BBA - Gene Regulatory Mechanisms 1865 (2022) 194779

Contents lists available at ScienceDirect

BBA - Gene Regulatory Mechanisms

journal homepage: www.elsevier.com/locate/bbagrm

Dealing with different conceptions of pollution in the Gene Regulation Knowledge Commons☆

Anamika Chatterjee [a,*], Tsjalling Swierstra [b], Martin Kuiper [c]

[a] Department of Philosophy and Religious Studies, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
[b] Department of Philosophy, Maastricht University, Maastricht, the Netherlands
[c] Department of Biology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

ARTICLE INFO

Keywords:
Pollution
Biocuration
Database

ABSTRACT

Current research of gene regulatory mechanisms is increasingly dependent on the availability of high-quality information from manually curated databases. Biocurators undertake the task of extracting knowledge claims from scholarly publications, organizing these claims in a meaningful format and making them computable. In doing so, they enhance the value of existing scientific knowledge by making it accessible to the users of their databases.

In this capacity, biocurators are well positioned to identify and weed out information that is of insufficient quality. The criteria that define information quality are typically outlined in curation guidelines developed by biocurators. These guidelines have been prudently developed to reflect the needs of the user community the database caters to. The guidelines depict the standard evidence that this community recognizes as sufficient justification for trustworthy data. Additionally, these guidelines determine the process by which data should be organized and maintained to be valuable to users. Following these guidelines, biocurators assess the quality, reliability, and validity of the information they encounter.

In this article we explore to what extent different use cases agree with the inclusion criteria that define positive and negative data, implemented by the database. What are the drawbacks to users who have queries that would be well served by results that fall just short of the criteria used by a database? Finally, how can databases (and biocurators) accommodate the needs of such more explorative use cases?

1. Introduction

According to Mary Douglas, the author of Purity and Danger [1], our ideas of purity and pollution are shaped by the normative values of the community to which we belong. Douglas suggests that these values provide us with the basic categories within which we classify our perceptions, and guide how we approach, process, and perceive new stimuli such as objects, behaviors, or practices. We use these categories as a scaffolding for the framework upon which all further perceptions will be assessed. They offer us a place to start from, an inoculum, to define what is normal, ordered, or *pure* within a context and consequently to identify what is abnormal, disordered, or *impure*. Stimuli that fit into an existing category are accepted as pure and those which do not are often rejected as polluted.

In the context of scientific research data, the core normative values of the scientific community provide these basic categories. Broad, overarching values such as trust, honesty, reliability, reproducibility, amongst others, offer members of this community a basic framework to distinguish between 'good' or pure science and 'bad' or polluted science. Practices that embody these values generate pure data, while those that do not result in polluted data. While the former increases the extent of knowledge uncovered by scientists, the latter misleads and corrupts further research efforts. For instance, knowledge that builds on data that is the outcome of false, fabricated, and biased research pollutes the vast pool of scientific knowledge. To safeguard purity, the scientific community relies on the process of peer review to thoroughly check the authenticity and quality of new research data before it is added to the existing body of scientific knowledge.

☆ This article is part of a Special Issue entitled: Curation of the Gene Regulatory Knowledge Commons edited by Dr. Colin Logie, Dr. Wyeth Wasserman and Dr. Julio Collado.

* Corresponding author.
E-mail address: achatterjee@ebi.ac.uk (A. Chatterjee).

https://doi.org/10.1016/j.bbagrm.2021.194779
Received 21 February 2021; Received in revised form 28 November 2021; Accepted 29 November 2021
Available online 28 December 2021

Additionally, scientific sub-communities rely on manually curated databases as a second safety checkpoint to minimize the entry of such pollution. In the case of the life sciences community, such databases form the backbone of most, if not all, research endeavors. They have grown to be rich, verified, and trustworthy sources of high-quality knowledge that enable further good science. Biocurators read through scholarly publications and organize the authors' claims into a computable form. They also check information for sufficient quality. Information that lacks descriptions sufficient to validate, organize and classify it, may be excluded. Similarly, data insufficiently supported by the evidence in the publication may be rejected. Biocurators thus play a crucial role in minimizing the impact of flawed data on further research. They not only transform knowledge into structured, organized, and well managed data, they also limit the spreading of polluted data.

However, the broad life sciences community contains different sub-communities that each adapt the basic categories offered by science's overarching values to their own specific needs. This results in frameworks for separating pure from polluted data that are to some extent local. Even though all databases share the quest for trustworthy information, each individual database also represents a distinct community with a distinct framework to assess data for purity. Biocurators judge what data fits their local framework and filter out the rest [2]. As a result, even if data is considered pure, based on general criteria, it may still be thrown out as polluted because of the local framework that determines its fate in that particular database.

These frameworks are made-up by the curation guidelines that guide the biocurators of a database in their assessment of data. Across the life sciences, including the gene regulation domain, which is the focus of our study, these guidelines represent heterogeneous criteria for how individual databases value coverage, methods, evidence, and data representation. For instance, a set of data that is considered sufficiently complete for a database that prefers a broad and lean coverage, may be rejected as insufficient or incomplete by one that values focused and deep curation. Similarly, while one database may accept a variety of data production methods, for the next, only a subset of these methods may be accepted as generating pure data. Because of these differences, one database may consider an observation to be significantly justified, while the other perceives it as insignificant. Similarly, one database may consider data to be neatly visually organized, labeled and classified, whereas another may decide that the same data does not meet these criteria. That is, while each database works to weed out insufficiently supported, or negative data, what qualifies as such may differ from one database to the next [2].

In addition to the variation between biocuration practices across databases within a domain, variation also exists at the level of database use. Different use cases may reflect varying demands and needs, which can translate into different criteria for distinguishing pure from polluted information. Databases apply criteria to classify pure and impure data. But users may have a need that leads them to adopt a threshold that can be more, or less restrictive than that of the database. That is, *what is pure according to the criteria of the database may be different from what is pure for the use cases.* These users may be better served if additional information which further qualifies the 'pure' and also the 'impure' data, is made available as 'metadata' that can be used for custom filtering. Alternatively, they would need to invest their own resources into curating information that fits their specific use.

There exist differences in criteria, although usually subtle, for separating polluted from pure data, both between databases and between ways that a database can be used (its use cases). In this paper we focus on the relation user-database, and ask: *what are the drawbacks for users of use cases when their purity criteria differ from the local criteria applied by a database? What role can databases and biocurators play in alleviating these drawbacks?*

In section one we introduce social anthropologist Mary Douglas, who has proposed that what is considered 'dirt [pollution] lies in the eye of the beholder' (p11, [1]). In other words: pollution is a *socially constructed category*. In section two we apply her theory to data purity and pollution in the life sciences and to manually curating data quality. In section three we draw attention to the differences in conceptions of pollution at different levels of a community. We first discuss how Douglas' theory has been previously applied to the different assessments of data by two databases in the gene regulation domain. We then extend this analysis to explore how pollution gets defined, in two different use cases of the same database: the *cautious* use case and the *greedy* use case. These use cases are different in the way they deal with false positives and false negatives. We present two examples where such differences occur and discuss the drawbacks to the users of greedy use cases. In section four we offer suggestions about how biocurators and databases could help users of greedy use cases with a relatively minor modification of their current practices.

2. Approach

This article presents findings generated within the COST Action 15,205: Gene Regulation Ensemble Effort for the Knowledge Commons (GREEKC, www.greekc.org). Within the gene regulation domain, efforts to standardize the diverse data management practices have been undertaken by the Gene Regulation Consortium (GRECO, www.thegreco.org). GRECO obtained funding from COST to bring its European members together in the GREEKC initiative so they could collectively align their respective data management practices. GREEKC has focused on developing the Gene Regulation Knowledge Commons (GRKC) (Kuiper et al., this issue), which is built on a standard framework for knowledge management.

While the term 'data' has various connotations, here, we use it to refer to what biocurators curate from scientific literature. The authors of Understanding the Knowledge Commons [3] distinguish between data, information and knowledge as "..data being raw bits of information, information being organized data in context, and knowledge being the assimilation of the information and understanding of how to use it." (p6) Yet the terms 'data', 'information' and 'knowledge' have often been used interchangeably within the GREEKC community. So, for purposes of brevity, we adhere to the term 'data' instead of 'information' or 'knowledge'. Additionally, although biocurators also annotate large sets of functional genomics data, this has not been a part of our analysis and we have exclusively focused on data that biocurators derive from traditional scholarly publications.

The efforts of GREEKC have themselves been the subject of study by an interdisciplinary collaboration between researchers from biology and humanities in Crossover Research 2.0 - Well constructed Knowledge Commons (CR2). CR2 focused on the integration of a Responsible Research and Innovation (RRI) approach towards the development of the knowledge commons for the life sciences (of which the GRKC is a part). RRI focuses on increasing the societal relevance and value of innovation processes [4]. The aim is to engage the diverse stakeholders involved in research and innovation efforts in a dialogue to anticipate, reflect and deliberate upon the potential roadblocks they may encounter during such collaborative endeavors. Members of CR2 have used semi-structured interviews as well as ethnography through participant observation to engage with those involved in GREEKC to explore how RRI can be integrated into collaborative research endeavors with diverse stakeholders [5,6].

The primary research method within this study has been the use of semi-structured interviews [7–9]. This technique draws from two other methods of qualitative research - surveys and focus groups. Surveys involve a set of close-ended questions that are used to cover a large number of respondents. Focus groups, on the other hand, involve a small group of respondents in a detailed discussion around a set of open-ended questions with common themes. Like surveys, semi-structured interviews are based on one-on-one interactions with respondents. Yet, similar to focus groups, these interactions are conversations that are set around open-ended questions. Here, an interview guide presents an

outline of the topics that are to be discussed. This flexible approach allows for a malleable interaction that takes into consideration the background, experience, and expertise of the respondent. Additionally, this technique offers respondents from a small community the confidentiality that is absent in focus groups, allowing them to freely express themselves on sensitive and delicate topics within a small community.

With this methodology, AC interacted with biocurators from the gene regulation community. This offered the respondents an environment to freely discuss and comment upon what, according to them, were polluting practices adopted by other members of the community. The sample selected for these interviews included 16 biocurators within the Gene Regulation domain (for reasons of anonymity, the names of these respondents are not disclosed). The interview guide (attached) was customized and adapted around five main themes to conduct more efficient interactions with each respondent keeping in mind the database they were affiliated with. The interviews were recorded, and the recordings were transcribed following the completion of all interviews. The transcripts were then coded and categorized. A thematic analysis of the narrative accounts was conducted to identify the common, as well as distinct themes that were revealed through the conversations. Following the analysis, the transcripts were anonymized and submitted with restricted access, details of which are available (Supplementary information) from the Norwegian Centre for Research Data (NSD).

3. The theory of pollution

In this section we explore Mary Douglas' proposal that purity and dirt are socially constructed concepts [1]. She establishes that both concepts are relative to each other, as well as to the context within which they are being assessed. The context, she explains, is informed by the social and cultural values of the community to which one belongs. These values offer community members a set of basic, foundational categories to classify what they perceive. These categories provide a common normative framework that allows members of a community to determine what belongs within a context and what does not, but also whether a stimulus (object, practice, or behavior) is pure or polluting. As she explains, it is not an object itself that is pure or polluted. That depends on the normative framework that an observer applies.

"Dirt is matter out of place," (p37) writes Mary Douglas in her analysis of the different conceptions of pollution: "Shoes are not dirty in themselves, but it is dirty to place them on the dining-table; food is not dirty in itself, but it is dirty to leave cooking utensils in the bedroom, or food bespattered on clothing; similarly, bathroom equipment in the drawing room; clothing lying on chairs; out-door things in-doors; upstairs things downstairs; under-clothing appearing where over-clothing should be, and so on." (p38). Whether these objects - shoes, food, bathroom equipment, clothes, out-door, or upstairs things, etc., are seen as dirty depends on the normative order of the settings within which they are observed. An object that appears as an oddity or a misfit in its surroundings, is seen as defiling and polluting those surroundings. One that fits in, is believed to retain the purity thereof.

Judging something pure or polluted is according to Douglas inherently linked to the culture of the assessor. "Dirt is disorder..there is no such thing as absolute dirt; no single item is dirty apart from a particular system of classification in which it does not fit." (p xvii, preface). This system of classification or *schema* is molded by one's community and its values. "Culture, in the sense of the public standardized values of a community, mediates the experience of individuals. It provides in advance some basic categories, a positive pattern in which ideas and values are tidily ordered." (p40). The foundational structure of one's schema reflects the core values of the particular community to which one belongs. Members of the same community share common criteria to distinguish between normal and abnormal, between pure and dirty. Perception is always selective. Some stimuli are highlighted, others ignored. And people tend to highlight what is relevant within their categorical framework. "In perceiving we are building, taking some cues and rejecting others. The most acceptable cues are those which fit most easily into the pattern that is being built up. Ambiguous ones tend to be treated as if they harmonized with the rest of the pattern. Discordant ones tend to be rejected." (p45). This means that members of different cultural communities learn to see and assess the world differently. They also tend to have different perceptions of purity and pollution. If a new stimulus squarely fits one of the categories of the schema of A, A accepts it as 'normal'; even though the same stimulus may not fit into the schema of B, who will therefore reject it as creating disorder.

This plurality in ideas of dirt and pollution have since been explored in a variety of contexts. Douglas analyzed how the dietary taboos of traditional Hebrews depend on a culturally specific distinction between normal and anomalous – therefore dirty, polluted - animals. The role of cultural classifications of dirt has also been examined in decisions driving urban architecture and infrastructure [10]. The theory has subsequently been used to help identify the contexts in which sound is experienced as sound, or where it becomes noise [11]. Researcher have investigated how consumers distinguish between pure and polluted food, as well as how these distinctions vary across time and cultures [12]. Another study describes how different perceptions of 'dirt' in a material science laboratory shape different laboratory practices [13]. These studies explore the different contexts where cultures, communities and groups agree on their conceptions of pollution but only up to a point, from where they begin to diverge. In the following section we explore to what extent Douglas' theory can also be applied to the field of biocuration.

4. Data pollution in the scientific community

For the scientific community, polluting behaviors are those that undermine its core values of quality, reproducibility, replicability, and trustworthiness. Data that upholds these values is instrumental in further research endeavors. But when data that does not uphold these values is reused, it can pollute the outcome of future research. One example of data pollution is muddying the scientific literature with findings from poorly conducted research in which experimental methods and designs are not properly implemented; instruments are incorrectly calibrated; software is inappropriately used for analyses; or low-quality materials, reagents, or chemicals are used. Another example of pollution is when research results are fabricated or falsified to suggest findings that cater to goals other than improving the state of scientific knowledge. Such behaviors result in the addition of dubious, erroneous, incorrect, and misleading data to the common pool of research data generated by and for the scientific community.

The consequences of such practices are costly. They add to the large volume of scholarly publications that scientists need to go through before they can identify what they need [14–19]. They may also place their trust in and build their own research on research that cannot be reproduced, thus risking pursuing futile research avenues [20,21]. Additionally, low quality or downright fraudulent research is costly to governments, funding bodies, and the society at large, as such research is often funded by taxpayer's money [22]. Finally, the real-world applications of such polluted research may mislead health professionals and bring physical, as well as mental harm to their patients' health [23–26].

5. Biocuration and pollution

While the peer review process aims to weed out polluted data, biocuration acts as an additional checkpoint. Several studies have established that the peer review process is not airtight [27–30]. This problem is further exacerbated by the rise of predatory journals that masquerade as legitimate sources of published research [31,32]. Thus, biocurators work towards filtering out data that does not reflect the normative values of the scientific community. As the International Society for Biocuration [33] states, "[Biocurators]..strive to distil the current 'best

view' from conflicting sources" (p2). This distillation involves the identification of data that is worthy of being reused. Biocurators gather and read publications falling within the scope of their respective databases, extracting key results, observations, evidence, and knowledge claims. Importantly, they weed out knowledge claims lacking sufficient supporting evidence. Through these efforts they keep their databases pure and safe from polluted data.

5.1. Common criteria to identify pollution

It is clear how important and complex the task of a biocurator is. We now analyze in more detail the different levels at which biocurators apply quality control and the decisions they make to select valid data for their database.

From the interviews we learned that biocurators' decisions are guided both by general standards, and by curation guidelines specifically developed for their database. These standards and guidelines provide the biocurators with rules and criteria for assessing data for their database. One such criterion is how *complete* the supporting evidence is. Data can be rejected because it lacks the details necessary to complete an annotation entry in the database. This information can be the species within which the research was conducted: "I mean that's probably the biggest frustration. You can see it's a good paper but it's not clear what the species is." (BC01) Data lacking information on the construct and its origin can also be rejected: "First thing you sort of look for is species, origins, where they bought the constructs, and if that's not there or not obvious […] you have to ditch it." (BC05) Publications where the supplementary data are missing are also often rejected: "Often relevant data was in the supplementary material, but the journals didn't keep it because the paper is 10 years out and they just deleted the supplementary material." (BC01) And: "So we do lose information because a lot of papers we could curate, but we don't because information is missing from them." (BC05).

Another criterion for rejection of data is *obsolete terminology*. For instance, identifiers allow for easy and unambiguous distinctions between entities. Yet, when the identifiers of the newly curated data set do not match those used by the database, biocurators have to manually translate or map the data to incorporate it into the database. Such mapping may result in the loss of data or in pollution of the database. "I recently mapped a data set into ours and lost about 15-20% of that data because they had used obsolete gene IDs. So, it's data that gets painfully lost and if it's used it's probably misused and that to some extent is pollution." (BC06) According to this respondent, obsolete identifiers would have contributed to erroneous results, and thus to pollution.

A third criterion for data rejection is if it appears *noisy or faulty*. "Anything that gives us erroneous results or misleading results is pollution. This could be background noise that messes up the results; poor implementation of well-understood techniques; trying to use techniques too early on in their life cycle; the wrong interpretation of data; manufacturers cutting corners like the quality of antibodies or biologically inactive enzymes.." Additionally, knowledge claims not based on experimental evidence, i.e., not shown in figures or tables, is rejected. "We don't curate data that is not shown directly in the publication. So there needs to be a figure or a table." (BC06). Unsupported data could contribute to faulty results and is thus considered pollution.

The overarching criterion for biocurators is that the data agrees with the quality threshold of their database. "If we think a paper just generally doesn't look good, …, we're under no obligation to put it in the database. " (BC09) And: "If [a peer reviewed publication] is really unreliable, we just don't write it because it could create more confusion." (BC01) This filtration can result in the exclusion of entire data sets of data or of just the parts that are considered to be polluted. "We have to run checks and balances to see that it agrees with our philosophy and threshold level of quality…things in the data set that aren't meeting our threshold criteria we normally throw away. We take the good stuff and throw out the bad stuff. " (BC10).

Biocurators thus make crucial decisions in assessing whether a piece of data is true or false, enriching or polluting their database. Biocurators may contact the authors in order to fill in the blanks and provide crucial missing information. They strip the data of what their database schema identifies as noisy, unnecessary, or irrelevant information. Furthermore, they try to incorporate the most updated terminologies, nomenclature, and protocols to upgrade older data sets. With these efforts they protect their users from pollution.

Biocurators help users overcome the typical hurdles encountered when reusing data [34,35]. They ensure "accuracy, comprehensiveness, integration, accessibility, and reuse." (p1) [33] of data. The value they add has made curated databases the gold standard against which computationally generated, or text mined data sets are compared [5,36–40]. Such databases now constitute the backbone of research in the life sciences domain.

6. Plural constructions of pollution

In this section, we discuss how, in addition to adding a second layer of quality control to peer reviewed data, biocurators also assess if data belongs in their particular database or should be discarded as an anomaly.

How does Douglas' analysis of the social construction of dirt and pollution apply to how the life sciences community identifies and prevents pollution? We have seen in the previous section that biocurators abide by certain commonly shared values of the life sciences community, when assessing data for quality. These common values are reflected in the database curation guidelines. But the respective frameworks of different databases also represent local values. Frameworks become more defined and selective at different granularities of a community. The life sciences domain is split into a range of smaller communities across different model organisms, entities or processes being studied. Such communities can have different databases, each incorporating locally shaped norms that answer to the specific needs of the subcommunity.

This local framework shapes the biocurators' decisions about which data is sufficiently 'pure' to get included in their database.

7. Pollution pluralism across databases in the Gene Regulation domain

The definition of 'quality data' may vary across databases [2]. Different databases may apply different curation guidelines that reflect the different values of their respective communities. Data that fits neatly into one database may present an anomaly to another database. So, while the data itself may not be of poor quality, whether it is rejected or accepted also depends on what is locally perceived to be pure data.

In other words, where two databases draw the line between positives and negatives, varies. One database may prefer to use Entrez identifiers, whereas another database may prefer to use Ensembl identifiers. One database prefers deep coverage and requests more experimental details, whereas another database allows shallower curation and settles for less details for each data set. For instance, databases that place value on the cell type, specific constructs or vectors will classify data lacking these details as impure, yet for a database agnostic to such details, this data qualifies as pure. Two databases covering the same biological pathways may accept different levels of evidence as sufficient as proof for an observation. For instance, while one requires proof of a specific molecular activity within the pathway, the other database considers evidence implicating the gene product in a specific molecular pathway without this information still sufficient for annotation.

Additionally, databases also differ in the type of methods they approve and in how they represent curated data. For instance, while one database curating molecular interactions accepts computationally predicted results as pure, another rejects this as polluted. Similarly, while one database simplifies the data so that its users can easily grasp it,

another's norm dictates capturing the data as it is presented by the authors. Consequently, despite a significant overlap in the data they capture, their schemas do represent different values and hence different perceptions of data purity.

8. Pollution pluralism affecting different use cases of a database

In the previous section, we explained how the common framework of the life sciences community has differentiated to meet the needs of various subcommunities. We specifically discussed this differentiation in the domain of gene regulation. In this section, we extend this argument to different use cases. As biological databases are used for a variety of use cases, these use cases may place different demands on a database. The result returned from that database may meet the needs of users to different degrees, depending on where that database draws the line between positives and negatives (which will affect the false positives that they include and the false negative that they discard) and on the user's aims: to deal only with the most reliable data, or expect abundant results, including potential false positives, but with the benefit of also having access to false negatives. The additional work that a user needs to perform to deal with false positives and false negatives, we argue, is more extensive for users who set a lower threshold for the results.

Biological databases typically cater to diverse users. According to Fuchs et al. [41], this diversity reflects their specific use cases: "...some researchers are interested in general genome organization and would readily merge overlapping sequence database entries to remove 'redundancy', ignoring minor tissue- or strain-specific sequence differences that, for their purposes, are unimportant. Others are particularly interested in the small differences eliminated by such merging." (p4) In addition to this, Douglas et al. [42] studied the diversity across three genomics databases - DiscoverySpace, InnateDB, and WormBase - and found that the diversity in user communities result from a combination of technical and socio-cultural factors. These factors include the amount of trust in a database, its accessibility, how user-friendly its interface is, and the different tastes within the bioinformatics subcommunities. Furthermore, Costabile et al. [43,44] add that user diversity is also a function of the different expertise of users based on their specific ".. user skill, culture, knowledge...specific abilities (physical/cognitive), tasks, and context" (p3) [43].

So, the distinction that a database draws between positives and negatives, between pure and polluted, can be too restrictive or too lax for a particular use case. We saw that the schema of a database that determines whether data are seen as pure, is adapted to the specific needs of a database's community. This community further splits into diverse sub-communities who adapt the database schema to what they value. As a result, their distinctions between pure and polluted also differ from the normative distinction of the database. In the words of Fuchs et al., "...current databases restrict their users to a specific 'view of the database', but with increased complexity of research, the same view can no longer serve all needs." (p4) [41]. This also includes the tension between how biocurators perceive data purity and how the views of certain users may differ from this local norm.

9. Cautious and greedy use cases

Depending on what a user expects from a database query, we can distinguish *cautious* use cases and *greedy* use cases (Fig. 1). In a cautious use case a user wants to have only reliable results that do not need further checking. This may in practice mean that the use case demands more restrictive criteria concerning true positives than the criteria underlying the database. Such users would require additional evidence in support of a database claim, even though it has passed the scrutiny of a curator. They will have to do additional checking to eliminate results that for their use case would be considered false positives, even if this costs them in terms of data coverage. What the database accepts as pure could be polluted for such cautious users. On the other hand, a greedy use case satisfies users with broader, more inclusive criteria for accepting positive data. Such users opt for working with a larger volume of data even if it comes at the cost of having to deal with potential false positives. In their case, data that has been rejected from the database may still be supported by sufficient evidence to be considered a false negative. Consequently, what the database rejects as polluted could be pure for such greedy users.

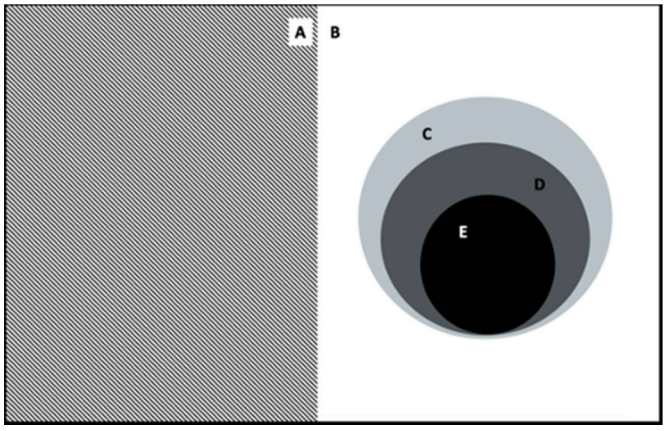

Fig. 1. Depiction of the different classifications of pure and polluted data between use cases ranging from very cautious to very greedy. The whole frame (A and B) depicts all published data within a biological domain; Part A represents data that is published but not yet curated; Part B indicates data that has gone through a curation process; Within part B, the white part depicts data that was rejected; the light grey segment C indicates data that is considered 'pure' by users of a Greedy Use case but was deemed 'polluted' and not incorporated by database D; D (dark grey) represents data that is pure for the database and pure for users of a greedy use case, but it may still contain some data that is considered 'polluted' by users of a cautious use case; and E (black) depicts data that is pure for all 3 groups. An example of such data, which has undergone additional selection to comply with a cautious use case, is described by Velthuijs et al., this issue of BBA [45] when mining the IntAct [46] and BioGRID [47] databases.

10. Examples of these use cases with different resources

10.1. Example 1. Searches for DNA binding transcription factors

A key item of knowledge for understanding gene regulatory mechanisms is the relationship of a DNA binding transcription factor (dbTF) with a target gene (TG). According to a recent estimate (Lovering et al., DbTF compendium, this BBA issue) there are 1457 human dbTFs that interact with regulatory sequences of their target genes and regulate the activity of RNA polymerase II-driven gene transcription. Knowing which dbTF is involved in the regulation of which target gene is a first step towards building a regulatory model of a gene regulatory process. Knowing the complete repertoire of dbTFs defines the search space that researchers must consider when analyzing how transcription factors determine expression profiles. As the dbTF compendium paper describes, the curators of this set developed a set of curation guidelines (Gaudet et al., this issue, Guidelines paper) which they subsequently followed to annotate a list of proteins with the highest possible likelihood of qualifying as a dbTF. The authors also acknowledge that some additional proteins may qualify to become members of this class if more data would become available. This approach serves the cautious use case. To serve the greedy use case, the user may check a supplementary file with all the human transcription regulators that were considered for inclusion, or they may check other resources (e.g., the TFcheckpoint database [48]). In doing so, they may take into account whatever information would be available for these proteins of interest, e.g. in the form of GO annotation terms and their evidence code, which may

indicate evidence obtained from computational predictions.

10.2. Example 2. Gene expression regulated by DNA binding transcription factors

Much effort is being put in the assembly of curated resources describing the relationship between a dbTF and its target gene(s). Curated dbTF-TG relationships are for example available from databases such as TRRUST [49], TFactS [50], HTRIdb [51], SIGNOR [52], as well as IntAct [46]. Such resources offer high confidence descriptions of dbTF-TG interactions that better fit cautious use cases, where users are only interested in what has been approved by biocurators as adequately documented dbTF-TG interactions.

While these databases serve as integral resources in research on dbTF-TG interactions, for the greedy use case, the contents of these databases may be limited. This is because although manual curation certainly provides the highest quality of data, the pace at which it proceeds leaves a gap in the coverage of literature on gene regulation. Although TRRUST currently contains a total of 7421 dbTF-TGs, far more dbTF-TG interactions have been published in the literature. Indeed, automated text mining is capable of surveying the entire MedLine resource for abstracts containing mentions of potential dbTF - TG interactions, and the ExTRI corpus (Vazquez et al., this issue) provides more than 10,000 new candidate dbTF-TG relationships for further analysis. Although a user needs to invest additional curation time to check the validity of these interactions, these data likely contain many instances of what may turn out to be useful information.

11. Drawbacks to greedy use cases for having few false positives or few false negatives

The above examples depict that a greedy use case incurs certain drawbacks that could be alleviated if information on the decisions underlying rejected data would be provided by databases. While there are several studies that have explored the intricacies of the curation process, what happens to data that is rejected in this process is still unclear [53,54]. According to BC02, data that has been screened but failed to meet the criteria is not tagged or labeled as such - "No.. we don't do that, where you can say I looked at this paper but don't bother". As a result, whereas in a cautious use case a user can rely largely on the strict curation criteria implemented by many databases, in a greedy use case the information about data that is rejected but which might still be interesting to pursue is not visible to users of the database. Such users then need to duplicate the biocurator's effort to find essentially the same information by themselves. In example 2, for instance, while the curation guidelines of the database explain the general reasons for the dismissal of data, the specifics of what data has been rejected have been internalized and are not accessible by users. It then takes an additional effort (such as ExTRI text mining) for users to identify the gap between dbTF-TG interactions that have been published and those that have been curated. Incidentally, the difference between both cohorts also includes interactions that have been identified and checked by biocurators, yet the value of this effort does not reach users of a greedy use case. It would be very useful if this user could easily identify *what* has been scrutinized by the biocurator and rejected.

The second drawback is the effort to identify *why* a specific data set has been rejected. While information regarding what data has been rejected is valuable to users of a greedy use case, knowing why this data has been rejected could further facilitate their task. Curators invest considerable time in assessing each data set prior to deciding its fate in the database. "So we internally try to make sure that it's trustworthy for the outside world who don't really know anything about these procedures. We might completely ignore data coming from an experiment, saying this just doesn't look sufficiently trustworthy. We don't capture it. And then we just leave it out, not necessarily flag this for to the user, because that would be just too much." (BC03) As a result, a significant part of the reasoning behind curatorial decisions does not meet the user's eye. However, in these steps information is considered that could be valuable for greedy use cases, for instance: what was missing to convince the curator about the validity of a claim such as showing valuable clues for users such as the identity of a paper and/or entities and relationships that users might check themselves; or alternatively, omit from analysis when they embark on de novo retrieval of information from published literature. Consequently, such users have to redo the work of the biocurators to identify why data has been assessed but excluded.

This is illustrated by example 1. Here, the GO dbTF catalogue offers detailed information on the dbTFs that have passed the scrutiny of the latest curation guidelines. It also presents information on candidates that have been excluded from the list of dbTFs, i.e., data that has been assessed by the curator and rejected. Yet, information on why a specific candidate has been rejected and placed into the rejected class is not provided in much detail. What the GO dbTFs compendium (Lovering et al., this issue of BBA) does indicate, however, is that if a candidate protein entry meets none of the seven applied inclusion criteria examined that would have led to inclusion on the dbTF list, it is rejected from the compendium. Hence, while it is relatively easy to identify what specific data has been rejected, the specific reason why a protein is rejected may thus be internalized and be the consequence of not meeting even one of the main criteria, namely (i) convincing published evidence for cis-regulatory region-mediated sequence-specific gene transcription modulation (ii) the presence of a characterized dbTF DNA binding domain instance (iii) the determination of DNA binding specificity by heterologously produced proteins, or (iv) being phylogenetically paralogous to established dbTFs. While this assessment guarantees the assembly of a dbTF compendium with the highest possible quality, recording more specific information about what information is not convincing could be instrumental to users of a greedy use case. Absence of evidence is not necessarily evidence of absence, and as a result, such users then need to reassess the rejected list based on their local criteria and reproduce the effort performed of the biocurators following their own inclusion/exclusion criteria.

12. How can databases help mitigate the drawbacks associated with greedy use cases?

We suggest that by making the process of how rejected data is handled visible to users, manually curated databases could serve a wider set of use cases without jeopardizing their local quality standards.

12.1. Proposal

Biocuration provides an indispensable resource for research in the life sciences. Whereas today's efforts in biocuration are primarily focused towards producing database entries ('positives') with as little contamination as possible (false positives), the efforts that are spent at sifting through vast volumes of scientific literature could produce even more useful results if information about declaring 'negatives', gathered in this process, is also archived, and made available to users.

We propose that the two examples could inspire biocurators to develop their own techniques to accommodate for a plurality of purity criteria. By adjusting how they handle data they reject and shifting their pollution control procedures to the front end, databases could offer their users information on what is *filtered out* in this process. First, a database could list all the papers that have been subjected to their curation process. This could be done, for instance, by tagging all publications in a triaged corpus with their status (curatable/non-curatable). Alternatively, biocurators could identify what evidence needed to satisfy their criteria is missing from the data by highlighting the specific parts of publications that have been excluded for not being sufficiently pure for the database. This would be valuable information in greedy use cases to allow users to decide to accept these less strong claims. Taking it a step

further, biocurators could attach the type of evidence that is needed for such knowledge claims to be accepted into the database (Gaudet et al., this issue; Lovering et al., this issue). This might even inspire users to follow up on this by producing additional experimental evidence.

An additional advantage of providing such information is that it sheds light on existing research niches where more knowledge is needed. As we have discussed above, curators thoroughly analyze a data set before they decide whether to reject it. Yet, by making a hard distinction between pure and polluted data, that which is *almost* pure is also often rejected. By explicitly labeling data that is insufficiently pure for the database, biocurators could indicate the research areas where evidence is lacking. Bringing forward this information would then be instrumental in motivating data producers to generate additional evidence for annotating this data.

Finally, reporting the literature that was considered for curation even when it was not included, and publishing curation guidelines and criteria, also increases the visibility of the intricacies of the curatorial process. While this does increase the annotation burden, it would do justice to the vast amount of work of the biocurator that today remains invisible to users [38,55,56]. As BC05 states, "They [users] are often not that aware of the level of detail and involvement that data curation requires. They are not aware of all the quality controls that we need to put in." This work also includes the time and effort that is spent in finding, ordering, and structuring data before a biocurator can take a decision whether to accept or reject it. To do so, they typically make notes, consider the available evidence, determine what evidence seems to be missing, pursue authors to fill in the blanks, follow up with journals to trace lost supplementary material - amongst other tasks. Yet, while these efforts and the resulting body of knowledge is valuable, users cannot access such information. For data that is assessed but rejected, the knowledge generated from the careful and scrupulous efforts of the curator remains underexploited. Bringing these aspects to the foreground also enables utilizing the outcome of an often-elaborate assessment of data that is subsequently excluded from the database.

Nevertheless, while this proposal is aimed at addressing the drawbacks faced by users of a greedy use case, we realize that there are also foreseeable disadvantages to our proposal. One such disadvantage is that making the reasons for rejection visible to users could compromise on the richness and honesty of biocurators' opinions. When such information is confined to the back end of the database, biocurators are allowed a certain level of anonymity. Yet, when accessible, decisions which result in the exclusion of an article from a database may be misconstrued as an assessment of poor quality of the article and, by extension, the authors, their university or even their country or continent. This proposal could contribute to potential tensions between biocurators and producers of data. Such collateral effects would need to be addressed by reiterating that pollution is a social and cultural and pragmatic construction which is relative to how a community perceives purity and order. In this case, this implies that a biocurator's decision to reject data depends on how well it fits the local schema of the database. As this schema differs across databases, as well as use cases, data considered polluted by a database could fit well into the schema of a use case.

13. Conclusion

So far, we have discussed Douglas' analysis of different conceptions of pollution as she presents it in *Purity and Danger: an analysis of the concepts of pollution and taboo*. We explored her arguments on the role played by the values and needs of communities in distinguishing purity and order from pollution and disorder. This distinction is constructed and context dependent.

We then explored the phenomenon of pollution in scientific research data. We began by briefly discussing the research practices that are known to pollute scientific research results and the possible effects of such pollution on new research. We then described the role of biocurators in tackling such forms of pollution. They carefully and often painstakingly reassess published data following the curation guidelines specific to their databases. We then discussed some general, overarching criteria that constitute guidelines used in all databases, to identify and curate only pure data or true positives, and to filter out polluted data or false positives.

Having identified these commonalities, we then turned to the differences. We explained that what is pure for one database may be impure for the next, as both databases adhere to different definitions of pollution. We extended this argument to explore the hypothesis that database users of two use cases - cautious and greedy - may also assess pollution differently than the database. When claiming that a piece of information is truly positive or truly negative, curation protocols use decision tree-like protocols that result either in a yes or a no, leaving no room for a 'maybe'. This 'maybe' class, we suggest, can be valuable as it may contain false negatives: a class of value that, based on current available data (from experiments or bioinformatics analysis), cannot be classified as positive. How this class is currently handled by databases presents certain drawbacks to greedy use cases.

We then asked how databases and biocurators could help users of the greedy use case in alleviating these drawbacks. We suggest that the key lies in databases disclosing the data they reject and the reasons for doing so. In this manner, databases could cater to a variety of conceptions of pollution without compromising its own local definition. Providing this information can also point users towards areas where knowledge is lacking. Users can then work on filling these knowledge gaps. Finally, implementing our proposal would also reveal the efforts of the curator to scrutinize data *before* deeming it to be polluted for the database.

In conclusion we return to the research questions. We first asked - *what are the drawbacks for users of use cases where the purity criteria differ from the local criteria of a database? U*sers of a greedy use case face a gap in the data that has been curated into the database and data that their use case needs because the conceptions of pollution vary between the database and that of the use case. We identified that the effort to bridge this gap demands a duplication of the biocurator's efforts by users of a greedy use case. Our second question was: *what role can databases and biocurators play in alleviating these drawbacks?* We propose that databases assist such users by offering them access to the details of the data that has been considered but rejected from the databases, and why. For the primary literature this is represented by part B on Fig. 1. While such information is available with the database, it is typically internalized and not accessible to users. While we acknowledge that this proposal would increase the burden of the already underfunded biocuration community, we propose that this could help mitigate the additional effort users of a greedy use case would have to invest in undertaking an ad-hoc information search and curation task.

CRediT authorship contribution statement

Anamika Chatterjee: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing. **Tsjalling Swierstra:** Conceptualization, Supervision, Writing – review & editing. **Martin Kuiper:** Conceptualization, Validation, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This publication is based upon work from COST Action CA15205: GREEKC, supported by COST (European Cooperation in Science and Technology), carried out with funding from NTNU-Health, Norway. We

would also like to acknowledge the interview respondents who have given their time and enriched this publication.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbagrm.2021.194779.

References

[1] M. Douglas, Purity and Danger: An Analysis of Concepts of Pollution and Taboo, Psychology Press, 2003 (216 p.).
[2] A. Chatterjee, Pure for Me or Impure for Us: Pollution in the Gene Regulation Knowledge Commons [Internet]. SocArXiv [cited 2020 Dec 31]. Available from: https://osf.io/preprints/socarxiv/yjhsa/, 2020.
[3] C. Hess, E. Ostrom, Introduction: an overview of the knowledge commons, in: Understanding Knowledge as a Commons: From Theory to Practice, MITP, 2007, pp. 3–26 [Internet]. [cited 2020 Jul 16]. Available from: https://ieeexplore.ieee.org/document/6284197.
[4] kamraro, Responsible Research & Innovation [Internet]. Horizon 2020 - European Commission [cited 2021 Jun 15]. Available from: https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation, 2014.
[5] S. Efstathiou, R. Nydal, A. Laegreid, M. Kuiper, Scientific knowledge in the age of computation: explicated, computable and manageable? Theor. Int. J. Theory Hist. Found Sci. 34 (2) (2019 Sep 25) 213–236.
[6] R. Nydal, G. Bennett, M. Kuiper, A. Lægreid, Silencing trust: confidence and familiarity in re-engineering knowledge infrastructures, Med. Health Care Philos. 23 (3) (2020 Sep 1) 471–484.
[7] D. Goodrick, P.J. Rogers, Qualitative data analysis, in: K.E. Newcomer, H.P. Hatry, J.S. Wholey (Eds.), Handbook of Practical Program Evaluation [Internet], John Wiley & Sons, Inc., Hoboken, NJ, USA, 2015, pp. 561–595 [cited 2021 Jun 14]. Available from: https://doi.org/10.1002/9781119171386.ch22.
[8] H.K. Mohajan, Qualitative research methodology in social sciences and related subjects, J. Econ. Dev. Environ. People 7 (1) (2018) 23–48.
[9] W.C. Adams, Conducting semi-structured interviews, in: Handbook of Practical Program Evaluation [Internet], John Wiley & Sons, Ltd, 2015, pp. 492–505 [cited 2021 Jun 14]. Available from: https://doi.org/10.1002/9781119171386.ch19.
[10] B. Campkin, Placing "Matter Out of Place": purity and danger as evidence for architecture and urbanism, Archit. Theory Rev. 18 (1) (2013 Apr 1) 46–61.
[11] H. Pickering, T. Rice, Noise as "sound out of place": investigating the links between Mary Douglas' work on dirt and sound studies research [cited 2021 Jun 29]; Available from: https://ore.exeter.ac.uk/repository/handle/10871/28334, 2017 Jul 4.
[12] K. Ditlevsen, S.S. Andersen, The purity of dirt: revisiting Mary Douglas in the light of contemporary consumer interpretations of naturalness, Purity Dirt. Sociology. 55 (1) (2021 Feb 1) 179–196.
[13] A Little Dirt Never Hurt Anyone:: Knowledge-Making and Contamination in Materials Science - Cyrus C.M. Mody, 2001 [Internet]. [cited 2021 Jul 12]. Available from: https://doi.org/10.1177/030631201031001002.
[14] D.A. Bray, Information Pollution, Knowledge Overload, Limited Attention Spans, and Our Responsibilities as IS Professionals [Internet], Social Science Research Network, Rochester, NY, 2007 Feb [cited 2020 Jul 16]. Report No.: ID 962732. Available from: https://papers.ssrn.com/abstract=962732.
[15] D.M. Levy, Information overload, in: The Handbook of Information and Computer Ethics [Internet], John Wiley & Sons, Ltd, 2009, pp. 497–515 [cited 2020 Jul 16]. Available from: https://doi.org/10.1002/9780470281819.ch20.
[16] P. Maes, Agents that reduce work and information overload, in: R.M. Baecker, J. Grudin, Buxton WAS, S. Greenberg (Eds.), Readings in Human–Computer Interaction [Internet], Morgan Kaufmann, 1995, pp. 811–821 [cited 2020 Jul 16]. (Interactive Technologies). Available from: http://www.sciencedirect.com/science/article/pii/B9780080515748500844.
[17] C. Castelluccia, M.A. Kaafar, Owner-centric networking (OCN): toward a data pollution-free internet, in: 2009 Ninth Annual International Symposium on Applications and the Internet, 2009, pp. 169–172.
[18] Kai-Yuan Cai, Chao-Yang Zhang, Towards a research on information pollution, in: 1996 IEEE International Conference on Systems, Man and Cybernetics Information Intelligence and Systems (Cat No96CH35929) Vol. 4, 1996, pp. 3124–3129.
[19] D.E. Nelson, Reducing Information Pollution in the Internet Age. Prev Chronic Dis [Internet] [cited 2020 Jul 16];4(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1832139/, 2006 Dec 15.
[20] M. Baker, 1,500 scientists lift the lid on reproducibility, Nature 533 (7604) (2016 May 1) 452–454.
[21] C. Fiala, E.P. Diamandis, Benign and malignant scientific irreproducibility, Clin. Biochem. (55) (2018 May 1) 1–2.
[22] M. Baker, Irreproducible biology research costs put at $28 billion per year, Nature 859 (2015 Jun 9) [Internet]. cited 2021 Jun 17, available from: https://doi.org/10.1038/nature.2015.17711.
[23] N. Levy, Taking responsibility for health in an epistemically polluted environment, Theor. Med. Bioeth. 39 (2) (2018) 123–141.
[24] R. Liu, S. Lundin, Falsified Medicines: Literature review, in: Work Pap Med Humanit [Internet], 2016 Jan 15 [cited 2021 Jun 17];2(1). Available from: https://journals.lub.lu.se/medhum/article/view/15308.
[25] M.S. Rahman, N. Yoshida, H. Tsuboi, N. Tomizu, J. Endo, O. Miyu, et al., The health consequences of falsified medicines- a study of the published literature, Tropical Med. Int. Health 23 (12) (2018) 1294–1303.
[26] H.Y. Vanderpool, G.B. Weiss, False data and last hopes: enrolling ineligible patients in clinical trials, Hast. Cent. Rep. 17 (2) (1987) 16–19.
[27] J. Bohannon, Who's afraid of peer review? Science 342 (6154) (2013 Oct 4) 60–65.
[28] F. Godlee, C.R. Gale, C.N. Martyn, Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: a randomized controlled trial, JAMA 280 (3) (1998 Jul 15) 237–240.
[29] M. Henderson, Problems with peer review, BMJ (340) (2010 Mar 15), c1409.
[30] S. Schroter, N. Black, S. Evans, J. Carpenter, F. Godlee, R. Smith, Effects of training on quality of peer review: randomised controlled trial, BMJ 328 (7441) (2004 Mar 20) 673.
[31] A.H.S. Kumar, Rise in polluters of scientific research: how to curtail information pollution (infollution), J. Nat. Sci. Biol. Med. 4 (2) (2013) 271.
[32] A.H.S. Kumar, Open review system: the new trend in scientific reviewing to improve transparency and overcome biasness, J. Nat. Sci. Biol. Med. 5 (2) (2014) 231–232.
[33] Biocuration IS for, Biocuration: distilling data into knowledge, PLoS Biol. 16 (4) (2018 Apr 16), e2002846.
[34] R. Drysdale, C.E. Cook, R. Petryszak, V. Baillie-Gerritsen, M. Barlow, E. Gasteiger, et al., The ELIXIR Core data resources: fundamental infrastructure for the life sciences, Bioinformatics 36 (8) (2020 Apr 15) 2636–2642.
[35] M.D. Wilkinson, M. Dumontier, IjJ Aalbersberg, G. Appleton, M. Axton, A. Baak, et al., The FAIR guiding principles for scientific data management and stewardship, Sci. Data 3 (1) (2016) 160018 (Mar 15. available from: https://doi.org/10.1038/sdata.2016.18.
[36] A. Bateman, Curators of the world unite: the International Society of Biocuration, Bioinformatics 26 (8) (2010 Apr 15) 991.
[37] S. Burge, T.K. Attwood, A. Bateman, T.Z. Berardini, M. Cherry, C. O'Donovan, et al., Biocurators and Biocuration: surveying the 21st century challenges, Database [Internet] (2012 Jan 1) [cited 2020 Jul 16]; Available from: https://doi.org/10.1093/database/bar059/429565.
[38] A.M. Gabrielsen, Openness and trust in data-intensive science: the case of biocuration, Med. Health Care Philos. (2020 Jun 10) [Cited 2020 Jul 16]. Available from: https://doi.org/10.1007/s11019-020-09960-5.
[39] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, et al., The future of biocuration, Nature 455 (7209) (2008 Sep) 47–50.
[40] S. Tripathi, S. Vercruysse, K. Chawla, K.R. Christie, J.A. Blake, R.P. Huntley, et al., Gene regulation knowledge commons: community action takes care of DNA binding transcription factors, Database [Internet] (2016 Jan 1) [cited 2020 Jul 16]. Available from: https://doi.org/10.1093/database/baw088.
[41] R. Fuchs, P. Rice, G.N. Cameron, Molecular biological databases — present and future, Trends Biotechnol. (10) (1992 Jan 1) 61–65.
[42] C. Douglas, R. Goulding, L. Farris, J. Atkinson-Grosjean, Socio-cultural characteristics of usability of bioinformatics databases and tools, Interdiscip. Sci. Rev. 36 (1) (2011 Mar 1) 55–71.
[43] M.-F. Costabile, D. Fogli, C. Letondal, P. Mussio, Piccinno° A., Domain-expert users and their needs of software development, in: HCI 2003 End User Development Session [Internet]. Crète, Greece, 2003 [cited 2021 Jun 28]. (Proceedings of the HCI 2003 End User Development Session). Available from: https://hal.archives-ouvertes.fr/hal-01299738.
[44] M.F. Costabile, D. Fogli, P. Mussio, A. Piccinno, End-user development: the software shaping workshop approach, in: H. Lieberman, F. Paternò, V. Wulf (Eds.), End User Development [Internet], Springer Netherlands, Dordrecht, 2006, pp. 183–205 [cited 2021 Jun 28]. (Human-Computer Interaction Series). Available from: https://doi.org/10.1007/1-4020-5386-X_9.
[45] Niels Velthuijs, Birgit Meldal, Quinte Geessinck, Pablo Porras, Yulia Medvedeva, Anatoliy Zubritskiy, Sandra Orchard, Colin Logie, Integration of transcription coregulator complexes with sequence-specific DNA-binding factor interactomes, Biochim. Biophys. Acta, Gene Regul. Mech. 10 (1864) [Cited 2021 October 1]: 194749. Available from: https://doi.org/10.1016/j.bbagrm.2021.194749.
[46] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, et al., IntAct: an open source molecular interaction database, Nucleic Acids Res. 32 (Database issue) (2004) D452–D455. Jan 1.
[47] Rose Oughtred, Jennifer Rust, Christie Chang, Bobby-Joe Breitkreutz, Chris Stark, Andrew Willems, Lorrie Boucher, et al., The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions, Protein Sci. 30 (1) (2021) 187–200. Available from: https://doi.org/10.1002/pro.3978.
[48] K. Chawla, S. Tripathi, L. Thommesen, A. Lægreid, M. Kuiper, TFcheckpoint: a curated compendium of specific DNA-binding RNA polymerase II transcription factors, Bioinformatics 29 (19) (2013 Oct 1) 2519–2520.
[49] H. Han, H. Shim, D. Shin, J.E. Shim, Y. Ko, J. Shin, et al., TRRUST: a reference database of human transcriptional regulatory interactions, Sci. Rep. 5 (1) (2015 Jun 12) 11432.
[50] A. Essaghir, F. Toffalini, L. Knoops, A. Kallin, J. van Helden, J.-B. Demoulin, Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data, Nucleic Acids Res. 38 (11) (2010 Jun), e120.
[51] L.A. Bovolenta, M.L. Acencio, N. Lemke, HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions, BMC Genomics 13 (1) (2012 Aug 17) 405.
[52] L. Licata, P. Lo Surdo, M. Iannuccelli, A. Palma, E. Micarelli, L. Perfetto, et al., SIGNOR 2.0, the SIGnaling network open resource 2.0: 2019 update, Nucleic Acids Res. 48 (D1) (2020 Jan 8) D504–D510.

[53] A.P. Davis, R.J. Johnson, K. Lennon-Hopkins, D. Sciaky, M.C. Rosenstein, T. C. Wiegers, et al., Targeted journal curation as a method to improve data currency at the Comparative Toxicogenomics Database, in: Database J Biol Databases Curation [Internet], 2012 Dec 6 [cited 2020 Dec 16];2012. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3515863/.

[54] T.C. Wiegers, A.P. Davis, K.B. Cohen, L. Hirschman, C.J. Mattingly, Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (CTD), BMC Bioinformatics 10 (2009 Oct 8) 326.

[55] S. Leonelli, Learning from data journeys, in: Data Journeys in the Sciences, Springer, Cham, 2020, pp. 1–24.

[56] M. Boumans, S. Leonelli, From Dirty Data to Tidy Facts: Practices of Clustering in Plant Phenomics and Business Cycles. Var Data Journeys, 2019.