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, USA

Abstract

Digital data is an important form of research product for which citation, and the generation of credit or recognition for authors, is still not well understood. The notion of *data credit* has therefore recently emerged as a new metric, defined and based on data citation theory.

Data credit is a real value that represents the importance of data cited by a paper or by another research entity. Credit can be used to annotate data contained in a curated scientific database, and used as a measure for the importance and impact of that data in the research world. As such, it is a new method that, together with traditional citations, helps recognize the value of data and its creators.

In this paper we explore the problem of Data Credit Distribution, the process by which credit is distributed to the database parts 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 widely-used curated scientific relational database. We define two new distribution strategies based on two forms of data provenance, why-provenance and how-provenance.

Using different distribution strategies, we show how credit can highlight frequently used database areas and how it be used as a new bibliometric measure for data and their corresponding curators. In particular, credit rewards data and authors based on their research impact, not merely on the number of citations. We also show how different distribution strategies, based on different types of data provenance, can vary in their sensitivity to an input tuple in the generation of the output data and reward input tuples 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 created and understood. They form a basis on which to give credit to authors, papers, and venues [20, 21, 58]. Citations are used, among other things, to decide on tenure, promotion, hiring, and funding of grants for researchers [22, 35, 40, 43].

Science and research are increasingly digital, and 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 [15, 41], and that data citations should contribute to the scientific reputation of researchers, scientists, data curators, and creators [4, 54]. It is also accepted that data citations should be counted alongside of traditional citations, and contribute to bibliometrics indicators [7, 48].

Many initiatives, at different levels, have been promoted to make data citation a reality. Scientific publishers, such as Elsevier, Springer and Nature, have been defining data policies and author guidelines to include data citations in the reference lists of published papers [20]. The European Commission has introduced the Open Research Data Pilot (ODP), whose aim is to improve and maximize the access and re-use of research data, together with an increase to the credit given to data creators and curators [52]. Initiatives such as FORCE11 and ESIP (Earth Science Information Partners) have collaborated on data and software citation principles and guidelines [28]. Other examples are the National Science Foundation (NSF), and the National Institute of Health (NIH) in the US [52].

Moreover, there are activities to promote and specify guidelines for data citations. A significant activity getting a broad adoption, is the Research Data Alliance (RDA), that produced a recommendation on citing specific subsets of dynamic data [51].While this approach provides reference and access to a precise subset of data, it does not address specific credit concerns for that subset, such as when different authors contribute to a larger collection [47].

A central problem in the data citation process 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 [9, 30]. However, even when correctly applied, data citations and the bibliometrics computed using them do not always correctly or completely 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) [29, 39, 57] 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 DCD 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.

In this paper, we consider data credit as a measure of value for data in a (curated) scientific database. Credit is a real value that 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 DCD 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. When a paper p_1 cites another paper p_2, a +1 citation "credit" is given to p_2, and to all its authors. It does not matter why or how paper p_1 cites paper p_2, [1] the result is always +1 to the citation count of p_2 and of its authors. A different credit distribution strategy can assign a quantity of credit to p_2, and its authors, that is *proportional* to the

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

role played by p_2 in p_1. Hence, we can weight the importance of the cited entities and assign credit according to their role.

2. Traditional citations are *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. In contrast, with data credit, we use 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. Citations, traditionally, are not.

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 [12, 14, 49]. RDBs are also frequently a test-bed for new methods that can be adapted to other databases, e.g., graphs or document databases. 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, (iii) tuples, and (iv) attributes. The ability to specify different levels of granularity in a relational database allows us to define the DCD problem at a particular level of granularity. In this paper, we focus on DCD at the tuple level.

The DCD process is summarized in Figure 1:

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

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

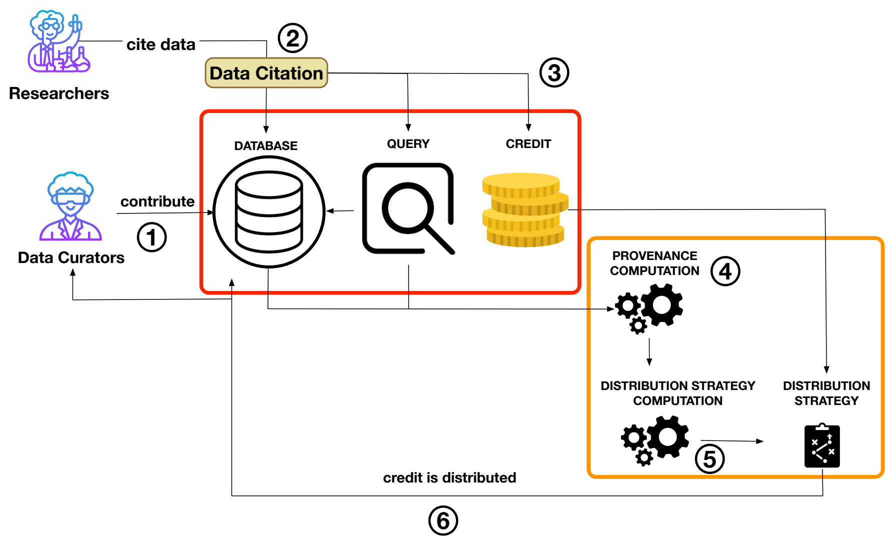

Figure 1: Overview of the credit distribution pipeline.

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 [13, 23, 32], 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 DCD 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 [25], 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 [25] by proposing two new DSs based on other forms of data provenance: why-provenance [13] and how-provenance [32]. 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 gives 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 [34], also known as GtoPdb[4], which contains expertly curated information about diseases, drugs, cellular drug targets, and

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

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

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 four sets of experiments. In the first one, real queries are extracted 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. In the fourth set of experiments we use both real and synthetic queries to assess the difference between traditional citation and the notion of credit distribution in terms of rewarding those responsible for the data, e.g. data curators.

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.
- A comparison between the behavior of traditional citations and data credit in rewarding data curators.

Outline. The rest of the paper is organized as follows: Section 2 presents the background and related work. Section 3 describes the GtoPdb use case we adopted. Section 4 briefly presents the forms of provenance used in the paper. Section 5 describes the credit distribution problem and the proposed distribution strategies. In Section 6 we present the experimental evaluation. Finally, Section 8 draws some conclusions and outlines future work.

2. Background

Data in Research. The world of research is rapidly transitioning towards the *fourth paradigm of science* [36], 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* [46], founded on methods and tools to share, discover, and access experimental data. The community has identified the FAIR principles (Findable, Accessible, Interoperable, and Reusable) [55], that should be enforced by every database. In particular, data should be accessible from the articles, journals, and papers that cite or use them [20]. 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* [37].

Data Citation: Principles and Motivations. Data Citation principles were proposed in [19], and later summarized and endorsed by the Joint Declaration of Data Citation Principles (JDDCP) [42]. The principles are divided into two groups [52]. 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 [52]:

- *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] [50] 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 Working Group are often relational.

In one of its most recent sessions [51], 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) [31]; 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) [26, 59].

More examples of work on data citation in relational databases are [2, 12, 24, 56]. 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.

---

[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`

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 [2, 10, 12] 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, [56] 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 papers" 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 [50]. 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 [38], the GtoPdb [34], and the VAMDC [59]. Every single Webpage is unequivocally identifiable and can be individually cited.

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 [39] 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 [39] credit is defined as a "quantity" that describes the importance of a research entity, such as papers, software, or data, mentioned in a citation. It also proposed the idea of a *distribution* of credit from research entities, such as papers or data, to other research entities through citations. Therefore, when talking about data credit, here we are focusing on two aspects of the topic: *credit computation*, the process in which the quantity of credit generated by the citation is computed, and *credit distribution*, the process by which credit is distributed and assigned to the responsible entities that contributed to the generation of the data being cited. In this paper we focus on the latter.

These two processes are 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. We add to this that the concept of credit can be built on top of the existing infrastructure handling traditional and data citations.

Katz [39] 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 [58] 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 [29] 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. [57] 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. [57], 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 [53].

Cheney et al. [17] presents four main types of data citation for DBMS: *lineage* [23], *why-provenance* [13], *how-provenance* [32] 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 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 drew from $\mathbb{N}$. This provenance also contains information on *how* the input tuples are used. For example, when two tuples are combined by a join, they are also combined in the polynomial by the $\cdot$ 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 [34] or GtoPdb[9]. GtoPdb is a well-known and well structured scientific relational database that contains expertly curated information about diseases, drugs 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

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

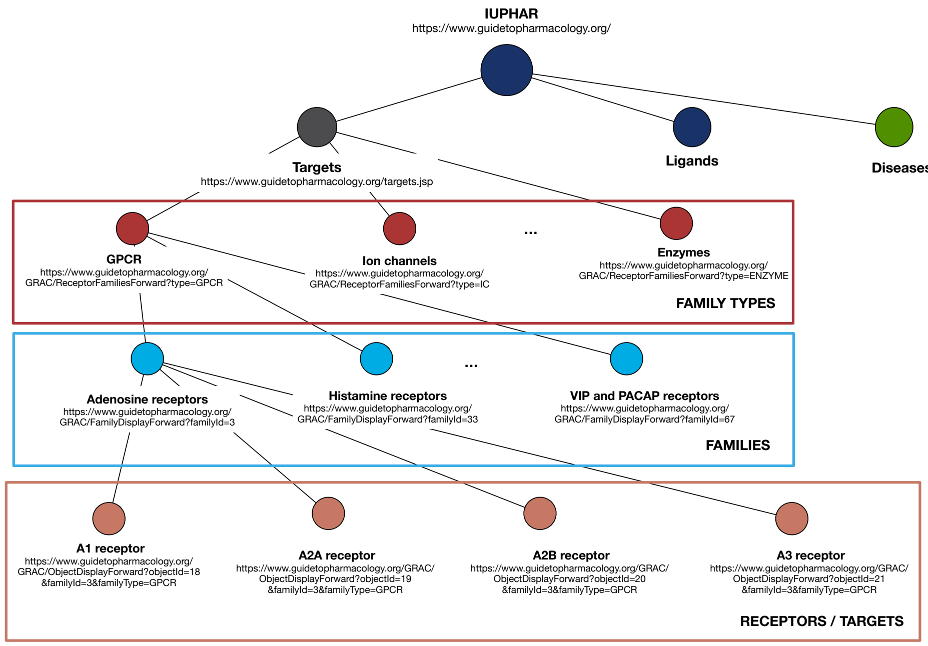

Figure 2: Partial map of the GtoPdb hierarchical structure grouping the targets into families and family types.

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

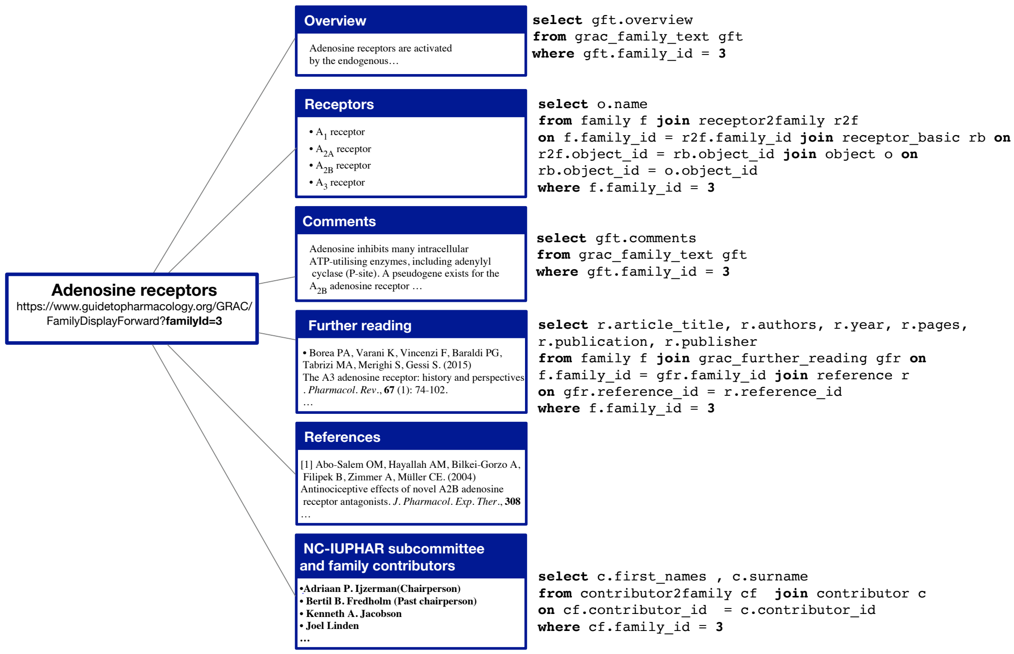

Figure 3: Basic web-page structure of "Adenosine receptors" family (ID 3), with queries used to retrieve the information contained in every section, except references.

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., [34]) 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.

In certain "lucky" cases, as with papers available in PDF and published

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.

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 "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.

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

4. Data Provenances

In this section, we present the three types of provenance used in this paper: lineage, why-provenance, and how-provenance. We also discuss of Causality and Responsibility that, even though are not forms of data provenance *per se*, they are still used as basis to define a DS.

4.1. Lineage

Lineage was first introduced by Cui et al. [23]. 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 contributed to its "production" [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 set, 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

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.

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 [32] 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. [32] 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 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).

4.4. Causality and Responsibility

A formal study of causality was initiated in [18, 33] and later expanded by Meliou et al. [44] to define the causes of answers and non-answers to queries. Causality is, more precisely, related to the provenance of a query result such as why-provenance. Causality adds information to the one already provided by the provenance.

In the following we define causality and responsibility as done in [44]. Differently from [44], we only focus on answers of a query, and not non answers, since they are not relevant in the context of this paper. Let $R_1, \ldots, R_k$ be the relation names of a standard relational schema, D be a database instance and q a conjunctive query. We also call $D^n \subseteq D$ the set of *endogenous tuples*, i.e. the tuples being actually considered to be possible causes of a query output; while $D^x = D - D^n$ is the set of *exogenous tuples*, the tuples being considered external, unconcerned factors, thus deemed not to be possible causes. This distinction between endogenous and exogenous tuple is application dependent, and it can be done by the user at query time. One example is with probabilistic databases with uncertain tuples, where erroneous data may be contained. By considering these uncertain tuples as part of the exogenous tuples dataset, we are factoring them out of the computation of causality.

Then, given a tuple $\bar{a}$ with the same arity as the query's answer, we write $D \vDash q(\bar{a})$ when $\bar{a}$ is an answer to q on D, and write $D \nvDash q(\bar{a})$ when $\bar{a}$ is a non-answer to q on D. Causality is defined as follows:

Definition 4.1. *Causality [44]*
Let $t \in D^n$ be an endogenous tuple, and $\bar{a}$ a possible answer for q. Then:

1. *t is called a* counterfactual cause *for $\bar{a}$ in D if $D \vDash q(\bar{a})$ and $D - \{t\} \nvDash q(\bar{a})$*
2. *$t \in D$ is called an* actual cause *for $\bar{a}$ if there exists a set $\Gamma \subseteq D^n$, called* contingency *for t, such that t is a counterfactual cause for $\bar{a}$ in $D - \Gamma$.*

id	name	responsibility
o_1	Dopamine Receptors	$f_1 : 1, c2f_1 : 0.5, c2f_2 : 0.5, c_1 : 0.5, c_2 : 0.5$
o_2	YANK Family	$f_4 : 1, c2f_4 : 1, c_1 : 1$

Table 5: Result of the example SQL query `Q1` with the corresponding responsibilities of the lineage tuples.

t is a *counterfactual cause* if, by removing it from the database, we remove $\bar{a}$ from the answer. Therefore, it can be fought as a tuple of the lineage which is fundamental for the presence of $\bar{a}$ in the answer. Vice-versa, t is an actual cause if it is possible to find a contingency set of tuples such that, if that set is removed, only then t becomes fundamental. In other words, when t is an actual cause, even if it was removed from the database, $\bar{a}$ would still be present in the result set thanks to the contingency set. Checking the causality degree of tuples is NP-complete in general [27], but Meliou et al. [44] proved that the causality of conjunctive queries may be determined in PTIME.

The notion of *responsibility* was first defined in [18], and it measure the degree of causality as a function of the size of the smallest contingency set. It allows to rank the tuples in a lineage based on their degree of causality in generating the output.

Definition 4.2. *Responsibility [44] Let $\bar{a}$ be an answer to a query q, and let t be a cause. The* responsibility *of t for the answer $\bar{a}$ is:*

$$\rho_t = \frac{1}{1 + min_{\Gamma}|\Gamma|}$$

where Γ ranges over all contingency sets for t.

As can be seen, a counterfactual cause will have the maximum responsibility of 1, while the bigger the minimum contingency of an actual cause, the smaller its responsibility since more tuples can still guarantee the presence of the answer $\bar{a}$.

While in general computing the responsibility is hard [18], Meliou et al. [44] showed that for each query without self-joins the responsibility is either computed in PTIME in the size of the database or checking if it has a responsibility below a given value is NP-hard.

As an example, consider Table 4, where we reported the tuples result of query `Q1` together with the tuples of their lineage accompanied with their responsibility values. With output tuple o_1, the tuple f_1 of the lineage is a counterfactual cause, since its contingency set is empty (when removed from the database, o_1 disappears from the result set). Consequently, its responsibility is 1. On the other hand, the other tuples of the lineage are all actual causes. c_1, for example, has as minimal contingency set $\{c2f_2\}$, and thus its responsibility is 0.5. For the output tuple o_2, all the tuples of the lineage are counterfactual causes, and thus they all have responsibility 1.

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)*** *[25]* *Given a query Q over I and $k \in \mathbb{R}_{>0}$, DCD is defined by the function $f_{I,Q} : TupleLoc \times \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ 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 [25]*
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 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

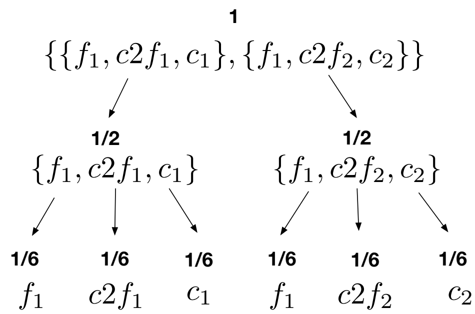

Figure 4: Distribution of credit using why-provenance-based DS for tuple o_1.

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.

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. This notation is also reported for easy reference in Table 6.

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 call e 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. te is a function that takes as input a tuple and a monomial,

Table 6: Notations used in Definition 5.4.

$\mathcal{H}$	provenance polynomial
M_i	a monomial in $\mathcal{H}$
t_j	a tuple in M_i
$c(\mathcal{H})$	sum of $\mathcal{H}$'s coefficients
$e(M_i)$	sum of M_i's exponents
$mc(M_i)$	M_i's coefficient
$te(t_j, M_i)$	exponent of t_j in M_i
$\gamma(t_j, \mathcal{H})$	set of monomials in $\mathcal{H}$ containing t_j

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

$$c(\mathcal{H}) = 4 \qquad e(M_2) = 7$$
$$mc(M_1) = 3 \qquad mc(M_2) = 1$$
$$te(c_2, M_2) = 3 \qquad \gamma(c_1, \mathcal{H}) = \{M_1\}$$
$$\gamma(f_1, \mathcal{H}) = \{M_1, M_2\}$$

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

and returns the exponent of the tuple in the monomial, if present; thus $te(c_2, M_2) = 3$. Finally, 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:

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

Going back to the example of Table 4, consider o_1 with provenance polynomial $f_1 c2 f_1 c_1 + f_1 c2 f_2 c_2$. 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

id	name
oxs_1	Dopamine Receptors

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

Table 7: 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.

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 $1/6$. Consider instead the example where the polynomial is $f_1^2 c2f_1 c_1 + f_1^2 c2f_2 c_2$ and let us focus on the first monomial. The monomial receives $1/2$ of the credit, then f_1 receives $1/4$ of this portion of credit, while the other two tuples receive $1/8$.

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 7, 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 shows the differences between the three DS for the tuple o_1 of Table 7. Subfigure 7.a uses lineage, sub-figure 7.b uses why-provenance, and sub-figure 7.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

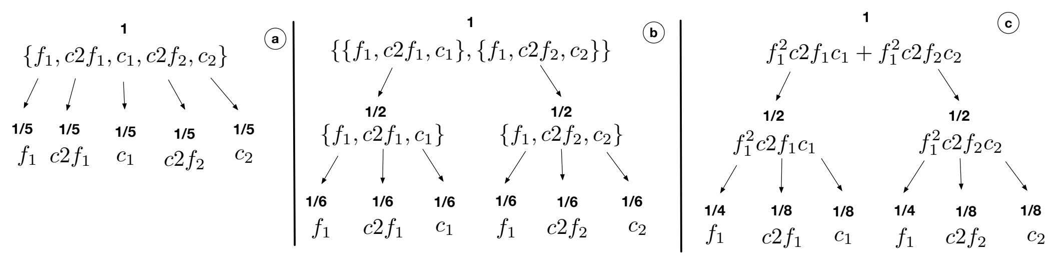

Figure 6: Comparison of different distributions strategies for tuple o_1 produced by query `Q2`.

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 [32], how-provenance is more general than why-provenance and lineage, in the sense that it contains more information.

5.5. Responsibility-based Distribution Strategy

As we described in Section 4.3, causality and responsibility are not new forms of data provenance, but rather new information that is added to the already available lineage. It is possible in fact to compute the causality of all the tuples of a lineage, distinguishing them between actual causes and counterfactual causes. Successively, it is also possible to compute their responsibility, which, by itself, can be envisioned as a form of credit and assigned to the corresponding tuples.

One first option to define a distribution strategy using responsibility is to simply assign the responsibility as credit of the single tuple. Using the example of Table 5, in the case of output tuple o_1, f_1 receives credit 1, the other tuples credit 0.5. This strategy however both generates the credit and gives it to the tuples. We want a DS that is also a function of the input credit value k in order to be comparable with the other strategies proposed so far.

Therefore, we define a new DS based on responsibility that is a function of the quantity of credit k_o that assigns to each tuple of the lineage a portion of this credit weighted by its normalized quantity of responsibility. Formally:

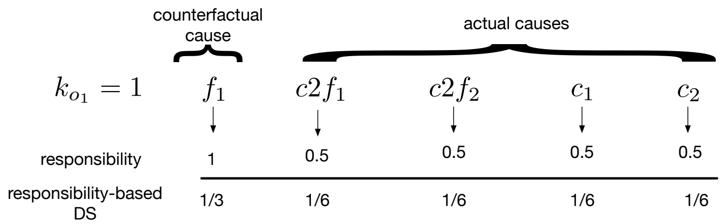

Figure 7: Example of distribution of credit using responsibility and normalized responsibility and the responsibility-based DS, assuming $k_o = 1$.

Definition 5.5. *Responsibility-based Distribution Strategy*
Let I be a database instance, Q a query over I, $o \in Q(I)$ an output tuple, $\mathcal{L}$ the lineage of o, and k_o the credit given to o. The credit given to tuple t in I is:

$$f_{I,Q}(t, k_o) = k_o \frac{\rho_t}{\sum_{t' \in \mathcal{L}} \rho_t'}$$

Note that only the tuples that belong to the lineage will receive a quantity of credit > 0. The more important the tuple, i.e., the higher its responsibility, the bigger the quantity of credit received.

Figure 7 shows the quantity of credit assigned to the tuples of the lineage of the output tuple o_1 of Table 5 when credit = responsibility and when we use the responsibility-based DS. with simple responsibility, we assign 1 to the only counterfactual tuple f_1, and 0.5 to the other tuples. Using the DS instead, and assuming that $k_{o_1} = 1$, f_1 receives credit 1/3, while the others receive credit 1/6. As we see, the DS in this case returns the same distribution obtained with why-provenance, as shown in Figure 6. This is not always the case though, as we show in the example of Section 6.2.

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 better 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

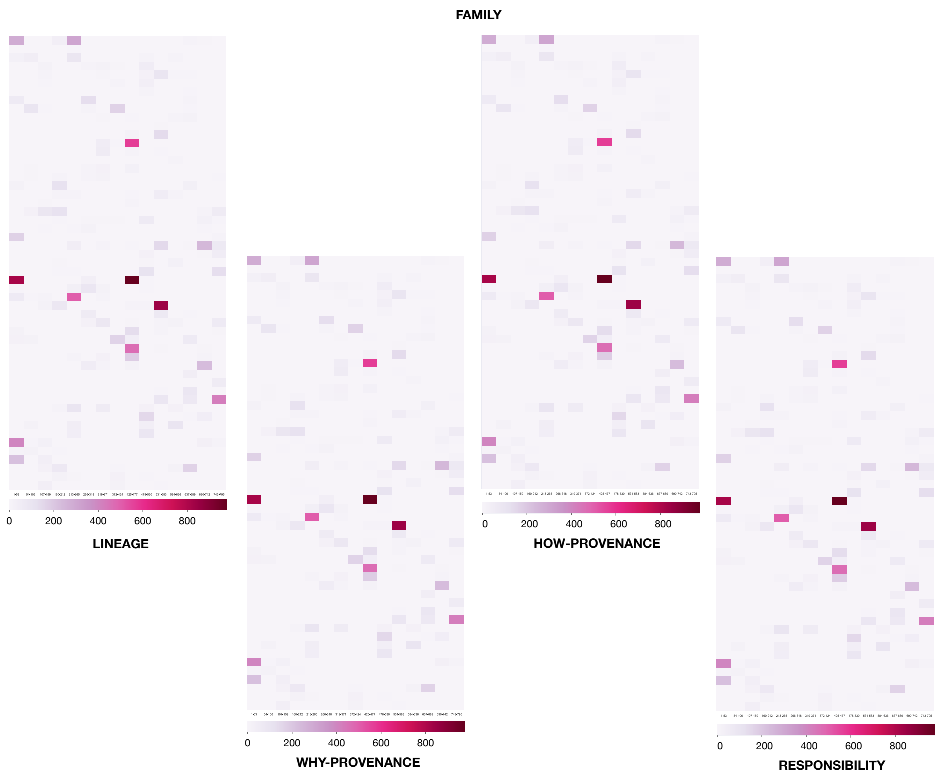

Figure 8: Comparison of four DS on the same table `family` using the distribution given by the queries retrieved from papers. Each cell is a tuple.

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.

All 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.[11]

6.1. Real-world queries

Examples of real queries are drawn from papers published in the British Journal of Pharmacology (BJP). [12] 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 [34] as of October 2020, and extracted all webpage URLs to GtoPdb contained within the paper.[13]

The queries that we inferred are those used to build target family webpages within GtoPdb. 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 changing the family id used in the query (for example, 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. Without loss of generality, we give each tuple in the output of a query a credit of 1.

Results. Figure 8 shows the heat-maps obtained by the distribution of credit according to the four 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. Each cell in a heat-map represents a tuple of the `family` table and the color indicates the amount of credit attributed to such tuple. It can be seen that the result of credit distribution over `family` is the same for all four 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 four strategies is that the queries are all simple SPJ queries, which use each table only once and

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

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

[13] 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.

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; (iii) a lineage that coincides with the witness in the basis, and (iv) all tuples are counterfactual causes. Hence, for these queries, the four 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
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 $\{\{c_{939}, cf_{496}\}\}$, its lineage is $\{c_{939}, c2f_{496}\}$, both these tuples are counterfactual causes and have responsibility 1. Therefore, the credit assigned to these tuples is 1/2 using all four 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 first three strategies and often times also with the fourth DS.

6.2. Synthetic queries

To simulate synthetic queries, we randomly generated provenance polynomials in which the coefficients and exponents could be greater than 1. The queries involve three GtoPdb tables: family, contributor2family, and contributor. The polynomials were generated as follows (in particular, every time we write "randomly", we mean using a uniform distribution): first, the number of monomials composing the polynomial is decided choosing randomly a number between 1 and 6. Then, we randomly choose a tuple from the tables family, one from the table contributor2family and one from table contributor, that are used as the monomial's variables. Again,

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

Causality: counterfactual causes: $\emptyset$,
actual causes: $\{f_1, f_5, c2f_1, c_1, c2f_1, c2f_2, c2f_{17}, c_1, c_2, c_{18}\}$
Responsibility:
$f_1 = \frac{1}{2}, f_5 = \frac{1}{2}, c2f_1 = \frac{1}{3}, c2f_2 = \frac{1}{3}, c2f_{17} = \frac{1}{2}, c_1 = \frac{1}{3}, c_2 = \frac{1}{3}, c_{18} = \frac{1}{2}$
Credit distribution:
$f_1 = \frac{3}{20}, f_5 = \frac{3}{20}, c2f_1 = \frac{1}{10}, c2f_2 = \frac{1}{10}, c2f_{17} = \frac{3}{20}, c_1 = \frac{1}{10}, c_2 = \frac{1}{10}, c_{18} = \frac{3}{20}$

Figure 9: Sample synthetic provenance polynomial (how-provenance) and corresponding why-provenance, lineage, causality and responsibility values, together with the corresponding credit distributions.

randomly, we choose a coefficient for this monomial (between 1 and 3) and an exponent for each tuple (between 1 and 4). For the next monomial, then, we decide if we want to keep the same tuple from the table family as first tuple of the new monomial. To do so, we generate a random number between 0 and 1. If the number is above 0.2, we change the family tuple.

An example can be found in Figure 9, which shows a sample synthetic provenance polynomial (the how-provenance), the corresponding why-provenance and lineage expressions, and the causality of the tuples of the lineage, together with their responsibility. The resulting credit distribution for each DS is shown after the provenance expression.

As an example of how the distribution strategies behave with these synthetic queries, consider tuple f_5 in Figure 9. This tuple receives the highest quantity of credit using responsibility-based distribution, and less credit using, in order, lineage, why- and how-provenance. This is because more information is available about the role of the tuple in the overall compu-

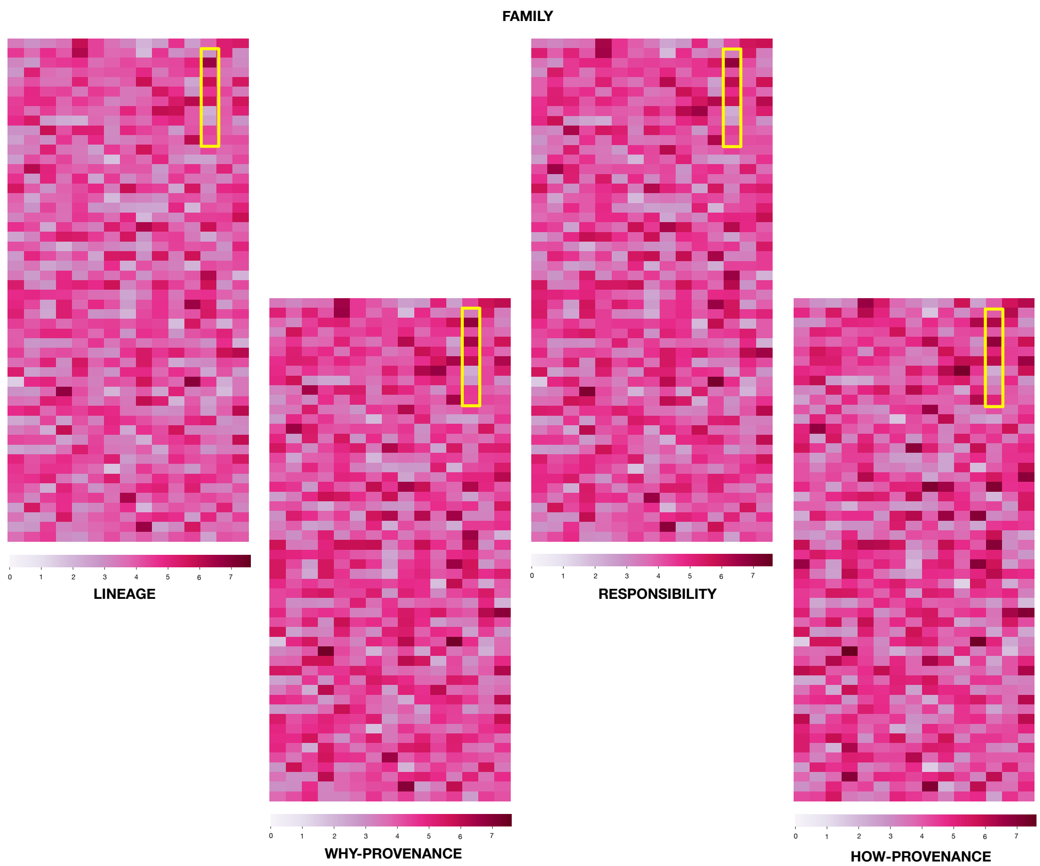

Figure 10: Comparison of three DS on the same table `family` after the distribution computed using 10K synthetic and randomly generated provenance polynomials. The tuples in the blue rectangles are used as example in the discussion connected to Figure 11.

tation. 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. Responsibility, on its part, can be seen as an enrichment of the information brought by lineage. It enriches the tuples of the lineage with a value providing us with a ranking describing the importance of tuples in generating the output. As such, the responsibility-based DS moves part of the credit to $f_1, f_5, c2f_17$ and c_18, since they are tuples that are more important than the others in generating the outputs. This notion of “importance” is connected to their corresponding minimal contingency sets. For example,

f_1 has as minimal contingency set (one of the many) $\{f_5\}$, with cardinality 1. On the other hand, c_1 has, as minimal contingency set (one of the many) $\{f_5, c_2\}$, with cardinality 2. This means that c_1 is "less important" of the tuples with minimal contingency sets of lower cardinality, and this is reflected on the different quantity of credit being distributed.

Despite being 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 10. We set the maximum value in the heat maps to the highest value reached by a tuple in all three distributions (i.e., 7.5).

As can be seen, the four strategies generate different credit distributions, indicated by the varying hues. However, there is a certain amount of consistency between them in that tuples which are highly rewarded by one strategy are also highly rewarded by the others. This shows that the four DSs consistently reward certain tuples more than others.

Note that lineage-based DS gives the least credit to tuples in the `family` table, indicated by an overall lighter hue. This is because the DS distributes credit equally 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. We also notice that the responsibility-based distribution strategy has a distribution that is quite similar to the one provided by why-provenance. It is often the case, for example when the witnesses of the why provenance share one common tuple, that the two distributions behave similarly. As a consequence, at times the generated polynomials are such that the two distributions behave in the same way, or very similarly.

We note that the lineage-based DS gives an average credit of 3.82 to each tuple in the table, while the DS based on why-provenance assigns 4.18 and the one based on responsibility 4.13. Moreover, lineage distributed a total of about 3121 units of credit to the `family` table, while responsibility assigned 3290 and why-provenance 3333.

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 on closer inspection, slight differences can be seen. This is because how-provenance also considers the frequency with 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 understand this difference, in the next subsection we consider the accrual of credit over time. In doing so, we will focus on the ten tuples shown within the large yellow rectangles in Figure 11. Each small rectangle within a large blue rectangle is a tuple, and we number them from 1 (top) to ten (bottom). These ten tuples were selected specifically because they allow us to see the evolution of the distribution of credit through time.

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 11.

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 those highlighted in the light blue boxes of Figure 10 from the `family` table.

The queries used are the same as the experiment of 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.

By the end of 1K queries, credit differentials between tuples as well as between strategies can be seen. For example, tuple 4 is usually rewarded the most credit by all three strategies. However, it receives the highest quantity of credit from the why-provenance-based strategy. Tuple 3 receives the highest

Figure 11: 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.

quantity of credit overall with how-provenance. This trend continues to the end of 2k queries. By the end of 5k queries, tuple 2 emerges with the highest value of credit for why- and how-provenance, a position which is strengthened by the end of 10k queries. This is because tuple 2 is used several times within queries being executed, which is rewarded strongly by why- and how-provenance but not taken into account in lineage.

While the relative value of credit "positions" of tuples within a DS strategy depends on what queries are being executed, the important thing to notice is the difference between the DSs over time: Overall, lineage gives far less credit to tuples in the `family` table than the other two strategies since credit is shared with tuples in other tables. However, the why- and how-

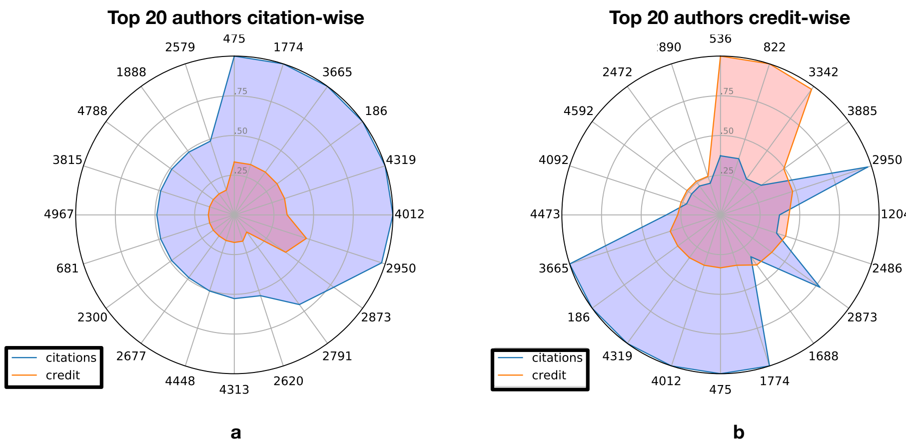

Figure 12: Radars presenting the top 20 authors citation-wise and credit wise, together with their (normalized between 0 and 1) values of citations and credit.

provenance-based strategies recognize the more important role being played by the `Family` tuples than those in the other tables. The differences between the why- and how-provenance-based DSs are also relatively minor (about plus or minus 0.2 out of 9.5) in most cases. However, there are certain situations in which the role of a tuple is particularly critical in a query, and in this case the difference in the value of credit assigned is notably higher for how-provenance. An example of this can be seen in tuple 9 of the 10k group of Figure 11.

To sum up, the DS based on lineage is sufficient to highlight which tuples in the database are used by a query, and distributes credit equally to these tuples. The resulting distribution rewards tuples that are used by more queries, but does not reward how many times tuples are used in the same query. However, a DS based on why- or how-provenance may be better if the queries are complex, since they reward more tuples that have a critical role in generating the output. In particular, these two DSs may be useful for finding "hotspots" in the database based on the role of tuples, with the how-provenance-based DS being preferable if a higher sensitivity to the role of a tuple in queries is required.

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 data authors and curators. Using both real-world and synthetic queries, we distribute credit to the authors responsible for the data under the different strategies. Our results show that credit rewards authors of data that is cited fewer times, but that has a higher impact on the query results.

To do so, we need to identify a set of authors and queries that cite data curated by them. Considering GtoPdb, each target family page has a list of curators, representing the people who are co-creators and curators of the data comprising the page. This list can be obtained using the last query shown in Figure 3. Each time a target family page is cited, we assign one *citation* to each author associated with the page. The authors also receive *credit* in the amount assigned to the data used by the query to construct the webpage, equally divided between the authors of the webpage.

Results: Real-world queries. As described in Section 6.1, we consider real-world queries taken from papers published in the BJP 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 used.

The results are shown in the radar plots of Figure 12, in which each number on the outer circle (e.g. 475, 1774 and 3665) represents an author (id) and the blue (red) line represents the normalized value of credit generated by citations (credit), respectively. The first radar plot, Figure 12.a, shows the top 20 authors in terms of *citations*, ordered in a clockwise direction, whereas Figure 12.b orders the authors based on *credit*. Comparing the author ids used in the outer circles of these two plots, it can immediately be seen that the "top authors" are very different using these two metrics, although there is some overlap (for example, authors 1774, 475, and 4012).

Diving a bit deeper to focus on the red and blue areas in each of the plots reveals that there is a significance difference between citations and credit: The top 20 authors in terms of citations do not have the highest values of credit (Figure 12.a). Conversely, the authors with the highest values of credit do not necessarily have a large number of citations (Figure 12.b). For example, author 536 has the highest value of credit, but is not even in the top 20 authors in terms of citations. This means that authors like 536, 822, and 3342 in Figure 12.b receive much more credit from their relatively few citations than authors like 475, who receives the largest number of citations. That is, the data underlying certain webpages is more "valuable" in terms of credit than a citation to the webpage.

The reason for the difference between citation and credit is partly due to

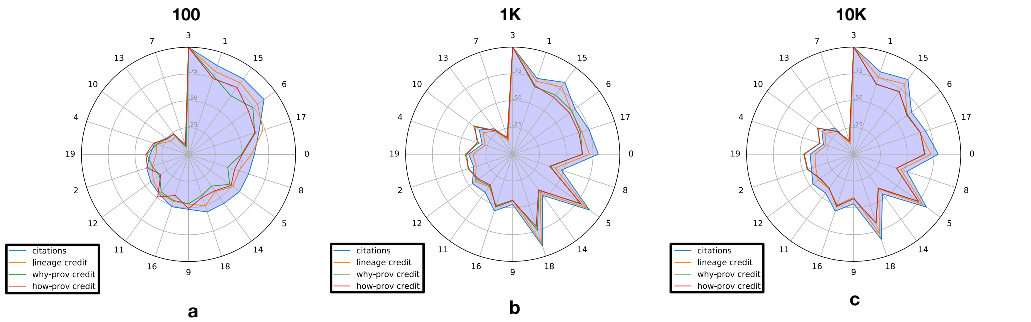

Figure 13: Radars presenting the 20 synthetic authors with corresponding citation and quantities of credit distributed through the 3 DS (all values normalized between 0 and 1) through different numbers of polynomials (respectively, 100, 1K and 10K). The order is the one defined by figure 1, i.e. descending order of citations obtained from 100 polynomials.

the experimental setup: Each output tuple carries a credit of 1, and there can be many tuples used to generate a webpage. Thus a webpage that is created from more tuples will have a higher credit value than one created from fewer tuples. Furthermore, authors who collaborated with fewer people will receive a biggest share of the equally divided credit. However, all authors will receive a citation of one.

Credit distribution therefore rewards authors differently than traditional citations: An author who has curated larger quantities of cited data and collaborated with fewer co-authors, will receive larger quantities of credit. Thus, credit rewards them for their larger contribution to the database.

Results: Synthetic queries. We produced 100, $1K$, and $10K$ batches of synthetic polynomials, as described in Section 6.2, and distributed credit through them to data. Since these polynomials are created by randomly selecting tuples from three tables, they usually correspond to a large set of authors who in reality did not collaborate. To make the size of the author set more realistic, we therefore created 20 synthetic authors, and randomly assigned one author to blocks of consecutive tuples in the database, with the size of each block varying between 10 and 40, to simulate different quantities of work performed by an author. Every time an author appears as curator of one or more tuples used in a polynomial, we assign them one citation. They also receive three kinds of credit, each one using a different DS.

Figure 13 shows three radar plots, one for each batch of synthetic polynomials. Each plot shows the top 20 authors in terms of citations (hence the authors and clockwise ordering is the same in each of the plots), and additionally shows the the normalized values of citation (blue line), lineage-based credit (yellow line), why-provenance-based credit (green line) and how-provenance-based (red line). As can be seen, given the synthetic nature of the queries, the correlation between the number of citations and the quantity of credit assigned to the authors appears to be a much stronger than with the real-world queries of Figure 12. In fact, for Figure 13.a 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. The credit distributed via lineage is closest to the number of citations (a linear correlation of 0.98, p value of 6.15e-16 in Figure 13.a), while the other two types of credit behave slightly differently (a linear correlation of around 0.95 in both cases in Figure 13.a). Similar observations can be made for Figure 13.b and 13.c.

What these figures show is that, in certain cases, authors who do not have a large number of citations receive more credit than others, as for example author 11 in Figure 13.a or author 19 in Figures 13.b and 13.c, especially when credit is distributed using how-provenance. This again shows how credit gives a different perspective on the role of data and authors by going beyond the limitations of traditional citations.

It is worth noting that, when scaling up to $1K$ and $10K$ polynomials, the credit distributions via why-provenance and how-provenance become almost identical (the linear correlation for the values of Figure 13.c is more than 0.99 with a p-value of 1.32e-32). This is consistent with what we observed in Figure 10.

7. Discussion

We note that, in our experiments, we always assumed that the credit carried by an output tuple is 1. Thus, each tuple in the output has equal importance. This in general may not be true, since different tuples in the output may have different weight, depending on the context of the citation. For example, data that is fundamental for the results of a paper may have more credit than data being cited as a reference. *Credit generation*, i.e. the process by which the credit of the output tuples is decided, is research problem with its own dignity and complexities, and we did not face it in this paper.

From the point of view of the model, even when the credit of the output tuples is different than 1, nothing needs to change in the models presented here, since they were defined for a generic value k. We note that, if the quantity of credit carried by an output tuple changes, as a consequence the final distribution will change, since certain tuples will be more "impactful" (i.e., distribute more credit) than others. With different quantities of credit, therefore, new results, different from the ones obtained in the previous sections, may be found. These results will depend on the nature of the context and the quantity of credit being considered.

8. Conclusions and Future Work

This paper defines two new distribution strategies based on why- and how-provenance, and compares them against the lineage-based distribution strategy defined in [25]. The first, why-provenance-based DS, 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, how-provenance-based DS, 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 how-provenance-based DS is more sensitive than the why-provenance-based DS 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 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, in both real-world and synthetic scenarios, credit rewards authors who contribute/curate data that has the highest impact, and therefore receives the biggest quantity of credit, and not necessarily the data with the highest citation count. 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 strategies to generate and distribute credit. In this paper we assumed that each output tuple carries credit 1. In more sophisticated scenarios we can employ different strategies to compute credit, that reflect the importance of cited data. Also, other, and more sophisticated strategies could also be used to decide how credit is distributed between the authors, beyond the uniform distribution used here, in a way to reflect the work performed by them on the cited data.

We will also explore new applications for 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* [45], which addresses the problem of reducing the vast – and rapidly expanding – amount of data that is being produced.

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.

Acknowledgement

The work was partially supported by the ExaMode project, as part of the European Union H2020 program under Grant Agreement no. 825292.

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] Chockler, H. and Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *J. Artif. Intell. Res.*, 22:93–115.

[19] 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.

[20] 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).

[21] Cronin, B. (1984). *The Citation Process. The Role and Significance of Citations in Scientific Communication.* London: Taylor Graham.

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

[23] 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.

[24] 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.

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

[26] 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.

[27] Eiter, T. and Lukasiewicz, T. (2002). Complexity results for structure-based causality. *Artif. Intell.*, 142(1):53–89.

[28] ESIP Data Preservation and Stewardship Committee (EDPSC) (2019). Data citation guidelines for earth science data, version 2. Version 2, Earth Science Information Partners.

[29] 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.

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

[31] 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.

[32] 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.

[33] Halpern, J. Y. and Pearl, J. (2013). Causes and explanations: A structural-model approach — part 1: Causes. *CoRR*, abs/1301.2275.

[34] 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.

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

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

[37] 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.

[38] 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.

[39] 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).

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

[41] 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.

[42] 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.

[43] 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.

[44] Meliou, A., Gatterbauer, W., Moore, K. F., and Suciu, D. (2010). The complexity of causality and responsibility for query answers and non-answers. *Proc. VLDB Endow.*, 4(1):34–45.

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

[46] 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.

[47] Parsons, M. A., Duerr, R. E., and Jones, M. B. (2019). The history and future of data citation in practice. *Data Science Journal*, 18(1).

[48] 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.

[49] 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, 6-9 October 2013, Santa Clara, CA, USA*, pages 307–312.

[50] 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.

[51] 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.

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

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

[54] 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.

[55] 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.

[56] 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.

[57] 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).

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

[59] 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.