InteractOA: Showcasing the representation of knowledge from scientific literature in Wikidata

Muhammad Elhossary [a] and Konrad U. Förstner [a,b,*]

[a] *Data Science and Services, ZB MED - Information Centre for Life Sciences, NRW, Germany*
E-mail: elhossary@zbmed.de
[b] *Institute of Information Science, TH Köln – University of Applied Sciences, NRW, Germany*
E-mail: foerstner@zbmed.de

Abstract. Knowledge generated during the scientific process is still mostly stored in the form of scholarly articles. This lack of machine-readability hampers efforts to find, query, and reuse such findings efficiently and contributes to today's information overload. While attempts have been made to semantify journal articles, widespread adoption of such approaches is still a long way off. One way to demonstrate the usefulness of such approaches to the scientific community is by showcasing the use of freely available, open-access knowledge graphs such as Wikidata as sustainable storage and representation solutions. Here we present an example from the life sciences in which knowledge items from scholarly literature are represented in Wikidata, linked to their exact position in open-access articles. In this way, they become part of a rich knowledge graph while maintaining clear ties to their origins. As example entities, we chose small regulatory RNAs (sRNAs) that play an important role in bacterial and archaeal gene regulation. These post-transcriptional regulators can influence the activities of multiple genes in various manners, forming complex interaction networks. We stored the information on sRNA molecule interaction taken from open-access articles in Wikidata and built an intuitive web interface called *InteractOA*, which makes it easy to visualize, edit, and query information. The tool also links information on small RNAs to their reference articles from PubMed Central on the statement level. *InteractOA* encourages researchers to contribute, save, and curate their own similar findings. *InteractOA* is hosted at https://tools.wmflabs.org/interactoa and its code is available under a permissive open source licence. In principle, the approach presented here can be applied to any other field of research.

Keywords: Wikidata, interactions, regulatory networks, citations

1. Background and related work

1.1. Knowledge graphs and Wikidata

The term "knowledge graph" was coined by Google in 2012[1]. Since then, knowledge graphs have attracted growing attention from researchers due to their robustness in many areas of science, besides application in numerous other fields [1]. Although several attempts have been made to define a knowledge graph, there is still no single, agreed-upon definition of what it entails [2]. As the name implies, a knowledge graph, also known as a semantic

*Corresponding author. E-mail: foerstner@zbmed.de.
[1] https://www.blog.google/products/search/introducing-knowledge-graph-things-not/

network, is a means of storing knowledge in a graph-based model. It is a structured data model that represents a network of real-world entities such as objects or concepts and illustrates the relationships between them. This information is usually stored in a graph database, and can be visualized in a graph structure. Knowledge graphs can be viewed as a network of nodes and edges, where nodes represent the entities and edges represent the relationships. They can store vast amounts of heterogeneous data in a structured manner and handle complex relationships, making them well suited for applications in various domains and research areas [3].

The Resource Description Framework (RDF) [4] is one model for implementing knowledge graph storage. The RDF specification[2] highlights the simplicity of the RDF data model, which is represented at its fundamental level in the form of subject-predicate-object triples. These triples can describe anything in a flexible and extensible way. To query these RDF triples stores, a query language known as SPARQL, which is similar to SQL, is used to retrieve and manipulate data stored in an RDF format.

Wikidata is one of the prime examples of knowledge graphs, and it combines two of their potential benefits: openness (as the content is published under the Creative Commons Zero licence[3]) and ease of editability [5, 6]. It is based on the Wikibase software, which allows RDF exports [7], and its content can be queried via the SPARQL endpoint known as the *Wikidata Query Service*. Wikidata is maintained by the Wikimedia Foundation, which aims to provide structured data about the world's knowledge and make it available for anyone to use and extend collaboratively. A number of projects are underway to analyse and increase the quality of Wikidata [8]. It is regarded as a key source of identifiers [9] and has tremendous potential which remains largely untapped [10].

1.2. Insufficient management of data, information, and knowledge in the life sciences

Data, information, and knowledge are being generated at an ever-increasing pace in the field of biology. This is due in large part to the widespread availability of high-throughput platforms in biological research, which can generate vast amounts of data, for example on genes, proteins, and other biological entities and their interactions. The ability to collect, organize, and analyse this data—and the information and knowledge derived from it—is crucial for future biological and biomedical research. While the FAIR principles are now widely acknowledged as a useful framework for managing data [11], a large fraction of the knowledge generated on the basis of this data continues to be stored in unstructured formats such as scholarly articles, which have formed the core of knowledge management in research for centuries. Moreover, a significant proportion of these articles are not available in a machine-readable formats (e.g. PDF), which in turn hinders the downstream information mining. Even the machine-readable formats (e.g. in HTML, and XML) continue to lack the semantic enrichment. Attempts have been made to address this [12, 13], but these have been largely ignored by the publishing industry and scholarly community. Recent projects such as the Open Research Knowledge Graph (ORKG) [14] have tried to overcome this hurdle by offering a semantic database to represent the research findings of publications in a separate machine-actionable form but are referencing only on the article level and not to exact passages. Furthermore, domain-specific biological knowledge graphs instances with SPARQL endpoints are also available such as UniProt [15] which is dedicated for storing information of proteins and IntAct [16] for molecular interactions – both also referencing only on the article level.

Alongside their unstructured representation in academic literature, data, information, and knowledge are also stored in biological databases. Even though these databases represent a valuable resource for larger or more specialized research communities, their development and maintenance is often limited to the lifetime of the respective research project. There are numerous cases of valuable databases that were created as part of such projects, but are no longer accessible. Examples of databases that cannot be accessed via their published URLs at the time of writing this article include BSRD [17] and sRNAdb [18], which are databases for bacterial small RNAs, as well as AANT [19], a database for amino acid-nucleotide interactions, and cpnDB [20], a database for bacterial Chaperonin sequences. Several studies have examined this issue. Their results show that more than 30% of bioinformatical web

[2] https://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
[3] https://creativecommons.org/publicdomain/zero/1.0/

services published over the last 23 years are currently unavailable, with lack of maintenance being the primary cause of this decay [21–24]. Additionally, restrictive licensing can make it difficult for researchers to access and (re-)use these resources, further hindering scientific advancement.

Another challenging aspect of knowledge management in research is the granularity of references. Claims are cited on the level of full articles, which makes it very time-consuming for readers to find the exact location of a particular statement in the referenced source. This poses a major obstacle to the verification and contextualization of such references by readers.

1.3. Wikidata as a knowledge-graph solution for biological data, information, and knowledge

In the life sciences and in other research fields, knowledge graphs such as Wikidata are used to represent and integrate information from a variety of sources, including genomic data, literature, and experimental results [25–27]. Knowledge graphs can be employed to model complex biological systems and processes, and to facilitate data mining and analysis. This may include the representation and integration of genomic data such as genes, proteins, and pathways. Furthermore, they can be used to model the interactions of biomedical entities, for example linking genes associated with antibiotic resistance in a pathogen [28].

Widespread use of knowledge graphs in the life sciences would facilitate the discovery of new biological insights and relationships through data mining and visualization, and improve the interpretation and understanding of biological data through the integration of diverse data sources. One example of the implementation of knowledge graphs for biological data is the Clinical Knowledge Graph (CKG) [29]. The CKG aims to assist in the delivery of personalized medical treatment by using machine-learning methods to mine data from heterogeneous domains. Wikidata pushes these capabilities further by facilitating the sharing and reuse of biological data through the use of standardized data formats and open-access solutions. Further examples of existing solutions include WikiGenomes [30], an openly editable knowledge graph for genomic annotations that is geared towards the molecular biology community, ChlamBase [31], which is a central access point for genomic and proteomic information specifically for the Chlamydia research community, and WikiPathways [32], an open, collaborative, community-based platform dedicated to the curation of biological pathways. A number of articles have been published that encourage the use of Wikidata-based solutions in biology, such as the Gene Wiki initiative [33]

1.4. Small RNA regulatory networks

In bacteria and archaea, gene expression is controlled by a variety of regulators. One class of regulators is the small RNAs (also known as non-coding RNAs) [34], which are not translated into proteins to perform their regulatory functions. This class of RNAs is responsible for vital regulatory roles in gene expression. These small RNAs are often expressed in their hundreds to control cellular functions such as the response to environmental changes, and each small RNA can influence the activity of multiple targets of proteins or messenger RNAs. Small RNAs regulate their targets through various mechanisms [35, 36] such as down-regulation (by disrupting mRNA translation through base-pairing to the ribosomal binding sites) and up-regulation (by inhibiting mRNA degradation). Despite their importance in all known bacterial and archaea species, knowledge of small RNAs is comparatively limited, and they often fail to be included in the creation of holistic models in systems biology.

The genomic locations of small RNAs are first identified through experiments and annotated in genomic reference sequences; then, the interactions between these annotated small RNAs and genes are computationally predicted and experimentally confirmed. The numerous interactions at the cellular level that occur under different environmental conditions between different regulators, such as small RNAs and their target genes, can be represented as a network. This network consists of nodes such as regulators and gene targets, with the edges between these nodes representing the identified interactions. Information on these interactions can be represented and visualized in network graphs to facilitate understanding of their complexity. Typically, researchers report their findings on bacterial small RNA regulatory networks in various, usually unstructured formats that lack a consistent standard. This data is presented

in the main body of articles, in supplementary materials such as spreadsheets, or in flat files. For a limited number of species, data is manually compiled by experts into web-based databases such as RegulonDB [37]. Without a uniform format for reporting such information, it is difficult to gain a comprehensive understanding of these regulatory networks.

2. Approach and Implementation

2.1. InteractOA as a Wikidata-based application for small RNA regulatory networks

With these needs and technological foundation in mind we modelled small RNA interactions in Wikidata and built an application called *InteractOA* (OA for Open Access as is builds on Open Access scholarly articles) that presents the data. The application features a graphical interface that allows users to query Wikidata for a chosen organism and displays the network of interactions or citation information at the statement level for all referenced interactions. The semantic nature of Wikidata, combined with its openness, can help to overcome many challenges in the management of data, information, and knowledge in gene regulation research. Due to its easy-to-use web interface, permissive license (CC0), already large corpus of data (including literature metadata) to which other data can be connected to, the availability of a SPARQL endpoint for querying data, a platform to create applications building upon the data (Toolforge[4]) as well as a strong (not only scientific) community backed by a sustainable organisation (Wikimedia Foundation) Wikidata was chosen as platform to implement this solution to compile extracted scholarly knowledge.

The Wikidata based approach of linking biological entities in order to represent regulatory networks is superior to other methods in terms of both its openness and sustainability. Since intracellular interactions are often complex and numerous, it is important to use a digital representation approach that is scalable, extensible, easily queryable, and usable. Wikidata possesses all of these properties.

2.2. Modelling small RNA interaction data in Wikidata

Typically, data related to genomes and genes including small RNAs are stored in text files such as GFF3 (General Feature Format)[5] files. These files can be obtained from public repositories like NCBI RefSeq [38]. Each line within a GFF file contains details about a single annotation, including the gene's location on the genome, its type, function, and other attributes. To model small RNAs and their interactions, we translated each gene annotation in these GFF files into Wikidata items. These items were then interconnected using specific properties to represent the interactions. Selected, widely used bacterial strains were used for this modelling process due to the abundance of their small RNA interaction data, including *Escherichia coli* strain *K-12 substr. MG1655* (NCBI genome assembly GCF_000005845.2, RefSeq accession NC_000913.3).

For the processing and integration of the data an ETL (extract-transform-load) workflow was established (see Figure 1). The first step was to obtain genomic annotation data for these organisms from the NCBI RefSeq repository. Next, a Python tool was developed to automate the process of importing annotation records from GFF files and linking them to Wikidata entries using defined IDs corresponding to specific strains. To do this, the tool extracts information from each annotation record, generates Wikidata item definitions (including name, synonyms, and description), and handles communication with Wikidata's API to create new non-redundant Wikidata items. These newly created Wikidata items were then expanded by importing the remaining annotation data, including entry type (e.g. protein or RNA), genomic location, and external gene identifiers. This process was based on selected Wikidata items and properties defined for the biology domain, which were collected from Wikidata's listings for life

[4] https://wikitech.wikimedia.org/wiki/Portal:Toolforge
[5] http://www.ensembl.org/info/website/upload/gff3.html

Table 1
Table of properties and items used in the modelling of the annotations, interactions, and citations of small RNA interactions.

ID	Name	Type	Usage / Description
P703	found in taxon	Property	used to link entities of annotations to a certain Wikidata item that represents an organism at the strain level
P31	instance of	Property	assigns the type of GFF entry e.g. gene or ncRNA to a Wikidata item
P644	genomic start	Property	assigns the genomic start location of an annotation entry to a Wikidata item
P645	genomic end	Property	assigns the genomic end location of an annotation entry to a Wikidata item
P2548	strand orientation	Property	assigns the genomic strand of an annotation entry to a Wikidata item
P688	encodes	Property	to link a gene (Wikidata item) to its product (Wikidata item), e.g. protein, or RNA
P702	encoded by	Property	vice versa of P688
P361	part of	Property	used to describe a partial product of a gene, like genes that splice to multiple mRNAs
P527	has part(s)	Property	vice versa of P361
P351	Entrez Gene ID	Property	assigns an identifier for an annotation entry originated from NCBI Entrez database
P2249	RefSeq genome ID	Property	assigns an identifier qualifier the start, end, strand of an annotation entry originated from NCBI RefSeq database
P2393	NCBI locus tag	Property	assigns an identifier for an annotation locus tag originated from NCBI
P637	RefSeq protein ID	Property	assigns an identifier for protein GFF entry originated from by NCBI
P128	regulates (molecular biology)	Property	links 2 annotation entities based on the interaction of any type if type is unspecified
P3777	antisense inhibitor of	Property	similar to P128, when interaction type is known as inhibition by antisensing
P3771	activator of	Property	similar to P128, when interaction type is known as up-regulation by activation
P3774	blocker of	Property	similar to P128, when interaction type is known as mRNA translation blocking
P3773	antagonist of	Property	similar to P128, when interaction type is known as antagonizing
P3772	agonist of	Property	similar to P128, when interaction type is known as agonizing
P248	stated in	Property	used to add reference about an interaction claim, which is a Wikidata item that represents a scholarly article
P1683	quotation	Property	used to add quote statement to the reference about an interaction claim
P932	PMCID	Property	used to add the PubMed Central identifier for the article of reference about an interaction claim
Q427087	non-coding RNA	Item	class of RNA that is not translated into proteins
Q423832	antisense RNA	Item	RNA molecules hybridizing to complementary sequences in either RNA or DNA, altering the function of the latter
Q201448	transfer RNA	Item	adaptor molecule composed of RNA
Q285904	Transfer-messenger RNA	Item	bifunctional RNA that has properties of a tRNA and an mRNA
Q424665	signal recognition particle	Item	protein-RNA complex facilitating translocation of proteins across membranes
Q1012651	ribonuclease P activity	Item	catalysis of the endonucleolytic cleavage of RNA, removing 5' extra nucleotides from tRNA precursor.
Q11053	RNA	Item	family of large biological molecules
Q7187	Gene	Item	basic physical and functional unit of heredity
Q22809680	forward strand	Item	forward oriented strand in a double-stranded DNA molecule
Q22809711	reverse strand	Item	reverse oriented strand in a double-stranded DNA molecule
Q215980	ribosomal RNA	Item	RNA component of the ribosome, essential for protein synthesis in all living organisms
Q277338	pseudogene	Item	functionless relative of a gene

sciences[6]. See Table 1 for a list of items and properties used here.

[6] https://www.wikidata.org/wiki/Wikidata:List_of_properties/natural_science#Wikidata_property_related_to_biology

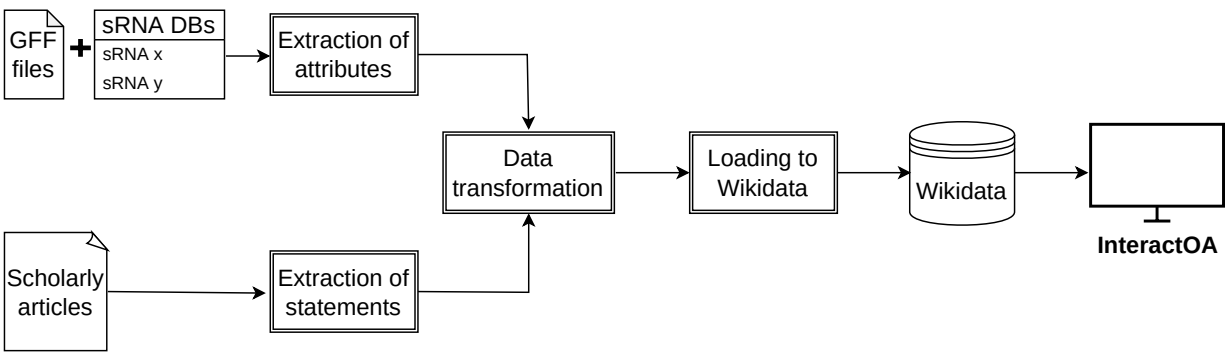

Fig. 1. Schema of the Extract-Transform-Load workflow used for integrating the data into Wikidata and presentation by *InteractOA*.

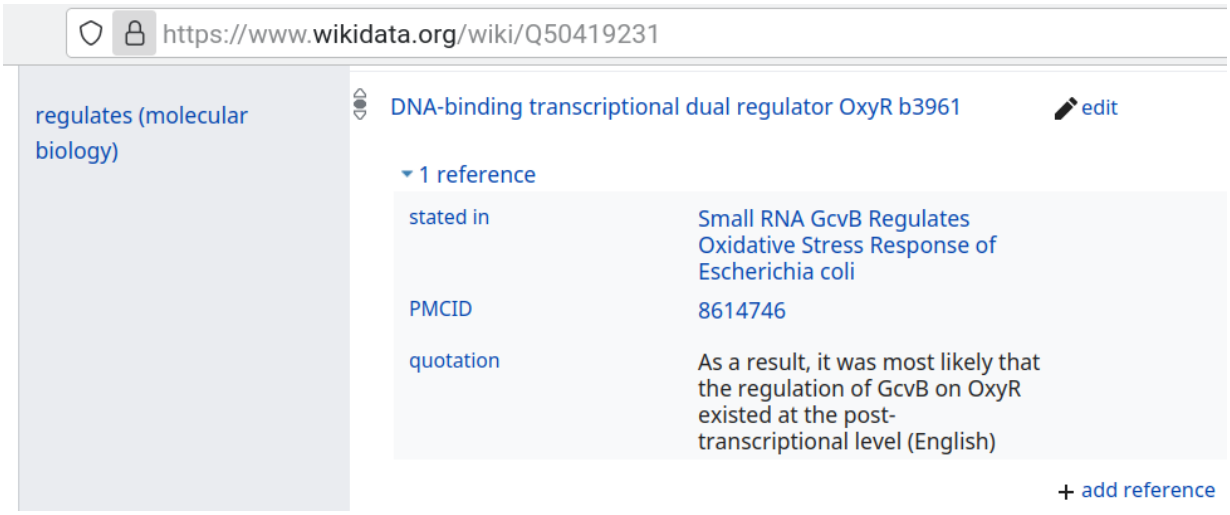

Fig. 2. The sRNA encoded by gcvB (Q50419231) as an example of a citation at the statement level.

Next, the Wikidata items representing RNAs were linked to model the interactions. To do this, interaction data for numerous small RNAs was manually obtained from research articles, RegulonDB [37], and the Staphylococcal Regulatory RNA Database" (SRD) [39]. For each of the selected interactions taken from the databases, the underlying statements were manually extracted from the corresponding article and the statements collected in a dedicated file. The compiled information from this manual curation as well as the data automatically extracted from the GFF files was imported to Wikidata using a dedicated Python tool which first extracts the interaction from input files and then maps the names of interaction partners in the parsed file to the respective pre-imported annotations in Wikidata. The item of the pre-imported annotation is retrieved by the tool using a query template to get the corresponding item's ID. After that, the tool links the Wikidata item representing RNAs by using the selected properties. For example, the property "antisense inhibitor of" (P3777) was used to link the interaction between sRNA *omrA* (Item ID Q50419343) and gene *csgF* (Item ID Q23087296). For an example of such a link, see Figure 2. If the type of interaction has no corresponding property in Wikidata, the tool falls back to the more generic property (P128 regulates). To showcase the implementation until the point of writing this article, 776 small RNA annotations were imported and 253 connections between small RNAs and their targets were contributed specific to 3 species: *Escherichia coli*,

Vibrio Cholerae, and *Staphylococcus aureus*.

Both the Python tools used to import the data were built upon Wikidataintegrator[7] [25] and pywikibot to facilitate interaction with the Wikidata API during querying and importing. They are available on GitHub[8], and they are archived in Zenodo[9].

2.3. Statement-level citations stored in Wikidata

A significant proportion of the sRNA interactions modelled using the method described above appear in open-access articles or other articles freely accessible on PubMed Central[10]. For several of these interactions, the statements describing the specific interactions were extracted manually. The Wikidata items of the source article, the PubMed Central ID and statements (using the "quotation" property P1683) were added as interaction properties. For example, as shown in Figure 2, the property "stated in" (P248) was used with the Wikidata item that corresponds to a journal article entitled "Small RNA GcvB Regulates Oxidative Stress Response of Escherichia coli" [40] (Q115652789) in order to link the interaction to the article in which it was mentioned. Moreover, the PubMed Central ID of this article was linked with the property "PMCID" (P932), and the statement mentioning the interaction was also linked using the "quotation" property (P1683). This method of storing source statements of sRNA interactions in Wikidata makes it easy to check and correct the features of interaction models.

2.4. InteractOA as a front end to relevant Wikidata items

The web front end *InteractOA* was developed to facilitate interactive exploration of the data stored in Wikidata, as described above. *InteractOA* is implemented in Python and the Flask web framework. Its code is available on GitHub[11] and its releases are archived at Zenodo[12]. Wikidata features a query service[13] that enables users to enter SPARQL queries. The service generates tables as well as various types of interactive visualizations, including bar charts and in this case, most importantly, network plots. The web front end uses these Wikidata capabilities to visualize the regulatory interaction as an interactive network plot (see Figure 3 for an example). The interface allows users to customize their queries using filters and to search using keywords without requiring any technical knowledge. Once the user has selected the desired filters and keywords, *InteractOA* sends the generated SPARQL query to the *Wikidata Query Service* and displays the results. The full landscape of small RNAs is displayed if no filters or keywords were used. Figure 3 shows an example network of three small RNAs and their interactions with other protein-coding genes based on a limited set of locus tag IDs as keywords.

Additionally, *InteractOA* provides a tabular view for all the extracted statements in Wikidata for a strain selected by the user and presents them in a searchable table (see Figure 4 for an example). Its search function can filter results by several criteria, for example by partners of interactions or by type of interaction. This solution enables users to combine multiple statements from several studies at once and very likely shortens the time needed to consult previous research on individual small RNAs. Moreover, the user can open the scholarly article from which the statement originated, with the statement itself highlighted (see Figure 5 for an example). *InteractOA* functions as an interface between the user and the Wikidata endpoint, with no data storage taking place on the platform itself. Essentially, *InteractOA* operates as a middle layer implemented in the Python web framework flask[14], designed to facilitate dynamic data querying based on predefined SPARQL query templates. These templates contain parameters

[7] https://github.com/SuLab/WikidataIntegrator
[8] https://github.com/foerstner-lab/GFF_to_Wikidata_importer https://github.com/foerstner-lab/sRNA_Interactions_to_Wikidata_Importer
[9] https://zenodo.org/record/7638542 and https://zenodo.org/record/7638552
[10] https://pubmed.ncbi.nlm.nih.gov/
[11] https://github.com/foerstner-lab/InteractOA
[12] https://doi.org/10.5281/zenodo.7638558
[13] https://query.wikidata.org/
[14] https://flask.palletsprojects.com/

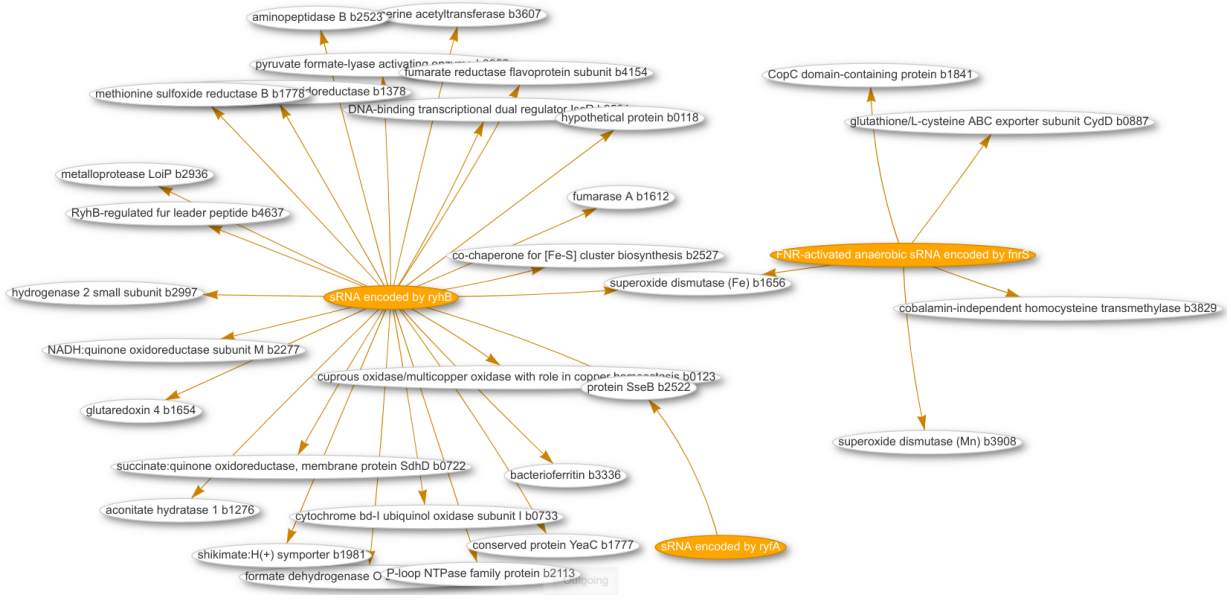

Fig. 3. Screenshot of a network visualization for 3 small RNAs (orange labels) and their interactions with the gene targets (white labels).

that correspond to specific fields and filters on the front end. When a user selects an organism and applies filters, the app processes these inputs and tailors the query from the template accordingly. In the case of visualizing small RNA networks, the specific query is channelled via URL parameters to be executed on Wikidata's query service. An iframe component on the webpage then employs this URL to display the results in the form of a network plot. For the highlighting of statements on small RNA interactions in scholarly articles, the app employs the user's selected organism to pass it to the corresponding parameter in the query template. The Python module WikidataIntegrator[15] [25] is then utilized to execute the query, and the results are included in a tabular view which is then presented to the client. Before transferring the results to the user's web browser, they are enriched with links to the Wikidata items of the sRNAs as well as links to the corresponding article hosted at PubMed Central (PMC) where the selected statements can be highlighted. This colouring works by retrieving the HTLM content of the PMCs articles to the server's memory based on the PMC ID and wrapping the content to be highlighted with marking tags, which is then returned to the user's browser and rendered there with the changed background colour. This process allows for aiding users in identifying and focusing on the relevant information.

3. Discussion

Nowadays, much of the research data on which scholarly articles are based is deposited in a structured format in dedicated repositories, yet the actual insights and knowledge derived from this research are often only accessible in an unstructured format within the confines of the article text. In this work, we have presented a solution to this dilemma based on the open-source Wikidata knowledge graph. As shown here, Wikidata provides a structured way to store knowledge generated within specific fields of research, for example by interleaving items of biological entities with bibliographic information while also providing links to the exact statements that are the source of each knowledge item in the corresponding open-access articles. In this work, we have showcased this approach by modelling the regulatory networks of small RNAs in bacteria and built a dedicated web tool that makes it easy to explore and visualize the data stored in Wikidata. The chosen approach also includes granular referencing of knowledge

[15] https://github.com/SuLab/WikidataIntegrator

Fig. 4. Screenshot of all available statement level citations for a selected organism, which can be filtered with keywords.

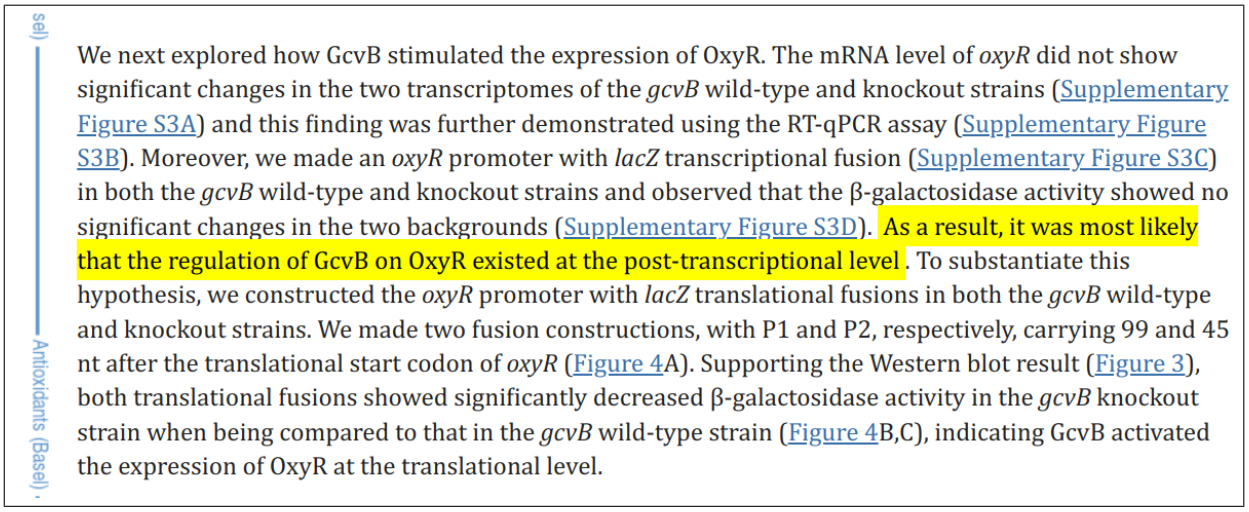

Fig. 5. The highlighted text in an article [40] is a statement level citation example for a claim about small RNA interaction.

sources. Storing the data in Wikidata ensures its long-term availability while opening up access to a large tool chain for imports, queries, and visualization. Moreover, it facilitates links to other relevant entities modelled in Wikidata.

Having demonstrated the usefulness of this approach to our own research field, we now intend to develop the application further to tap into its significant and, as yet, untapped potential. Currently, the steps required to extract statements from research articles are carried out manually. As the quantity of such manually curated article excerpts grows, we aim to train language models to assist with the human curators' extraction work. This text-mining-based approach will build upon related work conducted in our research group (Halder *et al.*, in preparation). For this

named-entity recognition (NER) of sRNA taken from databases followed by the extraction of potential interaction statements combined with a language model will generate the foundation for manual curation of the data. The curation will be conducted by collaborating microbiologist with an expertise of certain bacterial species, especially regarding their regulatory networks. For the curation, a web based tool developed by Halder *et al.* will be used. The curated data will then be integrated into Wikidata as described above and by that can be explored in *InteractOA*. The compiled data will also be used to improve the underlying language model and make the automatic extraction better. Besides this, there is an opportunity to directly incorporate other cellular interactions such as regulatory proteins, protein-protein interactions, and cellular sensing.

Despite the numerous useful features provided by Wikidata and its ecosystem of tools, there are challenges that should be considered when choosing a similar approach. Wikidata's API is comparatively slow, which makes the ingestion of larger data sets very time-consuming. Similarly, SPARQL queries are limited by the constraints of the *Wikidata Query Service*. This latter issue could be solved by working with full data dumps provided by Wikidata. Besides these technical issues, there is inevitably a risk of lower quality, or even significant vandalism, that comes with choosing an openly-editable form of data storage in which anybody can add, remove, or modify entries. Thanks to versioning, problematic edits do not pose a critical risk, however, and an interface to curate new edits could further address this issue.

One additional concern pertains to the availability of item classes and properties within Wikidata. The implemented model of *InteractOA* applies Wikidata's currently existing properties to the topic of small RNAs and their interaction partners. These item classes and properties despite their sufficiency for the current implementation are limited to relatively high-level descriptions such as "regulates" and "antisense inhibitor of". To further improve the model, more item classes and properties would need to be agreed upon for Wikidata, for example specifying if the regulation is positive or negative for the "regulates" property. There is also a need for further options similar to the "antisense inhibitor of" property and for other interaction types such as "promoter of", "cis-acting", or "trans-acting". Fortunately, there is currently an ongoing discussion on how to extend the pool of Wikidata properties for biological data[16].

Additionally, it has to be considered that currently only a fraction of the scholarly literature is covered by Wikidata[17] [41] (estimated 22.5 Million of 389 Million articles listed in Google Scholar) and that our suggested approach might have technical limitation in terms of scalability considering the current technical setup of Wikidata. In case this approach would be applied in large scale to numerous fields and millions of articles, the capacity and performance of Wikidata might not be sufficient. As solution for this would be domain specific Wikibase instances. This approach could also be helpful to reduce the friction of introducing new item classes and properties which is usually time-consuming as a consensus has to be found with all community members participating in the discussion and the overal state of the ontology be considered. Still, for the time being, a solution built on Wikidata lowers the access barriers to explore this approach by other research communities.

The here presented approach and its implementation stands at the intersection of Wikidata and biological scholars' communities to bridge the gap between them, and to bring the advantages of Wikidata to the forefront, promoting its use among biological scholars. Based on the showcase given by this, the prospective users would be encouraged to contribute, curate, and save their findings as a result of highlighting Wikidata as a readily achievable centralized solution for the preservation of their discoveries, as well a platform for promoting and citing their work. The introduced statement-level citation in *InteractOA* as a feature that could potentially simplify the process of locating and comparing information due to the amassed collective knowledge. Despite the challenges, we are convinced that the approach showcased in this project can be applied to numerous other communities, even those that require more complex data models. We hope the example provided here will motivate other research communities to make knowledge and its sources available in a more structured fashion.

[16] https://www.wikidata.org/wiki/Wikidata:WikiProject_Molecular_biology/Properties
[17] https://www.wikidata.org/wiki/Wikidata:Statistics

4. Acknowledgments

This work was supported by the Bundesministerium für Bildung und Forschung (BMBF, grant number 16OA031Z).

References

[1] P. Hitzler, A review of the semantic web field, *Communications of the ACM* **64**(2) (2021), 76–83. doi:10.1145/3397512.

[2] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G.D. Melo, C. Gutierrez, S. Kirrane, J.E.L. Gayo, R. Navigli, S. Neumaier, A.-C.N. Ngomo, A. Polleres, S.M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab and A. Zimmermann, Knowledge Graphs, *ACM Computing Surveys* **54**(4) (2021), 1–37. doi:10.1145/3447772.

[3] M. Kejriwal, Knowledge Graphs: A Practical Review of the Research Landscape, *Information* **13**(4) (2022), 161. doi:10.3390/info13040161.

[4] D. Brickley, R.V. Guha and B. McBride, RDF schema 1.1. W3C recommendation, *World Wide Web Consortium* **2** (2014).

[5] D. Vrandečić, Wikidata: a new platform for collaborative data collection, in: *Proceedings of the 21st International Conference on World Wide Web*, ACM, 2012. doi:10.1145/2187980.2188242.

[6] D. Vrandečić and M. Krötzsch, Wikidata: a free collaborative knowledgebase, *Communications of the ACM* **57**(10) (2014), 78–85. doi:10.1145/2629489.

[7] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez and D. Vrandečić, Introducing Wikidata to the Linked Data Web, in: *The Semantic Web – ISWC 2014*, Springer International Publishing, 2014, pp. 50–65. doi:10.1007/978-3-319-11964-9_4.

[8] K. Shenoy, F. Ilievski, D. Garijo, D. Schwabe and P. Szekely, A study of the quality of Wikidata, *Journal of Web Semantics* **72** (2022), 100679. doi:10.1016/j.websem.2021.100679.

[9] T.V. Veen, Wikidata - From "an" Identifier to "the" Identifier, *Information Technology and Libraries* **38**(2) (2019), 72–81. doi:10.6017/ital.v38i2.10886.

[10] M. Mora-Cantallops, S. Sánchez-Alonso and E. García-Barriocanal, A systematic literature review on Wikidata, *Data Technologies and Applications* **53**(3) (2019), 250–268. doi:10.1108/dta-12-2018-0110.

[11] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data* **3**(1) (2016). doi:10.1038/sdata.2016.18.

[12] D. Shotton, K. Portwin, G. Klyne and A. Miles, Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article, *PLoS Computational Biology* **5**(4) (2009), e1000361. doi:10.1371/journal.pcbi.1000361.

[13] A. Garcia, F. Lopez, L. Garcia, O. Giraldo, V. Bucheli and M. Dumontier, Biotea: semantics for Pubmed Central, *PeerJ* **6** (2018), e4201. doi:10.7717/peerj.4201.

[14] S. Auer, A. Oelen, M. Haris, M. Stocker, J. D'Souza, K.E. Farfar, L. Vogt, M. Prinz, V. Wiens and M.Y. Jaradeh, Improving Access to Scientific Literature with Knowledge Graphs, *Bibliothek Forschung und Praxis* **44**(3) (2020), 516–529. doi:10.1515/bfp-2020-2042.

[15] T.U. Consortium, UniProt: the universal protein knowledgebase, *Nucleic Acids Research* **46**(5) (2018), 2699–2699. doi:10.1093/nar/gky092.

[16] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler and H. Hermjakob, IntAct–open source resource for molecular interaction data, *Nucleic Acids Research* **35**(Database) (2007), D561–D565. doi:10.1093/nar/gkl958.

[17] L. Li, D. Huang, M.K. Cheung, W. Nong, Q. Huang and H.S. Kwan, BSRD: a repository for bacterial small regulatory RNA, *Nucleic Acids Research* **41**(D1) (2012), D233–D238. doi:10.1093/nar/gks1264.

[18] J. Pischimarov, C. Kuenne, A. Billion, J. Hemberger, F. Cemič, T. Chakraborty and T. Hain, sRNAdb: A small non-coding RNA database for gram-positive bacteria, *BMC Genomics* **13**(1) (2012). doi:10.1186/1471-2164-13-384.

[19] M.M. Hoffman, AANT: the Amino Acid-Nucleotide Interaction Database, *Nucleic Acids Research* **32**(90001) (2004), 174D–181. doi:10.1093/nar/gkh128.

[20] J.E. Hill, S.L. Penny, K.G. Crowell, S.H. Goh and S.M. Hemmingsen, cpnDB: A Chaperonin Sequence Database, *Genome Research* **14**(8) (2004), 1669–1675. doi:10.1101/gr.2649204.

[21] Á. Ősz, L.S. Pongor, D. Szirmai and B. Győrffy, A snapshot of 3649 Web-based services published between 1994 and 2017 shows a decrease in availability after 2 years, *Briefings in Bioinformatics* **20**(3) (2017), 1004–1010. doi:10.1093/bib/bbx159.

[22] J.D. Wren, C. Georgescu, C.B. Giles and J. Hennessey, Use it or lose it: citations predict the continued online availability of published bioinformatics resources, *Nucleic Acids Research* **45**(7) (2017), 3627–3633. doi:10.1093/nar/gkx182.

[23] S. Mangul, T. Mosqueiro, R.J. Abdill, D. Duong, K. Mitchell, V. Sarwal, B. Hill, J. Brito, R.J. Littman, B. Statz, A.K.-M. Lam, G. Dayama, L. Grieneisen, L.S. Martin, J. Flint, E. Eskin and R. Blekhman, Challenges and recommendations to improve the installability and archival stability of omics computational tools, *PLOS Biology* **17**(6) (2019), e3000333. doi:10.1371/journal.pbio.3000333.

[24] T.K. Attwood, B. Agit and L.B.M. Ellis, Longevity of Biological Databases, *EMBnet.journal* **21**(0) (2015). doi:10.14806/ej.21.0.803.

[25] A. Waagmeester, G. Stupp, S. Burgstaller-Muehlbacher, B.M. Good, M. Griffith, O.L. Griffith, K. Hanspers, H. Hermjakob, T.S. Hudson, K. Hybiske, S.M. Keating, M. Manske, M. Mayers, D. Mietchen, E. Mitraka, A.R. Pico, T. Putman, A. Riutta, N. Queralt-Rosinach, L.M. Schriml, T. Shafee, D. Slenter, R. Stephan, K. Thornton, G. Tsueng, R. Tu, S. Ul-Hasan, E. Willighagen, C. Wu and A.I. Su, Wikidata as a knowledge graph for the life sciences, *eLife* **9** (2020). doi:10.7554/elife.52614.

[26] A. Waagmeester, E.L. Willighagen, A.I. Su, M. Kutmon, J.E.L. Gayo, D. Fernández-Álvarez, Q. Groom, P.J. Schaap, L.M. Verhagen and J.J. Koehorst, A protocol for adding knowledge to Wikidata: aligning resources on human coronaviruses, *BMC Biology* **19**(1) (2021). doi:10.1186/s12915-020-00940-y.

[27] S. Bonner, I.P. Barrett, C. Ye, R. Swiers, O. Engkvist, A. Bender, C.T. Hoyt and W.L. Hamilton, A review of biomedical datasets relating to drug discovery: a knowledge graph perspective, *Briefings in Bioinformatics* **23**(6) (2022), bbac404. doi:10.1093/bib/bbac404.

[28] J. Youn, N. Rai and I. Tagkopoulos, Knowledge integration and decision support for accelerated discovery of antibiotic resistance genes, *Nature Communications* **13**(1) (2022). doi:10.1038/s41467-022-29993-z.

[29] A. Santos, A.R. Colaço, A.B. Nielsen, L. Niu, M. Strauss, P.E. Geyer, F. Coscia, N.J.W. Albrechtsen, F. Mundt, L.J. Jensen and M. Mann, A knowledge graph to interpret clinical proteomics data, *Nature Biotechnology* **40**(5) (2022), 692–702. doi:10.1038/s41587-021-01145-6.

[30] T.E. Putman, S. Lelong, S. Burgstaller-Muehlbacher, A. Waagmeester, C. Diesh, N. Dunn, M. Munoz-Torres, G.S. Stupp, C. Wu, A.I. Su and B.M. Good, WikiGenomes: an open web application for community consumption and curation of gene annotation data in Wikidata, *Database* **2017** (2017). doi:10.1093/database/bax025.

[31] T. Putman, K. Hybiske, D. Jow, C. Afrasiabi, S. Lelong, M.A. Cano, G.S. Stupp, A. Waagmeester, B.M. Good, C. Wu and A.I. Su, ChlamBase: a curated model organism database for the Chlamydia research community, *Database* **2019** (2019). doi:10.1093/database/baz041.

[32] M. Martens, A. Ammar, A. Riutta, A. Waagmeester, D.N. Slenter, K. Hanspers, R.A. Miller, D. Digles, E.N. Lopes, F. Ehrhart, L.J. Dupuis, L.A. Winckers, S.L. Coort, E.L. Willighagen, C.T. Evelo, A.R. Pico and M. Kutmon, WikiPathways: connecting communities, *Nucleic Acids Research* **49**(D1) (2020), D613–D621. doi:10.1093/nar/gkaa1024.

[33] S. Burgstaller-Muehlbacher, A. Waagmeester, E. Mitraka, J. Turner, T. Putman, J. Leong, C. Naik, P. Pavlidis, L. Schriml, B.M. Good and A.I. Su, Wikidata as a semantic framework for the Gene Wiki initiative, *Database* **2016** (2016), baw015. doi:10.1093/database/baw015.

[34] E.G.H. Wagner and P. Romby, Small RNAs in Bacteria and Archaea, in: *Advances in Genetics*, Elsevier, 2015, pp. 133–208. doi:10.1016/bs.adgen.2015.05.001.

[35] G. Storz, S. Altuvia and K.M. Wassarman, AN ABUNDANCE OF RNA REGULATORS, *Annual Review of Biochemistry* **74**(1) (2005), 199–217. doi:10.1146/annurev.biochem.74.082803.133136.

[36] M.G. Jørgensen, J.S. Pettersen and B.H. Kallipolitis, sRNA-mediated control in bacteria: An increasing diversity of regulatory mechanisms, *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1863**(5) (2020), 194504. doi:10.1016/j.bbagrm.2020.194504.

[37] A. Huerta, RegulonDB: a database on transcriptional regulation in Escherichia coli, *Nucleic Acids Research* **26**(1) (1998), 55–59. doi:10.1093/nar/26.1.55.

[38] N.A. O'Leary, M.W. Wright, J.R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C.M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V.S. Joardar, V.K. Kodali, W. Li, D. Maglott, P. Masterson, K.M. McGarvey, M.R. Murphy, K. O'Neill, S. Pujar, S.H. Rangwala, D. Rausch, L.D. Riddick, C. Schoch, A. Shkeda, S.S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R.E. Tully, A.R. Vatsan, C. Wallin, D. Webb, W. Wu, M.J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T.D. Murphy and K.D. Pruitt, Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation, *Nucleic Acids Research* **44**(D1) (2015), D733–D745. doi:10.1093/nar/gkv1189.

[39] M. Sassi, Y. Augagneur, T. Mauro, L. Ivain, S. Chabelskaya, M. Hallier, O. Sallou and B. Felden, SRD: a Staphylococcus regulatory RNA database, *RNA* **21**(5) (2015), 1005–1017. doi:10.1261/rna.049346.114.

[40] X. Ju, X. Fang, Y. Xiao, B. Li, R. Shi, C. Wei and C. You, Small RNA GcvB Regulates Oxidative Stress Response of Escherichia coli, *Antioxidants* **10**(11) (2021), 1774. doi:10.3390/antiox10111774.

[41] M. Gusenbauer, Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases, *Scientometrics* **118**(1) (2018), 177–214. doi:10.1007/s11192-018-2958-5.