ supports commenting on projects and datasets	☑ supports commenting on projects and datasets	☑ wiki-like project pages where collaborators can collaboratively edit documentation	☐ DNF	☐
European Data Portal	https://data.europa.eu/en		☑ standard	☑ Open Data Academy and links to external resources	☑ breadcrumbs navigation, multi-step navigation	☑ hub-search with OpenAPI specification, SPARQL search	☐ DNF	☑ simple Facebook, Instagram, Twitter, LinkedIn, Github links leading to institute's respective social pages; links for sharing datasets via social networks	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☐
Figshare	https://figshare.com/		☑ standard	☑ Figshare user guides and FAIR guides	☑ page headers	☑ Figshare OpenAPI and OAI-PMH protocol	☐ DNF	☑ In the footer simple facebook, twitter and vimeo links leading to insitute's social pages , share widget in the detail page	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☐
Harvard Dataverse	https://dataverse.harvard.		☑ standard	☑ documented curation services, multi-level curation guides and documentation	☑ breadcrumbs navigation, page headers	☑ DataCite REST API, integrations to GitHub, Dropbox, OSF	☐ DNF	☑ social media widgets for dataset sharing appear (facebook, twitter, linkedin)	☐ DNF	☐ DNF	☐ DNF	☑ interogating dataset using DataChat and Ask the Data	☐
Zenodo	https://zenodo.org/		☑ standard	☑ partially via metadata guidlines	☑ breadcrumbs navigation, page headers	☑ public REST API, Open Archives Initiative Protocol for Metadata Harvesting (OAI- PMH) GitHub Integration	☐ DNF	☑ partially: a twitter account handle for the repository in the contacts page	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☐
Statista	https://www.statista.com/		☑ standard	☐ DNF	☑ breadcrumbs navigation, page headers	☑ closed-source private search and data REST APIs for business customers for integration within products	☐ DNF	☑ social media widgets for sharing content, social media handles for following (facebook, twitter, linkedin, instagram)	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☐
Auctus Dataset Search	https://auctus.vida-nyu.or		☑ standard with detail being shown along side SERP	☐ DNF	☑ detailed page is merged with SERP showing both results and detail	☑ REST and Python APIs for search and data	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☐

DNF = Did Not Find

#			12	13	14	15	16	17	18	19	20	21
Cluster			Faceted Search					About Data and Provenance				
Name	Url	Feature	aborative tool integration	Topic Categorization	Geographic filter	Temporal Filter	Temporal Zooming	Data exclusion	format	metadata schema	size (in MB etc.)	license info	DOI / Persistant Identifier
Gesis	https://search.gesis.org/		DNF	☑ dynamic list of socio-economic topics, label-based filter	☑ country based geographic filtering	☑ collection year and publication year, simple sliders and input fields	☐ DNF	☐ DNF	☑ file extension, Data Type field, originating program	☑ dara, DDI, Dublin Core, DataCite	☑ in download file link	☑ optional field License in detail page	☑ DOI
London Data Store	https://data.london.gov.ul		DNF	☑ fixed list socio-economic topics, label-based filter	☑ list of locations and Map based interaction	☑ partial implementation in map based interacton	☐ DNF	☐ DNF	☑ Icon and file extension	☐ DNF	☑ next to file	☑ on detail page	☐ DNF
Dataset Search	https://datasetsearch.res		DNF	☑ fixed list (9) of scientific fields (topics) e.g. Social Sciences, Computing, Engineering	☐ DNF	☑ simple filter based on the last update: Past Month, Past Year, Past 3 Years	☐ DNF	☐ DNF	☑ file extensions	☑ schema.org (Dataset)	☑ when available included in size field in bytes	☑ optional metadata field	☑ optional DOI
ICPSR	https://www.icpsr.umich.e		DNF	☑ filters organized by 19 facets	☑ geographic filtering based on the available granularity (e.g. continents, regions, countries, states)	☑ available sorting by time relevance, simple year-based filter (from to), filter with recent releases	☐ DNF	☑ textual exclusion using minus sign followed by keyword ("-")	☑ file extensions, filtering by data format	☑ DDI 2.5, Dublin Core, DATS 2.2 (JSON), DCAT + guide	☑ some datasets feature tree-like document structure preview with file sizes	☑ on detail page, license agreement dialog when downloading	☑ DOI
Kaggle	https://www.kaggle.com/c		DNF	☑ filters organized by general and 6 computer science fields	☐ DNF	☑ simple filter by last dataset update, sorting results by date of update	☐ DNF	☐ DNF	☑ file extensions, filtering by data format	☑ Data Package specification (JSON), optionally described in Detail in Metadata panel	☑ file size in Search, SERP and Detail	☑ on detail page	☑ optional DOI
GenBank	https://www.ncbi.nlm.nih.g		DNF	☑ classification by organism, gene, molecule type, sequence type, and other structured fields	☐ DNF	☑ filter or search sequence records by release date, update date, or publication date using Entrez search	☐ DNF	☑ basic exclusion (e.g. NOT operator)	☑ FASTA and ASN.1	☑ ASN.1 GenBank flatfile, and FASTA structure	☑ only in Release notes and FTP site	☑ GenBank is openly available "with no restrictions"	☑ assigns own stable accession number
Inspire	https://inspirehep.net/		DNF	☑ Faceted filtering in search results via categories on the left sidebar.	☐ DNF	☑ date range filtering in search results	☑ partially using date ranges	☑ basic search exclusion operator - (NOT)	☑ file extensions	☑ ISO 19115/19139 and MARC (JSON Schema)	☐ DNF	☑ provides license and usage information in detail	☑ DOIs and HDL
Open Science Framework	https://osf.io/		DNF	☑ filtering by content type (registrations, preprints, projects, files)	☐ DNF	☑ filtering by date created or updated on projects and registrations	☐ DNF	☑ partially implemented: exclusion based on types of content (preprint files only)	☑ In detail tree-like directory browser shows file types (with icons) included in a project (detail page)	☑ Dublin Core	☑ tree-like directory browser shows MB or KB size of files (with icons) included in a project (detail page)	☑ license in the detail metadata side panel (also info when missing)	☑ DOI assigned to published dataset
European Data Portal	https://data.europa.eu/en		DNF	☑ in Search filtering based on theme/topic, category toggles, keywords in detail	☑ filtering based on country, region or administrative area in the EU	☑ date of dataset publication or update, letting users focus on current or historical datasets using date pickers or time period menus	☐ DNF	☐ DNF	☑ Format badge in Search for each dataset and in detail (CSV, Excel, JSON, TSV)	☑ DCAT-AP metadata schema (standard for European Data portals), Detail also supports downloading metadata in various formats (RDF, Turtle, Notation3, N-Triples, JSON-LD)	☐ DNF	☑ implicitly open data licenses approved by EU, included in metadata (not visible otherwise)	☐ DNF
Figshare	https://figshare.com/		DNF	☑ uses Australian and New Zealand Standard Research Classification (ANZSRC) categories and topics, keyword tagging is available in detail	☑ partially implemented via geospatial metadata search	☑ filter via date ranges and publication dates through advanced search	☑ date ranges	☐ DNF	☑ In detail tree-like directory browser shows file types included in the project	☑ DataCite schema	☑ In Detail tree-like directory browser shows sizes for individual files	☑ In Detail Page license section, Filtering based on license in Search Page	☑ All public items receive a DataCite DOI at publication
Harvard Dataverse	https://dataverse.harvard		DNF	☑ Classification by subject, controlled vocabularies of keywords and faceted search	☑ detailed geographic filtering in advanced search (country, state/province, city) as well as geospatial coverage	☑ filter based on publication year, in advanced search range based time search (start and end), distribution date	☑ partially by using date ranges	☑ partially using exclude subsets in advanced search	☑ In detail Files with extensions are listed in Files tab, any type of file, common tabular formats have additional features (CSV, Stata, SPSS, Excel, and RData up to 2.5 GB)	☑ DataCite, DDI, Dublin Core (support for Domain Specific Metadata)	☑ File sizes displayed in the Files tab next to given file name	☑ by default assigned CC0 Public Domain Waiver if no other is provided, can be filtered in search, displayed in detail	☑ optionally DOI, support for file-level DOIs
Zenodo	https://zenodo.org/		DNF	☑ Faceted filtering in search results via categories (subjects) on the left sidebar.	☐ DNF	☑ search based on publication date	☑ partially by using date ranges	☑ exclusion using regular expressions and missing values	☑ In detail in Files sections files with extensions are listed, tabular formats have additional features (CSV, excel)	☑ DataCite Metadata Schema, Dublin Core, MARCXM	☑ File sizes shown for uploaded files. Total record size limits default to 50GB	☑ mandatory at upload; displayed in Detail page in Rights section	☑ DOI automatically assigned on publication
Statista	https://www.statista.com/		DNF	☑ Advanced faceted search, categorizes data/statistics by clear topics and industries	☑ filtering by geographic regions if available for given query in the search page (panel select regional boost and rollout location filter)	☑ filtering by publication date and date ranges	☑ partially by using date ranges	☑ partially by using NOT operator	☑ In detail in Download section available data formats in tabular formats are displayed (CSV, Excel) unless protected by paywall	☐ DNF	☐ DNF	☑ communicated only to paying users	☐ DNF
Auctus Dataset Search	https://auctus.vida-nyu.or		DNF	☑ inferred data types and column names are used to categorize datasets, thematic tags are assigned	☑ spatial search via map bounding box, administrative area input	☑ filtering by fine-grained temporal resolution (year up to second)	☑ filtering by multiple fine-grained temporal intervals	☑ partially by excluding ommited temporal ranges in search	☑ file formats CSV, D3M are offered in detail page for download, (optionally JSON via API)	☑ Metadata extracted automatically by the profiler with summaries and data types described by JSON Schema	☑ size displayed next to dataset name in SERP and in dedicated section of detail page	☑ when available captured in metadata but not displayed	☐ DNF

DNF = Did Not Find

#			22		23		24		25		26		27		28		29		30		31		32	
		Cluster											Sense making											
Name	Url	Feature	last update		original purpose		how were data collected/derived		data processing steps		versioning information		key variables		textual summaries		statistical summaries		visual summaries		column-level summaries		sup	
Gesis	https://search.gesis.org/		☑	optional field with table version history overview	☑	textual description (abstract)	☑	metadata fields Sampling Procedure, Mode of Data Collection	☑	optionally described in supplemented documents (Codebook, Questionaire etc.)	☑	optional field with table version history overview	☑	available in results page and in detail; implemented direct search for variables	☑	optionally in supplemental materials	☐	DNF	☐	DNF	☐	DNF		
London Data Store	https://data.london.gov.u		☑	shown on RESP and detail	☑	not required, author dependent, textual description	☑	not required, author dependent, textual description example	☑	not required, author dependent, textual description	☐	DNF	☐	DNF	☑	not required, author dependent	☐	DNF	☑	interactive line charts and bar charts on home page only, Excel generated charts otherwise	☐	DNF	☑	
Dataset Search	https://datasetsearch.res		☑	Dataset updated field in detail	☑	author dependent textual description	☑	optionally includes field Measurement technique	☐	DNF	☑	simple filter based on last update: past month, past year, past 3 years	☑	optionally includes field variables measured	☑	not required, author dependent	☐	DNF	☐	DNF	☐	DNF	☑	
ICPSR	https://www.icpsr.umich.e		☑	published versions and date of last update available on the results page and detail, sorting by recent releases	☑	textual description (abstract) of originating paper, scope of the originating project, list of related publications and other purposes of the dataset	☑	filtering by collection method (survey, census, aggregate etc.), filtering by mode of data collection (face-to-face, CATI, CAPI, questionare)	☑	optionally in outlined in supplemented materials	☑	sorting by recent releases, versioning information in detail page	☑	implemented direct search by variables in search page, variables can be compared	☑	implemented for variables	☑	implemented in variable search (not individual datasets)	☑	implemented in variable search (not individual datasets)	☐	DNF		
Kaggle	https://www.kaggle.com/c		☑	in Detail in Expected Update Frequency panel in detail displays last update	☑	in Detail in textual description (About Dataset) of originating paper, scope of the originating project, list of related publications and other purposes of the dataset	☑	in Detail in Provenance panel, optional textual description	☑	steps in Notebooks if available, opitionally in About Dataset	☑	Expected Update Frequency panel in detail displays last update	☑	Available as Tags in Search and in Detail	☑	abstract, not required, author dependent	☑	In Detail in Data Explorer relative measurements for nominal data	☑	jupter notebooks visualizing data for linked datasets	☑	In Detail, each column has a corresponding visualization in an embedded Data Explorer	☑	
GenBank	https://www.ncbi.nlm.nih.		☑	submission and latest update dates	☑	in the record (e.g., gene description, organism, study), contextual info often included in annotations or source fields	☑	present in submission info	☑	automated/manual quality checks and curation steps for submissions, detailed steps sometimes in release notes	☑	versioning information, with version numbers; changes are tracked and detailed in the release notes	☑	critical fields: organism, gene name, sequence type, annotation, reference, and accession number	☑	textual descriptions and annotations	☐	DNF	☑	sequence visualizations if avaialable	☐	DNF	☐	
Inspire	https://inspirehep.net/		☑	creation, publication, and revision	☑	not required, textual description (abstract)	☑	in Detail optionally in abstract	☑	in Detail optionally in abstract	☑	unique DOIs displayed for different versions	☑	key variables under abstract in detail (defined in schema files)	☑	abstracts describing the data content, purpose and characteristics	☐	DNF	☐	DNF	☐	DNF	☐	
Open Science Framework	https://osf.io/		☑	displayed in detail for projects and files	☑	optional textual description (abstract)	☑	optional textual description (abstract)	☑	optional textual description (abstract) or supplemental Documents within the host project, included processing Scripts	☑	file versioning and histories managed by OSF	☑	partial implementatino: user-defined variable highlight	☑	textual descriptions and summaries in project descriptions	☐	DNF	☐	user-defined visual summaries only when generated by author in Documents	☐	DNF	☑	
European Data Portal	https://data.europa.eu/en		☑	displays last update or revision of datasets in the Catalogue Record field in detail	☑	issuing authority project description, optional described textually in abstract	☑	data from public sources (data by authorities), optionally intent described in metadata fields	☑	optionally intent described in metadata fields	☑	occasionally documented in metadata, last update or revision of datasets in the Catalogue Record field in detail	☑	optionally described variables in metadata	☑	abstracts describing the data content, purpose and characteristics	☐	DNF	☑	CSV data visualization tool in Detail page (table, barchart, linechart, donut etc.)	☐	DNF	☑	
Figshare	https://figshare.com/		☑	History section in Detail page displayes updates with commentary, versioning updates are timestamped and documented	☑	issuing authority project description, optionally descriptive metadata fields include description	☑	optionally in abstract	☑	optionally in abstract, included processing Scripts	☑	History section in Detail page displayes updates with commentary, versioning updates are timestamped and documented, versions linked via DOI	☑	optionally included in Abstract	☑	textual summaries in abstract provided by the authors	☐	DNF	☐	DNF	☐	DNF	☑	
Harvard Dataverse	https://dataverse.harvard		☑	Versions tab in Detail describes individual updates, also described in metadata tab	☑	optionally included in metadata, optionally documented in abstract by author	☑	optionally in abstract, optionally in specified in discipline specific metadata	☑	optionally in abstract, included processing Scripts	☑	Versions tab in Detail describes individual updates, also described in metadata tab	☑	automatic identification of key variables in tabular files	☑	textual summaries in abstract provided by the authors, optionally also in metadata	☑	computed and displayed in in the second-generation Data Explorer (mean, std dev, min, max, mean, count)	☑	summary info (bar chart) and frequency tables for discrete variables are available in the second-generation Data Explorer	☑	View Summary Statistics button in the second-generation Data Explorer	☑	
Zenodo	https://zenodo.org/		☑	publication date tracking, versioning with release notes, technical metadata modification tracking in detail page	☑	optionally included in metadata, optionally documented in abstract by author	☑	optionally in abstract, optionally in specified in discipline specific metadata	☑	optionally in abstract, included processing Scripts	☑	dataset versioning with new DOIs for updated versions and changelog/history metadata.	☐	DNF	☑	textual summaries in abstract provided by the authors, optionally also in metadata, keywords included	☐	DNF	☐	DNF	☐	DNF	☑	
Statista	https://www.statista.com/		☑	partially by indication of last update	☑	optionally included in abstract	☑	optionally in dedicated methodology section	☑	optionally in dedicated methodology section	☐	DNF	☑	inconsistently key variables and metrics mentioned on detail pages	☑	textual summaries and explanations with each statistic or dataset with key insights and analyst opinion in detail page	☐	DNF	☑	data visualized with interactive bar charts, line charts and maps	☐	DNF	☑	
Auctus Dataset Search	https://auctus.vida-nyu.or		☑	Last Updated Date in detail, captured by temporal metadata about dataset	☑	optionally included in metadata, optionally documented in abstract by author	☑	optionally in abstract	☑	optionally in abstract	☐	DNF	☑	Key columns and variable types (categorical, numerical, spatial) are automatically extracted by the profiler and displayed in SERP and in detail page	☑	metadata and profiler-generated summaries provide textual overview	☑	profiler generates statistics for columns (unique values, mean, variance), shown in detail page	☑	histograms and barcharts for suitable columns, geospatial area coverage in map bounding box when geospatial variable available	☑	detailed column-level textual summaries including variable type (unique values, mean, variance) as well as data types (Integer, Text, Enums, Nulls) in Detail page (section Dataset Sample)	☑	

DNF = Did Not Find

	#		33		34	35	36	37	38	39	40	41	42
	Cluster				Quality Indicators						Task Support		
Name	Url	Feature	ports spreadsheet view	horizontal scrolling	completeness	empty fields	empty headers	data quality issues and limitations description	Where Used before	toggleable quality indicators	side-by-side comparsion	combining of data	highlighting of common attributes
Gesis	https://search.gesis.org/		DNF	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☑ automatically referenced by infoLink or manually added	☐ DNF	☐ DNF	☐ DNF	☑ automatic detection of variables; direct search for variables
London Data Store	https://data.london.gov.ul		online Excel and Microsoft BI integration	☑ present in Excel and Microsoft BI integrations	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☑ author dependent, textual information with hyperlinks	☐ DNF	☐ DNF	☐ DNF	☐ DNF
Dataset Search	https://datasetsearch.res		built-in table previewer for CSV files	☑ supported by the built-in table previewer	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☑ referencing scholary articles from Google Scholar	☐ DNF	☐ DNF	☐ DNF	☐ DNF
ICPSR	https://www.icpsr.umich.e		DNF	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☑ references related publications	☐ DNF	☐ DNF	☑ partially implemented by visualizing variable frequency across datasets	☑ implemented search by variables
Kaggle	https://www.kaggle.com/c		built-in Data Explorer for spreadsheet-compatible formats (CSV, Excel)	☑ supported by the built-in table previewer	☑ optionally in Detail in the Coverage Information Panel	☑ optionally in Detail in the Coverage Information Panel	☑ optionally in Detail in the Coverage Information Panel	☐ DNF	☑ author dependent textual information in About Dataset, optionally in Detail in Provenance Panel, dataset usage statistics (number of time viewed and downloaded)	☐ DNF	☐ DNF	☐ DNF	☑ partially implemented, searching by common tags
GenBank	https://www.ncbi.nlm.nih.g		DNF	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☑ links to the paper used	☐ DNF	☐ DNF	☐ DNF	☐ DNF
Inspire	https://inspirehep.net/		DNF	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☐ DNF	☑ tracks paper citations, dataset usage statistics (number of time viewed and downloaded)	☐ DNF	☐ DNF	☑ partially implemented by integrative analysis of multiple datasets	☑ partially implemented, searching by common tags
Open Science Framework	https://osf.io/		built-in Data Explorer, previewing spreadsheets (CSV, Excel)	☑ supported by the built-in table previewer	☐ DNF	☐ DNF	☐ DNF	☐ in abstract author-dependent	☑ tracking of DOI and citations, dataset usage statistics (number of time viewed and downloaded)	☐ DNF	☐ DNF	☐ DNF	☑ partially implemented, searching by common tags
European Data Portal	https://data.europa.eu/en		built-in Data Explorer for spreadsheet-compatible formats (CSV, Excel)	☑ supported by the built-in table previewer	☑ Completeness assessed in quality section for every detail page implemented using Metadata Quality Assessment Methodology (MQA)	☑ Empty Fields assessed in quality section for every detail page implemented using Metadata Quality Assessment Methodology (MQA), encourages to explicitly mark fields as null or NA	☑ Empty Headers assessed in quality section for every detail page implemented using Metadata Quality Assessment Methodology (MQA)	☑ Issues and limitations assessed in quality section for every detail page implemented using Metadata Quality Assessment Methodology (MQA)	☑ datasets tracked via citations, optionally mentioned textually and in metadata, dataset usage statistics (number of time viewed and downloaded)	☐ quality flags are displayed but are not toggleable	☐ DNF	☐ DNF	☑ partially implemented, searching by common tags
Figshare	https://figshare.com/		built-in Data Explorer for spreadsheet-compatible formats (CSV, Excel)	☑ supported by the built-in table previewer	☐ DNF	☐ DNF	☐ DNF	☐ in abstract author-dependent	☑ author dependent, textual information with hyperlinks, dataset usage statistics (number of time viewed and downloaded)	☐ DNF	☐ DNF	☐ DNF	☑ partially implemented, searching by common tags
Harvard Dataverse	https://dataverse.harvard		built-in Data Explorer for spreadsheet-compatible formats (CSV, Excel), also second generation data explorer	☑ supported by the built-in table previewer	☐ DNF	☐ DNF	☐ DNF	☐ in abstract author-dependent	☑ datasets tracked via citations, optionally mentioned textually and in metadata, dataset usage statistics (number of time viewed and downloaded)	☐ DNF	☐ DNF	☐ DNF	☑ partially implemented, searching by common tags
Zenodo	https://zenodo.org/		built-in Data Explorer for spreadsheet-compatible formats (CSV, Excel) and textual formats (JSON, XML)	☑ supported by the built-in table previewer	☐ DNF	☐ DNF	☐ DNF	☐ in abstract author-dependent	☑ datasets tracked via citations, optionally mentioned textually and in metadata, dataset usage statistics (number of time viewed and downloaded)	☐ DNF	☐ DNF	☐ DNF	☐ DNF
Statista	https://www.statista.com/		built-in Data Explorer for spreadsheet-compatible formats (CSV, Excel)	☑ supported by the built-in table previewer	☐ DNF	☐ DNF	☐ DNF	☑ partially by explicit disclaimers about quality	☑ linked source is described in the sidepanel in detail page, further information is available for paying customers	☐ DNF	☐ DNF	☐ DNF	☑ partially implemented, filtering by common tags, displaying aggregated key figures based on searched data
Auctus Dataset Search	https://auctus.vida-nyu.or		built-in Data Explorer in detail page section Dataset Sample	☑ supported by the built-in table previewer	☑ partially: profiler extracts metadata by inspecting dataset for missing values	☑ Profiler inspects dataset for missing values displaying MissingData flag in the Detail Page	☐ DNF	☑ partially by mentioning MissingValues in data columns	☑ provided source of the data	☐ DNF	☐ DNF	☑ data augmentation via join and union searches, enabling combination of datasets by matching attributes	☑ Common attributes used internally for join search and union capabilities, intersections are shown in search results

A.2. Interview Consent

Information for participants and declaration of consent to participate in the study:

Towards Data Showcases Facilitating Data Reuse

Dear participant,

Thank you for your interest in participating in this study.

Your participation in this study is voluntary. You can refuse to participate at any time, without having to give a reason, or also withdraw your agreement to participate once the study has already started. There will be no negative consequences for you if you refuse to participate or if you withdraw from this study early.

This kind of study is necessary to gain new, reliable academic research results. However, your consent to participate in the study is an indispensable prerequisite for us to conduct this study. Please take time to read the following information carefully in addition to the explanatory talk, and do not hesitate to ask questions.

Before you decide whether or not to participate, it is important that you understand the purpose of the study, what your participation will involve, and any potential risks or benefits associated with it. Please read the following information carefully. If you have any questions or concerns, feel free to ask the investigator before making your decision.

Please only declare your consent with participation:

- if you have fully understood the type and procedure of the study,
- if you are willing to give your consent to participate, and
- if you are aware of your rights as a participant in this study.

1. **What is the purpose of this study?**

   The goal of this study is to collect insight from scientific data producers to inform design of data showcases.

2. **What is the procedure of the study?**

   If you agree to participate in this study, you will be interviewed online over Zoom platform. In the second half of the interview, you will be asked to share your screen with sent visuals. The interview will be recorded to ensure accurate data collection and analysis. The interview will last approximately fifty minutes.

3. **What are the benefits of participating in the study?**

   By participating in this study, you help us to identify requirements for a data showcase that may facilitate efficient data reuse.

4. **What are the possible risks of taking part in this study? Could participants experience any discomfort or other side effects?**

   The researchers are not aware of any possible risks associated with participation in this study.

5. **How will the data collected in this study be used?**

   The interview will be recorded to ensure accurate data collection and analysis. However, the identities of the participants will only be known to the researchers involved in this study, who are obliged to maintain secrecy. You will never be mentioned by name, without exception.

6. **Will there be any costs for the participants? Will they receive reimbursement or remuneration?**

   You will not incur any costs from participating in this study.

7. **Possibility to discuss further questions**

   The study coordinator (the interviewer) will be happy to answer any further questions you might have about the study.

Name(s) of the contact person(s):

[redacted] research group - [redacted]

Declaration of consent

I agree to participate in the study *Towards Data Showcases Facilitating Data Reuse.*

Hynek Zemanec provided me with clear and detailed information about the objectives, significance and scope of the study, as well as about the requirements resulting from my participation in the study both orally and in writing.

I have read this information text for participants and the declaration of consent, especially section 4 (regarding risks, discomforts or side effects). Hynek Zemanec answered all my questions sufficiently and in a comprehensible manner. I had enough time to decide whether I would like to participate in this study.

I will follow the instructions that are necessary for conducting this study. However, I reserve the right to end my voluntary participation at any time, without this being to my disadvantage. If I want to withdraw from the study, I can do so at any time within 6-month period after the interview has concluded by contacting persons listed in the contact list, either in writing or verbally.

I understand that taking part in the study involves audio recording which will be transcribed, analysed and then destroyed.

I understand that I will not be directly identified in any reports of the research.
I agree that my data are permanently saved electronically in pseudonymized form. Data that have not yet been pseudonymised are stored in a form that is only accessible to the project management and are secured in accordance with current standards.

I can request my data to be deleted within 6-month period after the interview has concluded by contacting any of the person listed in the contacts and without having to give a reason.

I have read and understood the information for participants. I had the opportunity to ask all the questions I was interested in. My questions were answered fully and in a comprehensible manner.

I have received a copy of this information for participants and declaration of consent.

[redacted]
[Date]

[Participant]
[Date]

A.3. Interview Protocol

Part	Type	ID	Question	Clarification	Objective
Introductory	**core**	**Q1**	**Can you briefly introduce your role and expertise in the dataset's scientific domain?**		**Establish scientific domain and background**
Introductory	**core**	**Q2**	**How do you present data?**	**in papers, conference presentation**	**Seeking for visualization methods of presenting research findings, tables, visualizations**
Introductory	probing	Q3	Do you use some specific visualization software?	e.g. Tableu	Identify software which handling and vis conventions can be further explored for further incorporation
Dataset	**core**	**Q4**	**Could you briefly describe the dataset?**	**about, features, collection methods**	**Gather basic dataset characteristics**
Dataset	probing	Q5	How would you describe this dataset to someone without technical knowledge?		Simplify dataset description
Dataset	**core**	**Q6**	**How would you categorize the dataset?**	**scientific domain, qualitative, quantitative**	**Determine dataset's categorization (scientific domain)**
Dataset	probing	Q7	Can you think of any other categorization that could fit the dataset?		Explore alternative dataset categorizations
Dataset	**core**	**Q8**	**Who else might benefit from using this dataset?**	**inside or outside the scientific domain**	**Identify potential beneficiaries of the dataset (adjusting to the needs of the group)**
Dataset	**core**	**Q9**	**Can you describe the context in which this dataset was formed?**	**processing methods, team members, financing, origin**	**Gather contextual information about the dataset**
Dataset	probing	Q10	Were there any issues during data collection or processing that you're aware of?		Identify potential challenges in data processing
Dataset	**core**	**Q11**	**Imagine you are searching for a similar dataset. What information would you need to use a similar dataset?**	**the same dataset as yours, key attributes you need to know about**	**Determine important factors in dataset selection**
Showcase	**core**	**Q12**	**Can you describe the following designs?**		**Understand initial impression of data showcase**
Showcase	probing	Q13	Is the purpose of those designs immediately clear to you?		Assess the clarity of the design's purpose
Showcase	probing	Q14	How do you figure out what those designs are meant for?		Explore strategies and steps for comprehending design
Showcase	**core**	**Q15**	**What is obvious and clear, unclear about the design?**		**Identify design elements that are immediately clear**
Showcase	probing	Q16	What stands out to you? (why)		Understand what are the most prominent elements on the first sight
Showcase	**core**	**Q17**	**Which elements help you understand the dataset's content?**		**Determine design elements contributing to clarity**
Showcase	**core**	**Q18**	**How would you try to access / download these datasets?**		**Understand ease of reuse of dataset from the design**
Showcase	probing	Q19	What data formats would you expect in the download dialog?	e.g. CSV, XML, JSON	identify expected data formats
Showcase	**core**	**Q20**	**What elements makes the dataset trustworthy or untrustworthy in your opinion?**		**Identify key factors contributing to trustworthiness**
Showcase	**core**	**Q21**	**As the dataset author, what are the most important elements to you?**		**Identify design elements of author's priority**
Showcase	probing	Q22	Are there any elements that you miss in the design?		identifying inconsistencies between priorities of data producer and the design
Showcase	probing	Q23	Would you describe it as overwhelming, adequate, or anything in between?	is the amount of information too much to comprehend	get the general feeling about the amount of presented information
Showcase	**core**	**Q24**	**Do you think you would be able to use this dataset based on the information shown here?**		**establish confidence of the interviewee in using the showcase for intended purpose**
Showcase	probing	Q25	Does the order of the information panels make sense to you?	the information panels below the chart that can be hidden	the established order of panels (most important first) is recognized as the same
Showcase	probing	Q26	Which one would you prefer? (why)	Is there a clear preference from for one or the other?	uncover motivations for subjective preference for given design
Showcase	**core**	**Q27**	**How would you improve the effectiveness of the prefered design?**	**effectivness of deciding to use the dataset**	**Gather suggestions for improving the design**
Showcase	probing	Q28	What do you find ineffective or confusing about the current solution?		get more insight on what is subjectively considered wrong
Showcase	**core**	**Q29**	**Do you have any other comment, thought or recommendation about the designs?**		**closing comments and thoughts**

A.4. Interview Coding Results

Solution	Code	Keyword	Theme	#	Interview_Pilot_Aug_11	Interview_1_aug_30	Interview_2_Sep_17	Interview_3_Oct_4
hide or omit the testimonials from the final app	Concerns	testimonials	Distrust in the testimonials	4	I'm not sure if we really need those to be there the whole time those testimonials basically down here	This is something that's very discussable because this can be, you know, both good to push work that has been done. If you read like a positive review, but this can also be a kind of ...this can bias how many people are going to use the dataset just because you know it's like ... maybe it's not very prettily prepared or so on and but the data itself is good if the review is negative because of the way it's prepared and then you have two or three negative reviews and anyone who encounters the dataset afterwards maybe. Like opposed to using it | It has been reviewed very positively, which as I said is like a double-sided thing.	And then testimonials. I'm not sure if this is necessary. | I'm thinking about the testimonials part. I see the rational. I mean, if it's a well organized data, it's nice that you can leave a comment. Or if it's not well organized, it's also good to leave a comment. But I'm, I'm just not sure. Yeah. How often would that be used, I guess. I mean, if it's, if it's a dataset that has a lot of good testimonials. Is it because of the structure of the way that the data is clear that is well organized, that it has all the crucial information, or is it just because of the content of the data, so that it just shows some interesting results. | So the people are leaving the testimonial because of how structured and organized and presented the details. Or is it because of the content that it for example, you know is presenting some controversial results or yeah, that's just my thought. I don't think it's right or wrong.	It could be impersonation, you know, like these reviews online.
	Potential Users	experts	Domain Experts	4	Another one could be for domain experts basically who work on a specific topic. We can build tools for them that they can more easily or more efficiently, basically, to analyze the dataset.	And we've discussed it with people from political science backgrounds and | Other than that, I mean, anyone who researches a topic similar to this one may be interested to look at it.	I would put it into the social sciences domain, specifically into communication science or of course, sometimes it's called media studies.	whole project was very interdisciplinary to begin with, so already primary investigators, they come from different fields. So, Bernhardt, for example, he is from economic sociology and from political science, social and other. I am from political science. Barbara, she's in health politics. And Hajo, he's in communications. | we also have people from economics. | We had people from working on psychology, for example health psychology, but also Medical or psychologists or, for example, whether the pandemic cause depression with young people and so on because of the lockdown because of school closures and so on. | We had people working on family sociology because also a lot of families had problems with school being clothes and so on.
	Positive Feedback	abstract	Information	4	I have to say I like A a bit more because I have all this information up here, which is really important for me. [abstract] | As I have already mentioned I guess a few times that the description up here is really, really nice. [	I haven't read what's written there, but it seems intuitive what's mentioned.	I really like the brief description of the data set on A1. The thing that it's visible right away. So you know it saves time. I don't have to click through to actually find out what's going on. | and then as I said on the left side I can see right away a brief description of the data set. Which I really appreciate. As I said, it's like when you have. An abstract of a paper. So it allows you to see right away whether this data set is something that is relevant to you or not. | A design that I really like that basically you have the short abstract that tells you what is the dataset about and then of course you have the name of the variables here on the graph and then I guess there is the toggle list with the features which basically describes what is the data set about. | on the A data set, I really like that you have the "about" the description, the brief description of the data set already up here	So, I like also better that you have the brief description up here. I think that's also helpful just the verbalized description uh. | So, I like this one better because it has a verbalized description just like in common language.
	Latent Need	features	Need for understanding features	3	I would have to look at the datasets so the data features that are in there and the level of detail that is in there. | I would have a short look on the features, basically, if there are the features there that I actually need, and if I find something interesting, then I would go into the table and really look at the data in the raw worksheet for instance	But yeah, these issues of naming they continue. | And I had like no proper documentation, so nothing to explain the variable names, the like the phrasing of survey items itself was not connectable easily with the variable name.	Then there is of course the information about the experimental condition that they were assigned to, their age, gender, education, political affiliation, so if there are more left or right, and then we also asked them about their experience with COVID and what we did is that we measured their attitudes towards COVID vaccination. | you need to make the data set as clean and as clear as possible, because of course if you are in the inside team that is working on the project, you understand which variables mean what.	
	Expectation	features	data feature description	3	So, what is really, really important for me and which I also mentioned in the in the interview before, the data features are really important for me.	And so a lot of transparency is good there, but I guess, usually the questions themselves and I think you can call it a catalogue, is enough and then the dataset in a way where it's well described and you can you can easily see what the variables mean and what they relate to yeah. | how were the items like if we're talking a survey dataset, I would want to know how was this tested? Is there like a quick description of the process beforehand? I mean obviously you need the things like that I just mentioned like the variable names explanations. You need to document the survey phrasing and like the order of segments within the survey and how they were displayed.		Because those attributes people do not use in their normal language in this very technical way, it's good to have them somewhere, but not for the way you know people answer in Google they write. Is it healthy to get vaccinated like this, kind of, you know, they get very verbalized question and is better.
	Expectation	download	Download button opens a dialog with multiple options to download	3	I would just click the download button and then I would hopefully get a window where I can select basically, if I wanted to have it in a CSV file, if I wanted to have it in a excel sheet | so different formats would be nice to select here. |	I imagine clicking on the download button and then I imagine something popping up and I'm going to be asked in which like which sort of file it's gonna ... | or maybe there's even like I click download and then I have the option to download the dataset, then I have the option to download information that's provided on the page as well, like some sort of other document like.		I would expect to happen something like. You need to register first, enter your e-mail address, blah blah blah, classified, have a click on user agreement, blah blah. | Download Button
	Expectation	download	Expected download formats	3	yes, because it really depends on where I'm using it, but CSV and excel sheet or for instance, uh yeah, would be the most important ones I guess.	I would hope for a CSV because I can work with a CSV.		Usually, you have multiple data formats being offered. So usually if this is social sciences usually would be offered Stata that you would be offered are usually would be also tabulated like CSV, CSV type of data set. So, I think this week will be the most conventional or sometimes SPSS as well.
	Concerns	spacing	Layout of the design is overwhelming	2	There is a lot going on. There's lots of text on this design.	here in the second one in B2 it's too scattered for me		
	Latent Need	statistics	Need for basic quantitative measures	2	If we have some quantitative data. Then I usually calculate the mean, the standard deviation. I look at the distribution of the quantitative data values. And for qualitative data, we usually do thematic analysis where we really have text. We try to come up with some themes and then attribute those themes to the text to really see if there are some patterns in there.			Sometimes we also did other kind of analysis, sometimes we did. Maybe just showing mean levels changing. | ... we need to sort of compress it more so we just show the mean trend for a different category. Yes, so that was sort of univariate statistics that we the most of the time. | I mean in the block, we did mostly this univariate statistics like this graphical display of distribution and means central tendency and distribution.
	Potential Users	government	Public Policy and Government	2			I would think that somebody from the public health domain, so for example, somebody who is trying to design uh Health Communication campaign and they are looking for ways to make it more effective | even like members of government or even people who are concerned with PR because it was about how medical professionals communicate.	we wrote a lot of blog posts to disseminate the research to the to the general audience, and also then later I became a member of the government committee advising the government on COVID-19
	Expectation	abstract	Introductory description	2	I mean, yes, we have a description here which is really, really important, I would say. | So, the brief description up there really helps. And then I would look at the description, I would look at the features and then really look at the data if that looks promising to me. | Those key information is also really, really important, especially if I want to use the data or if I upload it and give it to someone else	I think UM, as I said, key information, very important also when dealing with dataset this ... you need this for anything, you need it to understand what you're looking at.		
	Expectation	credibility	author and author(s)'s credibility	2	So, the source is really, really important. The author, if it's a famous author.			as I said, like the way I think about it is we know the Community, we know the research community and this is the trust, yeah. So we know who is doing the research, who is doing the practice. So maybe like name of the author. Also, the project, the people involved, what the publication have been, uh, doing what the research has been showing and so on.
	Expectation	method	Description of the method	2			I would say that I would need to first to see the method in which the data was collected. To be able to infer what the study was actually about, if I can, you know, draw some causal inferences and stuff like that. | The second would be the framework where I would see specifically what were the, what were the trying to do. So what were the independent dependent variables, the mediators and the moderators. | So, when you asked me about what is the number one thing that I would like to know when I'm looking at the dataset is the method in which it was conducted.	it's very hard to understand what it does, how it has been compiled, what the quality of the and so on, so you know it's a complete nightmare.
	Familiar Software	Excel	Spreadsheet editors for data sheets	2	So, usually we store our data in data sheets like for instance with Microsoft Excel or with Google Sheets		It's usually done via Excel, yes.	
	Familiar Software	R, SPSS, Stata	R and SPSS for more sophisticated processing	2			Or you can also generate some graphs in R. | we work with either SPSS or R when we are handling our data and analyzing it.	but we also use Stata and we also use R
	Confusion	Data Nutrition Label	Unfailiarity with Data nutrition label	2	I'm not quite sure what this actually is: about humans, technical review, ethical review and update frequency. I'm a bit confused. Is this just a button where I can click on and get some more information or I'm not quite sure what this is.			there's also something, uh, like little signs probably meaning something. What to do or can be done with it, I don't know exactly. Some sort of labels that's or something. | I don't know about this little thing, so maybe they are not needed. [data nutrition label]
	Suggestion	features	Description of features	2	I just wanted to mention that it would be great also to have an explanation of the different data features, but I guess this is this right.	as I said, in my case it would be like catalogue of questions, and you can also download the dataset and you can also download a file where the variables are explained like depending on what you see. | Maybe I would like to have a bit more transparency here as to what the dataset has been structured like, but then again, I don't know who this is going to be presented to.		
	Suggestion	testimonials	Hiding testimonials under an information panel	2	maybe we can put it into such a section, which I can then open up if I need them. So, I'm not sure if I need this really all the time. [points to testimonials]			As I said, the testimonials I probably would not need for me, the testimonials would be replaced, sort of by the publications that have been done by the data set. If there's a publication like the High End publication in the very high quality journal that has been done with this data set. That would be what would be this data set credibility.
	Negative Feedback	testimonials	Distrust in the testimonials	2		review sections and so on are everpresent, but I don't know if this should be added to scientific material to be honest, in this way | Yeah, I mean, the reviews, they're cute, but you could definitely throw them out.		Yeah, I would not put any trust in this because this is completely. [referring to the testimonials] | I would say even seems sketchy to me because this is not Amazon. [Referring to the testimonials]
	Positive Feedback	Information panels	Hiding of the information panels	2	And all this information down there is nice, but I also like it that we can just hide it if it is not really important and if we just want to focus on the graph for instance. | But I like the fact that you can just hide it if you don't need it.	I like the part with the arrows and that you can unfold things and put them back up together, which is good because as I've already seen | At first impression, that's. There's a lot, a lot, a lot going on, so if you can unfold. And then put it back together, that's nice for. Like deciding what you actually need right now, when once you're looking at the thing.		
	Positive Feedback	Information panels	Percieved usefulness of information panels	2		Yeah, I mean out of this like out of A2, the first four displayed [are] definitely very helpful. [information panels]	It displays everything you might need and things that you don't immediately need, let's say, are hidden behind the toggle list, which I think it's a very elegant solution. So you can open whatever information is relevant to you and whatever you're looking for	

	Positive Feedback	Preference A	More organized look of the design A	2			I would say I like this more. My first impression … and yeah … This is just my opinion not supported by any elaborate reasoning, but I would say that I like the left design more in the sense that it is all you know, you have these. That it's easier. It looks more organized just to me, yeah…	so this design I like more (A) | So I definitely like the left design better than the right one, I think.
	Concerns	experts	Being concerned that only experts will be able to use the solution	1				And so yeah, so for me it's not confusing in any way, but I'm sort of you know I'm a data producer.
Adding conferences	Concerns	experts	Experts will not use the solution	1				And so, this is, I would say very much field of expertise in the field that you know what is in the field, right. And I sort of try to have my students by explaining this to them because by searching in the archives and such you will find nothing | We are serious researchers. We meet on conferences, you know, so that is.
provide the author based keywords	Concerns	keywords	Unhelpful keywords	1				Usually, those keywords are not useful at all. | Even for me, who has to assign those labels to our own data set. I don't find those labels very helpful.
	Concerns	download	Wish to download all files including supplementary material in bulk	1				many download services and it's. Completely moronic. You cannot emerge in it instead. Of just having to download folder with all the relevant files you have to download every file.
	Concerns	design	The nature of design might not reflect the end product	1		I just want to clarify, it's not that I do not like the handwriting, it's just it's very small like in the smaller parts, but there's a lot of it. It's hard to read for like titles. In particular for like key information features, it's very clearly readable. It's not like a of course, choice or something just when there's a lot of text, I find it, it's like a bit inaccessible.		
	Concerns	accessibility	The design might not be accessible across all cultures	1		if you want different people with different like backgrounds to use this one dataset, they're always going to have like different perspectives on it and different needs so the more people you get involved in, like a display side of dataset on the display set of dataset. The more likely it is that it's going to be helpful and like understandable to people with different backgrounds.		
	Concerns	share	Usefulness of share feature	1			I'm not so sure about the share button, whether it's that useful, but I mean I can imagine a case where I would like I would find this data set and I would like to share it with a colleague that I'm working with, maybe yeah.	
	Latent Need	credibility	Need for credibility indicators	1	If that really relates to my data, and of course the source, this is really important for me I have to know who created the data and those must be serious sources basically or yeah, they must be credible.			
	Latent Need	spacing	More space for Visual story telling	1			it's a visual presentation, we can also include some kind of graphs, but we try to limit those as well so …	
	Latent Need	date	When data collection took place	1			so the data set we talked about is from a study that we've conducted in 2021	
	Latent Need	quality	Data Quality Indicators	1			if you want to publish it in a good journal and then publish the data publicly, the data should be clean and understandable for everyone.	
	Latent Need	contact	Option for contacting the authors	1				we did a lot of media stuff, so a lot of journalists called us all the time because they wanted to know about what the data are showing, we wrote a lot of blog posts to disseminate the research to the to the general audience
	Latent Need	temporal	Need for temporal trends	1				in those contexts we often included the regression analysis, panel regression analysis for example, like 60 facts regression or difference in difference modeling. Such kind of thing because this panel data structure, it lends itself to dynamic analysis to study, how the individual changes overtime, because this it is better for cause of interest compared to if you do a cross-sectional analysis. | interviewed a sample of 1500 residents in Austria aged 14 and older and we asked the same people multiple times at multiple time points, so we first we put it March 2020 and then the same person were asked multiple points in time after each other. There are so many of the response are like agree disagree. So many times we use stacked bar graph just to show the descriptive … of the distribution, for example, how it changes over time. Yeah, so we show stack bar graph for wave one for wave 2 so people can see how the distribution changes over time
	Latent Need	localization	Information about the language of dataset	1				What we are currently doing is that we have translated everything to English, so far only the prerelease was available in English language
	Latent Need	category	Type of research	1				This is really for people who have a video research-oriented agenda
	Latent Need	update frequency	Need for update frequency of update and planned support	1				this so and this kind of fund brought us through the 1st wave, but it became pretty clear in the summer, that the pandemic might not be over, that we might want to continue data collection in the fall
	Latent Need	method	Design of the study	1				So this would include uh general information regarding the sampling design, regarding who conducted the field work, how long the interviews lasted, what kind of mode the interviews were conducted in whether the respondents were paid to participate, what kind of incentives were used, yeah, who funded the study, like all this kind of question. What kind of what was the goal of the research, what kind of questions were included what kind of variables are included and what, … how representative were the data, so how well they match with demographics or objective external measures. Whether any weights are available, if the weights were … how the weights were constructed, what kind of data protection measures were applied to the data set post data collection.
	Latent Need	visualization	Need to communicate data visually	1				Yeah, we also do visualization for the panel retention, not just the number, but also, you know, to make this all a little bit more accessible and a little bit more beautiful.
	Latent Need	clarity	Clear design	1				Much of this information also needs to be included in the data set, yeah, because also many data users they read nothing. They don't read the data paper, and they don't read the method report. So also, in the data set is also very important that for example the date of the interview. | This is an environment that has not been created for the general audience or something. This has been created for.
	Latent Need	citation	Citation of the dataset	1				We also included the citation of the data set how to cite also in the data set, because many data users, they never download the document in material. They only download the data set.
	Latent Need	funding	Funding Ackowledgment	1				You know the timing the funding, so we also include the funding node and all this stuff, funding acknowledgement as variable.
	Latent Need	download	Intuitive way to download	1				when I have my students, when I teach quantitative methods in the BA, the first session I make them learn how to download data sets. We spent an entire session 90 minutes only on downloading, because you cannot imagine. What kind of hurdles are in the way. | if you can go to this website and try to find the data, yeah, the data hosting, data set data and then you have just, you know this is a link made by where the data set is this data set. You know, where are they where? The guidelines you know, where are they?
	Latent Need	discoverability	Discoverability of the dataset	1				So what is very hard to do as well is if you don't know already what the data set is about to find datasets that are just there.
	Potential Users	public	Health public information	1	if you have for instance A COVID-19 dataset, we might want to produce visualizations to inform the public | So, we have an agenda and we have a goal to actually let the people know, for instance, that there are a lot of cases in a specific country or something like that. | Then we could visualize it and give it to the public, and this would really benefit the public. I would say, to really get informed about the topic.			
	Expectation	download	One click download	1				I think the most important thing for me actually also would be that it downloads everything combined, including the documentation at once. | So in my view it should be a single click and you should get all the relevant files.
	Expectation	testimonials	testimonial-based trustworthiness	1	It also is trustworthy, if I have some kind of testimonials down there where I really see that people have already analyzed the data or worked with the data and actually produced something with this data			
	Expectation	license	License	1	I have to include the license so that they know where they can actually use the data or for what purposes.			
	Expectation	date	Publication date	1	The publication date is also really important.			
	Expectation	software	Information about survey software used	1		I mean the maybe the survey software that was used, that's something that I haven't mentioned yet in the other thing that I'm preparing, maybe that's useful to add because as I just said to you, this is something that also messes, or not messes with, but like influences the way the data is structured and so on.		
	Expectation	Preference A	Preference for hidden panels	1		So maybe when I go on this design then I would like to have something like this maybe in the beginning. With just the key information and data features open, all the other ones closed because as I already mentioned | when I go on this design or when I visit this website, then I would prefer something like this and then that I can just open it myself to not get overwhelmed all the information.		
	Expectation	abstract	Preference for keeping the abstract visible	1		Uh, I wouldn't put it here in the about I would really put it below the title to really have a short description of the dataset.		
	Expectation	interactivity	Interactive chart	1		Then you have visualized display of the dataset and where it feels like I could click on things because it seems interactive. Where you could also change the way it's visualized to maybe gain other insights		

	Expectation	interactivity	Interactive information panels	1			And just as is the case for the visualized data in the graph part, you can also interact with the like the informative part that follows beneath it and you can unfold information to decide how much you want to have displayed at a time.		
	Expectation	testimonials	moderation of testimonials	1			Maybe more elaborate, or maybe it needs to be moderated by someone.		
	Expectation	use-cases	Use-case suggestions	1			Then I would want to see the data and like the possibilities of how to visualize interpret the data and then like what follows is more this like. What else to do with it? Or like what has happened to it before these are like more … these are questions that go further than the basic usage or understanding.		
	Expectation	interactivity	Information Panel Hiding	1				I think it's called toggle buttons which will. when I would click on them, they would reveal more content	
	Expectation	Preference B	Visibility of statistics	1				And I also think that this would be visible the whole time, not when I just select something, whereas here it looks like that if I select this, then I would see the selected part and the whole data set. [in former participant refers to statistics of the scatter plot on design B2, in latter participant refers to the same in design A2]	
	Expectation	paper	Source paper for the data	1				So what is very important for me is the link to the actual paper, because sometimes when you just see the data set you, … you might not be immediately able to say what exact purpose it had, so in that case, seeing the name of the paper.	
	Expectation	funding	Who funded the paper	1				it's very important to see, for example, who funded the research because as we know, this is something that might be reflected in, let's say, the results of the study. Or what is the license, so I imagine that this license, uh information tells me whether I can reuse it and how, in what ways.	
	Expectation	Preference B	More elaborate menu	1				For A I mentioned that I would like to see right away the things that are visible on B in the upper right corner. So not just have download and share but also have the other icons such as report error or contact the author help these kinds of things I think are missing and should be visible right away in A.	
	Expectation	Keywords	Keywords describing the dataset	1					Oftentimes there are keywords some somewhere. Yeah, I don't know if you don't see any keywords here, but often there are some keywords.
	Expectation	Journal	Inofrmation about Journal, peer review	1					This is also why I say, for example, with the data paper that provides visibility because once. It is in a journal. And gone through peer review. This has credibility.
	Familiar Software	Tableau	Tableu for discovery	1	And if we just want to have a quick look at the data, so to get a notion of what is actually in the data, we could also just input it into Tableau, which we use quite often, I would say to just investigate what is actually there in the dataset.				
	Familiar Software	Qualtrics	Qualtrics for surveys	1				we had around over 400 participants I believe and it was conducted via Qualtrics. So that is a software where you put on all your questions, all the survey logic, all the stimuli.	
	Familiar Software	Familiarity	Design is has common features recognized by a data expert	1					Well, so to me this kind of dashboards and data download pages are very familiar. So, for me there's nothing to figure out.	I've seen such presentations. You can go to surveys star; you can go to UK data archive. So usually, you have a list of data sets and maybe they may have a data set explorer or something like that so I would say we see such things or similar things quite conventionally now used.
	Confusion	Visualization	Centered visualization	1		I'm wondering a bit why the visualized part follows in the middle.			
	Confusion	wording	Confused about word 'caveat' and 'watchlist'	1		Because I don't know what the watch list is, but I guess I don't need that either because I don't know what it is and then makes cause you know.			
	Confusion	icons	Confusing symbol for update frequency	1		OK. It's actually positive update frequency at the technical level. OK. I mean, yeah, just maybe the choice of this symbol. Has been confusing up until you just asked me about it because I thought something was … you know, it looks like something's restricted or forbidden symbol, yeah.			
	Confusion	citation	DOI assignment to paper or dataset	1				I'm not sure if it's the one the data set was published or the paper was published, because then we have DOI.	Is it the publication date of the paper or is it the publication date of the dataset and the same goes for the DOI. Here is it the DOI for the paper or for the data set or yeah.
	Confusion	statistics	What statistics are shown, dataset or subset, preferred B	1				So here when you have the select feature, uhm, it's not entirely clear to me in either of these designs whether I'm able to see the statistics for the whole dataset only when I select a portion of the data, or whether it's visible at all times, especially here.	Yeah, but this is something that's a bit unclear to me because, I'm not sure if I would know right away that I have the option to select some portion of the data.	then I said that I am not sure here about the select function because you see here you see selected and then the whole data set right below it. I think it makes more sense and looks more organized the way it's done in B where you have the statistics, the descriptive statistics about the whole data set here on the left and then the selected portion here on the right
	Confusion	speeddial	Unsure about the function of the speeddial	1					I was wondering with this sign, with this circle, with the plus, so [I] was not entirely clear because here it is… So, the plus, but this is to unfold all the boxes and not entirely sure
	Suggestion	preferences	Personalized preferences for the author	1	It would be nice, maybe, I don't know, if there's some user management that I can just select the things that I want to have open so that I can really personalize this website for me.				
	Suggestion	information panels	Reordering of information panels	1	Yes, I would reorder it. I would have the key information first and then right after the key information, the features. So, I would put the features here. And maybe only put those two in the first row, because for me those are the most important ones and all the other ones could be closed. [points to Key Information and Features | points to second place in the first row of the information panels]				
	Suggestion	spacing	Top and bottom padding of the visualization	1	Maybe there could be a little bit space more space here and here to not have it so aggregated or yeah comprised basically. [top and lower padding of the visualization]				
	Suggestion	axis	Axis selection next to axis	1	Uh, here we have the different axis. I'm not sure if we really need the feature selection here. It would just be nice to have maybe here uh. This dropdown menu, just put it here basically so that I can really select it on the axis. Then you don't need the boxes here. [refers to dropdowns for changing the axis label | next to chart axis]				
	Suggestion	information panels	Some information panels before visualization	1		if I was going to start working with this dataset, I think personally I would need these information that are featured below the visualization first before the graph would make sense to me. | I would put the graph beneath, let's say. OK, I'm looking at a three … beneath key information features. Yeah, beneath that I would put it.			
	Suggestion	testimonials	Moderation of testimonials	1		Maybe more elaborate, or maybe it needs to be moderated by someone. [referring to testimonials section]			
	Suggestion	abstract	No need or abstract	1		You don't need the brief description part at the beginning, beneath the name of the dataset because the key information part will probably, I mean, I haven't read it, but I'm expecting the same information there.			
	Suggestion	Preference A	Some elements from B but mostly prefered A	1		I would prefer the way it's done in A like this and then I would change it to be like it was done in B, you know. [B centered header]			
	Suggestion	spacing	More space around elements	1		And also I don't know like if I imagine this on the web page right now and like not just a sheet of paper because for me, OK, for me it's like a sheet of paper like you have an A4 sheet of paper and you just gonna use the entire space that you have like I would prefer to have it arranged in a way where maybe the header gets smaller and you have like more space on the sides and then it branches out again and you have the graphics displayed and then maybe the little squares with the informative parts they're like arranged, and maybe even more of a fun way. And you can add like a graphic next to it or something.		
	Suggestion	Description	Adding as much information as possible	1				And generally, I think that if you can add as many details about the data set as possible, which I would think is something that we would find for example here in the key information that also makes the data set, you know credible.	
	Suggestion	Method	Number of participants/size of the sample N	1				So, for example here I would like to know how many participants did they have, or how it was conducted.	
	Suggestion	Download	Bigger download button	1					I think the download could be even bigger. | your download button can even be bigger.
	Suggestion	Download	Avoid redirection when clicking on download	1					So this could also go elsewhere, because when I access it, I usually just want to download it. So I wouldn't really need any of this.
	Suggestion	Visualization	Add option to hide visualization	1					the dashboard, it could also be possibly hidden even

	Suggestion	Visualization	Start the with the signature piece of visualization	1					the dashboard should not be empty, but it really should show you like one signature piece or something you know, so that there is already a graph that looks nice. Yeah. So, to really to showcase to make it nice. | I would make. Sure, that this looks beautiful.
	Suggestion	citation	Click to export to the citation manager	1					For example data verse now puts a citation there, but at the end of the citation there's a strange number. You cannot copy it straight to a citation manager, and so on. So if you have just one button, it exports to your citation manager, that would be just such a significant improvement.
	Suggestion	similar datasets	Display status of datasets	1					Maybe this is in one of the things like the similar data, similar data set if you would just display. Little pictures of the publication with the status of that would be beautiful.
	Negative Feedback	data	Unable to locate tabulated raw data	1		I mean the only thing that irritates me is that I get like I see no option right now to see the dataset itself, because it's already visualized.			
	Negative Feedback	design	missing color	1		I miss color			
	Negative Feedback	design	Intrinsic design choices	1		I mean I guess I said a few things already and this like the visualize, this space is going to be it's going to lead you in some direction already. It's not very free and you know objective. You're … you make, making a choice already.			
	Negative Feedback	download	Overall would not use the design, only searching for download	1					this design. Would not be useful for me to, you know, so for me it would be OK. Nice. Nice download. OK. Yeah. So, yeah. So, I'm not interested.…
keep panels closed on startup	Negative Feedback	information panels	Information overload with the unfolded designs	1					So I mean with these unfolded designs, I think there's definitely too much information.
	Negative Feedback	Preference A	Dislike of the centered layout	1	Everything is in the center and this looks a bit weird to me.				
	Negative Feedback	Preference A	Dislike of abstract within the information panel	1			in the B design it's here. I would change that. [Abstract in B in an information panel below visualization]		
	Negative Feedback	Preference B	All menu options should be visible by default	1			Same with the error thingy. So in this regard I like to be one more because on the on the A1 I would need to click on the plus. And then yeah, I would see it too, but it doesn't seem entirely intuitive to me.		
	Negative Feedback	spacing	Dislike of the design compactness	1		now it feels very static and clustered like you know you had this sheet and you're just going to use all of it			
	Positive Feedback	Visualization	Selection of the data feature in scatter plot	1	And here we can also select the data feature. OK. That's nice.				
	Positive Feedback	Caveats	Important information about dataset caveats	1	We have caveats watch list. Which I find really, really important because the first already says that this dataset is partially incomplete, and this is really important to know. | I mean, I could also look at the raw data and see that there are data points missing, but it's nice to have it explicitly stated basically, in this watch list				
	Positive Feedback	Visualization	Visualization of the dataset that can be easily accessible	1	The graph really stands out for me the diagram. I really like it. I would really love to have such a website, for instance. | Where we have all the datasets in there with all this information that is actually there and to just browse through different datasets and look at the datasets more closely.				
	Positive Feedback	Preference A	Preference for justification to the left	1	And I like that it's not centered as much as in the right design B. I don't know here the space is used better I would say so, having it on the left side. It's nice.				
	Positive Feedback	Spacing	Liking the compactness	1	I really like how compact the design actually is.				
	Positive Feedback	Concentration	Handwritten style forces user to concentrate on the content	1		I don't know why that is, but like the handwritten style of it is like the first thing that captures my attention also because. It decreases the readability a little of. The text parts. And like in this description part, I will have to focus quite a bit to read what's going on. | this handwritten style is quite charming, especially in the visualization part, and like the arrows, because it makes everything a bit rougher.			
	Positive Feedback	Download	Downloading and sharing	1		It's quite nice because to know where you can download and share. It's always important.			
	Positive Feedback	Preference B	Centered header	1		I may prefer the header in B1,			
	Positive Feedback	Preference A	Design	1		and I would prefer A1 actually. Looks more clean.			
	Positive Feedback	similar datasets	Liking similar datasets	1		And there's similar datasets. That's interesting. | And … similar datasets I … if I didn't have as much work to do on this project as I have, this is something I may research possibly and add, but for time reasons I wouldn't do it right now, but I like the idea of adding this.			
	Positive Feedback	interactivity	Interactivity of the graph	1		I mean the graph is nice to play around, it's kind of inviting, so it makes sense to put it right, beneath the first descriptive part.			
	Positive Feedback	features	Overview of features information panel	1		So features, a list of features seems good. I mean, I haven't thought of this display structure in a manner where everything is going to be one page because I'm working with different files. But and an overview on features is something that … I'm doing something similar I guess, but it's not in the same structure.			
	Positive Feedback	Versioning	Versioning history information panel	1		The versioning history is something that I like which I haven't documented at all. So, in my case, and since I didn't consider this throughout working on the dataset while I see the use of it now, it's something I'm not prepared for to have, but I very much appreciate versioning history.			
	Positive Feedback	Preference B	Centered justification	1		OK, for the first page I am going to stick with B because I like the header centered and the things that give you more information arranged next to it in like a … In the symmetric.			
	Positive Feedback	Clarity	Intuitive navigation and use of the design	1		I mean, what's positive about this mockup, I feel like I would know where to click on. You know which symbols are interactive.			
	Positive Feedback	Preference B	Items of the menu visible straight away	1			OK, so this is the first thing that comes to my mind is that I really like to see these things right away, for example, contact order because this is something that you will need to if you find some data set that you would like to reuse or you have some questions about the data set, the contact author, uh icon should be right away visible. [points to design B icons at the top right] | I also like the, of course, the download and share here again, I am more inclined to the B one because it also right away includes other icons such as contact author, which I find immensely valuable and useful, as it is something that every researcher who is interested in doing something with this data will need.		
	Positive Feedback	Data Nutrition Label	Overview of the dataset with data nutrition label	1			I also really, really like that you included these in both design the ethical and technical review update frequency. These things are really important if you want to reuse data, you need to make sure that the data that you are. Trying to access were for example, collected in ethically accepted manner so that some ethical institution approved the collection of the data. | And the second thing that stands out to me are the data nutrition (labels), I think. It was called data nutrition labels, which I find very useful, and I don't think I've seen this before.		
	Positive Feedback	Visualization	Different visualizations for the same dataset	1			What I also really like is this function that I can switch between different views. | this is something really nice that will first allow me to estimate whether this data set is something that is useful for me that I can reuse.		
	Positive Feedback	Citation	Included information how to cite	1			Whereas if you want to use somebody else's data set, there is always the question how to cite it. So I really appreciate that there is the information on how to do this.		
	Positive Feedback	Information panels	Order of information panels is intuitive	1			I think in the A design it makes way more sense and, yeah, I think it makes complete sense in terms of how useful this information is so first, you have the key information, which is something that everyone is going to need.		
	Positive Feedback	download	Big download button	1					Yes, as needs to be big so I like the left design better because it's big. [referring to the size of the download button, experienced data producers don't go, searching for the dataset]
	Positive Feedback	Visualization	Interesting signature Visualization shown first to capture attention	1					So, this already seems like an interesting piece because there is a strong correlation here and so on. Yeah. So, this seems already like a showcase piece. If there's just a random actually graph that can actually turn people off. | I would say this is already quite beautiful because it is a strong correlation or something, but it would be to be picked in such a way that it's really something that showcases the data set.

A.5. User Testing Consent Form

Information for participants and declaration of consent to participate in the study:

Towards Data Showcases Facilitating Data Reuse

Dear participant,

Thank you for your interest in participating in this study.

Your participation in this study is voluntary. You can refuse to participate at any time, without having to give a reason, or also withdraw your agreement to participate once the study has already started. There will be no negative consequences for you if you refuse to participate or if you withdraw from this study early.

This kind of study is necessary to gain new, reliable academic research results. However, your consent to participate in the study is an indispensable prerequisite for us to conduct this study. Please take time to read the following information carefully in addition to the explanatory talk, and do not hesitate to ask questions.

Before you decide whether or not to participate, it is important that you understand the purpose of the study, what your participation will involve, and any potential risks or benefits associated with it. Please read the following information carefully. If you have any questions or concerns, feel free to ask the investigator before making your decision.

Please only declare your consent with participation:

- if you have fully understood the type and procedure of the study,
- if you are willing to give your consent to participate, and
- if you are aware of your rights as a participant in this study.

1. **What is the purpose of this study?**

   The goal of this study is to collect insight from scientific data producers to inform design of data showcases.

2. **What is the procedure of the study?**

   If you agree to participate in this study, you will be interviewed online over Zoom platform. In the second half of the interview, you will be asked to share your screen with sent visuals. The interview will be recorded to ensure accurate data collection and analysis. The interview will last approximately fifty minutes.

3. **What are the benefits of participating in the study?**

   By participating in this study, you help us to identify requirements for a data showcase that may facilitate efficient data reuse.

4. **What are the possible risks of taking part in this study? Could participants experience any discomfort or other side effects?**

   The researchers are not aware of any possible risks associated with participation in this study.

5. **How will the data collected in this study be used?**

   The interview will be recorded to ensure accurate data collection and analysis. However, the identities of the participants will only be known to the researchers involved in this study, who are obliged to maintain secrecy. You will never be mentioned by name, without exception.

6. **Will there be any costs for the participants? Will they receive reimbursement or remuneration?**

   You will not incur any costs from participating in this study.

7. **Possibility to discuss further questions**

   The study coordinator (the interviewer) will be happy to answer any further questions you might have about the study.

Name(s) of the contact person(s):

[redacted]

Declaration of consent

I agree to participate in the study *Towards Data Showcases Facilitating Data Reuse.*

Hynek Zemanec provided me with clear and detailed information about the objectives, significance and scope of the study, as well as about the requirements resulting from my participation in the study both orally and in writing.

I have read this information text for participants and the declaration of consent, especially section 4 (regarding risks, discomforts or side effects). Hynek Zemanec answered all my questions sufficiently and in a comprehensible manner. I had enough time to decide whether I would like to participate in this study.

I will follow the instructions that are necessary for conducting this study. However, I reserve the right to end my voluntary participation at any time, without this being to my disadvantage. If I want to withdraw from the study, I can do so at any time within 6-month period after the interview has concluded by contacting persons listed in the contact list, either in writing or verbally.

I understand that taking part in the study involves audio recording which will be transcribed, analysed and then destroyed.

I understand that I will not be directly identified in any reports of the research.
I agree that my data are permanently saved electronically in pseudonymized form. Data that have not yet been pseudonymised are stored in a form that is only accessible to the project management and are secured in accordance with current standards.

I can request my data to be deleted within 6-month period after the interview has concluded by contacting any of the person listed in the contacts and without having to give a reason.

I have read and understood the information for participants. I had the opportunity to ask all the questions I was interested in. My questions were answered fully and in a comprehensible manner.

I have received a copy of this information for participants and declaration of consent.

[redacted]
[date]

[Participant]
[date]

A.6. User Testing Session Protocol

Link	https://vida.vda.univie.ac.at/

- [ ] Before we begin, I would like to once again thank you for participating in this user testing session. I want to once again reassure you that your data will be anonymized. That includes everything that could potentially identify you, so feel free to speak freely.
- [ ] Can you confirm one more time that you have given your permission to record this session?
- [ ] To reiterate on the instructions that you have recieved in the consent form that you signed. I will send you in the moment a link to webpage. Once you do, I will ask you to share your screen and open it.
- [ ] Let us quickly verify that you can share your screen. Can you quickly share your screen?
- [ ] Everything clear so far?

	Part	#	ID	Question	Objective	success criteria	Points	Difficulty	Notes
☐	INTRO	1	IN-01	*Would you briefly introduce yourself?* *I am interested in your education and your profession and qualifications.*	demographics: age, education, profession	-	-	-	
☐	INTRO	2	IN-02	*Have you ever worked with a dataset?* *NO: What do you imagine under the term "dataset"?*	dataset profficiency, publishing dataset	-	-	-	
☐	INTRO	3	IN-03	*Have you ever used some tools that visualize data?* *YES: Which one?*	visualization literacy	-	-	-	
	Thank you for your responses, now let's proceed to the main part. [**send link** + **ask to share the screen**]								
☐	Q1	4	Q1-04	*Can you summarize your initial impressions of the page?* *Is there anything that does not makes sense?*	check if the cognitive load when coming to the page is improved	-	-	-	
☐	Q1	5	Q1-05	*Is there anything that makes you trust or distrust this dataset?*	percieved trustworthiness of the platform	-	-	-	
☐	From now on I will give you tasks to carry out. I.e. If I ask you a question I would like you to show me how you would accomplish a given task.								
☐	TG	6	TG-06	*How do you download the dataset?*	can download dataset	Download Button → Consent → Download Dataset			
☐	TG	7	TG-07	*How would you download all the available material at once?*	can download all materials	Download Button → Consent → Download Everything			
☐	TS	8	TS-08	*Can you identify the author and their institution?*	can find the author(s)	identifies **description (bottom)** or **key information** panel			
☐	TG	9	TG-09	*Where would you search for explanation of the columns in the table?*	can find the **features** panel	points to/opens **features** panel			
☐	TS	10	TS-10	*What is the description of the **shares** column / variable / feature?*	identifies **shares** variable in the **features** panel	reveals and reads **shares** variables in **Features** panel			
☐	TS	11	TS-11	***[Table]*** *In the table, can you find the column **shares**?* *Try to sort the records (individual rows) so that they are displyed from the largest to the smallest based on number of **shares**?*	can use horizontal scrolling in **table** can sort column	recognize horizontal scrolling in the table sort column by clicking on the arrow next to it			
☐	TG	12	TG-12	*Where would you search for what this dataset is about?*	can identify introduction as description	points to introduction			
☐	TS	13	TS-13	*Can you identify the method used to compile (gather) the dataset?*	can identify method in the **description**	searches in **description**			
☐	TG	14	TG-14	*If you decided to use this dataset, how would you attribute (credit) this dataset?*	can identify **cite this dataset**	finds **cite this dataset**			
☐	TS	15	TS-15	*This dataset has some shortcomings (problems). Can you identify them?*	can find **to consider** panel	finds **to consider panel**, notices **data nutrition label**			
☐	Let us turn our attantion to the tab **Discover Groups**. You can always look at the question mark symbol which tells you what the given visualization is about and what tools it has. But in the interest of time. I will point out that you can change the variables using the selects fields and it changes the data displayed.								
☐	TS	16	TS-16	***[Discover Groups]*** *How many **likes** has a **Twitter** post that has the most comments?*	can effectively use scatter plot	**3499**, uses tools demonstrate ability to make use of the tools			
☐	TS	17	TS-17	***[Group Distribution]*** *What is the **median** of **likes** on Wednesday (which you can find in Categorical Value post_day)?*	can effectively use Bar Chart	**2757** (change categorical value to **post_day**, y-axis to **median**)			
☐	TS	18	TS-18	***[Group Distribution: Single Value** enabled**]*** *What **post_id** had the most **shares**? How much?*	can effectively use **single value** mode in Bar Chart	**id:47**, **shares:993** and uses sort feature			
☐	TS	19	TS-19	***[Compare Trends 1/2]*** *Try to restrict the period of time to 30th of January until 1st of July.*	understands and can use the slider in Line Chart	correctly adjusts the time slider from **1/30/2023 13:30** to **7/1/2023 14:45**			
☐	TS	20	TS-20	***[Compare Trends 2/2]*** *At what date & time of this time interval was there a post with the most **shares**?*	understands and can use tools of Trends	**6/12/2023 9:30:00** or uses tools to narrow down the search			
☐	TS	21	TS-21	***[Explore Proportions]*** *For each **platform** (Instagram, Facebook, Twitter), what are the major subcategories of **sentimental_score** (positive, negative, neutral) when considering **comments**?*	understands and can operate Treemap	**Instagram** (**negative**): **18.2 %** **Facebook** (**positive**): **15.4 %** **Twitter** (**positive**): **12.3 %**			
☐	TG	22	TG-22	*Imagine you have your own dataset and want to create a page like this. There is an option to do it here. How would you do it?*	can find create dataset	**Floating Action Button (Menu)** → **New Dataset Button** or **Help** → **create your own dataset showcase** (link)			
☐	Q2	23	Q2-23	*Did you miss anything in the platform?*		-	-	-	
☐	Q2	24	Q2-24	*How does this platform compare to other similar tools you have used?* *Did you find the navigation of the platform intuitive?*	how similar is it to other tools	-	-	-	
☐	Q2	25	Q2-25	*How would you rate your overall experience with this tool, and why?* *Would you use it?* *In what context would you use it?*	percieved usefulness of the tool	-	-	-	