Relationships Are Complicated! An Analysis of Relationships Between Datasets on the Web

Kate Lin(✉), Tarfah Alrashed(✉), and Natasha Noy(✉)

Google Research, Google, San Francisco, USA
{kateslin,tarfah,noy}@google.com

Abstract. The Web today has millions of datasets, and the number of datasets continues to grow at a rapid pace. These datasets are not standalone entities; rather, they are intricately connected through complex relationships. Semantic relationships between datasets provide critical insights for research and decision-making processes. In this paper, we study dataset relationships from the perspective of users who discover, use, and share datasets on the Web: what relationships are important for different tasks? What contextual information might users want to know? We first present a comprehensive taxonomy of relationships between datasets on the Web and map these relationships to user tasks performed during dataset discovery. We develop a series of methods to identify these relationships and compare their performance on a large corpus of datasets generated from Web pages with `schema.org` markup. We demonstrate that machine-learning based methods that use dataset metadata achieve multi-class classification accuracy of 90%. Finally, we highlight gaps in available semantic markup for datasets and discuss how incorporating comprehensive semantics can facilitate the identification of dataset relationships. By providing a comprehensive overview of dataset relationships at scale, this paper sets a benchmark for future research.

Keywords: Web datasets · dataset relationships · semantic markup

1 Introduction

As the world becomes increasingly data-driven, researchers rely on open data to answer scientific questions and to understand complex phenomena [45]. This reliance on data has led to dataset publication becoming the norm in many scientific disciplines [39]. Unlike scientific publications, however, datasets are not static and standalone entities: dataset providers publish new versions of datasets as the data evolves, researchers produce new datasets by combining existing datasets, and meaningful subsets of large datasets may gain a life of their own. When a user chooses a dataset for her work, these distinctions become critical. For instance, when reproducing results from a publication, we must identify which specific dataset snapshot the authors used. When evaluating the

G. Demartini et al. (Eds.): ISWC 2024, LNCS 15231, pp. 47–66, 2025.
https://doi.org/10.1007/978-3-031-77844-5_3

trustworthiness of a dataset available on multiple platforms, users may want to choose the repository that they trust the most. If a scientist wants to compare slices of a large dataset, she wants to ascertain that these slices come from the same snapshot of the larger dataset. Therefore, understanding the semantics of relationships between datasets can be just as important as understanding other metadata about them.

How can we identify relationships between datasets? The most straightforward method is to look at information provided by dataset publishers. The industry standard for describing semantic metadata for any Web content (including datasets) is through structured markup in `schema.org` [16]. Standards like `schema.org` and W3C DCAT [2] provide means to identify pages containing datasets as well as the semantics of relationships between them. In addition to these Web standards, approaches such as datasheets [14] provide mechanisms to describe dataset origins, biases, and recommended usage. These methods of providing additional context for datasets enable publishers to link versions of the same dataset to one another, link a dataset in one repository to the original dataset in another repository, or to declare that one dataset is based on another. However, semantic markup is often unreliable and incomplete, [4,19,27] and only a small fraction of dataset metadata on the Web contains values for properties that link them to other datasets [8]. Furthermore, dataset authors often update or restructure datasets without providing notice or documentation [20,41,43]. Finally, current markup frameworks do not fully capture the variety and nuance of dataset relationships.

To illustrate the richness of dataset relationships, consider the collection of datasets provided by the United States Census Bureau. This collection captures a wide variety of measures (e.g., income data) over many decades at various levels of granularity, such as national, statewide, county-wide, and so on. On the Web, one can find various slices of this large dataset that may be relevant in a specific context: for example, there may be a dataset containing income data for California in 2008 or a dataset containing income data for the entire US in the same year. Each of these datasets is a subset of the larger 2008 Census dataset, but researchers may publish them in different Web sites and contexts. Figure 1 presents another example—a collection of datasets published by the US National Oceanic and Atmospheric Administration (NOAA). We have found more than 140 forms of this dataset on the Web, including annual, monthly, and daily sets. Many variations have multiple versions, and some of these datasets are subsets of larger ones. Additionally, many of the datasets in this collection are replicated across various Web sites. Critically, the information that helps us understand these relationships often must come from *metadata*: the data itself may not have enough context for us to understand the provenance or coverage of the datasets.

In this paper, we explore the relationships between datasets from the perspective of users who want to discover and analyze datasets. Rather than define these relationships in an abstract way, we take a user-centric view and ground the relationships in user tasks performed during dataset discovery. We analyze a large dataset corpus, generated from dataset pages on the Web with

`schema.org/Dataset` markup, to identify these relationships between datasets. Our evaluation focuses specifically on provenance-based relationships, which we can infer from metadata. There is a large body of work (Sect. 2) that infers relationships from data. Our focus on metadata and specifically on provenance-based relationships complements related work.

Specifically, we make the following contributions in this paper:

- We define a taxonomy of relationships between datasets on the Web and ground it in essential user tasks that rely on understanding these relationships (Sects. 3 and 4). To our knowledge, this taxonomy is the most comprehensive in the literature to date.
- We propose and compare several methods for identifying dataset relationships (Sects. 5). We show that machine-learning methods that use dataset metadata achieve a multi-class classification accuracy of 90%, outperforming `schema.org` and heuristics-based methods (Sect. 6).
- We analyze the prevalence of provenance-based relationships in a corpus of 2.7 million datasets on the Web. We found that 20% of datasets have at least one relationship with another dataset (Sect. 6).
- We present recommendations for enhancing dataset metadata, facilitating the discovery of more relationships between datasets (Sect. 7).
- We publish a collection of 2.7 million dataset pages with their basic metadata, along with the connections and interrelationships among these datasets.[1]

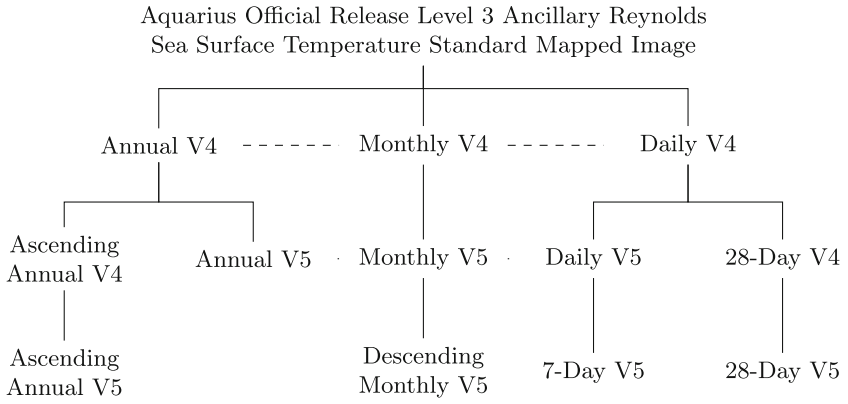

Fig. 1. The *"Aquarius Official Release Level 3 Ancillary Reynolds Sea Surface Temperature Standard Mapped Image"* dataset has annual, monthly, and daily variants with multiple versions. Variants are derived from each other and can have different replicas (e.g., *"Annual V4"* on three sites) and reconfigurations (ascending, descending). (See, for example, dozens of versions of this dataset in Google Dataset Search: https://bit.ly/3IrdEb2)

[1] https://figshare.com/articles/dataset/Metadata_for_Datasets_and_Relationships/22790810.

2 Related Work

Our research focuses on relationships between datasets on the Web. In this section, we examine approaches to tackling this topic, including exploring dataset provenance, understanding dataset evolution, identifying conceptually similar or joinable datasets, and linking datasets to scientific publications.

Research reproducibility benefits from understanding dataset provenance and relationships between datasets. Herschel et al. [18] proposed user-focused methods for capturing and analyzing provenance, emphasizing relationship presentation. Klump et al. [22] introduced a contextual framework and versioning principles surpassing simple revisions. Rauber et al. [33] identified subsets of large datasets supporting research findings, while Silvello [37] detailed citing linked open data subsets. Rauber et al. [32] stressed the importance of exact version citation for research reproducibility.

Semantic vocabularies like `schema.org` [16], DCAT [2], and VoID [3] enable dataset providers to specify relationships between datasets (e.g., `schema.org/isBasedOn` for derivations, `schema.org/sameAs` and `void:Linkset` for replicas, DCAT's `dct:isVersionOf` for versioning). However, these standards lack a comprehensive vocabulary for diverse dataset relationships. The PROV Data Model [6] offers a standardized framework for describing provenance information between entities, including `wasDerivedFrom` for dataset derivation. Indeed the PROV notion of derivation can form the basis for some of the relationships that we discuss later. However, because datasets and their relationships are not the focus of PROV, it does not capture many of the dataset-specific relationships, such as subsets, slices, or datasets that can be integrated.

Research in data evolution addresses challenges of detecting, tracking, and explaining dataset changes over time. Roussakis et al. [35] introduced a flexible framework for dynamic dataset analysis, offering various granularity levels and rich visualizations. Umbrich et al. [41] quantified change frequency in linked open data (LOD) for improved understanding of dataset dynamics. Shraga and Miller [36] proposed a semantic data versioning method using explanations to aid users in comprehending dataset changes.

Other LOD research focuses on recommending interconnected datasets. Lopes et al. [30] proposed collaborative filtering and content-based methods using dataset metadata, both validated through real-world evaluation. Leme et al. [25] suggested using core datasets for interlinking recommendations, demonstrating effectiveness on a large-scale dataset. Ellefi et al. [7] introduced an intensional approach based on conceptual similarity. The LOD research focuses on linking entities within datasets [1,9,13,42], complementary to our focus on complete datasets rather than individual datapoints.

In addition to finding conceptually similar datasets, researchers emphasize identifying joinable datasets to enrich information without foreign keys [12]. Zhu et al. introduced Josie to find joinable tables using overlap set similarity search [44]. Dong et al. developed a clustering-based method for grouping joinable tables [10]. Recent work includes DeepJoin, a deep learning model for efficient joinable table discovery [11]. These efforts streamline dataset discovery for data analysts.

Efforts to link datasets with scientific publications are significant. The Research Graph dataset [5] consolidates publication, dataset, and software details into a unified knowledge graph, simplifying research output discovery. Ayush et al. [38] applied natural language processing to extract data-related information from publications and match it to online datasets. Google Dataset Search [28] connects datasets with scholarly articles referencing them, enhancing scientific discovery by simplifying dataset finding for researchers.

The highlighted research emphasizes the importance of dataset relationships in scientific research. Our aim is to offer a broader framework for future research by identifying more relationships, evaluating their impact, and thus advancing the development of better dataset management tools and techniques.

3 Grounding Dataset Relationships in User Tasks

The data ecosystem relies on a continuous cycle of discovery, analysis, and sharing, necessitating a nuanced understanding of dataset relationships. To ground our taxonomy of dataset relationship, we begin by defining the tasks users undertake during data discovery and sharing.

Finding Datasets: The proliferation of data on the Web has complicated dataset discovery through traditional search engines [23]. Users face challenges in sorting through vast amounts of data and navigating diverse search criteria, especially when their intent varies. For instance, reproducing an experiment from a paper demands sorting through dataset versions, formats, and sources [22], while augmenting a dataset requires finding compatible data in structure, schema, and topic. Understanding dataset relationships facilitates efficient and accurate dataset selection, aiding in reliable data-driven decision-making.

Evaluating Dataset Trustworthiness: Evaluating whether to use a dataset involves an assessment of dataset trustworthiness [15,18]. Unlike research publications, datasets published on the Web rarely undergo peer review. As a result, users must rely on dataset attributes and metadata as proxies for dataset trustworthiness. User-experience studies reveal users weigh data provider, format, prior usage, and update frequency [15]. Hence, comprehending dataset provenance, identifying citations, and locating reliable sources aid in evaluation.

Citing and Referencing Datasets: Noy et al. [29] note that well-described datasets drive new research. Citing datasets like papers encourages better data collection and curation [18]. Proper citation requires persistent identifiers, metadata, and accurate provenance descriptions, including version number, source, and whether it is a subset of another dataset [22]. Proper dataset identification promotes transparency and collaboration, enhancing research quality.

Curating Datasets: Dataset curation involves collecting, organizing, and maintaining datasets from diverse sources to ensure availability for users. The goal is to create high-quality datasets beneficial for researchers, developers, and users. Curators must understand a dataset's relationships with others, including versions, replicas, and usage in research or projects [14].

While the list of user tasks in this section is by no means exhaustive, it demonstrates the need for improving our understanding of the relationships between datasets, capturing these relationships explicitly, and using these relationships to improve researchers' experiences with data.

4 Defining Dataset Relationships

We base our categorization of dataset relationships on the analysis of user tasks in dataset discovery (Sect. 3), prior research (Sect. 2), and observations from analyzing a large corpus of datasets on the Web. Specifically, we collected a corpus of datasets by relying on `schema.org/Dataset` markup (Sect. 6.1).

We broadly group dataset relationships on the Web into provenance-based and non-provenance-based relationships. *Provenance-based relationships* are relationships between the datasets that share a common original dataset, such as being derived or modified versions of the same original dataset. *Non-provenance-based relationships* involve connections between datasets based on content, topic, or task rather than their origin. For each relationship, we highlight which of the tasks from Sect. 3 it is particularly useful for.

4.1 Provenance-Based Relationships

We define a dataset on the web $D = (P, O, S, W)$ as consisting of a set of data points P, origin of dataset O (i.e., primary data-collection process), schema S, and a web site that hosts the dataset W. Note that in the context of dataset discovery, we may not have complete information about a given dataset and, in particular, may not know origin or schema.

We define schema S for the most common dataset types in our corpus. Relational (tabular) datasets are characterized by a header row with column names, data types, and constraints such as primary keys. Document datasets, such as PDFs, are defined by collections of key-value pairs, arrays, or nested documents. Structured datasets, like JSON, are defined by the structure and properties of elements, including data types, relationships, and constraints. RDF data is defined by RDF Schema.

Replica: Datasets $D_1 = (P_1, O_1, S_1, W_1)$ and $D_2 = (P_2, O_2, S_2, W_2)$ are *replicas* of each other iff their underlying data and origin are identical, but they are hosted on different sites: $P_1 = P_2$, $O_1 = O_2$, $S_1 = S_2$, and $W_1 \neq W_2$.

In today's Web data ecosystem, a common pattern is one repository aggregating datasets from multiple other repositories. For instance, the European data portal (europa.eu) offers access to datasets from member state data portals. In the United States, an open data site for a state government might include datasets from county data repositories, which in turn may include datasets from individual towns. Therefore, datasets from local governments (e.g., county) are *replicated* in the state repository. Identifying and grouping replicas of datasets in a dataset-discovery context helps users easily locate datasets and provides

choices of sources. It also enables users to obtain data from their most trusted source when available from multiple sources. For example, a user may trust their local government site more and opt to retrieve the dataset from there.

Version and Revision: Datasets $D_1 = (P_1, O_1, S_1, W_1)$ and $D_2 = (P_2, O_2, S_2, W_2)$ are *versions* of each other iff $P_1 \approx P_2$, $O_1 = O_2$, $S_1 \approx S_2$. W_1 may or may not be the same as W_2. This relationship captures evolution of a dataset over time, where changes between subsequent versions are usually relatively small to the size of the dataset.

Published datasets resemble software more than research publications, as they continue evolving after release. Evolution can range from error corrections and data adjustments to continuous updates with new observations. The Research Data Alliance Data Versioning Working Group stipulates that any alteration to a dataset forms a new version that authors must identify, encompassing minor changes such as data additions/removals [32]. However, authors often label only significant checkpoints as new versions, thus, we refer to stable, labeled releases as versions. Minor, unlabeled changes are revisions. For instance, if a dataset covers sales data from January to November and is updated with December data and give it a new label, they create a new version of a dataset. Converting Fahrenheit to Celsius in a weather dataset is typically a revision, not a new version. Identifying all dataset versions is crucial for research reproducibility, data quality assessment, and maintaining transparency and proper attribution when citing the data source.

Subset: A dataset $D' = (P', O', S', W')$ is a *subset (or slice)* of $D = (P, O, S, W)$ iff $P' \subset P$, $O' = O$, $S' \subset S$. W' may or may not be the same as W. There is usually an extraction function $F(x)$ that determines which data points from P are in P': $P' = \{x \in P | F(x) = true\}$

A dataset subset is a smaller, more focused set of data extracted from a larger dataset, published independently. The subset typically contains data selected based on specific criteria, like time period, geographical region, or variables. For example, a dataset containing weather information for a country may have subsets for specific regions or time periods. Researchers using a subset of a dataset in their work benefit from transparency, accuracy, and appropriate attribution.

Note that there is a complementary **superset** relationship. For simplicity, we refer only to the subset relationship in the paper.

Derivation: A dataset $D' = (P', O', S', W')$ is a *derivation* from a collection of datasets $\Delta = \{D_1, \ldots, D_n\}$ iff there exists a derivation function $M(x_1, \ldots, x_m)$ that transforms, combines, or otherwise manipulates data points for datasets in Δ. Thus, O' is different from $O_1, ..., O_n$. Schemas and web sites may or may not be the same.

A healthy data ecosystem enables users to build new datasets from published ones. A dataset can be derived from one or more datasets as a result of transforming, aggregating, or otherwise manipulating existing datasets. Examples of

derived datasets include summaries, aggregations of multiple datasets, and variables created by combining or transforming existing variables. Understanding which datasets served as input for a given dataset can help users evaluate trustworthiness of datasets and understand whether a given dataset has properties they are looking for. For dataset curation in particular, it is usually not sufficient to specify a dataset is derived from another dataset; rather, dataset authors must provide details on changes and modifications to ensure users understand implications and could reproduce the dataset[2].

Variant: Datasets $D_1 = (P_1, O_1, S_1, W_1)$ and $D_2 = (P_2, O_2, S_2, W_2)$ are *variants* of each other iff $P_1 \cap P2 \approx \emptyset$, $O_1 = O_2$, $S_1 = S_2$. W_1 may or may not be the same as W_2.

Consider two weather datasets covering different regions of the country and different years; the two datasets use the same schema and were collected in the same way. Teams in these regions may have collected these datasets independently or may have generated them by creating subsets of a larger dataset. The variant relationship captures the link between these two "sibling" datasets. Formally, two datasets are variants of each other if they share the schema, origin, and collection methods but differ in coverage along some dimension. This dimension is often temporal or spatial: the same statistics may be collected and published annually over multiple years, for example. These annual datasets would be variants of each other. Identifying dataset variants allows users to compare and analyze datasets to identify patterns and trends that may be obscured if we look only at one dataset in isolation. By comparing variants of datasets, users gain a more comprehensive understanding of the phenomena that the data represents and make more informed decisions.

4.2 Non-provenance-Based Relationships

With the wealth of datasets on the Web, users can gain useful insights from serendipitous relationships between datasets. These post-hoc relationships can be based on metadata, dataset usage, or similarity in content. Many of these relationships are context-dependent: a specific dimension makes sense in the context of a specific user task. We define several such relationships in this section. This list is not exhaustive as we cannot predict all possible uses of datasets.

Topically Similar: Datasets D_1 and D_2 are *topically similar* iff their topical similarity score exceeds a defined threshold θ: $Sim(D_1, D_2) \geq \theta$, where Sim represents a similarity function that yields a value within the range of 0 (indicating no similarity) to 1 (reflecting perfect similarity). The choice of threshold θ and

[2] Technically, we can consider a subset to be a derived dataset. We distinguish between the two relationships in our taxonomy; for derived datasets, their authors applied some processing or analysis to the data, while for subsets they simply selected the data from an existing dataset.

similarity function Sim relies on the context and desired similarity level. The function $Sim(D_1, D_2)$ can be implemented using methods like cosine similarity, Jaccard index, or others, depending on the nature of the datasets at hand.

Topically similar datasets cover the same subject or capture similar topics along the dimensions relevant to the user context. For instance, datasets covering ocean temperature and salinity can be topically similar when understanding the effects of climate on the oceans. In a different context, a dataset of stock prices might be compared to a benchmark index to assess performance; thus these two datasets would be topically similar.

Task-Similar: Datasets D_1 and D_2 are *task-similar* if $Sim(T(D_1), T(D_2)) \geq \theta$, where $T(D_1)$ and $T(D_2)$ are the tasks for which the two datasets have been designed, and Sim is a similarity function that yields values between 0 (indicating no similarity) and 1 (reflecting perfect similarity). The selection of the threshold θ and similarity function Sim depends on the specific context and the degree of similarity one aims to capture.

Dataset metadata may include not only intrinsic properties of a dataset but also a collection of tasks that a dataset may be best suited for or that dataset creators had in mind. Task-similar datasets share similarities in the tasks or problems they are used for. Datasets created for similar tasks allow for comparison and benchmarking of different algorithms or models, aiding in the evaluation and selection of the best model for a given task. For example, Human3.6M and KITTI datasets are each used for video prediction, yet their subjects vary drastically: humans versus cars.

Integratable: Datasets D_1 and D_2 are *integratable* if they share schema or content enabling the integration. Integratable datasets, D_1 and D_2, are *joinable* when their attribute sets have a non-empty intersection, serving as common attributes or foreign keys: $A(D_1) \cap A(D_2) \neq \emptyset$, where A represents the attribute set (data fields or columns) in the dataset. Integratable datasets, D_1 and D_2, are *unionable* when their attribute sets have a non-empty intersection $A(D_1) \cap A(D_2) \neq \emptyset$, and they share similar schemas $S_1 \approx S_2$, and the overlap between their data points is insignificant: $P_1 \cap P_2 \approx \emptyset$.

Integratable datasets share schema or content, allowing their combination. Datasets are *joinable* if they share common attributes or foreign keys, like traffic patterns and accident reports linked by location and time. Datasets are *unionable* if they capture similar data about complementary concepts, such as weather patterns in different cities. Unionable datasets differ from variants as they do not have the same schema or collection methodology.

4.3 Discussion

A dataset can have *multiple relationships* with other datasets. For instance, a national dataset of education statistics with state-level information can be a

source for multiple state-specific datasets. The state-specific datasets are *variants* of each other, but all the state-specific datasets are the *subsets* of the national dataset. Two datasets can have multiple relationships with each other. For example, a Monthly dataset and a Daily dataset in Fig. 1 are *variants* of each other. However, if the Monthly dataset was created by aggregating the Daily datasets, it is also *derived from* the Daily datasets.

Certain relationships are *bidirectional*, while others are *directional*. For instance, if datasets D_1 and D_2 are replicas, then D_1 is a replica of D_2 and D_2 is also a replica of D_1 (bidirectional $D_1 \leftrightarrow D_2$). However, if D_1 is a subset of or derived from D_2, it implies that $D_1 \rightarrow D_2$ is true, but not $D_2 \rightarrow D_1$.

Dataset providers can explicitly capture some relationships that we identify in this section. Specifically, `schema.org` supports two of these relationships: *replica* (`schema.org/sameAs`) and *derivation* (`schema.org/isBasedOn`). For other provenance-based relationships, we can rely on analyzing metadata (e.g., *versions*, *variants*) or the data itself (e.g., *integratable* datasets). *Topical and task similarity* relationships depend on the user context. Because `schema.org` metadata is not always reliable [4,19,27], combining it with metadata analysis helps identify relationships.

In our empirical analysis, we prioritize metadata, deferring relationships reliant on data or user context for future exploration. Moreover, extensive research exists on deriving dataset relationships directly from the data itself [10,21,44].

5 Empirical Analysis Methods

Earlier research has shown that `schema.org` metadata is not always reliable [4,19,27]. Furthermore, the markup for dataset relationships is extremely incomplete [28]. Thus, in addition to extracting relationships from `schema.org` markup, we propose a series of automatic approaches to infer dataset relationships. We focus specifically on evaluating the value of metadata (not data); thus, we concentrate on provenance-based relationships.

We evaluate four methods: First, we extract relationships directly using `schema.org`. Second, we develop a set of heuristics tailored to each relationship type. Heuristics-based approaches are usually efficient to implement. Finally, we propose two machine-learning-based approaches: a classical ML approach consisting of a gradient boosted decision trees classifier and a generative AI approach using a LLM-based classifier. Each of these models represents a larger class of methods that can be used to tackle this problem setting. Section 6 compares the accuracy of these approaches on a large ground-truth set.

5.1 Semantic Markup Analysis

We use the `schema.org` relationships that metadata explicitly captures: *replica* (`schema.org/sameAs`) and *derivation* (`schema.org/isBasedOn`). We consider datasets A and B to be replicas if one contains the DOI or URL of the other

in its `sameAs` field. Dataset A is considered derived from dataset B if dataset A contains the DOI or URL of dataset B in its `isBasedOn` field.

5.2 Heuristics-Based Methods

We define a set of heuristics based on regularities observed by analyzing a large corpus of metadata for datasets on the Web. All comparisons in this section use normalized names and descriptions; specific rules are tailored to each relationship type. Two datasets are **replicas** if their normalized names and descriptions are exact matches or one is a non-trivial prefix of the other. Two datasets are **versions** if their normalized names are the same except for the version number, which we extract using a regular expression. Two datasets are **variants** if their normalized names are the same except for months or dates, which we extract using a regular expression. Two datasets are also variants if their prefixes before a common delimiter are the same, but suffixes are different. Here we consider two dimensions for variants: temporal and spatial.

Directional relationships require complex rules to identify. We identify **subsets** by splitting dataset names into prefixes and suffixes based on a common delimiter. Dataset A is a subset of Dataset B if the name prefix of A and B are exact matches and only Dataset A has a suffix. Additionally, Dataset A is a subset of Dataset B if the two dataset names are the same after extracting a year or month from Dataset A but not from Dataset B. To identify the **derived** relationship, we use observed text patterns in the corpus, such as "analysis of." Dataset A is derived from Dataset B if the processed names are the same after removing these patterns from Dataset A but not B. Table 1 shows an example of one of the heuristics used to identify the subset relationship.

Table 1. An example of a heuristic to discover subsets in our corpus. We consider D_a to be a subset of D_b

ID	Dataset Name	Prefix	Suffix
D_a	"Survey of Earned Doctorates - 2019"	"Survey of Earned Doctorates"	"2019"
D_b	"Survey of Earned Doctorates"	"Survey of Earned Doctorates"	""

5.3 Gradient Boosted Decision Trees Based Classification

We used the ydf-implementation of GradientBoostedTreesLearner in TensorFlow2 [17] to train a GBDT-based multi-class classifier using manually annotated examples described in Sect. 6.1. We trained the model with a batch size of 128, a local growth method to optimize a cross entropy loss function, a random sampling method, a maximum depth of 4, a sparse oblique splitting method [40] and a shrinkage of 0.0887 as set by a hyperparameter sweep.

5.4 LLM-Based Classification

We fine-tuned the t5x [34] implementation of the T5.1.1 large-language model [31] to perform a multi-class classification task using the same manually annotated training examples described in Sect. 5.3. We used a batch size of 64 with 1,050,000 training steps and a learning rate of $1e-3$ set by a hyperparameter sweep.

6 Evaluation and Results

To understand which approach works best in practice, we compared the performance of the four methods from the previous section on manually annotated ground truth data. We then apply the best-performing method to a large corpus of datasets on the Web in order to understand the prevalence of different provenance relationships between datasets.

6.1 Training and Evaluation Data

Dataset Corpus. We generated a corpus of dataset metadata by crawling the Web to find pages with `schema.org` metadata indicating that the page contains a dataset. We then selected a subset of citable datasets: we categorize a dataset as *citable* if it has a persistent de-referencible identifier, like a digital object identifier (DOI). This corpus includes 2.7 million dataset-metadata entries.

Ground Truth. To generate ground truth for training and evaluation, the paper authors manually labeled 2,178 dataset pairs. The labelers had access to all metadata fields for these datasets, not just their names and descriptions.

We observed that some relationships (e.g., replica) are much more common than others (e.g., subset). Thus, our goal when generating a set of pairs for manual labeling was to ensure that this set likely had examples of all the relationships. We used the following procedure: We randomly sampled 125 *seed datasets* from the corpus. Each seed dataset must have between 1 and 10 replicas, as identified by a heuristic-based method (Sect. 5). Note that we required that a dataset *has* a replica but not that the replica is included in the sample. We observed that datasets with replicas are also likely to have other related datasets. Seed datasets also had to come from hosts with more than 30 datasets. We limited seed datasets to a maximum of 2 datasets per host to ensure diversity of the set. We projected the metadata of the datasets in our sample and our corpus into the NewsEmbed [26] embedding space (an embedding space that we found worked particularly well for datasets). Given a seed dataset S and its 20 nearest neighbors $\mathcal{D} = \{D_1, ..., D_{20}\}$, we label manually each relationship $\{\langle S, D_1\rangle, \ldots, \langle S, D_{20}\rangle\}$ and a random sample of 20 relationships between $\langle D_x, D_y\rangle$ where $D_x \in \mathcal{D}$ and $D_y \in \mathcal{D}$.

The number of subset and derived relationships in a random sample was still very low. Typical data interpolation techniques were not possible because subset

and derived relationships are non-reflexive. To address the sparse label space, we added government datasets used in Show US the Data Kaggle Competition [24]. We observed that these datasets were more likely to have subsets or derivations.

For machine-learning based methods, we used 70:15:15 split for training, validation, and evaluation data. For other methods, we used the same evaluation data as for the ML-based methods (the 15% of the labeled pairs).

6.2 Results

Table 2 presents the comparison of evaluating the four methods from Sect. 5. Our experiments find that `schema.org` metadata alone is insufficient for identifying relationships between datasets, even for the two types for which `schema.org` exist (replica and derived): Indeed, *no* pairs of datasets in our random sample had an explicit relationship defined between them.

Table 2. Precision (P), recall (R), and F1 scores for each method and relationship type. The scores for the method with the top F1 score for each relationship is bolded.

Relationship	Schema.org			Heuristics			GBDT			T5		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Replica	0.00	0.00	0.00	1.00	0.35	0.51	**0.97**	**0.95**	**0.96**	0.92	0.92	0.92
Version	N/A	N/A	N/A	0.96	0.53	0.68	0.92	0.80	0.86	**0.87**	**0.87**	**0.87**
Subset	N/A	N/A	N/A	0.49	1.00	0.65	**1.00**	**0.80**	**0.89**	0.82	0.90	0.86
Derived	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.33	0.44	**0.80**	**0.67**	**0.73**
Variant	N/A	N/A	N/A	1.00	0.50	0.67	**0.90**	**0.85**	**0.87**	**0.81**	**0.93**	**0.87**
None	0.33	1.00	0.49	**1.00**	**0.80**	**0.89**	**0.85**	**0.93**	**0.89**	**0.94**	**0.85**	**0.89**

Heuristics-based methods perform reasonably well for certain relationship types, such as *none* (i.e., there is no relationship between two datasets) and *version*. These methods often have low recall because they are brittle: small perturbations in names or descriptions of datasets significantly affect their performance. The GBDT Classifier and T5-Based Classifier both perform quite well and have similar F1 scores for all relationships except *derived*. For the *derived* relationship, which is more semantically complex, the T5-Based classifier outperforms the GBDT Classifier.

Figure 2 compares the overall accuracy of the methods. We opt to calculate accuracy as opposed to macro-average precision and recall because there exists a large imbalance between the prevalence of each relationship on the Web.

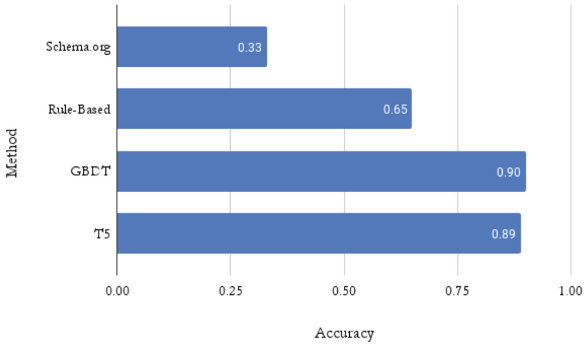

Fig. 2. The overall accuracy for each method type.

6.3 Corpus-Level Analysis

We analyzed the corpus of 2.7 million citable datasets (Sect. 6.1) using the GBDT classifier because it had the highest overall accuracy. Because classifying all of the $O(N^2)$ pairs of datasets is extremely expensive computationally, we clustered datasets first and then classified pairs within each cluster. Specifically, we used the NewsEmbed embedding space and classified a relationship between each dataset and its 20 nearest neighbors. This process gave us 42.4 million unique dataset pairs for the GBDT classifier to classify.

Out of 2.7M datasets in the corpus, 20.1% had at least one relationship with another dataset, and 22% had multiple relationships. Table 3 shows the distribution of the identified relationships, with the *replica* relation being the most prevalent. Only a handful of datasets had *derived* or *version* relation.

Table 3. The distribution of relationship in the corpus.

Relationship Type	Percentage
Replica	77.8%
Subset	14.2%
Variant	5.2%
Derived	1.8%
Version	1.0%

Datasets in each replica pair come from different sites by definition. In order to understand the publishing ecosystem better, we looked at whether or not pairs of datasets in other relationships come from the same repository or tend to be distributed. Out of 61,485 dataset subsets, 59% are from the same site as the parent dataset; the rest are from different sites. Conversely, the majority of variants, derived datasets, and versions exist on the same site: 79% of variants, 83% of derived datasets, and 97% of versions.

7 Discussion and Future Work

Our categorization of dataset relationships and corpus analysis highlighted both the complexity of and the need for identifying these relationships.

7.1 Relationships Are, Indeed, Complicated

Several previous works categorized dataset relationships (see Sect. 2). Our categorization in Sect. 4 complements this effort. Critically, the user-centric approach gave us a unique view. Consider, for example, the *subset* and the *integrated* relationships. In both cases, datasets share some of the content: with the subset relationship, they share some or parts of records. For datasets to be joinable, they must share a subset of foreign keys. However, if we approach the distinction from a user point of view, these two relationships are quite different. Usually we seek *subset* relationships for a manageable dataset slice, whereas for joinable datasets, we aim to expand existing data for new insights.

The user-centric view also helps us decide whether to include specific relationships in our taxonomy. For instance, we have not found the *replica* relationship in other categorizations. However, in the context of Web-based dataset discovery, users need to understand the relationships between original sources of datasets and repositories that aggregate datasets from multiple sources. Users then can choose a site that they find reliable and trustworthy to download the dataset.

Having the grounding in user tasks did not eliminate the need for difficult decisions. For example, determining topic similarity between datasets raises questions. While two datasets covering different weather aspects in the same region for the same timespan are clearly related, the situation differs with education outcome datasets. If one covers high schools and the other elementary schools, they may be considered topically similar only if the user's interest spans all school education. Topic similarity varies based on user task granularity.

In Google Dataset Search, we use several of the relationships from Sect. 4 directly in the tool. In dataset results, we group together *replicas* of a dataset giving the user an option to get the dataset from their preferred site. We also group versions and variants in order to simplify navigation and show the larger diversity of search results.

7.2 Semantic Markup for Dataset Relationships

Our analysis found that over 20% of datasets have at least one relationship with another dataset. These relationships are not captured by `schema.org` metadata. While some researchers noted the low quality of semantic markup (e.g., [4]), we found it to be extremely incomplete. If `schema.org/sameAs` relationships were accurate and complete, our replica-identification mechanism would not be needed. The metadata connecting datasets is often incomplete or inaccurate for several reasons. First, the community can improve and expand `schema.org/Dataset` properties, as some definitions are vague. For example, `isBasedOn` can capture subset, revision, or version relationships. `Schema.org`

also lacks relationships like linking to a previous version. Second, our analysis, supported by [18,22,45], highlights the need for data sharing best practices, including publishing datasets with digital object identifiers, linking datasets to papers, and capturing provenance information in metadata. Additionally, the lack of tools utilizing metadata may deter authors from providing accurate metadata; developing such tools could incentivize authors to improve metadata accuracy.

7.3 Limitations in Future Work

We analyzed millions of datasets, optimizing the process to find pairs of datasets to label. The ground truth datasets were limited to pairs that are neighbors in an embedding space, focusing on datasets with similar names and descriptions. Our methods may miss relationships when dataset names change significantly, although minor changes like adding acronyms should not be affected.

We used a few metadata fields to infer relationships, with future analysis planned to explore the impact of fields like authors, providers, and explicit temporal or spatial coverage values.

We derived our relationships using the corpus of datasets on the Web. These relationships likely paint a different picture for datasets in specific dataset repositories (e.g., Figshare, Zenodo), and also among the datasets that constitute the linked open data (LOD) cloud. Understanding how these relationships apply to the datasets in LOD will enable us to highlight similarities and differences between linked data and more traditional data.

Identifying provenance-based relationships lays the groundwork to studying data quality and trustworthiness of data changes on the Web, aiding users in finding reliable data sources and identifying information gaps.

Finally, understanding non-provenance relationships is the first step in helping users find the right data for their tasks. Users often seek data to complete specific tasks (e.g., training an ML model for weather prediction). Understanding dataset relationships helps us better assist users in finding the necessary data.

8 Conclusion

Understanding relationships between datasets is crucial for extracting valuable insights that can drive innovation and positively impact various domains. Using even simple analysis methods, we can see that datasets on the Web are connected in many different ways. However, while this analysis can help in identifying some of the relationships, research communities must develop best practices that encourage dataset authors to specify metadata. Overall, our paper sets a benchmark for future research and highlights the importance of understanding dataset relationships for scientific research and decision-making processes.

Supplemental Material Statement: We publish metadata and dataset relationships for the collection of 2.7 million dataset pages analyzed in Sect. 6 on Figshare (see Footnote 1). While the source code for the methods outlined in Sect. 5 references proprietary libraries and therefore cannot be released, we have reproduced the overall code logic in pseudocode, also available at the aforementioned link. The base t5x implementation described in Section 5.3 can be found on GitHub.[3] *The GBDT training algorithm described in Sect. 5.4 can be found on GitHub.*[4]

References

1. Achichi, M., Bellahsene, Z., Ellefi, M.B., Todorov, K.: Linking and disambiguating entities across heterogeneous RDF graphs. Web Semant. **55**(C), 108–121 (2019). https://doi.org/10.1016/j.websem.2018.12.003
2. Albertoni, R., Browning, D., Cox, S.J.D., González-Beltrán, A.N., Perego, A., Winstanley, P.: Data catalog vocabulary (DCAT) - Version 2, W3C recommendation (2020). https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/
3. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets. In: LDOW (2009)
4. Alrashed, T., Paparas, D., Benjelloun, O., Sheng, Y., Noy, N.: Dataset or not? a study on the veracity of semantic markup for dataset pages. In: Hotho, A., et al. (eds.) ISWC 2021. LNCS, vol. 12922, pp. 338–356. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88361-4_20
5. Aryani, A., et al.: A research graph dataset for connecting research data repositories using RD-Switchboard. Sci. Data **5**(1), 1–9 (2018). https://doi.org/10.1038/sdata.2018.99
6. Belhajjame, K., et al.: PROV-DM: the Prov data model. W3C Recommend. **14**, 15–16 (2013)
7. Ben Ellefi, M., Bellahsene, Z., Dietze, S., Todorov, K.: Dataset recommendation for data linking: an intensional approach. In: Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Ponzetto, S.P., Lange, C. (eds.) ESWC 2016. LNCS, vol. 9678, pp. 36–51. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-34129-3_3
8. Benjelloun, O., Chen, S., Noy, N.: Google dataset search by the numbers. In: Pan, J.Z., et al. (eds.) ISWC 2020. LNCS, vol. 12507, pp. 667–682. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-62466-8_41
9. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far, 1 edn., pp. 115–143. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3591366.3591378
10. Dong, Y., Takeoka, K., Xiao, C., Oyamada, M.: Efficient joinable table discovery in data lakes: A high-dimensional similarity-based approach. In: 2021 IEEE 37th International Conference on Data Engineering (ICDE), pp. 456–467. IEEE (2021). https://doi.org/10.48550/arXiv.2010.13273
11. Dong, Y., Xiao, C., Nozawa, T., Enomoto, M., Oyamada, M.: Deepjoin: joinable table discovery with pre-trained language models (2023), https://arxiv.org/abs/2212.07588

[3] https://github.com/google-research/t5x.
[4] https://github.com/google/yggdrasil-decision-forests.

12. Fan, G., Wang, J., Li, Y., Miller, R.J.: Table discovery in data lakes: state-of-the-art and future directions. In: Companion of the 2023 International Conference on Management of Data. SIGMOD '23, New York, NY, USA, pp. 69–75. Association for Computing Machinery (2023). https://doi.org/10.1145/3555041.3589409
13. Färber, M., Lamprecht, D.: The data set knowledge graph: creating a linked open data source for data sets. Quantitative Science Studies **2**(4), 1324–1355 (2022). https://doi.org/10.1162/qss_a_00161
14. Gebru, T., et al.: Datasheets for datasets. Commun. ACM **64**(12), 86–92 (2021). https://doi.org/10.48550/arXiv.1803.09010
15. Gregory, K., Groth, P., Scharnhorst, A., Wyatt, S.: Lost or found? discovering data needed for research. Harvard Data Sci. Rev. (2020). https://doi.org/10.1162/99608f92.e38165eb
16. Guha, R.V., Brickley, D., Macbeth, S.: Schema.org: evolution of structured data on the web. Commun. ACM **59**(2), 44–51 (2016). https://doi.org/10.1145/2844544
17. Guillame-Bert, M., Bruch, S., Stotz, R., Pfeifer, J.: Yggdrasil decision forests: a fast and extensible decision forests library. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD '23, New York, NY, USA, pp. 4068–4077. Association for Computing Machinery (2023). https://doi.org/10.1145/3580305.3599933
18. Herschel, M., Diestelkämper, R., Ben Lahmar, H.: A survey on provenance: what for? what form? what from? VLDB J. **26**, 881–906 (2017). https://doi.org/10.1007/s00778-017-0486-1
19. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the pedantic web. LDOW **628**, 26 (2010)
20. Kery, M.B., John, B.E., O'Flaherty, P., Horvath, A., Myers, B.A.: Towards effective foraging by data scientists to find past analysis choices. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19, New York, NY, USA, pp. 1–13. ACM (2019). https://doi.org/10.1145/3290605.3300322
21. Khatiwada, A., Shraga, R., Gatterbauer, W., Miller, R.J.: Integrating data lake tables. Proc. VLDB Endowment **16**(4), 932–945 (2022). https://doi.org/10.14778/3574245.3574274
22. Klump, J., Wyborn, L., Wu, M., Martin, J., Downs, R.R., Asmi, A.: Versioning data is about more than revisions: a conceptual framework and proposed principles. Data Sci. J. (2021). https://doi.org/10.5334/dsj-2021-012
23. Koesten, L.M., Kacprzak, E., Tennison, J.F.A., Simperl, E.: 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. CHI '17, New York, NY, USA, pp. 1277–1289. ACM (2017). https://doi.org/10.1145/3025453.3025838
24. Lane, J., Gimeno, E., Levitskaya, E., Zhang, Z., Zigoni, A.: Data inventories for the modern age? Using data science to open government data. Harvard Data Sci. Rev. **4**(2) (2022). https://hdsr.mitpress.mit.edu/pub/g6e8noiy
25. Leme, L.A.P.P., Lopes, G.R., Nunes, B.P., Casanova, M.A., Dietze, S.: Identifying candidate datasets for data interlinking. In: Daniel, F., Dolog, P., Li, Q. (eds.) ICWE 2013. LNCS, vol. 7977, pp. 354–366. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39200-9_29
26. Liu, J., Liu, T., Yu, C.: NewsEmbed: modeling news through pre-trained document representations. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. KDD 2021, New York, NY, USA, pp. 1076–1086. ACM (2021). https://doi.org/10.1145/3447548.3467392

27. Meusel, R., Paulheim, H.: Heuristics for fixing common errors in deployed schema.org microdata. In: Gandon, F., Sabou, M., Sack, H., d'Amato, C., Cudré-Mauroux, P., Zimmermann, A. (eds.) The Semantic Web. Latest Advances and New Domains. pp. 152–168. Springer (2015). https://doi.org/10.1007/978-3-319-18818-8_10
28. Noy, N., Burgess, M., Brickley, D.: Google dataset search: building a search engine for datasets in an open web ecosystem. In: The World Wide Web Conference, pp. 1365–1375 (2019). https://doi.org/10.1145/3308558.3313685
29. Noy, N., Goble, C.: Are we cobblers without shoes? making computer science data fair. Commun. ACM **66**(1), 36–38 (2022). https://doi.org/10.1145/3528574
30. Rabello Lopes, G., Paes Leme, L.A.P., Pereira Nunes, B., Casanova, M.A., Dietze, S.: Two approaches to the dataset interlinking recommendation problem. In: Benatallah, B., Bestavros, A., Manolopoulos, Y., Vakali, A., Zhang, Y. (eds.) WISE 2014. LNCS, vol. 8786, pp. 324–339. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11749-2_25
31. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1) (2020)
32. Rauber, A., Asmi, A., van Uytvanck, D., Proell, S.: Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC) (2015). https://doi.org/10.15497/RDA00016
33. Rauber, A., Asmi, A., van Uytvanck, D., Pröll, S.: Identification of reproducible subsets for data citation, sharing and re-use. Bull. IEEE Technical Committee Dig. Lib. **12**(1) (2016). https://doi.org/10.5281/zenodo.4048304
34. Roberts, A., et al.: Scaling up models and data with t5x and seqio. J. Mach. Learn. Res. **24**(377), 1–8 (2023). http://jmlr.org/papers/v24/23-0795.html
35. Roussakis, Y., Chrysakis, I., Stefanidis, K., Flouris, G., Stavrakas, Y.: A flexible framework for understanding the dynamics of evolving RDF datasets. In: Arenas, M., et al. (eds.) ISWC 2015. LNCS, vol. 9366, pp. 495–512. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25007-6_29
36. Shraga, R., Miller, R.J.: Explaining dataset changes for semantic data versioning with Explain-Da-V (technical report) (2023). https://doi.org/10.48550/arXiv.2301.13095
37. Silvello, G.: A methodology for citing linked open data subsets. D-Lib Magazine **21**(1/2), 1505–1524 (2015). https://doi.org/10.1045/january2015-silvello
38. Singhal, A., Srivastava, J.: Research dataset discovery from research publications using web context. In: Web Intelligence, vol. 15, pp. 81–99. IOS Press (2017). https://doi.org/10.3233/WEB-170354
39. Tedersoo, L., et al.: Data sharing practices and data availability upon request differ across scientific disciplines. Sci. Data **8**(1), 192 (2021). https://doi.org/10.1038/s41597-021-00981-0
40. Tomita, T.M., et al.: Sparse projection oblique randomer forests. J. Mach. Learn. Res. **21**(104), 1–39 (2020). https://doi.org/10.48550/arXiv.1506.03410, http://jmlr.org/papers/v21/18-664.html
41. Umbrich, J., Hausenblas, M., Hogan, A., Polleres, A., Decker, S.: Towards dataset dynamics: change frequency of linked open data sources. In: Proceedings of the WWW2010 Workshop on Linked Data on the Web (LDOW2010) (2010)
42. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Discovering and maintaining links on the web of data. In: Bernstein, A., et al. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 650–665. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04930-9_41

43. Zhang, A.X., Muller, M., Wang, D.: How do data science workers collaborate? roles, workflows, and tools. Proc. ACM Hum.-Comput. Interact. **4**(CSCW1) (2020). https://doi.org/10.1145/3392826
44. Zhu, E., Deng, D., Nargesian, F., Miller, R.J.: JOSIE: overlap set similarity search for finding joinable tables in data lakes. In: Proceedings of the 2019 International Conference on Management of Data. SIGMOD '19, pp. 847–864, New York, NY, USA. ACM (2019). https://doi.org/10.1145/3299869.3300065
45. Zuiderwijk, A., Shinde, R., Jeng, W.: What drives and inhibits researchers to share and use open research data? A systematic literature review to analyze factors influencing open research data adoption. PLoS ONE **15**(9), e0239283 (2020). https://doi.org/10.1371/journal.pone.0239283