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. Introduction

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 [2]. Although several attempts have been made to define a knowledge graph, there is still no single, agreed-upon definition of what it entails [3]. As the name implies, a knowledge graph, also known as a semantic network, is a means of storing knowledge in a graph-based model. It is a structured data model that represents a

*Corresponding author. E-mail: foerstner@zbmed.de.

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[4].

The Resource Description Framework (RDF) [5] is one model for implementing knowledge graph storage. The RDF specification [6] 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 [7]) and ease of editability [8, 9]. It is based on the Wikibase software, which allows RDF exports [10], 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 [11]. It is regarded as a key source of identifiers [12] and has tremendous potential which remains largely untapped [13].

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 [14], most 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. Although scholarly articles are now available in digital formats (e.g. in HTML, PDF and XML), they continue to lack semantic enrichment. Attempts have been made to address this [15, 16], but these have been largely ignored by the publishing industry and scholarly community. Recent projects such as the Open Research Knowledge Graph (ORKG) [17] have tried to overcome this hurdle by offering a semantic database to represent the research findings of publications in a separate machine-actionable form.

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 life 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 [18] and sRNAdb [19], which are databases for bacterial small RNAs, as well as AANT [20], a database for amino acid-nucleotide interactions, and cpnDB [21], a database for bacterial Chaperonin sequences. Several studies have examined this issue. Their results show that more than 30% of bioinformatical web services published over the last 23 years are currently unavailable, with lack of maintenance being the primary cause of this decay [22–25]. 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 mostly 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 [26–28]. 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 [29].

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) [30]. 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 that prioritize data preservation. Further examples of existing solutions include WikiGenomes [31], an openly editable knowledge graph for genomic annotations that is geared towards the molecular biology community, ChlamBase [32], which is a central access point for genomic and proteomic information specifically for the Chlamydia research community, and WikiPathways [33], 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 [34]

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) [35], 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 [36, 37] 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 [38]. Without a uniform format for reporting such information, it is difficult to gain a comprehensive understanding of these regulatory networks.

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

The semantic nature of Wikidata, combined with its openness and preservation capabilities, can help to overcome many challenges in the management of data, information, and knowledge in gene regulation research. This method

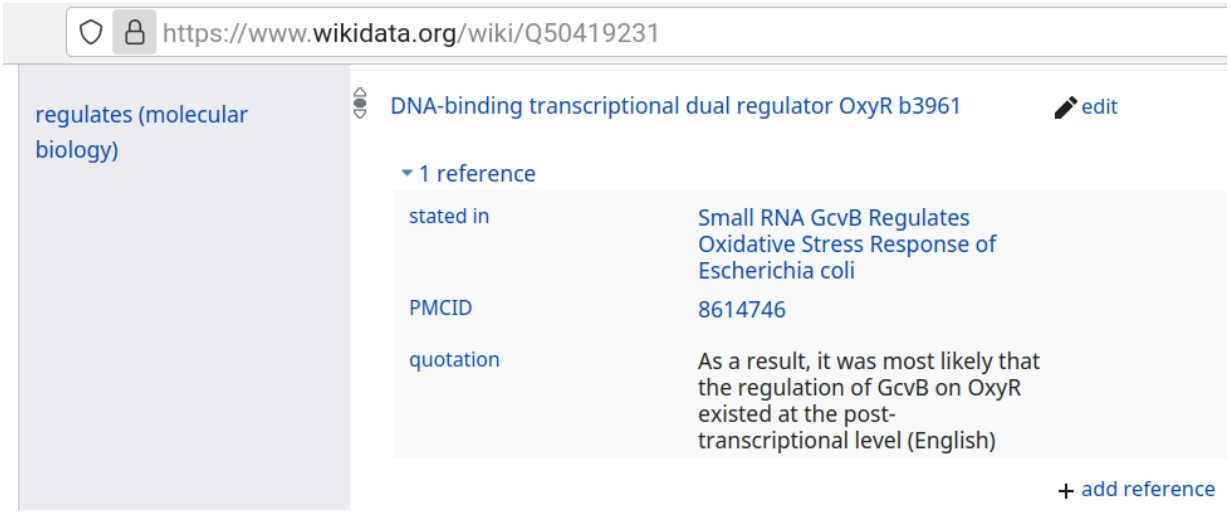


Fig. 1. The sRNA encoded by gcvB (Q50419231) as an example of a citation at the statement level.

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.

For this study, we modelled small RNA interactions in Wikidata and built an application called *InteractOA* on top of it. The application features a user-friendly graphical interface that allows users to query Wikidata for a chosen organism and display the network of interactions or citation information at the statement level for all referenced interactions.

2. Results

2.1. Modelling small RNA interaction data in Wikidata

The modelling of small RNAs and their interactions was conducted by translating gene annotations from GFF (General Feature Format) files into Wikidata items and then linking them via 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).

The first step was to obtain genomic annotation data for these organisms from the NCBI RefSeq repository [39]. 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 sciences [40]. See Table 1 for a list of items and properties used here.

Next, the Wikidata items representing RNAs were linked to model the interactions. To do this, interaction data for numerous small RNAs was obtained from numerous research articles, RegulonDB [38], and other sources. The

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

compiled information was submitted 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

Interactions and References

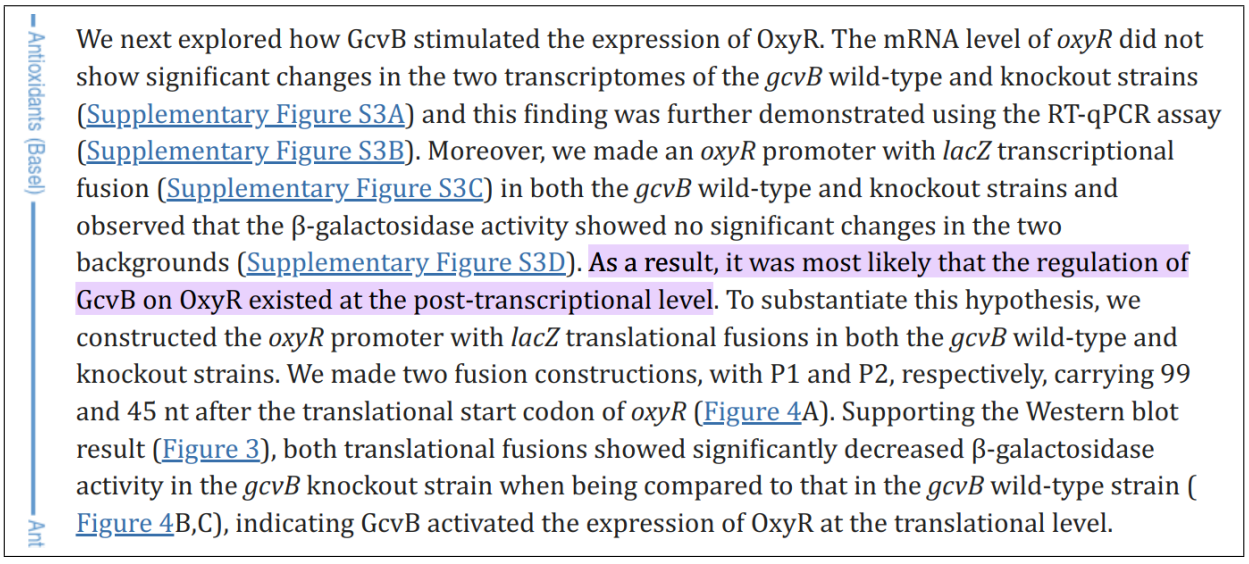
Referenced items: Escherichia coli str. K-12 substr. MG1655

Search: b4443

#	sRNA	sRNA synonyms	Type of Regulation	Target Gene	Quote	Quote from	Wikidata
5	sRNA encoded by gcvB	gcvB, b4443, ECK2804, IS145, JWR0247, psrA11	antisense inhibitor of	serine/threonine:Na(+) symporter b3089	We compared RNA isolated from a wild-type strain and a gcvB deletion strain grown to mid-log phase in Luria-Bertani (LB) broth by microarray analysis to identify any additional regulatory targets of GcvB. One potential target identified by microarray analysis was sstT, which encodes a Na+/l-serine and l-threonine transport protein.	PMC	
11	sRNA encoded by gcvB	gcvB, b4443, ECK2804, IS145, JWR0247, psrA11	antisense inhibitor of	branched chain amino acid/phenylalanine ABC transporter periplasmic binding protein b3460	among the top candidate targets for the sRNA GcvB were mRNAs gltI, livJ, livK, ytfT, aroP and argT, all genes encoding periplasmic transport proteins.	PMC	
12	sRNA encoded by gcvB	gcvB, b4443, ECK2804, IS145, JWR0247, psrA11	regulates (molecular biology)	DNA-binding transcriptional dual regulator OxyR b3961	As a result, it was most likely that the regulation of GcvB on OxyR existed at the post-transcriptional level	PMC	
14	sRNA encoded by gcvB	gcvB, b4443, ECK2804, IS145, JWR0247, psrA11	antisense inhibitor of	oligopeptide ABC transporter periplasmic binding protein b1243	The specific repression of dppA::gfp and oppA::gfp by pPLgcvB was evident from strongly reduced colony fluorescence of these strains on agar plates (Fig. 2B), which established that GcvB regulates dppA and oppA in the 5′ mRNA region.	PMC	

Showing 1 to 12 of 12 entries (filtered from 53 total entries)

Fig. 2. Screenshot of all available statement level citations for a selected organism, which can be filtered with keywords



Antioxidants (Basel)

We next explored how GcvB stimulated the expression of OxyR. The mRNA level of *oxyR* did not show significant changes in the two transcriptomes of the *gcvB* wild-type and knockout strains (Supplementary Figure S3A) and this finding was further demonstrated using the RT-qPCR assay (Supplementary Figure S3B). Moreover, we made an *oxyR* promoter with *lacZ* transcriptional fusion (Supplementary Figure S3C) in both the *gcvB* wild-type and knockout strains and observed that the β-galactosidase activity showed no significant changes in the two backgrounds (Supplementary Figure S3D). As a result, it was most likely that the regulation of GcvB on OxyR existed at the post-transcriptional level. To substantiate this hypothesis, we constructed the *oxyR* promoter with *lacZ* translational fusions in both the *gcvB* wild-type and knockout strains. We made two fusion constructions, with P1 and P2, respectively, carrying 99 and 45 nt after the translational start codon of *oxyR* (Figure 4A). Supporting the Western blot result (Figure 3), both translational fusions showed significantly decreased β-galactosidase activity in the *gcvB* knockout strain when being compared to that in the *gcvB* wild-type strain (Figure 4B,C), indicating GcvB activated the expression of OxyR at the translational level.

Fig. 3. The highlighted text in an article (Ju et al.) is a statement level citation example for a claim about small RNA interaction

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 1. If the type of interaction has no corresponding property in Wikidata, the tool falls back to the more generic property (P128 regulates).

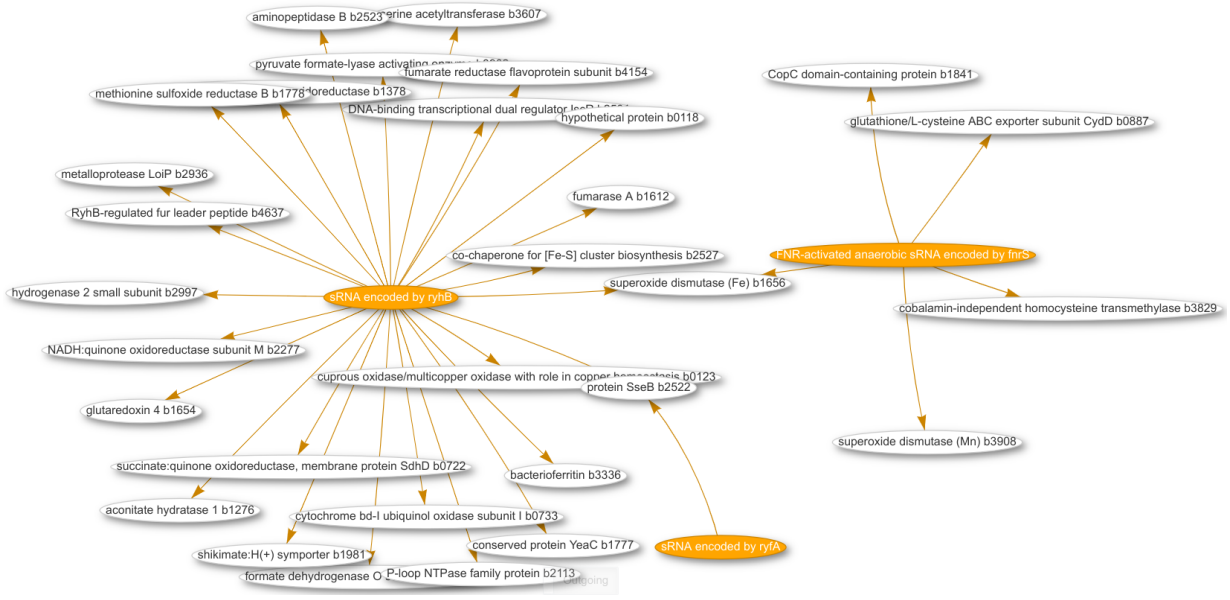


Fig. 4. Screenshot of a network visualization for 3 small RNAs (orange labels) and their interactions with the gene targets (white labels)

Both the Python tools used to import the data were built upon Wikidata Integrator [41] and pywikibot to facilitate interaction with the Wikidata API during querying and importing. They are available on GitHub and at (https://github.com/foerstner-lab/GFF_to_Wikidata_importer) and (https://github.com/foerstner-lab/sRNA_Interactions_to_Wikidata_Importer), respectively. Besides this they are archived in Zenodo at https://zenodo.org/record/7638542 and https://zenodo.org/record/7638552, respectively.

2.2. *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 [42]. 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 1, 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" [43] (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.

2.3. *InteractOA as a front end to relevant Wikidata items*

The web front end *InteractOA* was developed to facilitate user-friendly interaction with the data stored in Wikidata as described above. *InteractOA* is implemented in Python and the Flask web framework. Its code is available on GitHub (https://github.com/foerstner-lab/InteractOA) and its releases are archived at Zenodo https://doi.org/10.5281/zenodo.7638558. Wikidata features a query service [44] 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 4 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 4 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 that queries all the extracted statements in Wikidata for a strain selected by the user and presents them in a searchable table. 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 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 3 for an example).

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 Wikibase knowledge base. Wikibase knowledge graphs provide 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 paper, 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 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 general usefulness of this approach, 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). 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.

Another issue is the question of which item classes and properties are made available by Wikidata. The implemented model of *InteractOA* applies Wikidata's currently available properties to the topic of small RNAs and their interaction partners. These item classes and properties 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 [45].

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.

4. Acknowledgments

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

References

[1] Google, Introducing knowledge graph, 2012, https://www.blog.google/products/search/introducing-knowledge-graph-things-not/.
[2] P. Hitzler, A review of the semantic web field, *Communications of the ACM* **64**(2) (2021), 76–83. doi:10.1145/3397512.
[3] 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.
[4] M. Kejriwal, Knowledge Graphs: A Practical Review of the Research Landscape, *Information* **13**(4) (2022), 161. doi:10.3390/info13040161.
[5] D. Brickley, R.V. Guha and B. McBride, RDF schema 1.1. W3C recommendation, *World Wide Web Consortium* **2** (2014).
[6] W3C, RDF specifications and concepts, https://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.
[7] Creative Commons, CC0 1.0 - Public Domain Dedication, https://creativecommons.org/publicdomain/zero/1.0/.
[8] 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.
[9] D. Vrandečić and M. Krötzsch, Wikidata: a free collaborative knowledgebase, *Communications of the ACM* **57**(10) (2014), 78–85. doi:10.1145/2629489.
[10] 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.
[11] 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.
[12] 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.
[13] 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.
[14] 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.
[15] 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.
[16] 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.
[17] 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.
[18] 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.
[19] 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.

[20] M.M. Hoffman, AANT: the Amino Acid-Nucleotide Interaction Database, *Nucleic Acids Research* **32**(90001) (2004), 174D–181. doi:10.1093/nar/gkh128.
[21] 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.
[22] 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.
[23] 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.
[24] 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.
[25] 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.
[26] 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.
[27] 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.
[28] 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.
[29] 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.
[30] 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.
[31] 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.
[32] 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.
[33] 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.
[34] 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.
[35] 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.
[36] 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.
[37] 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. https://linkinghub.elsevier.com/retrieve/pii/S1874939919302688.
[38] 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.
[39] 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.
[40] Wikimedia foundation, List of properties/natural science, 2019, https://www.wikidata.org/wiki/Wikidata:List_of_properties/natural_science#Wikidata_property_related_to_biology.
[41] Andrawaag, Sebotic, G. Stupp, Myst, C.T. Hoyt, T. Arrow, Konstin, Jleong-Ndn, D. Fdez, D. Mietchen, D. Raison, A. Su, R. Shigapov, T. Putman, A. Mignone, C. Wilson, D. Vrandečić, D. Espenschied, J. Tu, , Jasper, J. De Dauw, J. Samuel, M. Schubotz, N. Finberg, I. Leturia and Paul-Olivier Dehaye, SuLab/WikidataIntegrator: Add precision to the publication helper function, Zenodo, 2022. doi:10.5281/ZENODO.7041852. https://zenodo.org/record/7041852.
[42] National Institutes of Health, PubMed Central, https://pubmed.ncbi.nlm.nih.gov/.
[43] 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.
[44] Wikimedia foundation, Wikidata query service, 2017, https://query.wikidata.org/.

[45] Wikimedia foundation, Proposed properties for biological data, 2019, https://www.wikidata.org/wiki/Wikidata:WikiProject_Molecular_biology/Properties.