Credit Distribution through Data Provenance in Relational Scientific Databases

Dennis Dosso[a], Susan B. Davidson[b], Gianmaria Silvello[a]

[a] *Department of Information Engineering, University of Padua, Italy*
[b] *Department of Computer and Information Science, University of Pennsylvania, United States*

Abstract

In the current world of research data is a fundamental method to disseminate scientific knowledge, to determine scholarship, and to provide credit and recognition to the authors of research endeavors. However, issues like data citation, handling and counting the credit generated by such citations are still open research questions.

In this context, data credit has recently emerged as a new measure of value, defined and built on top of the data citation theory. Data credit is a real value that represents the importance of data cited by a paper, or by another research entity. As such, credit can be used to annotate data contained in curated scientific databases, and it can be considered as a measure for their importance and impact in the research world. As such, it is a new method that, together with traditional citations, helps to recognize the value of data and its creators in a world more and more dependent on data.

In this paper we explore the problem of Data Credit Distribution, the process by which credit is divided and assigned to the data in a database that are responsible for the production of data being cited by a research entity.

We adopt as use case the IUPHAR/BPS Guide to Pharmacology (GtoPdb), a curated and well-known scientific relational database. We define two new distribution strategies, functions that perform this task, based on two form of data provenance, why-provenance, and how-provenance.

Using different distribution strategies, we show how credit can highlight areas of a database that are frequently used, and how it can work as a new bibliometric measure for data and their corresponding curators. Credit in particular rewards data and authors based on their research impact, and not merely on the number of citations. Also, we show how different distribution strategies, based on different types of data provenance, can be more sensible to the role of an input tuple in the generation of the output, and thus rewarding it differently.

Keywords: Data Citation, Data Credit

1. Introduction

Citations are an essential component of scientific research, enabling research products to be found as well as the relationships between research products to be understood. They form a basis on which to give credit to authors, papers, and venues [55, 19, 20]. Citations are used, among other things, to decide on tenure, promotion, hiring, and funding of grants for researchers [41, 21, 32, 38].

Nowadays, science and research are increasingly digital. There are numerous curated databases that are at the core of scientific research efforts [12]. It is therefore generally accepted that data must be cited and citable [39, 15], and that data citations should contribute to the scientific reputation of researchers, scientists, data curators, and creators [4, 50]. It is also accepted that data citations should be counted alongside of traditional citations, and contribute to bibliometrics indicators [7, 44].

A central problem in data citation is how to attribute credit to data creators and curators [11]. How to handle and count the credit generated by data citation, and how it contributes to traditional and new bibliometrics, are long-standing research issues Garfield [28], Borgman [9]. However, even when correctly applied, data citations and the bibliometric computed using them do not always correctly reward the creators of data used in a database. Data, in fact, is often cited at the "database level" or the "webpage level". In the first case, the whole database is cited and therefore all credit goes to the key personnel of the database. In the second case, the database has a website with webpages that can be individually cited. The webpages use data extracted from the database, which is aggregated by topic and built to resemble a traditional research paper. Often the creators and curators of the webpage's data are not credited or only marginally credited for their work [3].

Recently, the concepts of *data credit* and *Data Credit Distribution* (DCD) [26, 36, 54] have emerged, built on top of methodologies for data citation. Data credit is a value that is computed based on the importance of the data being cited in a paper, and represents the impact of the data on the citing paper. The Data Credit Distribution problem consists of distributing this credit to elements in the databases in the citation graph that are responsible for the generation of the data being cited. The goal of DCD is to improve and expand the reach of data citation, rather than being an alternative to it. This means that to employ DCD techniques, we need data citations in some form.

[37] defined credit as a "quantity" that describes the importance of a research entity, such as papers or data mentioned in a citation, and proposed the idea of a *distribution* of credit from research entities, such as papers or data, to other research entities through citations. This can be done by exploiting the structure of the *citation graph*, a directed graph whose nodes are publications and edges are citations. This graph is the model at the core of systems such as Google Scholar and the Web of Science. Zeng et al. [54] and Fang [26] further explored this concept by defining frameworks for the computation and distribution of credit between papers, authors, and data used by papers in the citation graph.

In this paper, we consider data credit as a data value measure in a (curated) scientific database; credit can be assigned to data of any kind and at any level of granularity. Therefore the concept of "data" is left intentionally vague, although in this paper we focus on relational databases. Credit is a positive *real* value, acting as a proxy for the value of data based on the measure of citations, accesses, clicks, downloads, or other surrogates for data use. We call Data Credit Distribution the process, method, or algorithm used to assign credit to a given datum or dataset.

The DCD problem differs from the traditional citation setting since:

1. In a traditional setting, when a paper cites another paper, a +1 "credit' is given to the cited paper (and to its authors). It does not matter why or how paper p_1 cites paper p_2[1], the result is always +1 from p_1 to p_2 and thus a +1 to the citation count of the authors of p_2. With a different credit distribution strategy, the "value" given to the cited entity can be *proportional* to the role played in the citing entity. Hence, we can weigh the importance of the cited entities and assign credit according to their role.

[1] Note that there is vast research on this topic and many alternative proposals, but none of them currently work at a large scale.

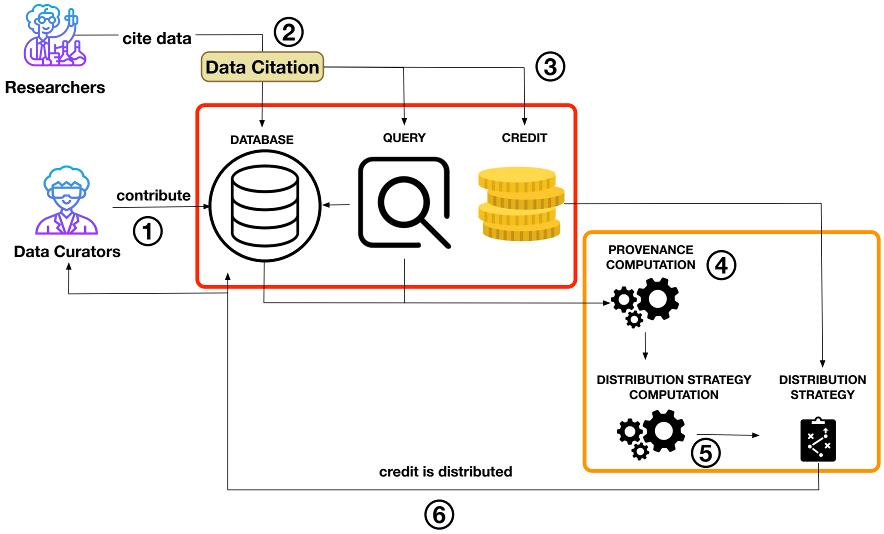


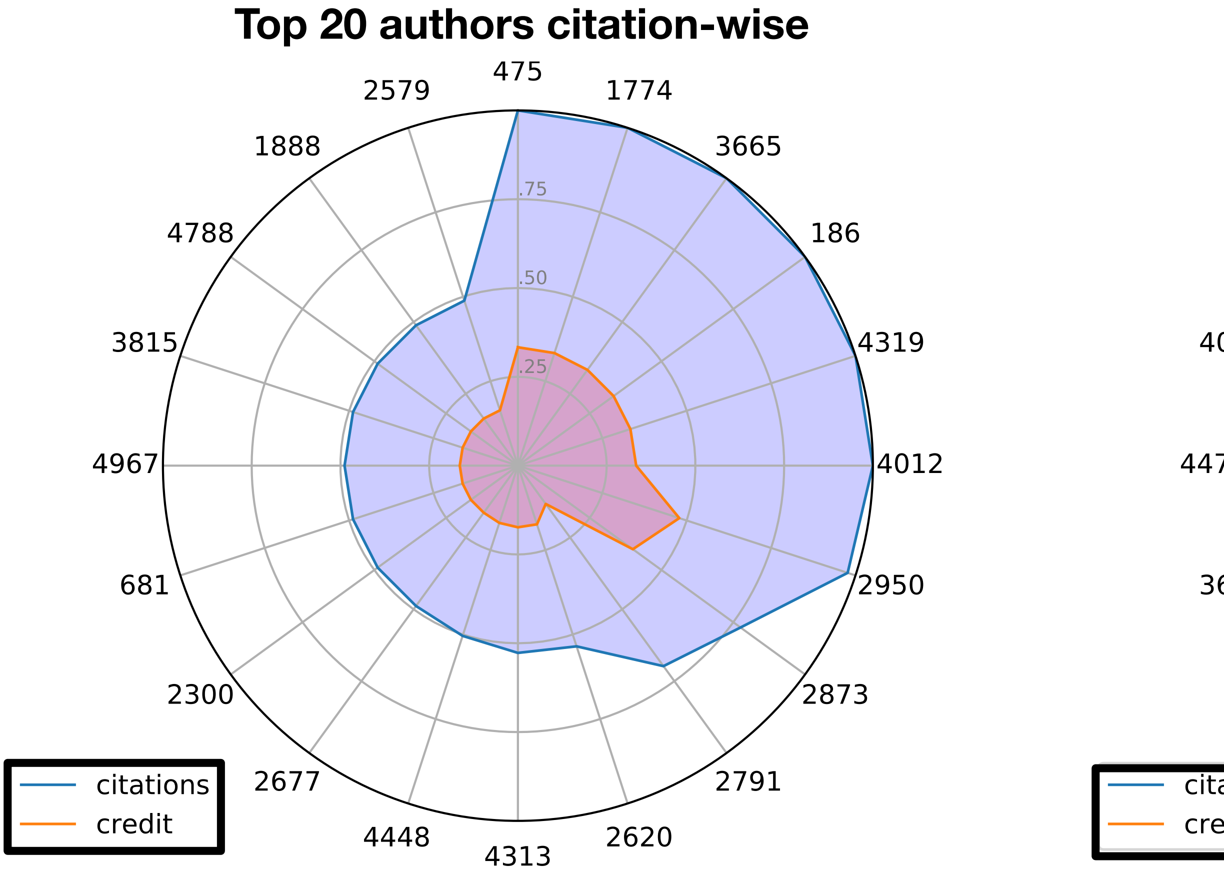
Figure 1: Overview of the credit distribution pipeline.

2. Traditional citations are considered to be *atomic*. A citation from p_1 to p_2 can never be broken into pieces and assigned in part to p_2 and in part to other papers or data that contributed to p_2. This is due to the intrinsic difficulty in grasping the role and "weight" of the other papers and data, and in automating the credit assignment process. In contrast, we consider data credit to be a *non-atomic* real value, which can be divided and distributed to multiple components of a database.
3. Credit can be *transitive*, that is, it can be propagated through one cited entity to other entities cited by it that contributed to its content.

We study the DCD problem in the context of relational databases (RDBs) since they are widely used [2] and are the main focus of current work in data citation methods [14, 12, 45]. RDBs are also frequently a test-bed for new methods that can be adapted to other databases, e.g., graphs or document databases. Furthermore, the "portions" of data in an RDB that can be credited can be defined at different levels of granularity, in particular: (i) the whole database, (ii) tables, and (iii) tuples.

The DCD process is summarized in Figure 1:

[2] The "relational database market alone has revenue upwards of $50B" [1].


Top 20 authors citation-wise
475
1774
3665
186
4319
4012
2950
2873
2791
2620
4313
4448
2677
2300
681
4967
3815
4788
1888
2579
.75
.50
.25
citations
credit
40
447
36
cita
cre

a

Step 1 Scientists and experts contribute the curated information contained in a scientific database. These are called the "Data Curators".

Step 2 Other researchers use the data in their research, and when possible, cite them.

Step 3 The citation to the data generates credit, that can be used as a proxy for the impact of the data on the citing paper. This credit is represented as a real value $k \in \mathbb{R}_{>0}$.

Step 4 Given the database instance I and the query Q, it is possible to compute the *data provenance* of $Q(I)$. The provenance of $Q(I)$ is a form of metadata that describes the generation process undertaken by Q, and the data used in I to generate the output [17]. Many different notions of provenance have been proposed in the literature for data in database management systems [22, 13, 30], describing different kinds of relationships between data in the input and the output of a query. As reported in [17], these provenances have been used in several applications beyond giving information on how queries work, for example, annotation propagation and the view update problem. In this paper, we consider three types of provenance: lineage, why-provenance, and how-provenance.

Step 5 Provenance is input to the CDC problem, whose aim is to compute the *Credit Distribution Strategy* (CDS, also referred only as Distribution Strategy, DS). The CDS is a function that distributes k to the data in the input database I, and is defined on the basis of citation policies decided at the database administration level or at the domain community level. In this paper, since we base CDS on data provenance, we describe three CDS, each one based on a different form of provenance.

Step 6 Once the CDS is computed, it is used to distribute the given credit k to the parts of the database that are responsible for the generation of $Q(I)$. Transitively, this credit is also divided and given to the corresponding authors of those data.

This paper expands our recent work in [24], which addressed the problem of how to reward data and data curators who are typically overlooked in current citation systems. In that work, we first defined the problem of DCD in relational databases, and proposed a viable Distribution Strategy (DS) based on *lineage*, which is the simplest form of *data provenance*. The lineage of a tuple t in the output $Q(I)$ is defined as the set of all and only the tuples in the database instance I that are "relevant" to the production of t, that is the tuple that are used by Q in the production of t. The lineage-based strategy equally redistributes the credit k to the tuples in the lineage set, thus each tuple receives credit $k/|L_t|$, where L_t is the lineage set of t.

One may argue that this DS is too simplistic, since lineage only tells the relevant tuple used to produce the output, and does not convey any information about their role or importance in the query. Therefore, one may desire to give more credit to the tuples that are more relevant or *essential* to the production of the output, i.e. those tuples that, if removed, would prevent the output tuple from appearing in the final result, or those tuples used more than once by the query.

Therefore, in this paper, we expand the ideas in [24] by proposing two new DSs based on other forms of data provenance: why-provenance [13] and how-provenance [30]. We compare them with the lineage-based solution, and discuss why one may be preferred to another depending on the application and its goals. In particular, we show that why-provenance and how-provenance are more sensitive to the *role* of a tuple in a query, i.e. how many times the tuple is used and how it is used. The DS based on why-provenance give more reward to tuples that are essential to the production of the result set, whereas the DS based on how-provenance also takes into consideration the different ways that a tuple is used.

For evaluation, we use a well-known curated database, the IUPHAR/BPS[3] Guide to Pharmacology [31], also known as GtoPdb[4], which contains expertly curated information about diseases, drugs, cellular drug targets, and their mechanisms of action. We chose GtoPdb for two main reasons: (i) it is a widely-used and valuable curated relational database, (ii) many papers in the literature use, and cite its data (i.e., families, ligands, and receptors). Real queries used in papers can therefore be seen as data citations which, in turn, can be used to assign data credit.

We perform three sets of experiments. In the first one, real queries are ex-

[3] International Union of Basic and Clinical Pharmacology/British Pharmacology Society

[4] `https://www.guidetopharmacology.org/`

tracted from papers published in the British Journal of Pharmacology (BJP), that represent data citations to GtoPdb, and are used to distribute credit in the database using the three different provenance-based DSs. In the second and third experiment we analyse the behaviour of the different DS when complex citation queries are employed.

Contributions. Contributions of this work include:

- The definition of new distribution strategies for the problem of Data Credit Distribution, based on why-provenance and how-provenance;
- An in-depth analysis of the effects of credit distribution on real-world curated data and of the differences between the three proposed Distribution Strategies.

Outline. The rest of the paper is organized as follows: Section 2 presents the background and related work. Section 3 describes the use case we adopted. Section 4 briefly presents the forms of provenance used in the paper. Section 5 describes the problem of DCD and the proposed DS. In Section 6 we present the experimental evaluation. Finally, Section 7 draws some conclusions and outlines future work.

2. Background

Data in Research. As described by Jim Gray in his last talk [33], the world of research is rapidly transitioning towards the *fourth paradigm of science*, that is, data-intensive scientific discovery, where data are important for scientific advances as well as for traditional publications [6].

The scientific community is promoting an *open research culture* [43], founded on methods and tools to share, discover, and access experimental data. The community has identified the FAIR principles (Findable, Accessible, Interoperable, and Reusable) [52], that should be enforced by every database. In particular, data should be accessible from the articles, journals, and papers that cite or use them [19]. Aspects such as the need for the *reproducibility* of experiments through the used data; the *availability* of scientific data; the *connections* between data and the scientific results are all needed aspects for the fourth paradigm, and are all relevant to the domain of *data citation* [34].

Data Citation: Principles and Motivations. Data Citation principles were first described in detail in [18], and later summarized and endorsed by the Joint Declaration of Data Citation Principles (JDDCP) [40]. The principles are divided into two groups [48]. The first one contains principles concerning the role of data citation in scholarly and research activities such as the (i) *importance* of data (why data citation is important and why data should be considered as first-class citizens); (ii) *credit* and *attribution* to the creators and curators of the data; (iii)*evidence*; (iv) *verifiability*; and *interoperability*, with these last three requiring data citation methods to be flexible enough to operate through different communities. The second group defines the main guidelines to establish a data citation systems, and contains principles such as the (i) *unique identification* of the data being cited; (ii) *(open) access* to data; (iii) guarantee of *persistence* and *availability* of citations even after the lifespan of the cited entity; the (iv) *specificity* of a citation, i.e. it must lead to the data set originally cited.

It is possible to outline six main motivations for data citation [48]:

- *Data attribution*: identify the individuals that should be credited for data with variable granularity.
- *Data connection*: connect papers to the data being used.
- *Data Discovery*: citations helps to find data records and subsets that would be otherwise not findable via search engines.
- *Data Sharing*: share data obtained by researchers within the whole community.
- *Data Impact*: highlight the results obtained in writing papers using specific data, the frequency and modality data were used.
- *Reproducibility*: data citation greatly impacts the reproducibility of science [5]. Many authoritative journals ask to share data and provide valid methodologies to reproduce experiments.

2.1. Data Citation in Relational Databases

In this paper, we develop our methods and experiments on relational databases. RDBs have been the main target of data citation methods since the surge of the data-centric research paradigm. The RDA “Working Group on Data Citation: Making Dynamic Data Citable"[5] [46] has been working in the last years on large, dynamic, and changing datasets. The working group has finished the development of its guidelines and has now moved on into an adoption phase. The datasets considered by the WG are often relational.

In one of its most recent sessions [47], the Working Group (WG) on Data Citation reported that there are various implementations of its guidelines for Data Citation on MySQL/Postgres relational databases. Some of these databases are: DEXHELPP[6] (Social Security Records); NERC (ARGO Global Array); EODC (Earth Observation Data Centre) [29]; LNEC (River dam monitoring); MDS (Million Song Database) [8]; CBMI[7] (Center for Biomedical Informatics); VMC (Vermont Monitoring Cooperative); CCA[8] (Climate Change Center Austria); VAMDC (Virtual Atomic and Molecular Data Center) [25, 56].

More examples of work on data citation in relational databases are [12, 53, 2, 23]. The website `https://fairsharing.org/` keeps a long updated list of curated and scientific databases (many of which are relational or graph-based) following FAIR guidelines. These databases are citable since they are compliant with the most recent guidelines, and they are in the vast majority of cases accessible via dynamically created Webpages. In all these databases is, therefore, possible to implement DCD on top of the existing infrastructures for citing data.

Data citation techniques are primarily applied to relational databases because of their diffusion and also because the portions of data that are to be cited are easily identified: the whole database, a relation, a tuple, or even an attribute. Many papers [10, 12, 2] consider more complex citable units, recognizing that often the *views* of a database are the ones to be cited. Generally, a *view* is a query on the database. To this end, [53] suggested decomposing the database in a set of views, where each view is associated with its citation.

At present, the most common practices to cite databases include:

1. A database cited as a whole, even though only parts of the databases are used in the papers or datasets. Alternatively, the so-called "data pa-

[5] `https://www.rd-alliance.org/groups/data-citation-wg.html`

[6] `http://www.dexhelpp.at/`

[7] `https://medicine.missouri.edu/centers-institutes-labs/center-for-biomedical-informatics`

[8] `https://ccca.ac.at/startseite`

pers" can be cited, being traditional papers that describe a database [16]. In this case, all the credit from the citations goes to the database administrators or to the authors of the data papers.

2. Subsets of data, obtained by issuing queries to a database, are individually cited. This is the solution adopted by the *Resource Data Alliance* (RDA) working group on Data Citation [46]. In this case, the credit generated from citations can be distributed among the contributors of the portions of data being cited, and/or to the database administrators.
3. The database is accessible via a series of Webpages that arrange the content of the database by topic or theme. Examples in the life science domain include the Reactome Pathway database [35], the GtoPdb [31], and the VAMDC [56]. Every single Webpage is unequivocally identifiable and can be individually cited.

Despite all the research efforts dedicated to the study and promotion of data citation, none of the largest citation-based systems, such as Elsevier Scopus, Web of Science, Microsoft Academia, or Google Scholar, consider scientific datasets as citable objects in academic work. Clarivate Analytics Data Citation Index (DCI) [27] is an exception, since its infrastructure tracks data usage in scientific domains and provides the technical means to connect datasets and repositories to scientific papers. However, DCI considers only citations to (previously registered and approved) databases as a whole and does not count citations to database portions such as views, tables, or tuples.

2.2. Data Credit

Data credit is related to data citation: they both aim to recognize the work of data creators and curators. Data credit can therefore also be seen as a by-product of data citation, since credit attribution is impossible without the presence of data citations.

Katz [36] suggests the need for a *modified citation system* that includes the idea of *transient* and *fractional credit*, to be used by developers of research products as software and data. In the paper two considerations are made: (i) research objects such as data and software are currently not formally rewarded or recognized by the community; (ii) even in traditional papers, the contribution of each author to the work is hard to understand, unless explicitly specified in the paper. This is even more true for data, where different groups of people work on the same database.

In [36] credit is defined as a "quantity" that describes the importance of a research entity, such as papers, software, or data, mentioned in a citation. We add that the concept of credit can be built on top of the existing infrastructure handling traditional and data citations. Katz [36] further explores the idea of a *distribution* of credit from research entities (i.e., papers and data) to other research entities through citations that connect them. Thanks to traditional citations and now also to data citations, this distribution is finally possible, at least between papers and data. Some problems related to traditional citations can thus be solved by citations:

1. Credit rewards research entities that to date are not (formally) recognized (a goal shared with data citation).
2. Credit can reward authors *proportionally* to their role in generating the entity. The more an author contributes to a paper, the more credit is given to him. Zou and Peterson [55] work on something similar with their zp-index, which includes in its formulation the position (and thus the role) of a publication author to represent its impact in the work itself.
3. Credit can be *transitively* channeled through a chain of papers citing each other, thus enabling the rewarding of older papers that are no more cited, since other papers summarize or report their content but are nevertheless crucial in a research area for the influence of their content.

Fang [26] presents a framework to distribute the credit generated by a paper to its authors and to the papers in its reference list in a transitive way. Let us consider the *citation graph* as the graph where the nodes are papers and the links are the citations among them. In this graph, every paper is a source of credit, which is then transferred to the neighboring nodes. The quantity of credit received by each cited paper depends on its impact/role in the citing paper. So far, this theoretical framework is limited to papers, but it can be easily extended to a citation graph including both papers and data.

Zeng et al. [54] proposes the first method to compute credit within a network of papers citing data. Adopting a network flow algorithm, they simulate a random walker to estimate a score for each dataset, leveraging real-world usage data to compute the credit. This is the first step towards an automatic credit computation procedure. This proposal is, however, limited to assigning credit to whole datasets, and it does not deal with the granularity of data. It does not work to assign credit to a single research entity within a dataset.

Differently from Zeng et al. [54], we do not treat the credit computation process, but we focus on the distribution process.

2.3. Data Provenance

To distribute credit, we base our methods on *data provenance*. Data provenance is information that describes the origin and the process of creation of data. It can also be seen as metadata pertaining to the derivation history of the data. It is particularly useful to help users to understand where data are coming from, and the process they went through. Data citation and data provenance are closely linked [3] since both are forms of annotations on data retrieved through queries. Data provenance has been widely studied in different areas of data management. In this paper, we focus on provenance for database management systems (DBMS). For further details on data provenance, please refer to surveys like [17] and [49].

Cheney et al. [17] presents four main types of data citation for DBMS: *lineage* [22], *why-provenance* [13], *how-provenance* [30] and *where-provenance* [13].

Let us start with the first three provenances. Given a database instance I, a query Q, and the result $Q(D)$, consider one tuple t of the output. Its provenance is information about its generation through the tuples of the input that are used by Q. Different types of provenance convey different levels of information. Since these three provenances are computed for each tuple of the output, they are also referred to as *tuple-based*.

Lineage is somehow the simplest among the forms of provenance. It has been defined in different ways [17], but it can be thought of as the set of all the tuples that are used in some way by the query to produce the output tuple, the ones that are somehow *relevant* to its generation.

The definition of why-provenance is based on the notion of *witness set*. A witness is a set of relevant tuples that guarantees the existence of t in $Q(D)$. The lineage is therefore an example of a witness. The why-provenance of a tuple t is a peculiar set of witnesses – described in [13] – that are computed from the query, called *witness basis*. A witness basis may be composed of more than one witness. Therefore, the why-provenance contains more information than the lineage, since it describes *alternative* ways in which the same output may be generated.

The how-provenance takes the form of a polynomial, called *provenance polynomial*, where the variables are taken from the set of identifiers of the tuples (provided that each tuple in I has an identifier) and the coefficients are taken from $\mathbb{N}$. This provenance also contains information on *how* the input

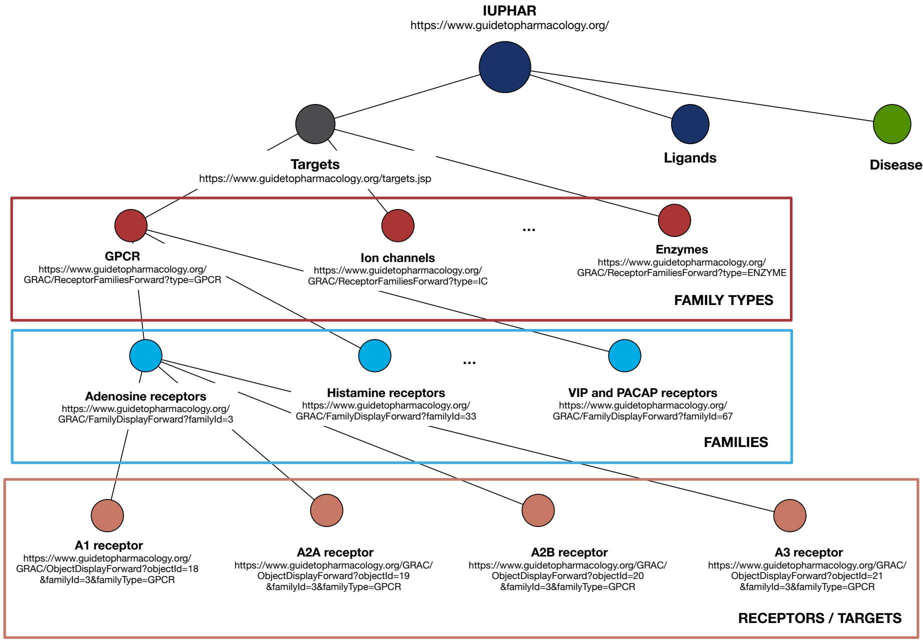

IUPHAR
https://www.guidetopharmacology.org/
Targets
https://www.guidetopharmacology.org/targets.jsp
Ligands
Diseases
GPCR
https://www.guidetopharmacology.org/
GRAC/ReceptorFamiliesForward?type=GPCR
Ion channels
https://www.guidetopharmacology.org/
GRAC/ReceptorFamiliesForward?type=IC
...
Enzymes
https://www.guidetopharmacology.org/
GRAC/ReceptorFamiliesForward?type=ENZYME
FAMILY TYPES
Adenosine receptors
https://www.guidetopharmacology.org/
GRAC/FamilyDisplayForward?familyId=3
Histamine receptors
https://www.guidetopharmacology.org/
GRAC/FamilyDisplayForward?familyId=33
...
VIP and PACAP receptors
https://www.guidetopharmacology.org/
GRAC/FamilyDisplayForward?familyId=67
FAMILIES
A1 receptor
https://www.guidetopharmacology.org/
GRAC/ObjectDisplayForward?objectId=18
&familyId=3&familyType=GPCR
A2A receptor
https://www.guidetopharmacology.org/GRAC/
ObjectDisplayForward?objectId=19
&familyId=3&familyType=GPCR
A2B receptor
https://www.guidetopharmacology.org/GRAC/
ObjectDisplayForward?objectId=20
&familyId=3&familyType=GPCR
A3 receptor
https://www.guidetopharmacology.org/GRAC/
ObjectDisplayForward?objectId=21
&familyId=3&familyType=GPCR
RECEPTORS / TARGETS

Figure 2: Partial map of the GtoPdb hierarchical structure grouping the targets into families and family types.

tuples are used. For example, when two tuples are combined by a join, they are also combined in the polynomial by the · operator. When two or more tuples become equivalent due to a union or a projection, the corresponding monomials are combined by the + operator.

It has been shown in [17] that the how-provenance is the more general and informative of the three, containing the other two.

Where-provenance, differently from the other three, is *attribute-based*, so we do not take it into account in this work since we consider the tuple as the finest citable unit.

3. Use Case: GtoPdb

As use case we refer to the IUPHAR/BPS Guide to Pharmacology [31] or GtoPdb[9]. GtoPdb is a well-known and well structured scientific relational database that contains expertly curated information about diseases, drugs

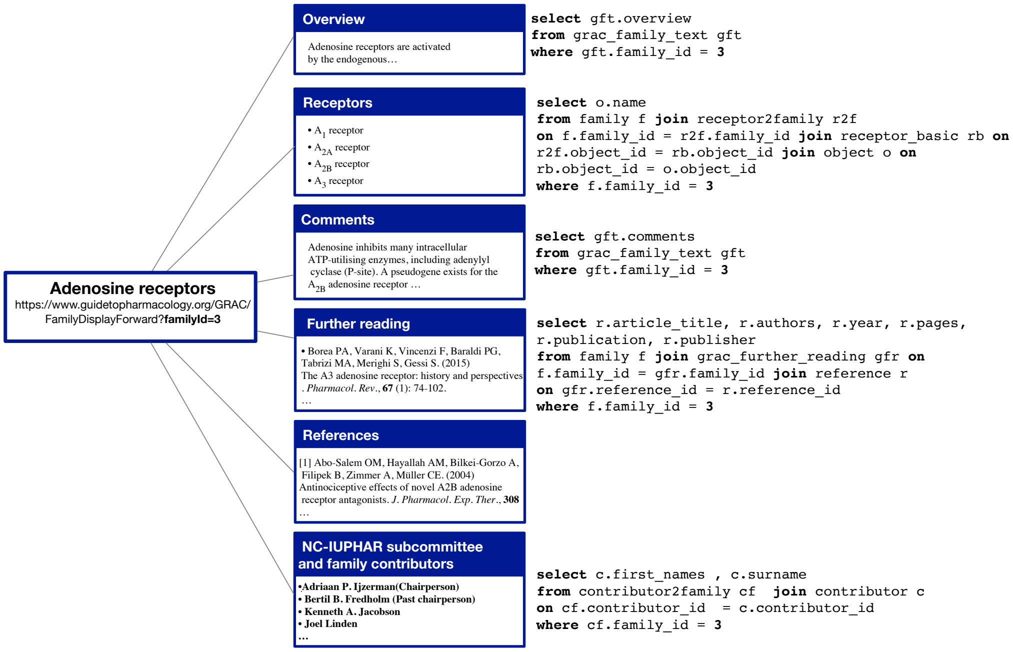
[9] https://www.guidetopharmacology.org/

in clinical use, their cellular targets, and the mechanisms of action on the human body. It is curated and maintained by the GtoPdb Committee, and by 96 subcommittees, comprising 512 scientists collaborating with in-house curators who draw the information contained in the database from high-quality pharmacological and medicinal chemistry literature. Roughly 1000 researchers from all over the world have contributed to the database, and the curators wanted to give recognition to these contributors. This led to some early work on data citation [10].

GtoPdb is relational, but its logical structure is hierarchical as shown in Figure 2. The information contained in the database is also organized into webpages focused on specific diseases, targets or ligands, and families for easier access by users. As depicted in Figure 2, the database can be thought of as a tree where the root is the database; the first level consists of all targets, ligands, and diseases; and the lower levels consists of specific targets, ligands and diseases. In this paper, we focus on targets; thus at the third level in the figure we show examples of family types, at the fourth level we show specific families of targets (a finer level of granularity), and finally, at the last level, the single targets (also known as receptors).

GtoPdb provides access to the webpages corresponding to all these nodes through URLs. The webpages corresponding to target families all present a similar structure, as shown in Figure 3 for the "Adenosine receptors" family. Each page has an *Overview*, a brief text describing the content of the page; a list of *Receptors* comprising the family; a section of *comments* about the family; the *References*, a list of the papers consulted by the curators of the page, similar to a reference list of a paper; the *further reading* list, reporting papers that an interested reader may want to consult to obtain more insight on the family; and a final section called *How to cite this family page*, containing text snippets useful to cite the specific page or the whole database. Figure 3 shows the SQL code that retrieves the information used to build the corresponding sections (apart from the References section). Therefore, each family page can be considered a full-fledged traditional publication, consisting of title, authors, abstract (the overview), content, and references.

In practice, many papers in the literature only reference GtoPdb (the root) without including a reference to the specific page being cited. That is, they only cite a paper describing GtoPdb as a whole (e.g., [31]) and refer to targets, ligands, diseases, etc. only by name. Thus, citations to specific families are *de-facto* "hidden" to citation systems such as Google Scholar, and useless for the computation of bibliometrics.



Adenosine receptors
https://www.guidetopharmacology.org/GRAC/FamilyDisplayForward?**familyId=3**
Overview
Receptors
Comments
Further reading
References
NC-IUPHAR subcommittee and family contributors

Adenosine receptors are activated by the endogenous…

```
select gft.overview
from grac_family_text gft
where gft.family_id = 3
```

- A_1 receptor
- A_{2A} receptor
- A_{2B} receptor
- A_3 receptor

```
select o.name
from family f join receptor2family r2f
on f.family_id = r2f.family_id join receptor_basic rb on
r2f.object_id = rb.object_id join object o on
rb.object_id = o.object_id
where f.family_id = 3
```

Adenosine inhibits many intracellular ATP-utilising enzymes, including adenylyl cyclase (P-site). A pseudogene exists for the A_{2B} adenosine receptor …

```
select gft.comments
from grac_family_text gft
where gft.family_id = 3
```

- Borea PA, Varani K, Vincenzi F, Baraldi PG, Tabrizi MA, Merighi S, Gessi S. (2015) The A3 adenosine receptor: history and perspectives . *Pharmacol. Rev.*, **67** (1): 74-102.
…

```
select r.article_title, r.authors, r.year, r.pages,
r.publication, r.publisher
from family f join grac_further_reading gfr on
f.family_id = gfr.family_id join reference r
on gfr.reference_id = r.reference_id
where f.family_id = 3
```

[1] Abo-Salem OM, Hayallah AM, Bilkei-Gorzo A, Filipek B, Zimmer A, Müller CE. (2004) Antinociceptive effects of novel A2B adenosine receptor antagonists. *J. Pharmacol. Exp. Ther.*, **308**
…

- **Adriaan P. Ijzerman(Chairperson)**
- **Bertil B. Fredholm (Past chairperson)**
- **Kenneth A. Jacobson**
- **Joel Linden**

…

```
select c.first_names , c.surname
from contributor2family cf  join contributor c
on cf.contributor_id  = c.contributor_id
where cf.family_id = 3
```

Figure 3: Basic web-page structure of “Adenosine receptors” family (ID 3), with queries used to retrieve the information contained in every section, except references.

In certain “lucky” cases, as with papers available in PDF and published in the British Journal of Clinical Pharmacology [10] (BJCP), when a family, ligand, receptor name, etc. are used, they have a hyperlink pointing to the corresponding webpage in GtoPdb. Therefore, the citations to the families can be detected and counted using the URLs reported in the papers. However, these citations to GtoPdb webpages are not counted as such by citation systems, so they are not converted into credit for curators and collaborators.

For our running example, consider Table 1. This simplified version of GtoPdb illustrates three tables: `family`, `contributor` and `contributor2family`. The first table, `family`, has tuples representing families with three attributes: the id of the family, its name, and type. Table `contributor` consists of people who have helped generate the data of the database. The third table, `contributor2family`, serves as a link between the families and the people who contributed to them. For instance, “John Smith” ($\mathtt{c}_1$) contributed to

[10] https://bpspubs.onlinelibrary.wiley.com/journal/13652125

family

id	name	type
f_1	Dopamine Receptors	gpcr
f_2	Bile Acid Receptor	gpcr
f_3	FAK Family	enzyme
f_4	YANK Family	enzyme

contributor2family

id	family_id	contributor_id
$c2f_1$	f_1	c_1
$c2f_2$	f_1	c_2
$c2f_3$	f_2	c_3
$c2f_4$	f_4	c_1

contributor

id	Name	Country
c_1	John Smith	UK
c_2	Jim Doe	UK
c_3	Hans Zimmerman	Germany
c_4	Roberta Rossi	Italy

Table 1: Example of a database consisting of three tables. `family` includes some receptor families in the database; `contributor` contains the name and country of contributors; `contributor2family` connects contributors to the families they contributed to.

"Dopamine Receptors" ($\mathtt{f_1}$) as well as to the "YANK Family" ($\mathtt{f_4}$). We use this example throughout the rest of the paper. In particular, we are using the `id` attribute of the tables as *provenance token* of its corresponding tuples, that is, as a symbol that serves to identify a tuple when talking about provenance.

4. Data Provenances

In this section, we present the three types of provenance used in this paper: lineage, why-provenance, and how-provenance.

4.1. Lineage

Lineage was first introduced by Cui et al. [22]. Given a database instance I and query Q, lineage associates with each tuple $o \in Q(I)$ the set of tuples in the input that helped "produce" it [17]. As an example, consider the following SQL query `Q1`, applied to the database described in Table 1, that asks for the names of families curated by researchers based in the United Kingdom (UK):

```
Q1: SELECT DISTINCT f.name
FROM family AS f JOIN contributor2family AS c2f
ON f.id = c2f.family_id
```

```
    JOIN contributor AS c ON c2f.contributor_id = c.id
    WHERE c.country = 'UK'
```

id	name	lineage
o_1	Dopamine Receptors	$\{f_1, c2f_1, c_1, c2f_2, c_2\}$
o_2	YANK Family	$\{f_4, c2f_4, c_1\}$

Table 2: Result of an SQL query applied to the database instance in Table 1, which asks for the names of families curated by a researcher based in the UK. Attribute `id` is not part of the output and was added to succinctly identify each tuple as provenance token. Each tuple is also annotated with its lineage.

Table 2 shows the query result, which consists of two tuples. We add an extra attribute `id` so that we can easily refer to each result tuple. The lineage for tuple o_1 is the set $\{f_1, c2f_1, c_1, c2f_2, c_2\}$, since the tuple f_1 was joined with $c2f_1$ and then with c_1, and was also joined with $c2f_2$ and c_2. No other tuple is used in the database to produce o_1. For tuple o_2 the lineage is $\{f_4, c2f_4, c_1\}$. Lineage is defined for each tuple of the output, and can differ between tuples.

4.2. Why-Provenance

Why-Provenance was first defined in terms of a deterministic semistructured data model and query language [13]. While why-provenance can be defined in many ways, we refer to [17], where it is expressed in terms of the relational model using the relational algebra.

In particular, while lineage aims to find all and only the tuples in the input relevant to the production of an output tuple, why-provenance aims to find sub-instances of the input that "witness" a part of the output. Given a tuple t in the query's output, a *witness* is any sub-instance of the database that produces t. In particular, the whole database and the lineage of t are both witnesses of t. Since the definition of witness allows for the presence of "irrelevant" tuples, the set of all witnesses is finite (since the database instance I is finite), but it is potentially exponentially large [17].

Buneman et al. [13] defined the why-provenance of an output tuple t in the result $Q(I)$ as a special *subset* of the set of witnesses called the *witness basis*. The witnesses of the basis depend on Q; thus, each basis's size is bounded by the size of Q. The witnesses of the basis exclude tuples that are irrelevant to t being produced by Q, and thus the basis tends to be very small compared to the set of all possible witnesses [17]. The witnesses are also *minimal*, in the sense that if one tuple is removed from one of these witnesses, it cannot produce the output.

id	name	why-provenance
o_1	Dopamine Receptors	$\{\{f_1, c2f_1, c_1\}, \{f_1, c2f_2, c_2\}\}$
o_2	YANK Family	$\{\{f_4, c2f_4, c_1\}\}$

Table 3: Result of a SQL query applied on the database of Table 1 with the why-provenance of the corresponding results.

In a sense, each witness in the witness basis captures one possible way in which the query can generate the output. To better understand this, consider the example in Table 3, where each tuple in the result of query `Q1` is annotated with its why-provenance.

The why-provenance of output tuple o_2 has only one witness, which coincides with its lineage. This happens because there is only one way this output tuple can be produced, i.e., for tuple f_4 to be joined with $c2f_4$ and c_1. On the other hand, o_1 has a witness basis with of two witnesses, since there are two possible ways in which the query can generate o_1. One possibility is that f_1 is joined with $c2f_1$ and c_1 (the first witness), and the second possibility is that f_1 is joined with $c2f_2$ and c_2 (the second witness). This means that to generate o_1, it is sufficient that only one of the two witnesses is present in the input database.

4.3. How-Provenance

While why-provenance describes the source tuples that witness an output tuple in the result of the query, it leaves out information about how the source tuples are used. How-provenance was therefore defined in [30] to capture this information using a *semiring* algebraic structure, and is a form of provenance that takes the form of a *polynomial*.

The key idea in Green et al. [30] is to use the two operators $+$ and $\cdot$ to represent two basic transformations that source tuples undergo as a result of applying a relational query to a database [17]. Two tuples may either be joined together, as an effect of a join (represented with the $\cdot$ operator) or merged via union or projection (represented with the $+$ operator).

Table 4 shows a simple example in which the two output tuples of our running example are annotated with their respective how-provenances. Tuple o_2 was produced through the join among the input tuples $f_4, c2f_4$, and c_1. The three provenance tokens are, therefore "multiplied" together. The case of

id	name	how-provenance
o_1	Dopamine Receptors	$f_1 \cdot c2f_1 \cdot c_1 + f_1 \cdot c2f_2 \cdot c_2$
o_2	YANK Family	$f_4 \cdot c2f_4 \cdot c_1$

Table 4: Result of the example SQL query `Q1` with the corresponding how-provenances of the output tuples annotated.

o_1 is slightly more complex. This tuple, as already discussed, can be obtained through two different joins. The two monomials composing the polynomial represent these two alternatives. They correspond, in a way, to the witnesses of the why-provenance of o_1. The $+$ operator represents the fact that the two monomials describe alternative derivations. The output tuple is the result of a merge of two distinct tuples after the projection on the attribute `name`. This merge is due to the fact that the result of a relational algebra expression is always a *set* of tuples, which corresponds to the presence of the `DISTINCT` operator in an SQL query. This simple example gives the basic idea behind how-provenance and how it allows us to track the operations that produced an output tuple.

Provenance polynomials may also have monomials whose exponents and/or coefficients are greater than one, for example, $3f_1 \cdot c2f_1 \cdot c_1 + f_1 \cdot c2f_2^3 \cdot c_2^3$. This is a polynomial of a tuple produced by a query where the result of the join between the tuples f_1, $c2f_1$, and c_1 is produced three times and then merged (e.g. as the result of a union), and the tuples $c2f_2$ and c_2 are used three times in the operation described by the second monomial (e.g., with nested queries).

5. Credit Distribution and Distribution Strategies

We now give formal definitions of data credit and Data Credit Distribution (DCD), and present three different Distribution Strategies (DSs) based on the forms of provenance discussed earlier: Lineage-based DS, Why-Provenance-based DS, and How-Provenance-based DS. We also show how these strategies distribute credit in the IUPHAR example discussed earlier.

5.1. Data Credit and Data Credit Distribution

Given a database instance I, a *recipient of credit* is a unit of information within I. In the case of relational databases, recipients may be (i) the whole database; (ii) a table; (iii) a tuple; or (iv) an attribute.

Data credit is a value $k \in \mathbb{R}_{>0}$. Every recipient in a database is annotated with a quantity of credit as a proxy for its importance. In this paper, we focus on *tuples* as recipients of credit.

Given a *distribution strategy* (DS), *Data Credit Distribution* (DCD) takes a database instance I, quantity of credit k, and query Q over I, and splits k among the recipients of credit in I.

In the following, we use the notation in Cheney et al. [17]: Given an instance I, a *tuple location* (R, t) is a tuple t in relation R. With reference to the running example, `(family,` $\langle f_1$`, Dopamine Receptors, gpcr`$\rangle$`)` is the tuple location of the first tuple in the `family` relation. The set of all tuple locations in I is called *TupleLoc*. We use this to formally define DCD at the *tuple level*.

Definition 5.1. ***Tuple Level Data Credit Distribution (DCD)*** *[24]* *Given a query Q over I and $k \in \mathbb{R}_{>0}$, DCD is defined by the function $f_{I,Q}$: $TupleLoc \times \mathbb{R}_{>0} \rightarrow \mathbb{R}_{\geq 0}$ such that $f_{I,Q}(t,k) = h$ where $0 \leq h \leq k$ and $\sum_{t \in TupleLoc} f_{I,Q}(t,k) = k$. The function f_{IQ} is the distribution strategy (DS).*

As we can see, the DS is a function that annotates each tuple in the database with a real value, which is a fraction of the given quantity k. The only constraint is that the sum of the credit annotations on tuples must be k, i.e. that no credit is generated or destroyed during the distribution. Given I and Q, many different DSs may be defined as long as they sum up to k.

In what follows, we use information provided by data provenance to define distribution functions. For simplicity, we assume that the credit k is distributed equally across the set of output tuples (i.e. the result of a query), and discuss how the credit of one output tuple o, k_o, is distributed across the instance I.

5.2. A Lineage-based Distribution Strategy

In the lineage-based distribution strategy, each tuple in the output of a query distributes credit equally to each input tuple that appears in its lineage. More formally:

Definition 5.2. *Lineage-based Distribution Strategy [24]*
Let I be a database instance, Q a query over I, $o \in Q(I)$ an output tuple and k_o the credit associated to o. Let L be the lineage of o and t be a tuple in I,

then t receives credit equal to:

$$f_{I,Q}(t, k_o) = \begin{cases} 0 & \textit{if } t \notin L \\ \frac{k_o}{|L|} & \textit{if } t \in L \end{cases}$$

Note that lineage-based DS distributes credit only to input tuples that have a role in creating o by the query Q, and that each receives an equal share of credit via o. Thus, the more tuples in a lineage set, the less credit each tuple receives.

As an example, consider the output tuples of Table 2, and assume that each output tuple has credit $k_o = 1$. The lineage of the first tuple, o_1, is the set $\{f_1, c2f_1, c_1, c2f_2, c_2\}$. Therefore, each tuple in this set receives credit 1/5. The other tuples of the database receive zero credit. The lineage of the second output tuple is $\{f_4, c2f_4, c_1\}$, therefore each of these tuples receives credit 1/3.

At the end of the process, tuples f_1, $c2f_2$ and c_2 each receive credit 1/5, tuples f_4 and $c2f_4$ receive 1/3, while tuple c_1 receives 8/15. Note that if a tuple appears in more than one lineage set, then it will accumulate credit from the distribution associated with each one of these sets, implying that it has a more significant role in the context Q, as is the case with c_1 in this example.

Not all of the tuples in the lineage of an output tuple are necessary to be present at the same time for the output tuple to appear in the query results. For example, if the database only had the set of tuples $\{f_1, c2f_1, c_1\}$ or the set $\{f_1, c2f_2, c_2\}$, the existence of o_1 would still be guaranteed. In other words, while f_1 is always needed for o_1 to appear in the output, only one of the sets of tuples $\{c2f_1, c_1\}$ and $\{c2f_2, c_2\}$ is required. One could therefore argue that it would be more fair for f_1 to receive more credit than the other four tuples, given its role in producing o_1.

This highlights one limitation of the lineage-based DS: while able to find all and only the relevant tuples of the output, it does not distinguish the *importance* of tuples in the query computations. We therefore present two other, more sophisticated, forms of distribution strategies based on why- and how-provenance.

5.3. A Why-Provenance-Based Distribution Strategy

The distribution strategy based on why-provenance first equally distributes the credit k_o among the witnesses of the witness basis for o, and then equally

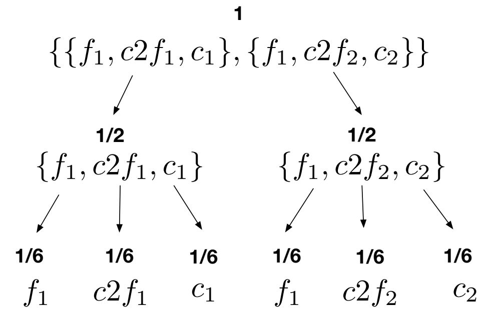


Figure 4: Distribution of credit using why-provenance-based DS for tuple o_1.

divides the credit of a witness among the tuples in the witness. Since a tuple may appear in more than one witness, it will receive more than one portion of credit from the same distribution. More formally:

Definition 5.3. *Why-Provenance-based Distribution Strategy*
Let I be a database instance, Q a query over I, $o \in Q(I)$ an output tuple and k_o the total credit associated to o. Let $\mathcal{W} = Why(Q, I, o)$ be the witness basis of o according to Q and I, and $W \in \mathcal{W}$ be a witness.

Then tuple t in I receives credit equal to:

$$f_{I,Q}(t, k_o) = \frac{k_o}{|\mathcal{W}|} \sum_{W \in \gamma(\mathcal{W}, t)} \frac{1}{|W|}$$

where γ is a function which returns all witnesses W in which t appears:

$$\gamma(\mathcal{W}, t) = \{W \in \mathcal{W} : t \in W\}$$

Figure 4 shows the distribution of credit with why-provenance-based DS for tuple o_1. The credit is first equally divided between the two witnesses, so that both receive credit 1/2. The credit is then further divided among the tuples in each witness. Since each witness has three tuples, each tuple in a witness receives 1/6 of credit. At the end of the distribution, f_1 receives a total credit of 1/3, and the other tuples receive 1/6 each. This distribution better reflects the role of f_1 in the generation of o_1 since, as discussed earlier, it is the only mandatory tuple for o_1 to appear in the output; only one of the two other pairs of tuples are necessary for o_1 to appear in the result.

This example illustrates that why-provenance can better reward input tuples depending on their role. Tuples that appear in more than one witness are rewarded more than others.

$$\mathcal{H} = \underbrace{3f_1 \cdot c2f_1 \cdot c_1}_{M_1} + \underbrace{f_1 \cdot c2f_2^3 \cdot c_2^3}_{M_2}$$

$$\begin{array}{cc} c(\mathcal{H}) = 4 & c(M_2) = 7 \\ mc(M_1) = 3 & mc(M_2) = 1 \\ e(c_2, M_2) = 3 & \gamma(c_1, \mathcal{H}) = \{M_1\} \end{array}$$

$$\gamma(f_1, \mathcal{H}) = \{M_1, M_2\}$$

Figure 5: Illustration of notation used to define the how-provenance based DS in Definition 5.4.

5.4. A How-Provenance Based Distribution Strategy

How-provenance conveys more information than why-provenance since it not only captures what tuples are relevant to the output and in which combination, but also how they are used. The "how" is captured through the provenance polynomials.

The how-provenance-based DS therefore first distributes the credit to the monomials of the polynomial accordingly to the weight represented by their coefficients, then to the tuples of each monomial accordingly to the weights represented by their exponents.

To define the DS more formally, we introduce some notation and illustrate it using the provenance polynomial $\mathcal{H}$ shown in Figure 5.

We call c the function that, given a polynomial, returns the sum of the coefficients of the polynomial; thus $c(\mathcal{H}) = 3{+}1 = 4$. We use the same name for the function that, given a monomial, returns the sum of its exponents; thus $c(M_2) = 1{+}3{+}3 = 7$. mc is the function that takes as input a monomial and returns its coefficient. e is a function that takes as input a tuple and a monomial, and returns the exponent of the tuple in the monomial, if present; thus $e(c_2, M_2) = 3$. γ takes as input a tuple and the whole polynomial, and returns a set containing the monomials containing that tuple, if present in the polynomial; thus $\gamma(f_1, \mathcal{H}) = \{M_1, M_2\}$.

Definition 5.4. *How-Provenance-Based Distribution Strategy*
Let I be a database instance, Q a query over I, $o \in Q(I)$ an output tuple, $\mathcal{H}$ be the provenance polynomial for o, and k_o the credit given to o. The credit given to tuple t in I is:

$$f_{I,Q}(t, k_o) = \frac{k_o}{c(\mathcal{H})} \sum_{M \in \gamma(t,\mathcal{H})} mc(M) \frac{e(t, M)}{c(M)}$$

id	name
oxs_1	Dopamine Receptors

lineage	why-provenance	how-provenance
$\{f_1, c2f_1, c_1, c2f_2, c_2\}$	$\{\{f_1, c2f_1, c_1\}, \{f_1, c2f_2, c_2\}\}$	$f_1^2c2f_1c_1 + f_1^2c2f_2c_2$

Table 5: Result of query `Q2` applied on the database of Table 1 and its different provenances. The reported numbers are the credit distributed through the process.

Going back to the example of Table 4, consider o_1 with provenance polynomial $f_1c2f_1c_1 + f_1c2f_2c_2$. The how-provenance-based DS firstly divides the credit between the two monomials. Since the coefficients of each monomial are 1, the credit is split in half. If they were, for example, 1 and 2 respectively, 1/3 of the credit would go to the first monomial, and 2/3 to the second. Since in our example each variable has exponent 1, the credit is further divided equally among the three variables. Thus, at the end of the computation, f_1 receives 1/3, and the other tuples receive 1/6. If, for example, the first monomial was $f_1^2c2f_1c_1$, then the portion of credit of this monomial would be divided in this way: 1/2 to f_1 and 1/4 to each of the other two tuples.

In this specific example, the how-provenance-based DS has the same outcome as the one based on why-provenance. We therefore consider another query over GtoPdb, `Q2`, that asks for the families of type `gpcr` that have as contributor a researcher located in the UK:

```
Q2: SELECT DISTINCT F.name
FROM family as F JOIN
(SELECT DISTINCT f.name AS name
FROM family AS f JOIN contributor2family AS c2f ON f.id = c2f.family_id
JOIN contributor AS c ON c2f.contributor_id = c.id
WHERE c.country = "UK") AS R ON F.name = R.name
WHERE F.type = "gpcr"
```

The result of `Q2` is shown in Table 5, and consists of one tuple, annotated with each of the three provenances. As can be seen, lineage and why-provenance are identical to those of the tuple o_1 in the previous example. The how-provenance, however, is different since tuple f_1 is used twice: first in the join of the inner query, and second in the join of the outer query. This information is lost in the first two forms of provenances since they are sets, but it is captured in how-provenance through the use of the operator '$\cdot$'.

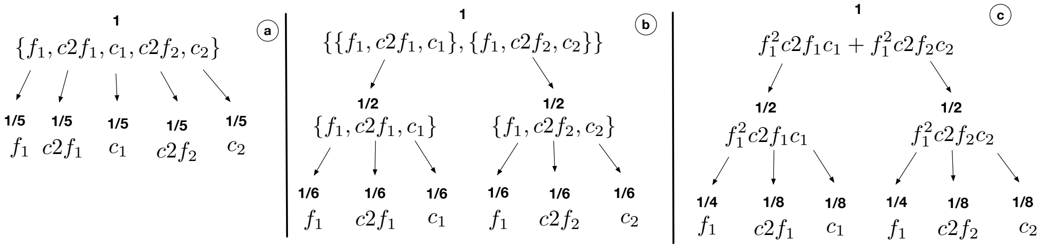


Figure 6: Comparison of different distributions strategies for tuple o_1 produced by query Q2.

Figure 6 shows the differences between the three DS for the tuple o_1 of Table 5. Subfigure 5.a uses lineage, sub-figure 5.b uses why-provenance, and sub-figure 5.c uses how-provenance. The DS based on the provenance polynomial gives credit 1/2 to f_1, and 1/8 to the other tuples. This is reasonable since Q2 relies on f_1 even more than Q1 does. The distribution based on how-provenance can reward f_1 more, showing that how-provenance is even more sensitive to the tuples' role in a query than why-provenance. This is a direct consequence of the fact that, as proven in [30], how-provenance is more general than why-provenance and lineage, in the sense that it contains more information.

6. Experimental Evaluation

To understand the trade-offs between these Distribution Strategies (DSs), we perform four sets of experiments using queries over target families presented on the GtoPdb website. The first set of experiments use real queries extracted from citations to GtoPdb published in the British Journal of Pharmacology. The second set uses synthetically produced provenance polynomials, corresponding to more complex queries, in order to highlight the differences between the DSs. The third set of experiments considers the accrual of credit over time by the three strategies, again using synthetic queries. The fourth set of experiments shows how the DSs compare to traditional citations in giving credit to data curators using both real and synthetic queries. We close by discussing relative execution times of the three strategies.

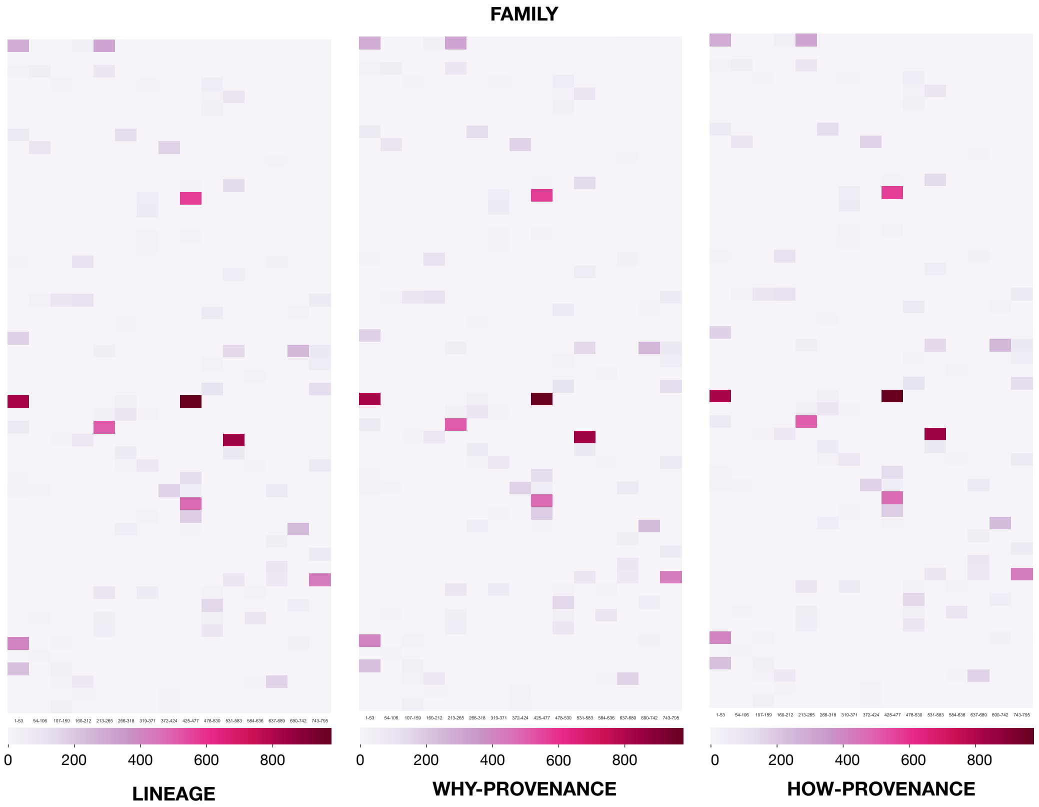


Figure 7: Comparison of three DS on the same table `family` using the distribution given by the queries retrieved from papers.

6.1. Real-world queries

Examples of real queries are drawn from papers published in the British Journal of Pharmacology (BJP) [11]. Each time a paper in this journal cites a webpage from GtoPdb, it reports the URL of the page. From this URL, the query used to obtain the webpage data can be determined. We considered all 889 papers in BJCP citing the IUPHAR/BPS Guide to pharmacology [31] as of October 2020, and extracted all webpage URLs to GtoPdb contained within the paper[12].

[11] `https://bpspubs.onlinelibrary.wiley.com`

[12] The IUPHAR/BPS Guide is a journal that describes the structure and evolution of GtoPdb. At the time of writing, it had received more than 1200 citations on Google Scholar.

There are eight target family types presented on the GtoPdb website: *GPCR*, *Ion channels*, *NHRs*, *Kinases*, *Catalytic receptors*, *Transporters*, *Enzymes* and *Other protein targets*.

The queries that we inferred are those used to build target family webpages. An example was given in Figure 3, where we show how the structure of the "Adenosine receptors" family can be mapped into queries over the underlying database. In GtoPdb, all target family pages share a similar structure; the only difference is that individual sections, such as "contributors" or "further readings", may be absent. Therefore, the same queries can be used to build all of the target family pages by simply changing the family id used in the query (in Figure 3, it is 3). Note that the queries are fairly simple SQL queries, and fall into a class called "select-project-join" or "SPJ" queries. A total of more than $12K$ different queries were built in this way.[13] Without loss of generality, we give each tuple in the output of a query a credit of 1.

Results. Figure 7 shows the heat-maps obtained by the distribution of credit according to the three different DS on one of the tables in the underlying database, `family`, which is often joined with other tables in the database to build the webpages. It can be seen that the result of credit distribution over `family` is the same for all three strategies. The same result is also obtained with the other tables of the database used by the queries shown in Figure 3.

The reason why credit distribution is the same for all three strategies is that the queries are all simple SPJ queries, which use each table only once and do joins on key attributes. Under these conditions, each tuple of the output presents: (i) a how-provenance that is a single monomial with coefficient 1 and exponent 1 in each variable; (ii) a why-provenance with only one witness; and (iii) a lineage that coincides with the witness in the basis. Hence, for these queries, the three DSs behave in the same way: credit is uniformly distributed among the tuples present in each provenance.

To illustrate this, consider one of the queries in Figure 3 which is used to build the output webpage:

```
Q3: SELECT c.first_names, c.surname
FROM contributor2family AS cf JOIN contributor AS c ON
cf.contributor_id = c.contributor_id
```

[13] For reproducibility purposes, the code we used for our experiments and all queries are available here: `https://bitbucket.org/dennis_dosso/credit_distribution_project`.

How-provenance: $3f_1^3c2f_1^2c_1^2 + 2f_1c2f_2^3c_2^3 + 4f_5c2f_{17}^4c_{18}^3$
Credit distribution:
$f_1 = \frac{59}{315}, f_5 = \frac{1}{18}, c2f_1 = \frac{2}{21}, c2f_2 = \frac{2}{15}, c2f_{17} = \frac{2}{9}, c_1 = \frac{2}{21}, c_2 = \frac{2}{15}, c_{18} = \frac{1}{6}$

Why-provenance: $\{\{f_1, c2f_1, c_1\}, \{f_1, c2f_2, c_2\}, \{f_5, c2f_{17}, c_{18}\}\}$
Credit distribution:
$f_1 = \frac{2}{9}, f_5 = \frac{1}{9}, c2f_1 = \frac{1}{9}, c2f_2 = \frac{1}{9}, c2f_{17} = \frac{1}{9}, c_1 = \frac{1}{9}, c_2 = \frac{1}{9}, c_{18} = \frac{1}{9}$

Lineage: $\{f_1, f_5, c2f_1, c_1, c2f_1, c2f_2, c2f_{17}, c_1, c_2, c_{18}\}$
Credit distribution:
$f_1 = \frac{1}{8}, f_5 = \frac{1}{8}, c2f_1 = \frac{1}{8}, c2f_2 = \frac{1}{8}, c2f_{17} = \frac{1}{8}, c_1 = \frac{1}{8}, c_2 = \frac{1}{8}, c_{18} = \frac{1}{8}$

Figure 8: Sample synthetic provenance polynomial (how-provenance) and corresponding why-provenance and lineage expressions with deriving credit distributions.

```
    WHERE f.family_id = 3
```

`Q3` returned 10 tuples from the version of GtoPdb used. The first tuple, `<Bertil B., Fredholm>`, has $c_{939} \cdot c2f_{496}$ as its provenance polynomial. c_{939} represents the provenance token of a tuple in `contributor`, and $c2f_{496}$ the provenance token of a tuple in table `contributor2family`. The why-provenance of this tuple is $\{\{c_{939}, cf_{496}\}\}$ and its lineage is $\{c_{939}, c2f_{496}\}$. Therefore, the credit assigned to these tuples is 1/2 using all three DS. This happens for all the tuples in the output of each query of GtoPdb, thus making the distributions equivalent over all outputs.

However, this is not the case with more complex queries. As we showed in the previous section, when two or more tuples are merged as a result of a projection or union, the credit distributions will differ between the three strategies.

6.2. Synthetic queries

To simulate synthetic queries, we randomly generated provenance polynomials in which the coefficients and exponents could be greater than 1 over three GtoPdb tables: `family`, `contributor2family`, and `contributor`. An example can be found in Figure 8, which shows a sample synthetic provenance polynomial (the how-provenance) and the corresponding why-provenance and lineage expressions. The resulting credit distribution for each DS is shown after the provenance expression.

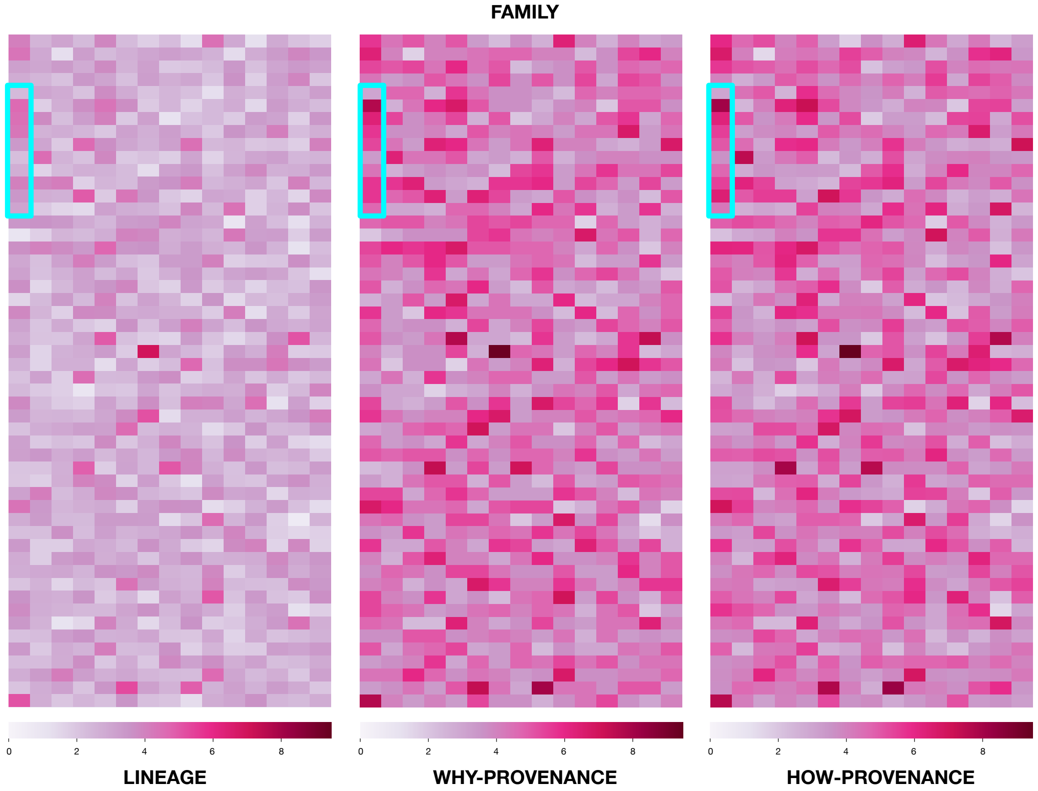


Figure 9: Comparison of three DS on the same table `family` after the distribution computed using 10K synthetic and randomly generated provenance polynomials. The tuples in the blue rectangles are used as example in the discussion connected to Figure 10.

As an example of how the distribution strategies behave with these synthetic queries, consider tuple f_5 in Figure 8. This tuple receives the highest quantity of credit using lineage-based distribution, and less credit using why- and how-provenance because more information is available about the role of the tuple in the overall computation. Generally speaking, the more complex the distribution (the most complex being how-provenance), the more credit is given to tuples which are more frequently used, and thus have a higher impact in producing the output tuple.

Although synthetic, these provenance polynomials represent realistic queries. The polynomials can be obtained by any nested query with join and union operations that use the same tuple multiple times (in which case the exponents are bigger than 1), and the same combination of operations more than once (in which case the coefficients of monomials are bigger than 1).

Results. The results of credit distribution on the `family` table using 10K randomly generated synthetic provenance polynomials are shown in Figure 9. We set the maximum value in the heat maps to the highest value reached by a tuple in all three distributions (i.e., 9.4).

As can be seen, the three strategies generate significantly different credit distributions indicated by the varying hues. We note that, however, there is a consistency in how credit is distributed between the tuples, i.e. tuples that are highly rewarded by one strategy are also highly rewarded by the others, and vice-versa. This shows that the three DS consistently reward certain tuples more than others.

In particular, lineage-based DS gives the least credit to tuples in the `family` table, indicated by an overall lighter hue. This is because the DS equally distributes credit to all tuples appearing in the lineage. Since these queries also use two other tables, credit is distributed to tuples in those tables.

Moving to why-provenance based DS, we see that more credit is given to tuples in the `family` table than with the previous strategy. This is because the DS considers the different ways that a tuple is used, e.g. in joins with other tuples. If the same tuple is present in more than one witness, it will draw more credit and take it from other tuples in the witness basis. In this case, tuples in `family` drew more credit, taking it from tuples in the other two tables, due to the role that `family` tuples played in the queries that were executed.

Finally, consider the how-provenance-based DS heat-map. As with why-provenance, more credit is typically given to tuples in `family` compared to lineage-based DS since it recognizes the role of these tuples in the queries, and the overall hue is deeper. The two distributions appear similar, although, to a closer inspection, it is possible to note some smaller differences between the two distributions. This is because this DS also considers the frequency by which tuples are used, not only the ways in which they are used. Therefore, although the overall distribution is similar, there are small differences due to the presence of exponents and coefficients in the provenance polynomials, influencing the distribution of credit.

To better show this difference, let us consider the ten tuples within each large blue rectangle (each small rectangle within the large blue rectangle is a tuple). We will number them from 1 (top) to ten (bottom), and we use

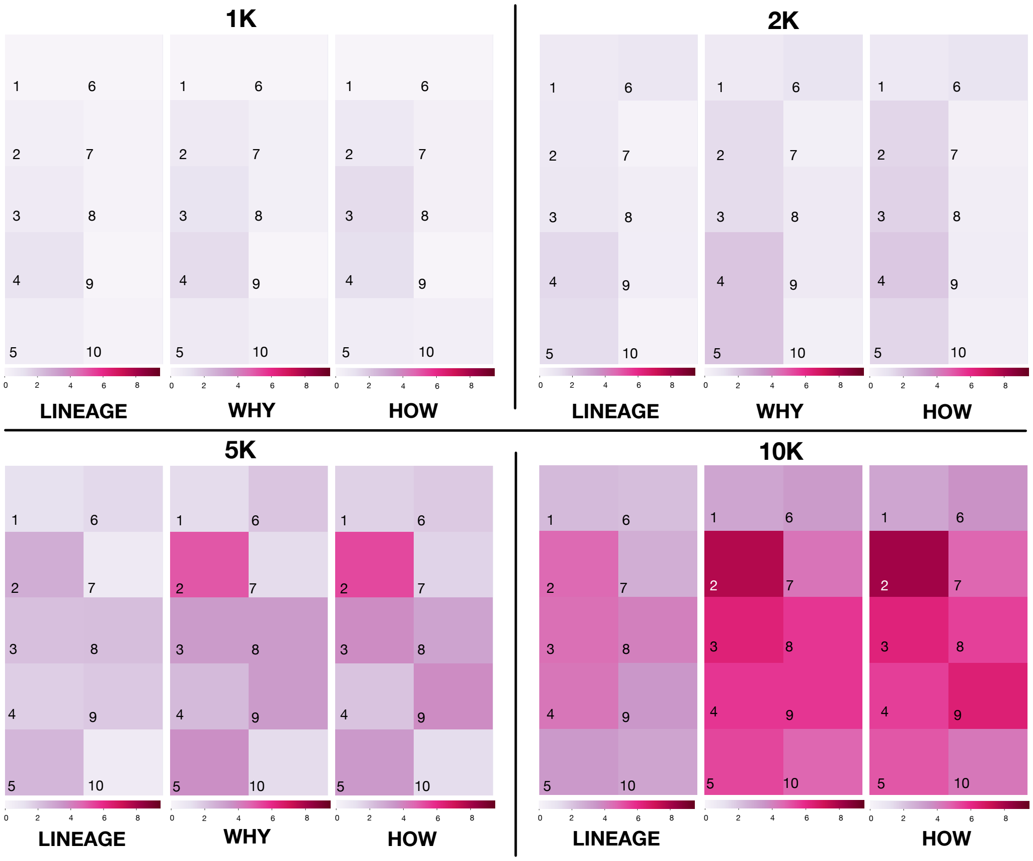


Figure 10: Comparison of the distribution of credit performed by the three DSs on a subset of 10 tuples taken from the `family` table, simulating the passing of time. The number at the top of each group of heat-maps represents the number of queries.

them in the discussion in the next section.

6.3. Credit accrual over time

Since credit accrues over time, we simulate the passage of time by varying the number of queries executed, and look at the "snapshots" of credit for each of the strategies using synthetic queries. The results are shown in Figure 10.

In this figure, four groups of heat-maps are shown. Each group represents a "snapshot" taken after 1K, 2K, 5K and 10K provenance polynomials have been considered for credit distribution. The ten tuples in each heat-map are from the `family` table, and are the ones highlighted previously in the blue boxes of Figure 9.

The queries used are the same as the experiment reported in the previous section. The range of credit in each map goes from 0 (no credit) to 8 (the maximum quantity of credit reached on one of the tuples of the considered window at the "snapshot" with 10K queries). The color hue of the legend, as can be seen, still ranges from 0 to 9.5.

Focusing on the 1K and 2K groups, we see that credit distribution by the three DS are very similar. Still, there are small differences. We note that, in the 1K group, tuple 4 is the one with the highest value of credit with all 3 strategies. Moving to 2K group, it is still the one with the highest value of credit, although it presents the highest value with the why-provenance DS. The other two strategies rewarded it less. Also, why-provenance and how-provenance rewarded more tuples 2 and 3 than the strategy based on lineage. Tuple 5 appears to be more rewarded by why-provenance, and less by how-provenance and even less by lineage. This shows, even with these lower values of polynomials, that the strategies may differ and reward certain tuples more than others, and vice versa. We see the tendency of the lineage DS to reward the tuples in this table less than the other strategies, since it does not take into consideration their importance. Instead, the DS based on why-provenance rewards more tuples like 4 and 5 (values 2 in both cases). The same can be said of the strategy based on how-provenance. However, in this case, tuples 4 and 5 are rewarded a little less (with credit values of 1.9 and 1.5 respectively). This is due to the fact that how-provenance contains more information. Thus, this DS rewarded more other tuples in the other used tables. Viceversa, tuples 2 and 3 are rewarded more by how-provenance DS (values 1.5 and 1.6) that the why-provenance DS (value 1.3 in both cases), due to the fact that their roles in the polynomials are more important.

Moving to the 5K group, we see how credit was accumulated on the tuples. Now tuple 2 is the one with the biggest quantity of credit in this window. This shows how credit is able to track how the importance of tuples changes over time. In this group we see of it is more evident the difference between the distribution based on lineage and the other two strategies. The why-provenance and how-provenance based DSs appear to work similarly, that is to give similar values of credit to the same tuples. We can still see differences, for example on tuples like 8 and 6, that are more rewarded by the DS based on lineage.

Similar observations can be seen for the 10K group. We see how tuple 2 is still highly rewarded by all three provenances. In the case of lineage, however, it is at the same level with tuples 3, 4, and 8, while the other two strategies reward it the most. Once again we see the DSs based on why- and how-provenance operate similar distributions (we still note differences of few decimals between the values assigned to the tuples). However, it is still possible to see how tuples like 9 are more rewarded by one DS, in this case the how-provenance one, than the other. This shows how the last two DS operate in a similar way. The differences between the credit assigned to the same tuple is of few decimals between the two strategies in most of the cases. However, there are certain situations when the role of a tuple is particularly critical in a query. This information is captured by the provenance polynomial, and this is why in certain cases the differences in the credit assigned to one tuple is notably different between the two strategies.

To sum up, the DS based on lineage is sufficient when a user only wants to highlight the tuples of the database used by a query. This strategy equally distributes the credit to the tuples of the lineages. The resulting distribution rewards more the tuples that are used by more queries, but does not reward how the tuples are used in the same query.

For this reason, a user may want (depending on the nature of the queries) to use DS based on why-provenance and how-provenance. Using the why-provenance and how-provenance DS, it is possible to change the distribution of credit to the tuple, rewarding more the tuples that have a more critical role in generating the output. The distribution based on how-provenance is most of the times similar to the one based on why-provenance. However, in certain instances, it differ from the other one due to the specific role of a tuple, an information that is present in the provenance polynomial but lost in the witness basis. These two DSs can be preferred when the user aims to find "hotspots" in the database based on the tuples' role, and the DS based on how-provenance can be chosen when the user wants the higher sensibility to the role of a tuple.

6.4. Credit vs Citations

In the last set of experiments, we compare traditional citations to the proposed credit distribution strategies to see the difference in reward for authors, including data curators.

Each target family page in GtoPdb has a list of curators, that represents all the people who worked on the page and that can be considered as co-authors of that page. It is possible to obtain the list of one page authors using the last query shown in Figure 3. As set of authors, we consider all the curators of one target family in GtoPdb. Each time data are cited, we

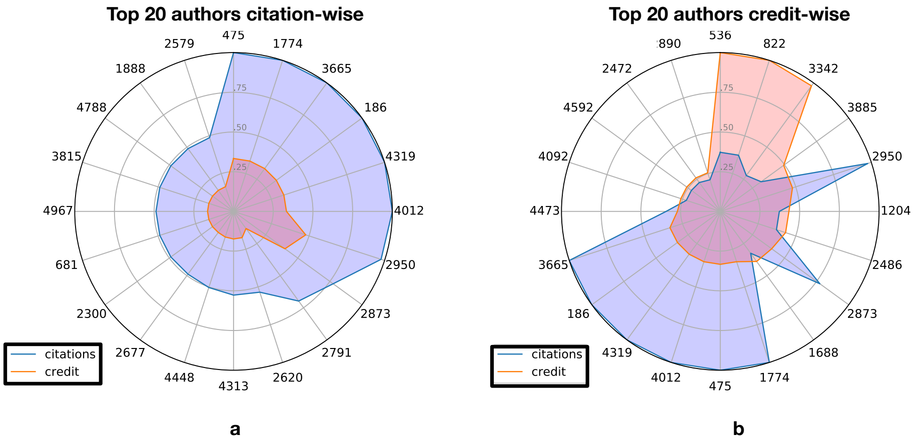


Figure 11: Radars presenting the top 20 authors citation-wise and credit wise, together with their (normalized between 0 and 1) values of citations and credit.

assign one citation to each author associated to that data. Each time some data, e.g. a tuple, receives credit, we equally divide and transitively assign that quantity of credit to the authors of that data.

Results: Real-world queries. As described in Section 6.1, we consider real-world queries, taken from papers published in the BJCP which reference webpages in GtoPdb. Since for these queries there is no difference in the distribution of credit between the three DS, only one value for credit is given in Figure 11. As we said, each time a webpage is cited, the authors of that webpage receive one citation, and also receive a quantity of credit that is equal to the credit assigned to the data and generated from the citation, equally divided among them.

The results are shown in the radar plots of Figure 11, in which each number on the outer circle (e.g. 475, 1774 and 3665) represents an author (id) and the blue (pink) line represents the normalized value of credit generated by citations (credit), respectively. The first radar plot, Figure 11.a, shows the top 20 authors, ordered in a clockwise direction based on citations, whereas Figure 11.b, orders the authors with the highest value of credit. As we see from Figure 11.a, the top 20 authors that present the highest values of citations do not also have the highest values of credit. Vice-versa, as seen in Figure 11.b, the authors with the highest values of credit do not necessarily also have the highest values of citations, i.e. the two sets of authors are not equal (although there is partial overlapping between the two, e.g. authors 1774, 475, 4012, etc.).

From this figure we understand that it is not the case that authors with the higher number of citations also present the highest values of credit. This is due to the fact that certain citations are more "valuable" in terms of credit, i.e. an author receive more credit from his/her citations, even if fewer, than other authors. This, in turns, happens because certain citations generate more credit than others. So, for example, author 536 presents the highest value of credit, although he is not even in the top 20 authors in terms of citations. This means that he receives much more credit from his relatively few citations than author 475.

Given how we prepared our experiments, the citations whose query produce more tuples, are also the ones that generate more credit, since we assumed that each output tuple carries credit 1. Thus, the authors of the data returned by these queries are the ones that will receive more credit from these citations. Also, the authors that collaborated with fewer people will also receive a biggest share from the equal subdivision of credit.

This shows how credit distribution actually rewards differently authors than traditional citations. An author that curated larger quantities of cited data, and together with fewer co-authors, will receive bigger quantities of credit. Thus, credit can reward him for his more relevant "contribution" to the database.

In other settings, where the quantity of credit can be computed considering different criteria, such as the impact of the cited data in the citing paper, credit distribution can even better help to identify the data and the authors that contributed more to the scientific impact in the context defined by the considered citations.

Results: Synthetic queries. We produced 100, $1K$, and $10K$ synthetic polynomials and distributed credit through them to data. *** I don't understand the rest of this paragraph. ARe you saying you can't figure out the creators/curators of the data? But you were able to do that in earlier experiments, so why not now? I'm missing something in the setup. *** Since these polynomials correspond to queries whose authors are not easily identifiable, we created 20 "synthetic" authors, and we randomly assigned one author to each tuple in the database. The authors receive "blocks" of consecutive tuples, with each block of the size varying between 10 and 40 to simulate different quantities of "work" performed by an author.

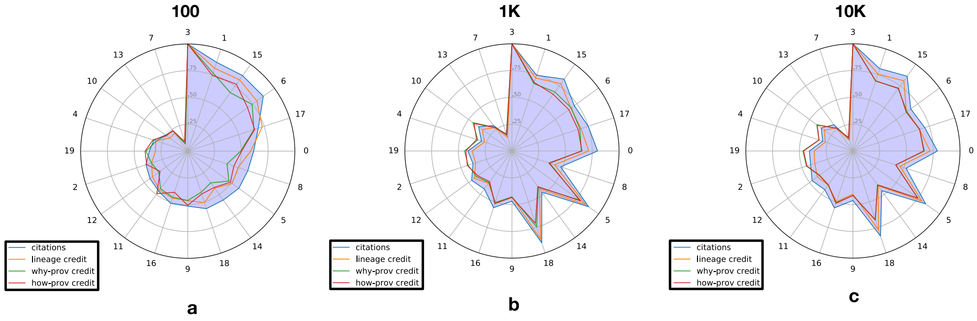


Figure 12: Radars presenting the 20 synthetic authors with corresponding citation and quantities of credit distributed through the 3 DS (all values normalized between 0 and 1) through different numbers of polynomials (respectively, 100, 1K and 10K). The order is the descending one of the citations of the authors with 100 polynomials.

Every time an author appears as curator of one or more tuples used in a query, we assigned one citation to that author. He also receives three kinds of credit, the ones assigned to his tuples through the three different DSs.

Figure 12 reports the three radar plots that are a consequence of the distribution of credit and citations performed and described above with the different quantities of polynomials. Figure 12.a reports the radar plot obtained with 100 polynomials, showing the normalized values of the citation and types of credit assigned to each author. As we see, given the synthetic nature of these queries, the correlation between the number of citations and the quantity of credit assigned to the authors appears to be a much stronger with respect to the case with the real-world queries (the linear correlation between the citation number and all three types of credit is always above 0.95 with p values in the order of 1e-11). Nonetheless, it is still possible to observe how credit does not always exactly follow the citations. The credit distributed via lineage is the one that follows closer the number of citations (a linear correlation of 0.98, p value of 6.15e-16), while the other types of credit behave slightly differently (a linear correlation of around 0.95 in both cases).

Similar observations can be made for Figure 12.b and 12.c, where we kept the order of authors as found in Figure 12.a.

What appears from these figures is that, in certain cases, authors that do not have the highest values of citations receive more credit than others, as for example author 11 in Figure 12.a, or author 19 in Figures 12.b and 12.c, with credit distributed with how-provenance-based DS.

This once again shows how credit allows us to gain a different perspective on the role of data and authors by going beyond the limitations of traditional citations.

It is worth pointing out that, when scaling up to $1K$ and $10K$ polynomials, the distributions performed via why-provenance and how-provenance become almost equivalent. We can note that, although not exactly overlapping, the values of credit assigned to the authors by those DS become quite similar with these higher quantities of polynomials, suggesting a sort of equivalence between the two DSs in this case, at least in the task of rewarding authors (the linear correlation for the values of Figure 12.c is more than 0.99 with a p-value of 1.32e-32).

Since in these experiments we assumed that each output tuple carries credit 1, the queries that return outputs with more tuples also generate more credit. In Figures 11 and 12 the authors that curated bigger bulks of data also receive higher quantities of credit. In more complex and sophisticated scenarios, where different strategies may be implemented to decide the generated quantity of credit to be distributed, new factors beyond the only "quantity" of curated data can be factored in in rewarding data curators. The result will be a distribution of credit that represents even better the actual work and worth of data curators.

6.5. Execution time

# of polynomials	lineage	why-prov.	how-prov.
100	226.6 ms	192.0 ms	185.5 ms
200	431.2 ms	392.2 ms	403.2 ms
500	1.013 s	934.2 ms	881.8 ms
1K	2.041 s	1.934 s	1.744 s
2K	3.773 s	3.491 s	3.510 s
5K	8.992 s	8.653 s	8.889 s
10K	17.10 s	16.84 s	16.84 s
20K	34.59 s	35.30 s	39.70 s
100K	3.289 min	3.442 min	3.652 min
1M	35.91 min	34.87 min	37.91 min

Table 6: The times required to perform the three DS for different number of synthetic polynomials.

*** Where is this table? Table 5 does not show times, and I**

can't find one that does. * Table 5 shows the time required to calculate the credit distribution for the three strategies. As can be seen, the execution time grows linearly with the number of polynomials that are submitted to the system. With a high number of polynomials (1M), the time required by the DS based on lineage and why-provenance is lower than the time needed for the DS based on how-provenance. This is due to the more significant number of operations required to calculate the how-provenance DS and distribute the portions of credit to be assigned to the different tuples. We note that, since we created these polynomials on-the-fly, these values do not include the time required to compute the provenances. Therefore, limited to the time required to distribute credit, the three DS are equivalent in terms of performances. The first differences can be seen only with high number of polynomials, when lineage and why-provenance may be preferred if there are no requirements to assign credit with the strategy implemented by the how-provenance-based DS.

All the experiments were carried out on a MacBook Pro with a 2.4 GHz processor Intel Core i5 quad-core and 8 GB of memory at 2133 MHz. Code was written in Java, supported by a PostgreSQL database.

7. Conclusions

This paper defines two new distribution strategies based on why- and how-provenance, and compares them against the lineage-based distribution strategy defined in [24]. The first DS, based on why-provenance, uses the concept of a witness, and gives more credit to tuples that appear in more than one witness. In this way, tuples that are more important to the query and are used in different ways are rewarded more. The second DS, based on how-provenance, considers the frequency with which a tuple or combination of tuples is used in the query through the information contained in a provenance polynomial. In this case, the distribution is even more sensitive than the first to the role and importance of tuples.

To show the differences between the three DSs, we performed extensive experiments based on GtoPdb, a curated scientific relational database, using both real and synthetic queries. In the first set of experiments, we used select-project-join (SPJ) queries extracted from citations to webpages in GtoPdb found in papers published in the British Journal of Pharmacology. Using these "real" queries, we distributed credit to tuples in different tables of the database, highlighting tuples that were more frequently used. We showed that, with these queries, the three strategies produce the same distribution. This is because the SPJ queries were fairly simple, and did not use self-joins. Therefore the formulas underlying the different DSs had the same output.

In the second set of experiments, we synthetically produced more complex provenance polynomials, corresponding to more complex synthetic queries, that resulted in exponents and coefficients in the provenance polynomials that were greater than (or equal to) 1. These experiments highlighted the differences between the three DSs. While the DS based on lineage rewards all the tuples used by a query equally, the strategy based on why-provenance gives more credit to tuples that are more critical to the query. In particular, why-provenance consider the different ways in which a tuple is used in a query. How-provenance is even more sensitive to the tuple's role: it also considers the frequency with which a tuple or a set of tuples is used.

In the third set of experiments, we showed how the differences between the DS are compounded over time, i.e. when more and more queries are processed by the system.

In the fourth set of experiments we compared traditional citations to authors to the credit accrued to them via the DSs. We showed how, both in the real-world and synthetic scenarios, credit rewards authors who contribute/curate data with the highest quantities of credit, and not necessarily the data with the highest citation count. *** I don't really understand this point. *** In this sense, credit appears to be an useful new measure to discover data and their corresponding curators that have a high impact in the research world, even when they are cited few times or do not appear at all in the data that are cited (i.e. the case of data used to build the output of a query but that is not visualized in the output itself).

In future work, we plan to explore different applications of credit over relational databases. One example is *data pricing*, which gives a price to a query submitted by a user who wants to buy the produced information. Currently, a commonly strategy used for data pricing is based on query rewriting: A database stores a set of views with their price. When a new query arrives, the system rewrites it using the stored views to obtain a query price, a process that can be computationally expensive. We plan to distribute credit through carefully planned and representative queries, and use credit information to define a new, faster, and potentially more flexible pricing function.

Another application is *data reduction* [42], which addresses the problem of reducing the vast – and rapidly expanding – amount of data that is being produced. Ideas that are being explored include finding reduced and relevant data streams from the multi-gigabytes of data produced by big data systems every second, and dealing with the curse of dimensionality which requires unbounded computational resources to uncover actionable knowledge patterns [51]. *** I'm not sure what the previous sentence means! ***

Data credit can also address this problem, by helping find "hotspots" and "coldspots" of data. A hotspot is data in a database (e.g. a tuple) with a high quantity of credit, which is therefore valuable for the set of queries that execute frequently over the data and distribute the credit. On the other hand, a coldspot is data with a low quantity of credit, which is therefore considered less important and could be deleted or moved to cheaper and/or less efficient memory.

References

[1] Abadi, D., Ailamaki, A., Andersen, D., Bailis, P., Balazinska, M., Bernstein, P., Boncz, P., Chaudhuri, S., Cheung, A., Doan, A., Dong, L., Franklin, M. J., Freire, J., Halevy, A., Hellerstein, J. M., Idreos, S., Kossmann, D., Kraska, T., Krishnamurthy, S., Markl, V., Melnik, S., Milo, T., Mohan, C., Neumann, T., Chin Ooi, B., Ozcan, F., Patel, J., Pavlo, A., Popa, R., Ramakrishnan, R., Ré, C., Stonebraker, M., and Suciu, D. (2020). The seattle report on database research. *SIGMOD Rec.*, 48(4):44–53.

[2] Alawini, A., Davidson, S. B., Hu, W., and Wu, Y. (2017). Automating data citation in citedb. *PVLDB*, 10(12):1881–1884.

[3] Alawini, A., Davidson, S. B., Silvello, G., Tannen, V., and Wu, Y. (2018). Data citation: A new provenance challenge. *IEEE Data Eng. Bull.*, 41(1):27–38.

[4] Altman, M., Borgman, C. L., Crosas, M., and Martone, M. (2015). An Introduction to the Joint Principles for Data Citation. *Bulletin of the Association for Information Science and Technology*, 41(3):43–45.

[5] Baggerly, K. (2010). Disclose all data in publications. *Nature*, 467(7314):401–401.

[6] Bechhofer, S., Buchan, I. E., De Roure, D., Missier, P., Ainsworth, J. D., Bhagat, J., Couch, P. A., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Michaelides, D. T., Owen, S., Newman, D. R., Sufi, S., and

Goble, C. A. (2013). Why linked data is not enough for scientists. *Future Gener. Comput. Syst.*, 29(2):599–611.

[7] Belter, C. W. (2014). Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets. *PLoS ONE*, 9(3):e92590.

[8] Bertin-Mahieux, T., Ellis, D., Whitman, B., and Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, pages 591–596.

[9] Borgman, C. L. (2016). Data Citation as a Bibliometric Oxymoron. In Sugimoto, C. R., editor, *Theories of Informetrics and Scholarly Communication*, pages 93–116. De Gruyter Mouton.

[10] Buneman, P. (2006). How to cite curated databases and how to make them citable. In *18th International Conference on Scientific and Statistical Database Management, SSDBM*, pages 195–203. IEEE Computer Society.

[11] Buneman, P., Christie, G., Davies, J. A., Dimitrellou, R., Harding, S. D., Pawson, A. J., Sharman, J. L., and Wu, Y. (2020). Why data citation isn't working, and what to do about it. *Database J. Biol. Databases Curation*, 2020.

[12] Buneman, P., Davidson, S. B., and Frew, J. (2016). Why data citation is a computational problem. *Commun. ACM*, 59(9):50–57.

[13] Buneman, P., Khanna, S., and Tan, W. C. (2001). Why and where: A characterization of data provenance. In *Database Theory - ICDT 2001, 8th International Conference*, pages 316–330.

[14] Buneman, P. and Silvello, G. (2010). A rule-based citation system for structured and evolving datasets. *IEEE Data Eng. Bull.*, 33(3):33–41.

[15] Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., Ault, L., Bell, P., Bowie, R., Leadbetter, A. M., Lowry, R. K., Moncoiffé, G., Harrison, K., Smith-Haddon, B., Weatherby, a., and Wright, D. (2012). Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres. *International Journal of Digital Curation*, 7(1):107–113.

[16] Candela, L., Castelli, D., Manghi, P., and Tani, A. (2015). Data Journals: A Survey. *Journal of the Association for Information Science and Technology*, 66(9):1747–1762.

[17] Cheney, J., Chiticariu, L., and Tan, W. (2009). Provenance in databases: Why, how, and where. *Foundations and Trends in Databases*, 1(4):379–474.

[18] CODATA-ICSTI Task Group on Data Citation Standards and Practices (2013). *Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data*, volume 12.

[19] Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., and Simons, N. (2019). Bringing citations and usage metrics together to make data count. *Data Science Journal*, 18(1).

[20] Cronin, B. (1984). *The citation process. The role and significance of citations in scientific communication.* London: Taylor Graham.

[21] Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *JASIST*, 52(7):558–569.

[22] Cui, Y., Widom, J., and Wiener, J. L. (2000). Tracing the lineage of view data in a warehousing environment. *ACM Trans. Database Syst.*, 25(2):179–227.

[23] Davidson, S. B., Deutch, D., Milo, T., and Silvello, G. (2017). A model for fine-grained data citation. In *CIDR 2017, 8th Biennial Conference on Innovative Data Systems Research*. www.cidrdb.org.

[24] Dosso, D. and Silvello, G. (2020). Data credit distribution: A new method to estimate databases impact. *Journal of Informetrics*, 14(4):101080.

[25] Dubernet, M. L., Antony, B. K., Ba, Y. A., et al. (2016). The virtual atomic and molecular data centre (VAMDC) consortium. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 49(7):074003.

[26] Fang, H. (2018). A discussion of citations from the perspective of the contribution of the cited paper to the citing paper. *JASIST*, 69(12):1513–1520.

[27] Force, M., Robinson, N., Matthews, M., Auld, D., and Boletta, M. (2016). Research data in journals and repositories in the web of science: Developments and recommendations. *Bulletin of IEEE Technical Committee on Digital Libraries, Special Issue on Data Citation*, 12(1):27–30.

[28] Garfield, E. (1999). Journal impact factor: a brief review. *Can. Med. Assoc.*, 979-980.

[29] Gößwein, B., Miksa, T., Rauber, A., and Wagner, W. (2019). Data identification and process monitoring for reproducible earth observation research. In *2019 15th International Conference on eScience (eScience)*, pages 28–38. IEEE.

[30] Green, T. J., Karvounarakis, G., and Tannen, V. (2007). Provenance semirings. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 31–40. ACM.

[31] Harding, S. D., Sharman, J. L., Faccenda, E., Southan, C., Pawson, A. J., Ireland, S., Gray, A. J. G., Bruce, L., Alexander, S. P. H., Anderton, S., Bryant, C., Davenport, A. P., Doerig, C., Fabbro, D., Levi-Schaffer, F., Spedding, M., Davies, J. A., and Nc-Iuphar (2018). The IUPHAR/BPS guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. *Nucleic Acids Research*, 46(Database-Issue):D1091–D1106.

[32] Hartley, J. (2017). Authors and their citations: a point of view. *Scientometrics*, 110(2):1081–1084.

[33] Hey, T., Tansley, S., and Tolle, K. M. (2009). Jim Gray on eScience: a transformed scientific method.

[34] Honor, L. B., Haselgrove, C., Frazier, J. A., and Kennedy, D. N. (2016). Data citation in neuroimaging: proposed best practices for data identification and attribution. *Frontiers in neuroinformatics*, 10:34.

[35] Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G. R., Wu, G. R., Matthews, L., Lewis, S., Birney, E., and Stein, L. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(Database-Issue):428–432.

[36] Katz, D. (2014). Transitive credit as a means to address social and technological concerns stemming from citation and attribution of digital products. *Journal of Open Research Software*, 2(1).

[37] Katz, D. S., Hong, N., Clark, T., Fenner, M., and Martone, M. (2020). Software and data citation. *Computing in Science & Engineering*, 22 (2):4–7.

[38] Kosten, J. (2016). A classification of the use of research indicators. *Scientometrics*, 108(1):457–464.

[39] Lawrence, B., Jones, C., Matthews, B., Pepler, S., and Callaghan, S. (2011). Citation and Peer Review of Data: Moving Towards Formal Data Publication. *International Journal of Digital Curation*, 6(2):4–37.

[40] Martone, M. (2014). Joint declaration of data citation principles. *FORCE11. San Diego CA. Data Citation Synthesis Group.* `https://www.force11.org/datacitationprinciples`, online September 2020.

[41] Meho, L. I. and Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar. *Journal of the american society for information science and technology*, 58(13):2105–2125.

[42] Milo, T. (2019). Getting rid of data. *Journal of Data and Information Quality (JDIQ)*, 12(1):1–7.

[43] Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, M., Paluck, E. L., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E. J., Wilson, R., and Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242):1422–1425.

[44] Peters, I., Kraker, P., Lex, E., Gumpenberger, C., and Gorraiz, J. (2016). Research data explored: An extended analysis of citations and altmetrics. *Scientometrics*, 107(2):723–744.

[45] Pröll, S. and Rauber, A. (2013). Scalable data citation in dynamic, large databases: Model and reference implementation. In *Proceedings of the 2013 IEEE International Conference on Big Data*, pages 307–312. IEEE.

[46] Rauber, A., Ari, A., van Uytvanck, D., and Pröll, S. (2016). Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use. *Bulletin of IEEE Technical Committee on Digital Libraries, Special Issue on Data Citation*, 12(1):6–15.

[47] Rauber, A., Asmi, A., van Uytvanck, D., and Proell, S. (2015). Data citation of evolving data: Recommendations of the working group on data citation (wgdc). *Result of the RDA Data Citation WG*, 20.

[48] Silvello, G. (2018). Theory and practice of data citation. *J. Assoc. Inf. Sci. Technol.*, 69(1):6–20.

[49] Simmhan, Y., Plale, B., and Gannon, D. (2005). A survey of data provenance in e-science. *SIGMOD Record*, 34(3):31–36.

[50] Spengler, S. (2012). Data Citation and Attribution: A Funder's Perspective. In of Sciences' Board on Research Data, N. A. and Information, editors, *Report from Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop*, pages 177–178. National Academies Press: Washington DC.

[51] Ur Rehman, M. H., Liew, C. S., Abbas, A., Jayaraman, P. P., Wah, T. V., and Khan, S. U. (2016). Big data reduction methods: a survey. *Data Science and Engineering*, 1(4):265–284.

[52] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3.

[53] Wu, Y., Alawini, A., Davidson, S. B., and Silvello, G. (2018). Data citation: Giving credit where credit is due. In *Proceedings of the 2018 International Conference on Management of Data, SIGMOD*, pages 99–114.

[54] Zeng, T., Wu, L., Bratt, S., and Acuna, D. E. (2020). Assigning credit to scientific datasets using article citation networks. *Journal of Informetrics*, 14(2).

[55] Zou, C. and Peterson, J. B. (2016). Quantifying the scientific output of new researchers using the zp-index. *Scientometrics*, 106(3):901–916.

[56] Zwölf, C. M., Moreau, N., and Dubernet, M.-L. (2016). New Model for Datasets Citation and Extraction Reproducibility in VADMC. *Journal of Molecular Spectroscopy*, 327:122–137.