Maastricht University

Biological pathway abstractions

Citation for published version (APA):

Waagmeester, A. S. (2024). *Biological pathway abstractions: from two-dimensional drawings to multidimensional linked data*. [Doctoral Thesis, Maastricht University]. Maastricht University. https://doi.org/10.26481/dis.20240116aw

Document status and date:
Published: 01/01/2024

DOI:
10.26481/dis.20240116aw

Document Version:
Publisher's PDF, also known as Version of record

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy
If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Download date: 16 May. 2024

BIOLOGICAL PATHWAY ABSTRACTIONS: FROM TWO-DIMENSIONAL DRAWINGS TO MULTIDIMENSIONAL LINKED DATA

Andra Sachinder Waagmeester

BIOLOGICAL PATHWAY ABSTRACTIONS: FROM TWO-DIMENSIONAL DRAWINGS TO MULTIDIMENSIONAL LINKED DATA

Dissertation

To obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus, Prof. Dr. P. Habibović,
in accordance with the decision of the Board of Deans,
to be defended in public
on 2024-01-16, at 16:00

by

Andra Sachinder Waagmeester

Promotor
Prof. Dr. Chris T. Evelo

Copromotor
Dr. Susan L.M. Coort
Dr. Egon L. Willighagen

Assessment Committee
Prof. Dr. Michel J. Dumontier, (chair)
Prof. Dr. Ir. Andre L.A.J. Dekker
Dr. Rachel Cavill
Dr. Marco Roos, Leiden University Medical Center
Prof. Dr. Carole Goble, University of Manchester, United Kingdom

© Andra Waagmeester, Maastricht 2024.
All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior written permission of the author.

Cover Andra Waagmeester, 2024
Production Waagmeester 2024
ISBN 978-94-6473-361-7

To Nico Waagmeester

Contents

1

Introduction

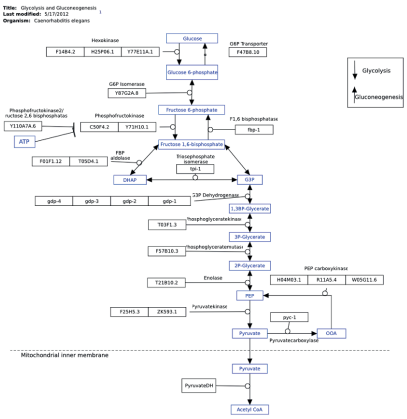
Every living organism is constantly interacting with its environment. To thrive and stay healthy the body needs to respond to these external inputs to maintain a healthy physical state. A state that needs to retain itself between a set of strict boundaries. For example, the human body has to keep a temperature between 36.3 and 37.3 degrees Celsius. Any deviation from these parameters puts the body in direct danger. The body responds to these deviations by sending signals throughout the body for an orchestrated response to work towards a return to the norm. An increase in body temperature above the viable temperature ranges could be such a response in itself.

This orchestration within the body happens at every level of the body, both at the macroscopic and microscopic levels. For this, a myriad of signal responses exist. There are the nervous system and the hormonal system, but also various chemical reactions to different metabolites, pathogens, and other chemical compounds that can present a danger from both inside and outside the body. For us to understand this complex chain of interacting processes we need to understand the dynamics of this concert. Molecular biologists use biological pathways as instruments to get a better understanding of the complexity behind the processes that keep us healthy or to understand where these processes get disrupted and cause disease. These pathways come in different forms. There are metabolic pathways that describe the exchange of energy from food to ATP and its release of energy when needed. Together with the metabolic pathways, there are roughly three main categories of biological pathways. These are

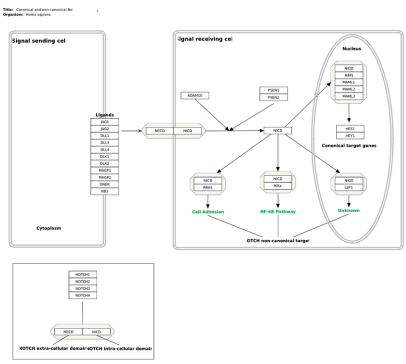
1. metabolic pathways (Figure 1.1 A) [1]
2. signal-transduction pathways (Figure 1.1 B) [2]
3. gene-regulatory pathways (Figure 1.1 C) [3]

More detailed classifications do exist [4], where these three are further classified in e.g. protein-protein interaction pathways, protein-compound interactions, and genetic interaction networks.

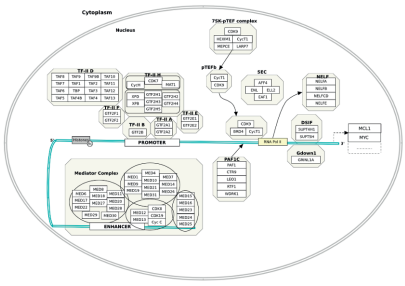
Biological pathways are thus instrumental in molecular biology as abstractions of what we already know about what happens in the cell and as such are driving the continuously increasing understanding of our knowledge of



A. metabolic pathway



B. signaling pathway



C. gene-regulatory pathways

Figure 1.1: The three main pathway categories

cellular processes. A pathway is usually depicted as a sequence of interactions between molecules in the cell leading to cell products or a change of state in the applicable cells, surrounding tissue, organs, or even the state of the species.

Initially, these pathways were primarily depicted as illustrations or graphical diagrams. Until recently, The Biochemical Pathways Wall Chart (Figure 1.2) was a common sighting on walls of many biomedical and biochemical departments. Due to the increased complexity of each revision, updates on this wall chart were released in editions published as books [1]. Currently, these pathways are also maintained online[1].

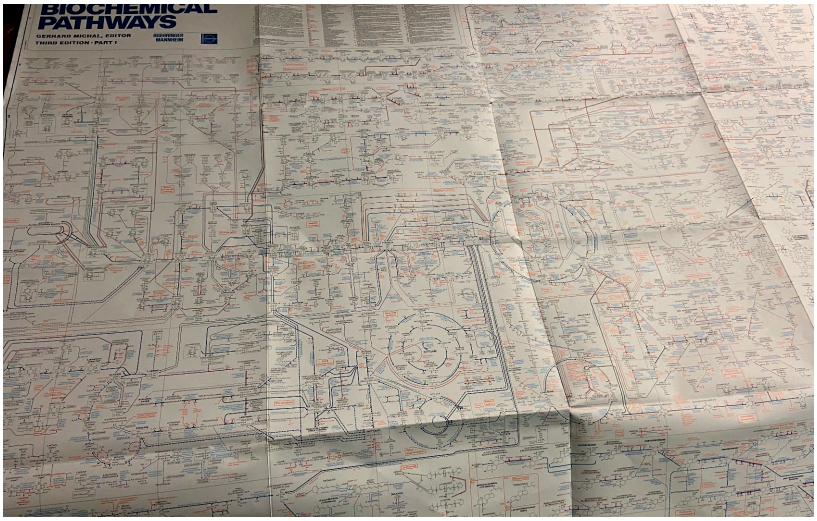


Figure 1.2: Biochemical Pathways Wall Chart by Gerhard Michal initially published by Boehringer Ingelheim and now continued by Roche.

Source: reddit (`https://www.reddit.com/r/chemistry/comments/faux7y/finally_got_a_copy_of_roche_biochemical_pathways/`)

The change in the medium on which these pathways were published also diverted the application of pathways in biomedical research. As graphical

[1] `http://biochemical-pathways.com/#/map/1`

diagrams pathways were - and remain - a platform for discussions among peers, or as illustrations in scientific publications and presentations. However, the application of pathways also diverted in different directions. While mathematical methods or machine learning can be helpful with structuring large datasets [5], getting an overview can be challenging. Here it helps to be able to create a diagram or road map. Frameworks are needed to align the existing knowledge in various biological databases with results from research. Pathways gained value in this role as research instrument by aggregating the knowledge captured in biological databases [4, 6], using declarative languages [7] such as BioPAX [8], SBML [9] and SBGN [10]. By expressing them in declarative languages pathways become machine-actionable [11] and when these declarative languages are used to also store the respective database identifiers of data sets from various research studies it becomes straightforward to align the results with the above-mentioned knowledge bases. Basically, pathways are then proxies between biological databases and study results.

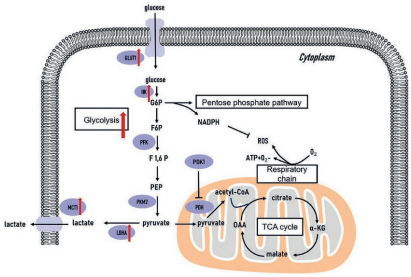
When the different parts of a pathway are annotated with external identifiers used in study results it is possible to project these values onto pathways creating visualizations, where colours can be used to show increased or decreased activity. For example, if a data set contains measurements of a studied gene expression and the data set contains identifiers to the genes under scrutiny, these values can be visualized by applying colour coding on a pathway. At different thresholds, different colours are used showing the magnitude of the gene expression at a certain location on the pathway.

For the first role - as illustrations in peer-to-peer discussions a graphical illustration suffice. However, to be machine-actionable, the pathways need to be in a format that computer programs can automatically process. Doing so on images from the literature is challenging, if at all possible. However, if the pathways are first stored in a set of instructions that drive the rendering of these pathways and the instructions also contain identifiers to various databases used in study results it is possible to project these study results onto those pathways. Figure 1.3 shows the different versions of the same pathway. Figure 1.3.1 is the original pathway as it was published in the scientific literature [12]. Figure 1.3.2. It Is the same pathway but described on WikiPath-

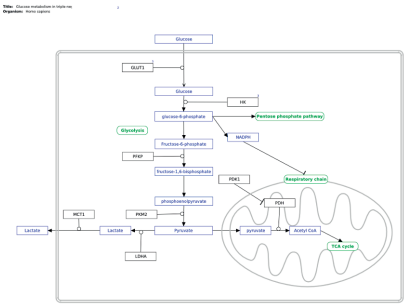
ways using a file format with instructions to render the pathway, making it a machine-actionable pathway. Figure 1.3.3 depicts (part of) the machine-actionable set of instructions to render the diagram. Each element of these pathways parts should contain mappings or citations to external databases or the literature. When the sample data comes with annotations using the same database citations, both the pathways and the sample data can be linked automatically. As such, Pathways become proxies between observational data, biological databases, and the scientific literature. Researchers can project their results on these diagrams providing a visual oversight of the process under scrutiny and pointers to the relevant literature. This has made pathways a valuable research instrument, especially in the context where molecular biology transformed from studying single genes, towards multi-omics studies where not single genes, but large sets of genes work together in orchestration in a certain context. This topic is discussed in more detail in Chapter 2 of this thesis.

Identifier mapping

While unified identifier schemes do exist and will be discussed in greater detail below, most biological databases do not have unified underlying identifier schemes. This means that similar concepts can have multiple identifiers. Different services and platforms exist to map between identifiers and identifier schemes. Notable examples are DICT [14], CRONOS [15], MatchMiner [16], AliasServer [17], PICR [18], Synergizer [19] and Ensembl BioMart [20]. BridgeDb [21, 22] builds on these identifier mappings to provide identity resolution services to be used in pathway tools to link between various identifiers used in pathways and data from study results. This allows the use of pathway analysis even if the identifier schemes are different between biological databases and study results.



An example pathway diagram as published in the scientific literature. [13]



The same pathway in WikiPathways where the diagram is built from the machine-actionable code below

```
<?xml version="1.0" encoding="UTF-8"?>
<Pathway xmlns="http://pathvisio.org/GPML/2013a" Name="Glucose metabolism in
triple-negative breast cancer cells" Version="20220316" Organism="Homo sapiens">
  <Comment Source="WikiPathways-description">Glucose metabolism in
  triple-negative breast cancer cells. The glycolytic pathway is significantly
  upregulated in triple-negative breast tumors. The genes coding the key
  glycolytic enzymes are overexpressed in triple-negative breast tumors. TCA,
  tricarboxylic acid cycle; G6P, glucose-6-phosphate; F6P, fructose-6-phosphate;
  F1,6P, fructose-1,6-bisphosphate; PEP, phosphoenolpyruvate; OAA, oxaloacetate;
  α-KG, α-ketoglutarate; GLUT, glucose transporter; HK, hexokinase; PFK,
  phosphofructokinase; PKM2, pyruvate kinase isozyme type 2; LDHA, lactate
  dehydrogenase A; MCT1, monocarboxylate transporter 1; PDK1, pyruvate
  dehydrogenase kinase 1; PDH, pyruvate dehydrogenase.</Comment>
  <BiopaxRef>d3e</BiopaxRef>
  <Graphics BoardWidth="1190.0" BoardHeight="896.6666666666666" />
  <DataNode TextLabel="Acetyl CoA" GraphId="a6d66" Type="Metabolite">
    <Graphics CenterX="878.0" CenterY="661.0" Width="90.0" Height="25.0" ZOrder="
    32768" FontSize="12" Valign="Middle" Color="0000ff" />
    <Xref Database="ChEBI" ID="CHEBI:15351" />
  </DataNode>
  <DataNode TextLabel="PDK1" GraphId="a8818" Type="GeneProduct">
    <Graphics CenterX="672.0833333333335" CenterY="512.4166666666667" Width="90.0"
     Height="25.0" ZOrder="32768" FontSize="12" Valign="Middle" />
    <Xref Database="Ensembl" ID="ENSG00000152256" />
  </DataNode>
  <DataNode TextLabel="Pentose phosphate pathway" GraphId="aeed8" Type="Pathway">
    <Graphics CenterX="833.5" CenterY="329.4166666666666" Width="199.0" Height="
    25.0" ZOrder="32768" FontWeight="Bold" FontSize="12" Valign="Middle" ShapeType
    ="RoundedRectangle" Color="14961e" />
    <Xref Database="WikiPathways" ID="WP134" />
```

The same pathway but now rendered in a machine-actionable format (GPML)

Figure 1.3: Three types of pathway representations

1.1 Pathway curation

With pathways being evolved into machine-readable objects that allow to align what we know about a given process with study results, the pathways need to be kept up-to-date with the scientific knowledge. To be constantly accurate, newly gained insights need to be included into perpetuity. This means that scientists need to assess novel knowledge from the scientific literature and the available scientific databases. This involves reading the related literature, finding the appropriate databases and harmonizing or unifying all the knowledge into the native format of the pathway databases being used. Tailoring biological knowledge into formats understandable by both humans and machines is called biocuration [23, 24]. Hence, Pathway curation is every human process that combines pathway knowledge into a format that allows both humans and machines to parse, assess and process them. In Chapter 2 describes how pathway curation is a process where the field of bioinformatics has a central role. Pathway curation tools are usually graphical design tools, where the different paths are drawn as diagrams and where annotations are being added as citations to the literature or the databases. These diagrams are then stored in a file format like for example GPML [25]. Pathway curation is quite implicit in this thesis, but forms the main rationale for the work done. Pathways form a relative novel research method in aligning what we know and what we see in new study results.

So to use biological pathways as a multi-omics research tool, we just need citable literature and easy accessible tools. Bioinformatics is the area of research that delivers and studies both the tooling as well as the actual databases needed to maintain the available knowledge for pathway curation.

Mining the literature

Peer-review literature remains the knowledge backbone of science. However, access to this total sum of knowledge captured in the literature can be challenging. Access guarded by sometimes hefty subscription fees limits the overall access. While, various funding agencies are more and more requiring open science policies [26] that actively promote the removal of access limitation

to the research community, understanding and assessing the related knowledge can still remain challenging due to the sheer size of the scientific literature. The amount of papers being written remains growing at a staggering pace [27, 28]. Current estimates indicate that the total body of the scientific literature doubles every 15 years [28]. Manually assessing this corpus of available scientific literature is challenging at best, maybe impossible. Different methods from the field of computer science exist to make that volume of available knowledge more accessible. Example fields from where the research community can use technologies are information retrieval [29], named-entity recognition [30] and other automatic information organization and retrieval methods [31]. In Chapter 3 of this thesis one such text-mining method is described. Here we automatically extracted terms from a manually curated set of on-topic relevant scientific articles and - in the process - enrich the corpus of topical scientific articles. We were able to identify candidate terms for inclusion in biological pathways.

1.1.1 Biological database and pathway resources

While scientific literature remains the prime source for scientific knowledge, progress in bioinformatics has also led to many different machine-readable and machine-actionable biological databases. "Nucleic Acids Research" (NAR) is a scholarly journal that publishes peer-review scholarly articles that report on the "results of leading-edge research into physical, chemical, biochemical and biological aspects of nucleic acids and proteins involved in nucleic acid metabolism and/or interactions". Since 1993 it publishes an annual, designated issue on the biological databases, where novel databases are announced and accelerating data sets are reported [32]. This series has already led to almost 1700 scholarly articles on biological databases. Pathway curators need tools and methods to efficiently process existing pathway knowledge in both the literature and the database space.

1.2 (Biological) structured data

Not only the sheer size of the volume of the available knowledge captured in the scientific literature and (biological) databases, curators also have to extract this knowledge through a myriad of different data formats [33, 34], webservices [35] or query languages [36]. The literature can be available in different forms. Articles printed on paper come to mind, but with the arrival of the internet some decades ago, came the opportunity to publish the literature in a digital form, allowing more automatic approaches in knowledge extraction.

Large-scale - automatic - knowledge extraction for pathway curation is hindered by its sheer size, varying data formats, a wide range of query language and legal constraints. Yet, corpora containing scientific literature exists. In this thesis an automatic extraction method is described to extract pathway parts from Pubmed. Pubmed is an online public corpus of abstracts from the biomedical literature. Although covering only a fraction of the full paper, the abstract do contain the main points covered. Using text mining, which is an umbrella term for all automatic processes making text from the literature machine-readable, it was possible to extract a set of gene products and metabolites from the literature in such a machine readable format and incorporate that in a metabolic pathway. This is described in Chapter 3.

While successful, text mining comes with limitations. First, there is the already mentioned limited availability, often only abstracts, but the written text also comes with ambiguity. The process that was described in 3 boiled down to breaking down the grammatical structures of text into its individual tokens and then counting the number of occurrences these tokens appeared. Analysis of the token frequency was sufficient to identify potential novel pathway parts. This still required substantial pathway curation, i.e. co-authors being domain experts in the field assessed the extracted pathway terms and decided whether or not a term was fit for inclusion in that pathway. Basically, they re-read the papers from which those terms came. So while text mining did allow them to address more literature than when those curators had to read all papers, also because they were able to access the full papers. These curators acted upon the tokens extracted from the text-mining approach, which were still expressed as natural language. In the next chapter (Chapter 4) the semantic web and linked

data are explored to replace the natural-language tokens with identifiers. This is to exclude the ambiguity that comes with natural language.

1.2.1 Introducing Linked Data/Semantic Web

The World Wide Web is an Internet standard for the linking and share of documents [37]. On the World Wide Web, a web address or Universal Resource Locator (URL) provides the means to locate and render documents and media on the Internet through a web browser. A URL has three distinct parts. The first part indicates the core protocol, being the HyperText Transfer Protocol [38] (http:// or https://), followed by either the Internet Protocol (IP) address or hostname, where the documents are located. Finally, the location of the webserver is given. With the ability to capture media files in HyperText Markup Language (HTML), the web has now matured into a globally spanning collection of documents and media files. However, at the receiving end of each URL, there is a document that still requires a person to assess the meaning of that document. Its end-user is a person.

The Semantic Web [39, 40] extends the World Wide Web (WWW), by following the same internet standards, to link data points. Where in the WWW each document has a URL on the Semantic Web, each concept is given a Unique Resource Identifier (URI) or Internationalized Resource Identifier (IRI), when the web addresses also use non-ASCII codes. Due to its inclusiveness, this document will use the term IRI since a URI is an IRI, but not vice versa. A URI/IRI also does not necessarily point to a given location, as a URL does. The URL specifically points to a physical location or a server. A URI/IRI points to a concept. By using the same URI/IRI resources across different resources, curators agree that the distinct resources capture (meta-) data about the same concept. Many life-science databases were early adapters of the semantic web notable examples are Bio2RDF [41], UniProt [42], the linked Cancer Genome Atlas [43], the Open Phacts initiative [44], the RDF platform of the European Bioinformatics Institute [45], PubChem [46] and ChEMBL [47]. Initiatives like DBPedia [48] and Wikidata [49] allowed linking other databases to align with the semantic web. Either by transforming structured data captured in the (English) Wikipedia (in the case of DBPedia)

or by directly transforming primary sources into a format compatible with the semantic web (in the case of Wikidata). In this thesis, in chapter 4 we describe the steps involved to present pathways captured in WikiPathways on the semantic web. I.e. building the infrastructure to export pathways as RDF. By doing so making the pathways captured in WikiPathways are linked to all the above-mentioned resources and beyond.

1.2.2 FAIR

The semantic web and its underlying triples and URI allow for schema-less storage of data. Knowledge can be easily integrated by simply concatenating different RDF files into a single file. However, by concatenating large collections of data into a single file, it becomes even more necessary to capture the provenance of the individual parts of the linked data cloud that forms the semantic web. I.e. which dataset came from where? In the last decade, the FAIR principle emerged to capture exactly this [50]. Data should be Findable, Accessible, Interoperable, and Reusable to facilitate the optimal application of data. Chapter 5 addresses FAIR. It also introduces Wikidata, which is the linked data resource of Wikipedia.

1.2.3 Wikipedia, WikiPathways, and Wikidata

Wikipedia is an online public crowd-sourced encyclopedia [51]. Started twenty years ago, currently, it expands to 300 languages and is still growing. Where legacy encyclopedias have large editor teams, Wikipedia does not. It is the community at large that maintains online knowledge. Its back end is called MediaWiki, which is a stack of different software components. The open and collaborative nature of Wikipedia inspired many other initiatives [52, 53]. WikiPathways which forms a core component of this thesis was also inspired by the success of Wikipedia [54]. WikiPathways was initially built by embedding a slim version of the PathVisio [55] pathway editor in the media wiki stack [56]. Where the initial MediaWiki version of Wikipedia captured the document in a relational database, Instead of storing documents,

WikiPathways captures the GPML in the MediaWiki database. The slim version of PathVisio is used to render these documents.

Wikidata emerged in 2012 as a linked-data back end of Wikipedia [49]. Due to the many language versions, each maintained by distinct and independent communities, knowledge does not always synchronize equally across different language versions. The same knowledge basically needs to be written and updated independently across all language versions. By providing a structured/linked database, different language Wikipedia communities can choose to base their articles on Wikidata. This means that once the information is updated in Wikidata, it is immediately ready for uptake in the different language versions of Wikipedia.

Initially envisioned as the linked database for Wikipedia, Wikidata now is also used extensively in other use-cases [57–60]. The data in Wikidata is available under an open and public license, which means that anyone can source data from Wikidata.

Wikidata offers biocurators the infrastructure to scale. With both text mining and aligning with the semantic web, substantial bandwidth from the curator remained needed. Wikidata does bring a crowd to the curation process. Not necessarily by helping in a specific curation task, but by curating each other's curation work on a public platform. Chapter 5 described Wikidata and its role in biocuration.

1.2.4 Schema descriptions

Different file formats (e.g. XML, JSON, CSV) exist to store data in a standard and machine-readable structure. RDF builds on those file formats to add a semantic layer to the data. When data is rendered in RDF it allows semantic interoperability between different (RDF) datasets [61]. s Having the knowledge in RDF makes not only the data machine readable, but also the underlying semantics explicit and machine-readable as well. Due to the unified way data is structured in RDF (ie. as triples), all data sets available in RDF can be integrated by simply combining the data sets in a single file/server. However, for humans to make sense of the knowledge, the data needed be assessed by

following the various links. This can be tedious at best but often impossible by the proverbial rabbit hole it quickly gets into. The last chapter of this thesis (Chapter 6) introduces Shape Expressions (ShEx) as a means to document the links and patterns rendered in RDF data [62–64]. ShEx is a formal language to describe data patterns in RDF. Having a formal language it is possible for data owners to describe the structure of their data. It also allows data users and UI developers to express expectations of the data they would like to ingest. Chapter 6 describes in detail a protocol for how a specific part of a knowledge resource relates to other parts of the semantic web.

This thesis addresses primarily how pathway curators can be helped by bioinformatics methods that transform the available knowledge in FAIR and thus maximally machine-readable knowledge, so that computer technology can ease the overall process of pathway curation. This is done by exploring a set of subsequent questions. These are:

1. Can text mining help in processing the ever-increasing body of scientific literature?

2. How do you align a pathway resource with the semantic web to allow straightforward access to other biological resources on the semantic web?

3. How do you align biological resources to Wikidata, which aligns with the semantic web?

4. How do you describe and document the relationships on the semantic web using a formal schema language?

1.3 Thesis outline

Pathway curation is the process that creates, improves, and maintains the pathway abstractions that describe cellular processes. Bioinformatics has a crucial role in the process of pathway curation to keep the growing total body of existing knowledge manageable. The process involved in pathway curation is

described in Chapter 2. Peer review of scientific literature remains the cornerstone of science. This means that most, if not all, scientific knowledge is captured in the literature. However, the volume of the total sum of scientific literature is increasing at an ever-increasing pace. Pathway curation through bioinformatics however requires structured data to manage the existing knowledge by automatically parsing this knowledge. Chapter 3 describes an experiment where text mining is applied to extract structured knowledge from the literature. This chapter extends a pathway on vitamin A metabolism by extracting components in this cellular process from the scientific literature. A seed corpus of well-annotated citations was used and by applying a similarity search, the seed corpus was extended with more citations of vitamin A metabolism. Documents from this extended corpus were homogenized into a single document, and named-entity recognition allowed identifying novel pathway terms. Although both the citations, as well as the results from the text-mining process, were available as structured data, the analysis and decisions still involved major engagement from domain experts, where they had to flesh through the proposed suggestion before being deemed relevant. This process does not scale, because the meaning of the terms identified still required major human assessment. Chapter 4 explores the value of the semantic web which, in contrast to the approaches explored in the previous chapters, explicitly captures the meaning and context of data and knowledge as structured data. Having WikiPathways available on the semantic web allows for more scalable integration of research data and knowledge. In this chapter, we review the process needed to make the knowledge available in WikiPathways explicit. Aligning WikiPathways with the semantic web remains an expensive process. Making the semantics explicit, and selecting the proper context descriptions remains a laborious process. To use the semantic web as a proxy of pathway knowledge that is scattered over the literature and biological databases requires that all is made explicit on the semantic web. For WikiPathways alone the initial development of the semantic web model was expensive, maintenance is a perpetual process. Chapter 5 describes the value of Wikidata as a publicly linked data source where, by crowdsourcing, the cost of maintaining pathway knowledge on the semantics can be less expensive. Finally, Chapter 6 describes a protocol to document choices made in designing the (implicit) schemas when a resource is aligned with the semantic

web.

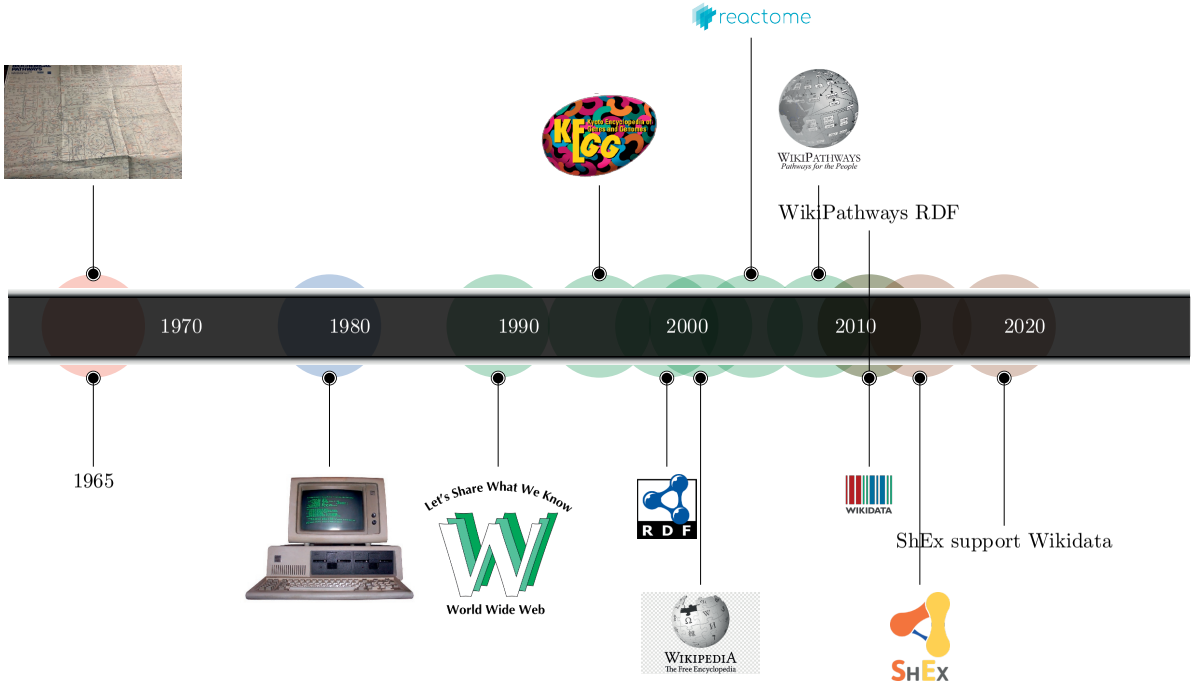


Figure 1.4: Timeline from the paper-based pathway diagrams towards linked and machine-readable pathway models.

References

[1] *Biochemical pathways: an atlas of biochemistry and molecular biology (ed. Michal, Gerhard)*.

[2] Eric Davidson and Michael Levin. "Gene regulatory networks". *Proceedings of the National Academy of Sciences*. 2005. 102 (14): pp. 4935–4935.

[3] Jason A Papin et al. "Reconstruction of cellular signalling networks and analysis of their properties". *Nature reviews Molecular cell biology*. 2005. 6 (2): pp. 99–111.

[4] Gary D. Bader, Michael P. Cary, and Chris Sander. "Pathguide: a Pathway Resource List". *Nucleic Acids Res*. 2006. 34 (Database issue): pp. D504–D506.

[5] David B. Allison et al. "Microarray data analysis: from disarray to consolidation and consensus". *Nat Rev Genet*. 2006. 7 (1): pp. 55–65.

[6] Michael P. Cary, Gary D. Bader, and Chris Sander. "Pathway information for systems biology". *FEBS Lett*. 2005. 579 (8): pp. 1815–1820.

[7] Lena Strömbäck and Patrick Lambrix. "Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX". *Bioinformatics*. 2005. 21 (24): pp. 4401–4407.

[8] Joanne S. Luciano. "PAX of mind for pathway researchers". *Drug Discovery Today*. 2005. 10 (13): pp. 937–942.

[9] Michael Hucka et al. "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models". *Bioinformatics*. 2003. 19 (4): pp. 524–531.

[10] Tobias Czauderna, Christian Klukas, and Falk Schreiber. "Editing, validating and translating of SBGN maps". *Bioinformatics*. 2010. 26 (18): pp. 2340–2341.

[11] Tomasz Miksa et al. "Ten principles for machine-actionable data management plans". *PLOS Computational Biology*. 2019. 15 (3): e1006750.

[12] Xiangyu Sun et al. “Metabolic Reprogramming in Triple-Negative Breast Cancer”. *Frontiers in Oncology*. 2020. 10 (428): .

[13] J. Malone et al. “Ten Simple Rules for Selecting a Bio-ontology”. *PLoS Comput Biol*. 2016. 12 (2): e1004743.

[14] Brad T Sherman et al. “DAVID gene ID conversion tool”. *Bioinformation*. 2008. 2 (10): p. 428.

[15] Brigitte Waegele et al. “CRONOS: the cross-reference navigation server”. *Bioinformatics*. 2009. 25 (1): pp. 141–143.

[16] Kimberly J Bussey et al. “MatchMiner: a tool for batch navigation among gene and gene product identifiers”. *Genome biology*. 2003. 4 (4): pp. 1–7.

[17] Florian Iragne et al. “AliasServer: a web server to handle multiple aliases used to refer to proteins”. *Bioinformatics*. 2004. 20 (14): pp. 2331–2332.

[18] Richard G Côté et al. “The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases”. *BMC bioinformatics*. 2007. 8 (1): pp. 1–14.

[19] Gabriel F Berriz and Frederick P Roth. “The Synergizer service for translating gene, protein and other biological identifiers”. *Bioinformatics*. 2008. 24 (19): pp. 2272–2273.

[20] Arek Kasprzyk et al. “EnsMart: a generic system for fast and flexible access to biological data”. *Genome research*. 2004. 14 (1): pp. 160–169.

[21] Martijn P. van Iersel et al. “The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services”. *BMC Bioinformatics*. 2010. 11 (1): p. 5.

[22] *BridgeDb: Human and SARS-related corona virus gene/protein mapping database derived from Wikidata*. BridgeDb. URL: `https://zenodo.org/record/3860798` (visited on 11/25/2020).

[23] D. Howe et al. “Big data: The future of biocuration”. *Nature*. 2008. 455 (7209): pp. 47–50.

[24] Sarah Burge et al. “Biocurators and biocuration: surveying the 21st century challenges”. *Database (Oxford)*. 2012. 2012 (0): bar059.

[25] Martijn P van Iersel et al. “Presenting and exploring biological pathways with PathVisio”. *BMC Bioinformatics*. 2008. 9: p. 399.

[26] *The EU’s open science policy*.

[27] Lars Juhl Jensen, Jasmin Saric, and Peer Bork. “Literature mining for the biologist: from information retrieval to biological discovery”. *Nat Rev Genet*. 2006. 7 (2): pp. 119–129.

[28] Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. “Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases”. *Humanities and Social Sciences Communications*. 2021. 8 (1): pp. 1–15.

[29] Archana Goyal, Vishal Gupta, and Manish Kumar. “Recent named entity recognition and classification techniques: a systematic review”. *Computer Science Review*. 2018. 29: pp. 21–43.

[30] Jung-jae Kim, Piotr Pezik, and Dietrich Rebholz-Schuhmann. “MedEvi: Retrieving textual evidence of relations between biomedical concepts from Medline”. *Bioinformatics*. 2008. 24 (11): pp. 1410–1412.

[31] Gerald Salton. “Automatic information organization and retrieval”. 1968. : .

[32] D. J. Rigden and X. M. Fernández. “The 2022 Nucleic Acids Research database issue and the online molecular biology database collection”. *Nucleic Acids Res*. 2022. 50 (D1): pp. D1–D10.

[33] Jon Ison et al. “EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats”. *Bioinformatics*. 2013. 29 (10): pp. 1325–1332.

[34] Pieter BT Neerincx and Jack AM Leunissen. “Evolution of web services in bioinformatics”. *Briefings in bioinformatics*. 2005. 6 (2): pp. 178–188.

[35] Jiten Bhagat et al. “BioCatalogue: a universal catalogue of web services for the life sciences”. *Nucleic acids research*. 2010. 38 (suppl_2): W689–W694.

[36] Kenneth Baclawski and Tianhua Niu. *Ontologies for bioinformatics (computational molecular biology)*. The MIT Press, 2005.

[37] Tim J Berners-Lee. “The world-wide web”. *Computer networks and ISDN systems*. 1992. 25 (4-5): pp. 454–459.

[38] *RFC2616: Hypertext Transfer Protocol–HTTP/1.1*.

[39] Tim Berners-Lee, James Hendler, and Ora Lassila. “The semantic web”. *Scientific American*. 2001. 284 (5): pp. 34–43.

[40] *The Semantic Web - Scientific American*. URL: https://www.scientificamerican.com/article/the-semantic-web/ (visited on 11/25/2020).

[41] François Belleau et al. “Bio2RDF: towards a mashup to build bioinformatics knowledge systems”. *Journal of biomedical informatics*. 2008. 41 (5): pp. 706–716.

[42] Nicole Redaschi and UniProt Consortium. “UniProt in RDF: tackling data integration and distributed annotation with the semantic web”. *Nature precedings*. 2009. : pp. 1–1.

[43] Muhammad Saleem et al. “Big linked cancer data: Integrating linked tcga and pubmed”. *Journal of web semantics*. 2014. 27: pp. 34–41.

[44] Antony J. Williams et al. “Open PHACTS: semantic interoperability for drug discovery”. *Drug Discov Today*. 2012. 17 (21-22): pp. 1188–1198.

[45] Simon Jupp et al. “The EBI RDF platform: linked open data for the life sciences”. *Bioinformatics*. 2014. 30 (9): pp. 1338–1339.

[46] Gang Fu et al. “PubChemRDF: towards the semantic annotation of PubChem compound and substance databases”. *Journal of cheminformatics*. 2015. 7 (1): pp. 1–15.

[47] Egon L Willighagen et al. “The ChEMBL database as linked open data”. *Journal of cheminformatics*. 2013. 5 (1): pp. 1–12.

[48] Sören Auer et al. “Dbpedia: A nucleus for a web of open data”. *The semantic web*. Springer, 2007, pp. 722–735.

[49] Denny Vrandečić. “Wikidata: a new platform for collaborative data collection”. Paper presented at: *Proceedings of the 21st International Conference on World Wide Web*. New York, NY, USA. Association for Computing Machinery. 2012. pp. 1063–1064.

[50] Mark D. Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. *Scientific Data*. 2016. 3 (1): p. 160018.

[51] Jim Giles et al. “Wikipedia rival calls in the experts”. *NATURE-LONDON-*. 2006. 443 (7111): p. 493.

[52] Barend Mons et al. “Calling on a million minds for community annotation in WikiProteins”. *Genome biology*. 2008. 9 (5): pp. 1–15.

[53] Paul Groth, Andrew Gibson, and Jan Velterop. “The anatomy of a nanopublication”. *Information Services & Use*. 2010. 30 (1-2): pp. 51–56.

[54] Alexander R. Pico et al. “WikiPathways: Pathway Editing for the People”. *PLOS Biology*. 2008. 6 (7): e184.

[55] Augustin Luna et al. “PathVisio-MIM: PathVisio plugin for creating and editing Molecular Interaction Maps (MIMs)”. *Bioinformatics*. 2011. 27 (15): pp. 2165–2166.

[56] Daniel J Barrett. *MediaWiki: Wikipedia and beyond.* ” O’Reilly Media, Inc.”, 2008.

[57] Thomas Pellissier Tanon et al. “From freebase to wikidata: The great migration”. Paper presented at: *Proceedings of the 25th international conference on world wide web*. 2016. pp. 1419–1428.

[58] Finn Årup Nielsen, Daniel Mietchen, and Egon Willighagen. “Scholia, scientometrics and wikidata”. Paper presented at: *European Semantic Web Conference*. 2017. pp. 237–259.

[59] Katherine Thornton et al. “Modeling the Domain of Digital Preservation in Wikidata.” Paper presented at: *iPRES*. Kyoto. 2017.

[60] Stacy Allison-Cassin and Dan Scott. “Wikidata: a platform for your library’s linked open data”. *Code4Lib Journal*. 2018. (40): .

[61] Stefan Decker et al. “The semantic web: The roles of XML and RDF”. *IEEE Internet computing*. 2000. 4 (5): pp. 63–73.

[62] Katherine Thornton et al. “Using shape expressions (ShEx) to share RDF data models and to guide curation with rigorous validation”. Paper presented at: *European Semantic Web Conference*. Slovenia. springer. 2019. pp. 606–620.

[63] Iovka Boneva et al. “Shape Expressions Schemas”. *ArXiv*. 2015. abs/1510.05555: .

[64] Slawomir Staworko et al. “Complexity and Expressiveness of ShEx for RDF”. Paper presented at: *ICDT*. Brussels. dblp. 2015.

[65] A. S. Waagmeester, T. Kelder, and C. T. A. Evelo. “The role of bioinformatics in pathway curation”. *Genes Nutr*. 2008. 3 (3-4): pp. 139–142.

[66] Andra Waagmeester et al. “Pathway Enrichment Based on Text Mining and Its Validation on Carotenoid and Vitamin A Metabolism”. *OMICS: A Journal of Integrative Biology*. 2009. 13 (5): pp. 367–379.

[67] Andra Waagmeester et al. “Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources”. *PLOS Computational Biology*. 2016. 12 (6): e1004989.

[68] Andra Waagmeester et al. “Wikidata as a knowledge graph for the life sciences”. *eLife*. 2020. 9 (e52614): e52614.

[69] Andra Waagmeester et al. “A protocol for adding knowledge to Wikidata: aligning resources on human coronaviruses”. *BMC Biology*. 2021. 19 (1): p. 12.

2

The role of bioinformatics in pathway curation

Adapted from: A. S. Waagmeester, T. Kelder, and C. T. A. Evelo. "The role of bioinformatics in pathway curation". *Genes Nutr*. 2008. 3 (3-4): pp. 139–142.

Abstract

Diagrams and models of biological pathways are useful tools in biology. Pathway diagrams are mainly used for illustrative purposes for instance in textbooks and in presentations. Pathway models are used in the analysis of genomic data. Bridging the gap between diagrams and models allows not only the analysis of genomics data and interactions but also the visualization of the results in a variety of different ways. The knowledge needed for pathway creation and curation is available from three distinct sources: databases, literature, and experts. We describe the role of bioinformatics in facilitating the creation and curation of pathways.

2.1 Introduction

Biological pathway diagrams are used to describe molecular biology processes in a graphical way. A pathway is a set of related reactions in a given context, i.e. glycolysis, Krebs cycle, or apoptosis [1]. Traditionally, pathway diagrams are used as representations of knowledge, such as used in textbooks or in discussions among scientists.

Recently, pathway representations gained momentum as a research instrument. High-throughput genomics experiments, such as DNA microarrays, present the researcher with volumes of research data that are often too large for manual assessment. Mathematical methods like clustering or principal component analysis can be used to structure large data volumes [2], but do not normally lead to increased understanding. Pathway diagrams can be used to present the outcome of such mathematical methods or genomics data directly. Biologically relevant changes are more visible when projected on pathway diagrams than when presented as large sets of tabular data.

This new role of the pathway representations in analysis creates specific requirements for their creation and curation. When a pathway diagram is used as an illustration, desktop publishing tools can be used to draw the diagram (Adobe Photoshop, Paintshop Pro, etc.). In this role, pathway entities and relations between entities have not to be made explicit, as long as the diagram can be interpreted by human assessment. When pathway diagrams are used as a research tool, they should be available in a computer-readable form. Not only every visible aspect of a pathway diagram needs to be made explicit, but also all relationships are needed for analysis. For instance, visible genes products and metabolites need to be connected to a database entry that can be used to link them to experimental data and reactions. Reactions between metabolites or gene products need to be treated as edges in an interaction network to allow network analysis.

The research field of bioinformatics has an active role in this part of pathway modelling and creation. In this paper, we emphasize on pathway models, which can be used in research tools in bioinformatics. We will distinguish between aesthetic pathway diagrams and technical pathway models. We will

also show why it could be advantageous to be able to combine both aesthetics pathway diagrams and computer-readable pathway models.

2.2 Pathway diagram formats

The visual style and information of pathway diagrams vary between different resources. Figure 7.1 contains four examples of pathway diagrams covering knowledge about the same topic. They are extracted from KEGG [3], WikiPathways [4], Biocarta (`http://www.biocarta.com`) and Metacore (`http://www.genego.com/metacore.php`). When one wants to project research data onto a pathway, the diagrams shown in Figure 2.1 will not suffice. In order to be suitable for data analysis, pathway diagrams need to be stored in a computer-readable format, such as XML. All four resources mentioned above have pathway repositories that contain both a graphical representation and at least logically structured data on the genes appearing in those pathways. WikiPathways and KEGG do this in an easily computer-readable and extendable XML format. The utilization of XML alone is not sufficient to allow computational analysis in a biological context. Scalable vector graphics (SVG), for example, is an XML format that only describes the graphical elements of an image, but not the biological meaning of these elements. A model of at least the biological entity types in a pathway is also required.

BioPAX [5] is an initiative started at the ISMB'02 Conference that aims at developing an exchange standard for facilitating the integration of pathway knowledge from various sources. BioPAX could be seen as the opposite pathway type of graphical pathway diagram. In contrast to the graphically oriented pathway diagrams, BioPAX focuses on capturing the pathway information in a non-graphical, highly structured form. However, the BioPAX model lacks support for storing any graphical information. Although it is a deliberate design choice, it limits the practical usability of BioPAX for the visual interpretation of genomics data.

To be able to both capture logical knowledge and graphical information we started the development of GPML (Genmapp, Pathway Markup Language)

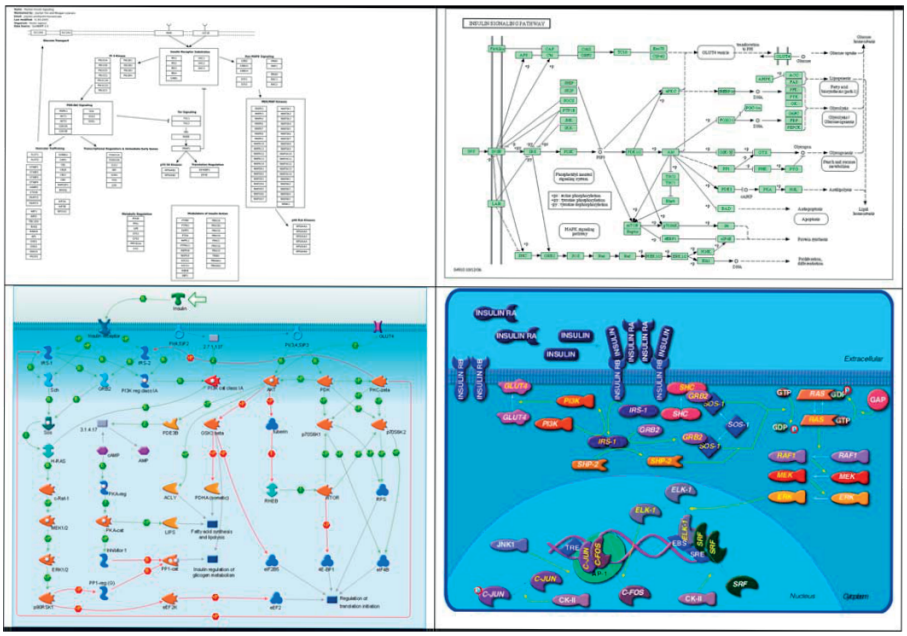


Figure 2.1: Four examples of pathway diagrams from different pathway resources, all capturing knowledge from a similar topic. From left to right, top to bottom: 1 WikiPathways, 2 KEGG, 3 Metacore, and 4 Biocarta

in collaboration with the Conklin Group at the University of California, San Francisco [6]. GPML is an XML implementation of the older Genmapp data format extended with explicit interactions and the possibility to add BioPAX elements like literature references.

2.3 The process of pathway curation

The application of pathway diagrams for data analysis requires the integration of knowledge from a wide variety of sources. Typically, a pathway diagram is created by integrating knowledge from biological databases, the scientific literature, and intrinsic knowledge from domain experts [7]. Each of these three sources presents specific challenges to extracting and integrating the knowledge into a pathway diagram. Part of this challenge is the exponentially growing amount of available information that is scattered over a wide variety of sources and knowledge domains. Pathguide (`http://www.pathguide.org`), a website that lists pathway resources, already lists 261 different databases (October 2008) containing pathway knowledge [1]. As Cary [8] points out, there is a need to be aware of potential biases when integrating data from biological databases. Biological databases are often constructed around a specific species or with a specific set of research questions in mind.

Presenting knowledge as easily readable text often means that computers have a problem interpreting it. An author would for instance try to avoid the utilisation of the same word more than once in the same paragraph, even when it denotes the same entity. This means that the actual lines containing crucial information have to be read and interpreted by humans before they can be available in a computer-readable form. As it is no longer possible to keep up-to-date with the exponentially growing amounts of literature, we need ”text-mining approaches” to find the important parts [9]. This provides specific challenges to the (bio) informatics community. New approaches need to be developed to automatically deal with text primarily intended for human interpretation. The requirement for text mining is not exclusive to the biology community; other disciplines in science are also facing an increase in scientific literature.

Although each discipline could benefit from general text mining solutions, the specific characteristics of the way literature is stored require specific solutions for each discipline. One benefit biomedical sciences have is the existence of Pubmed (`http://www.ncbi.nlm.nih.gov/pubmed/`), which provides us with a standard in the representation of scientific literature. Another specific characteristic of the current biology field is the vast amount of experimental data that is created in genomics. This directs the questions we want to see answered from text mining efforts. For instance, we do want to find relationships between genes influencing each other's regulation.

Integrating knowledge from different domain experts is probably the most challenging task, especially if more than one domain expert is involved. This process requires commitment from the domain experts and an extensive social network and social skills by the pathway diagram curator. In the next chapter, we will address this point more in detail.

2.4 The role of bioinformatics in pathway curation

These different challenges provide the opportunity for bioinformatics to have an active role in pathway curation.

An illustration of how bioinformatics can assist pathway curation is WikiPathways [4] (`http://www.wikipathways.org`). WikiPathways builds upon the concept of community-based curation of biological knowledge using a wiki [4, 10, 11]. A wiki is a website the content of which can be edited by users. A well-known example of a wiki is Wikipedia, an online encyclopedia where all content is created and maintained by its users. No formal approval of a human editor is needed. Still, recent studies showed that the quality of Wikipedia is similar to editor-based encyclopedias [12].

WikiPathways applies a similar approach to biological pathway information. Pathway diagrams on WikiPathways can be created and curated by every member of the scientific community. As mentioned earlier, integrating expert knowledge requires an extensive social network and commitment from

the community. Still, it requires a time-intensive procedure to collect knowledge from different domain experts involved. Using WikiPathways, domain experts from all over the world can directly collaborate on improving specific pathway diagrams. Research groups focusing on a specific research area can adopt a set of pathways and identify themselves as a community by creating a portal page. This makes it easier to capture and integrate information from different domains and increases the commitment of the community to pathway curation. Specific research communities can adopt a set of pathways, which will then be arranged into a portal.

2.5 Where aesthetic pathway diagrams meet pathway knowledge models

We can define a spectrum of pathway knowledge types with complete graphical pathway diagrams on one side, and BioPAX pathway models on the other tail. In between these extremes, we find a spectrum of resources capturing pathway knowledge in both graphical form and in a computer-readable format. Additionally, computational methods based on pathway diagrams can extract relevant parts out of large experimental data sets and improve statistical power.

As mentioned before, the different applications pose different requirements on the pathway representations. As an example, a pathway diagram needs to be primarily aesthetic and in a human-readable format. In this format, we can use colours and other graphical features to distinguish various entity types such as cell structures, genes, proteins, etc. Indeed pathways need to be ”beautiful”. By applying colours or other aesthetic features, researchers understand knowledge better. However, pathways diagrams, which are used in data analysis, need to be simple; they should allow an overlay of the experimental data that we want to show. Very colourful and detailed pathway representations as we see them for instance in Metacore, actually make it harder to understand data representations.

One would expect that typically bioinformatics have a role in the curation of pathway models for use in data analysis. For complete graphical pathways di-

agram any standard imaging program would be sufficient. We would advocate for bridging the two application areas by providing hybrid pathway representations. Having such hybrid pathway diagrams would eliminate redundancy. Such pathway diagrams would contain enough structured, computer-readable information to be used in data analysis with bioinformatics tools, as well as a clear and flexible graphical representation to aid human interpretation.

2.6 Conclusion

As bioinformaticians, we have taken an active role in facilitating community-based pathway curation. Data analysis on results from genomics studies requires accurate and complete pathway models and the corresponding diagrams need to be interpretable by scientists. Initiatives such as WikiPathways aim to collect and present pathway information that meets both requirements using a community-based curation approach and a graphically oriented pathway format, while at the same time facilitating the integration of knowledge from other sources such as databases and scientific literature.

References

[1] Gary D. Bader, Michael P. Cary, and Chris Sander. “Pathguide: a Pathway Resource List”. *Nucleic Acids Res*. 2006. 34 (Database issue): pp. D504–D506.

[2] David B. Allison et al. “Microarray data analysis: from disarray to consolidation and consensus”. *Nat Rev Genet*. 2006. 7 (1): pp. 55–65.

[3] H Ogata et al. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Res*. 1999. 27 (1): pp. 29–34.

[4] Alexander R. Pico et al. “WikiPathways: Pathway Editing for the People”. *PLOS Biology*. 2008. 6 (7): e184.

[5] Joanne S. Luciano. “PAX of mind for pathway researchers”. *Drug Discovery Today*. 2005. 10 (13): pp. 937–942.

[6] Martijn P van Iersel et al. “Presenting and exploring biological pathways with PathVisio”. *BMC Bioinformatics*. 2008. 9: p. 399.

[7] Michiel E. Adriaens et al. “The public road to high-quality curated biological pathways”. *Drug Discov Today*. 2008. 13 (0): pp. 856–862.

[8] Michael P. Cary, Gary D. Bader, and Chris Sander. “Pathway information for systems biology”. *FEBS Lett*. 2005. 579 (8): pp. 1815–1820.

[9] Lars Juhl Jensen, Jasmin Saric, and Peer Bork. “Literature mining for the biologist: from information retrieval to biological discovery”. *Nat Rev Genet*. 2006. 7 (2): pp. 119–129.

[10] Allison Doerr. “We the curators”. *Nature Methods*. 2008. 5 (9): pp. 754–754.

[11] Mitch Waldrop. “Big data: Wikiomics”. *Nature*. 2008. 455 (7209): pp. 22–25.

[12] Jim Giles. “Internet encyclopaedias go head to head”. *Nature*. 2005. 438 (7070): pp. 900–901.

3

Pathway Enrichment Based on Text Mining and Its Validation on Carotenoid and Vitamin A Metabolism

Adapted from: Andra Waagmeester et al. "Pathway Enrichment Based on Text Mining and Its Validation on Carotenoid and Vitamin A Metabolism". *OMICS: A Journal of Integrative Biology*. 2009. 13 (5): pp. 367–379.

Abstract

Carotenoid metabolism is relevant to the prevention of various diseases. Although the main actors in this metabolic pathway are known, our understanding of the pathway is still incomplete. The information on carotenoids is scattered in the large and growing body of scientific literature. We designed a text-mining workflow to enrich existing pathways. It has been validated on the vitamin A pathway, which is a well-studied part of the carotenoid metabolism. In this study, we used the vitamin A metabolism pathway as it has been described by an expert team on carotenoid metabolism from the European Network of Excellence in Nutrigenomics (NuGO). This workflow uses an initial set of publications cited in a review paper (1,191 publications), enlarges this corpus with Medline abstracts (13,579 documents), and then extracts the key terminology from all relevant publications. Domain experts validated the intermediate and final results of our text-mining workflow. With our approach, we were able to enrich the pathway representing vitamin A metabolism. We found 37 new and relevant terms from a total of 89,086 terms, which have been qualified for inclusion in the analyzed pathway. These 37 terms have been assessed manually and as a result, 13 new terms were then added as entities to the pathway. Another 14 entities belonged to other pathways, which could form the link of these pathways with the vitamin A pathway. The remaining 10 terms were classified as biomarkers or nutrients. Automatic literature analysis improves the enrichment of pathways with entities already described in the scientific literature.

3.1 Introduction

Carotenoids are natural nutritional compounds. Fruits, vegetables, and derived products constitute the major sources of these lipid-soluble pigments. There is strong evidence that high dietary intake of these compounds is beneficial for humans and that it is associated with a lower risk of developing cardiovascular diseases, cancers, and age-related eye diseases [1–3]. Several mechanisms have been described that explain at least in part the observed effects. Examples include [A] a role in the scavenging of free radicals, [B] the enhancement of gap-junction communication, [C] a role in immunomodulation, [D] reduction of the risk of site-specific cancer and heart disease, and [E] the protection of eye tissue [4]. Among the more than 600 identified carotenoids, around 10% are categorized as pro-vitamin A, meaning that they can be cleaved by the organism to produce retinal, the precursor of vitamin A (retinol). Harrison et.al. [5] estimated in 2005 that more than 100 million children worldwide suffer from vitamin A deficiency, and indicated that an increased understanding of the metabolic processes concerning the absorption of carotenoids could lead to better nutrition and thereby in general to an increase in the health of several populations [5]. Getting a profound understanding of the metabolic pathway of carotenoid metabolism is assisted by the integration of a variety of study results from different sources and biomedical research domains [6].

Pathways are abstract and functional representations of biological knowledge [7]. Our understanding of their functioning is essential in the analysis of genomic results [8] and, because such an approach requires accurate and up-to-date representations of pathways, this research work is ongoing. For the creation of such representations, the scientific curators (biologists and other domain experts) merge their expert knowledge with information from biological databases [9] and available scientific literature.

In biology, the Medline database is the primary resource of scientific literature. It contains more than 18 million references to scientific journal publications in the biomedical field, including author information and abstracts. The rate of growth of Medline is so high that scientists have big difficulties to keep up to date [10]. New approaches to filtering information from the scientific

literature are required. Traditional procedures for gathering documents from a corpus (i.e., a collection of documents. PubMed could be considered as a corpus of biomedical abstracts) rely on selecting a set of appropriate keywords with which one hopes to retrieve as many relevant publications as possible from the document collection at hand. Measures have been defined to monitor and optimize the retrieval for a given task to the selection of relevant documents [10]. Whenever the resulting set of references is too overwhelming, one coping strategy is to limit the search to review articles. To this end, PubMed, the Web-based interface to Medline, offers functionality to limit the searches to retrieve only reviews. By manually selecting appropriate reviews scientists acquire references to other relevant articles.

A literature search is optimal if both precision and recall are high. Precision is the number of relevant documents in all retrieved documents. The recall is the portion of retrieved documents out of all relevant documents in the corpus [11]. In other words, recall is the ratio between the true positives and the sum of the true positives and the false negatives. The precision is the ratio between the true positives and the sum of the true positives and the false positives. A literature search is optimal if the results are generated with high recall and high precision at the same time, which is difficult to achieve because often high precision hinders high recall and vice versa (see the explanations below).

Two factors have a strong influence on recall and precision. The first factor is the polysemy (a term denoting multiple meanings) of keywords used in user queries leading to their ambiguous interpretation in the text. Highly ambiguous terms decrease precision because the number of false positives could increase, that is, the more senses a term has the more diverse the final retrieval of documents (e.g., the term cancer denoting a disease and a species). The second factor is synonymy (the use of different terms to denote the same concept). If a concept is represented by more than one term, and neither the query system nor the user formulating the query considers all relevant synonyms to generate a complete query, this changes the recall due to the potential increase of false negatives. An example that illustrates the influence of the latter factor is the retrieval of documents describing MRI (magnetic resonance imaging) [12]. This imaging method is widely used in medicine. The

same underlying physical method is used in chemistry for spectral analyses but is here called NMR (nuclear magnetic resonance). Documents describing NMR could be relevant for someone interested in MRI. Unawareness of synonymous terms leads to missing relevant literature. PubMed uses thesauri like MESH to overcome some of these problems. The link between NMR and MRI will be found in MESH. But synonymy could be so subtle that it would require very big thesauri, and then again the excessive use of synonyms creates the risk of increasing ambiguity. Because of these limitations in the use of keywords as query terms and due to the increasing volume of scientific literature, we are prone to miss relevant literature.

Biological network and pathway models are needed to integrate different types of genomic results in a systems biology approach to understand for instance micronutrient health [13]. This approach is currently limited by the quality of the available biological pathways. To improve this, curators integrate knowledge from various scientific domains. Because each domain has specific sets of journals, this means that many journals need to be explored. So in pathway curation, the chance of missing important information is thus likewise increased. What we present here is an integrated approach where the pathway content is used to start a text-mining effort, and the results are represented in such a way that they can easily be integrated into the pathway itself by manual curation. In our study, we have compiled a text-mining workflow for finding potential pathway entities in the literature for a given context (using Vitamin A and carotenoids as an example). Other researchers have used text mining to build abstract representations of relations in biology. Text mining is also used to generate a controlled vocabulary of terms, which is a set of relevant words or phrases [12]. In a different study, the text-mining solution generated networks of related concepts [14]. Our approach is similar to those mentioned above but differs in the sense that we go one step further. Through a combined representation of the existing pathway and the newly found terms, we allow people knowledgeable in the domain under scrutiny to combine already formulated pathway knowledge with text-mining results. Thus, offering integration of automated and manual knowledge collection. In our study, we assessed the relevance of the identified terms in collaboration with domain experts and integrated a number of candidates into the latest pathway repre-

sentation. In the present study, we apply a semiautomatic approach to identify pathway entities in Medline abstracts. We started with a pathway developed for vitamin A metabolism, which is the best-understood sub-pathway of the carotenoid pathway (Fig. 3.1). The text-mining approach was initiated from a single review article with 1,191 references [15]. In the first step, a corpus of Medline abstracts was automatically generated based on these references. This corpus was subsequently analyzed for pathway-related terms. The final output is a ranked list of terms that have been validated by curators for their relevance and novelty to the pathway under scrutiny. Our approach improves the process of generating, enriching, and updating pathways.

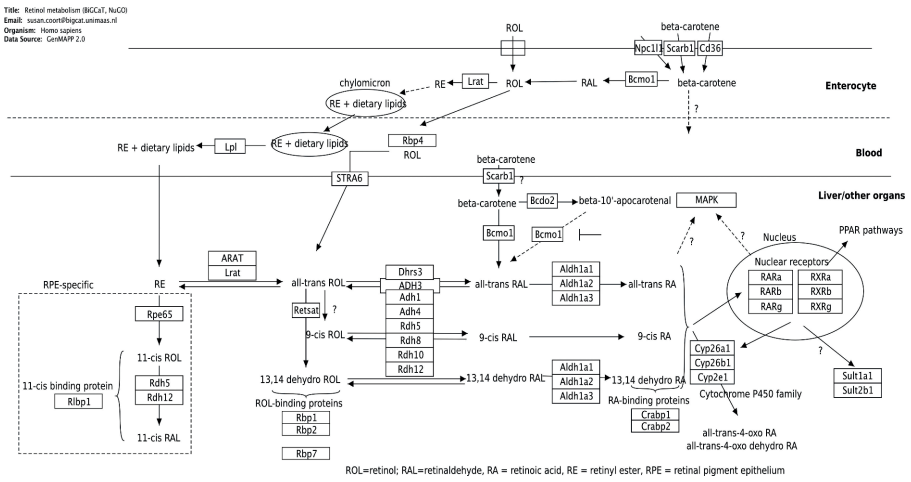


Figure 3.1: The metabolic pathway of vitamin A (retinol) metabolism constructed by a NuGO focus team on carotenoid metabolism (source: `http://www.wikipathways.org` November 2007)

3.2 Methods

Our text-mining solution makes use of the text-mining infrastructure at the European Bioinformatics Institute (EBI) where individual modules solve specific tasks (Whatizit, MedEvi) [16, 17].In addition to filtering out pieces of information, the solutions also provide explicit links from the identified informa-

tion to entries in biomedical data resources such as Uniprot [18], ChEBI [19], and the Human Metabolome Database [20]. Uniprot or the Universal Protein Resource, came into existence by merging Swiss-Prot, TrEMBL, and PIR. ChEBI is a database capturing chemical entities active in a biological context, and the Human Metabolome Database contains information on Human endogenous metabolites.

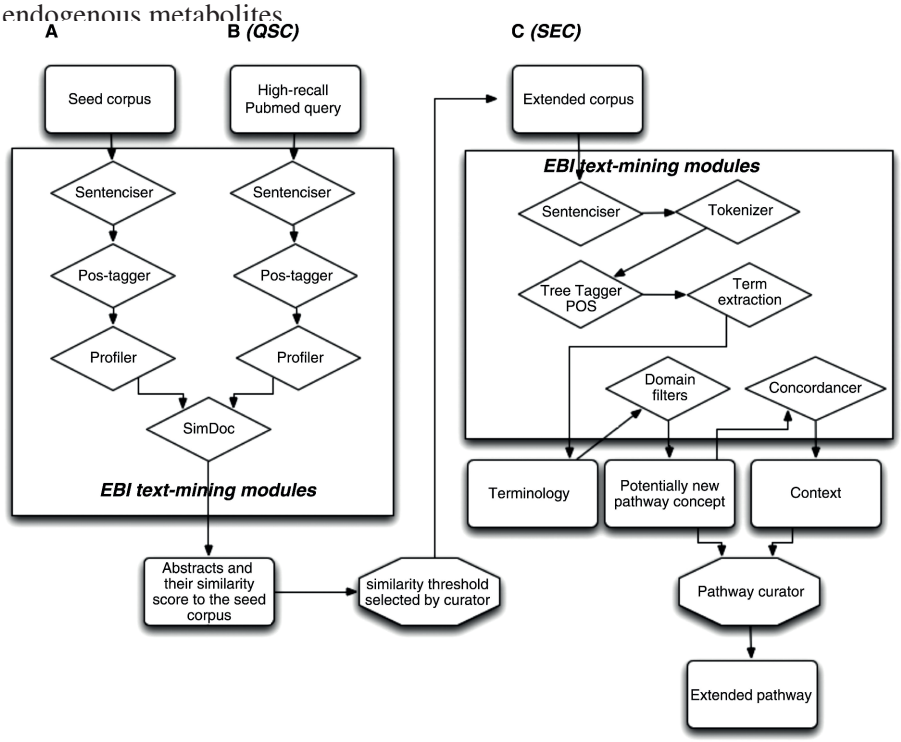


Figure 3.2: A two-step text-mining workflow for the enrichment of pathways. During the first phase a corpus of manually assessed abstracts (A) is extended in a semiautomatic manner with abstracts obtained from a high-recall query from PubMed (B). This extended corpus (C) is then input to the second phase in which terms are extracted. The resulting terminology comprises a potential pathway concept, which is available to pathway curators to extend a pathway.

3.2.1 Corpus generation

In the first step, we generated a seed corpus (SC) from the comprehensive review publication [15] that provides 1,191 references (see Fig. 3.2A). Then we generated a second corpus consisting of a large set of abstracts from Pubmed by using the query "carotenoids" (called "query selected corpus," QSC, see Fig. 3.2B). Each abstract from the QSC is assessed for relevance to our topic by measuring the similarity to the SC (see Fig. 3.2). The goal was to generate a large set of documents from the scientific literature that contains all relevant documents for the topic, that is, the carotenoid pathway, without imposing any bias on the selection. To this end the documents from the QSC are assessed against the documents from the SC to generate the "similarity extended corpus" (SEC, see Fig. 3.2C).

3.2.2 Relevance assessment of the QSC documents for the SC

For the SC all the titles and abstract texts were combined and processed together, whereas the documents from the QSC were processed individually. First, the sentences in the texts were marked up explicitly, and the part-of-speech information was added to the lemmatized tokens, that is, tokens represented in their dictionary form. Every document was transformed into a vector of its lemmatized tokens after removing non-informative tokens provided from a stop list, and the frequency of the lemmatized tokens in the document was kept to rank the tokens.

For every document we selected the most discriminative 20 terms according to the following procedure:

- For each lemmatized term found in a document Dunning's log-likelihood value has been computed. Typically, likelihood values are calculated in the context of hypothesis testing, which follows a binomial distribution; that is, true or false. The values are used to either accept or reject a null hypothesis. Dunning et al. adapted the log-likelihood ratio test to be applicable in a set with multinomial distributions. This adaptation makes the log-likelihood ratio test applicable in automatic text analysis. When counting word occurrences

in a document the resulting set follows a multinomial distribution. Dunning's log-likelihood value gives an indication that a term is informative in a given context.

- All terms are ranked according to their log-likelihood values where the highest log-likelihood value obtained is ranked at 1. The ranks are then used to compute the term weight according to the following formula: the inverse of [1 + log(rank)]. This transformation results in a smoother distribution of term weights for a given document.

- The same procedure is applied to generate the term vector of the SC. Because the SC of the carotenoid pathway contains 1,191 documents, a much larger vector containing the weights of 1,000 terms was used.

- For each document from the QSC the cosine similarity score is computed between the document's vector and the vector of the SC. Notice that the term weight can be equal to 0 whenever a term that is contained in the SC is not found in the document vector being tested for similarity. The documents from the QSC are ranked according to their decreasing cosine similarity, which indicates their relevance. The best relevant documents are added to the final SEC.

- A random set of abstracts from the QSC ranked according to their similarity score to the SC was evaluated for their relevance to the studied topic. Based on this assessment we derived a threshold for the similarity score. The selection of the threshold for similarity will be explained in detail in the results section. All abstracts with a similarity score higher than the selected threshold were added to the final SEC.

3.2.3 Extracting terminologies containing potential pathway entities

In the next step the documents from the SEC were processed (see Fig. 3.2C) to deliver a list of statistically significant terms, which are then filtered on relevance to the curated pathway. Potentially novel pathway entities are assumed

to belong to the following semantic types: (1) proteins, (2) genes, (3) enzymes, (4) chemical compounds representing metabolites, that is, chemicals being intermediate substances in metabolic processes, and (5) other chemical compounds. Our analysis was tuned to identify multi-word terms in the SEC because we applied part-of-speech patterns and the mutual information test that measures the association strength between the constituents of multi-word terms [21]. Recognition of multi-word terms in natural language is complex when the words are not consecutive in a sentence, and this may lead to the identification of shorter terms as sub-concepts. Presentation of results containing such a sub-concept to an expert is often enough to make him see the larger picture. In addition, we estimated the usage patterns for each term from the SEC, such as the evenness of a term's distribution in comparison to its relative frequency in the corpus. Evenness is measured with the Juilland Dispersion (JD). To this end the corpus is separated into 100 segments, that is, bins containing similar numbers of sentences. Then we compute the frequencies (X_t) of all statistically identified terms in all segments and determined the mean of their frequency over all segments [X = mean (X_t)] and the standard deviation[S = standard deviation (X_t, X)]. The JD is calculated as follows:

$$JD = 100 * (1 - \frac{V}{\sqrt{n-1}} \tag{3.1}$$

where (n = 100) is the number of segments in the SEC and V is the inverse of the mean of frequencies (X) normalized by the standard deviation (S):

$$V = \frac{S}{X} \tag{3.2}$$

A high dispersion value JD(t) occurs if a term t is frequent in almost all segments of the corpus. Comparing a term's relative frequency to its dispersion value could be a useful clue in the identification of potentially novel entities due to the following two interpretations:

- If the term is frequent and has a high dispersion value, then it can be assumed to denote a concept, which is likely to be very important in the

context of the pathway represented by the corpus. However, it is likely that such entities have already been included in the pathway representation, because they are mentioned consistently throughout the corpus, and thus were likely to also occur in the documents used to create the initial pathway.

- If the term is frequent but has a low dispersion value, then this signifies that a term is local to a smaller number of segments of the corpus. Given that we only accept pathway-relevant named entities such as UniProt entries, metabolites, and other chemical compounds as candidate pathway entities, a combination of high frequency and low dispersion could be indicative of a situation where the term is used in a few abstracts that report on important and new findings related to the pathway. Such entities are more likely to appear as new suggestions for pathway extension but still have to be evaluated for relevance.

3.2.4 Selecting the SC

In our efforts to develop a suitable text-mining approach for pathway generation, we focused on an important part of the carotenoid metabolism, the vitamin A metabolism. The reason for focusing on this part is twofold. It is the best-explored part of the carotenoid pathway and thus serves as a good benchmark in comparison to less understood pathways, and in addition, the review article [15] formed an ideal SC on gene regulation by retinoic acid. This review was an important source of information for the pathway editors from NuGO in that it listed 1,191 articles reporting relevant information to the vitamin A pathway. NuGO is the European Network of Excellence in Nutrigenomics (see `http://www.nugo.org`).

3.2.5 Increasing the recall

We have used the above-mentioned 1,191 references as an SC. We extended the SC to 13,579 Medline abstracts by setting the cutoff point for the similarity score at 0.07. The similarity threshold was obtained as a result of an integrated process using the curation capabilities of domain experts and the

text-processing capabilities of a novel text-mining solution. The three following steps were followed:

- we calculated the similarity score for an increasing number of potentially relevant abstracts (see Fig. 3.3). The different distributions in Figure 3 depict the similarity scores applied to term vectors of 30, 100, 500, or 1,000 terms, respectively. The introduction of more terms as vector elements decreased the similarity score of initially better-ranked documents and increased the similarity score of documents at lower ranks. We conclude that the increase in the term vector size leads to an increase in noise (i.e., terms that are not specific to the domain). We select the cutoff point at the value where the increase of terms in the vector leads to a decrease in the ranking for relevant documents and vice versa. This cutoff point was found to be in the range of 15,000 and 20,000 documents with similarity scores between 0.06 and 0.08.

- Three groups of documents were selected that were all represented by term vectors with 1,000 entries (Fig. 3.3). The first section was picked from the part of the distribution where the term vectors showed high similarity scores (above 0.15). The second part was picked from the region where the different graphs intersect with similarity scores between 0.06 and 0.08. The third part was selected from the tail of the graph, with similarity scores below 0.04. From the first group we selected 20 abstracts with a similarity score of just above0.15. For the second group, we selected 20 abstracts with a similarity score around the median value of 0.07 and for the last group 10 abstracts were selected with a similarity score below 0.04 and 10 with a similarity score around 0.01. This evaluation method is known as precision at rank [22].

- now, all the selected abstracts were presented to the domain experts from NuGO in random order. The abstracts were assessed for their relevance to the studied pathway and based on the evaluation by domain experts, the precision was calculated for each section (Table 3.1).

- The precision in the first section was 100%, 60% in the second, and 15% in the last one (Table 3.1).

Table 3.1: Precision at Rank Evaluation

A

Rank	Pubmed Id	Similarity score	Relevant
1141	7141883	0.154899	+
1142	10693163	0.154814	+
1143	1317113	0.154776	+
1144	7486470	0.154721	+
1145	8843984	0.154714	+
1146	9795972	0.154656	+
1147	438897	0.154653	+
1148	8504149	0.154643	+
1149	9212349	0.154601	+
1150	15735074	0.154554	+
1151	16614418	0.154553	+
1152	7346779	0.154543	+
1153	8823154	0.154535	+
1154	9108943	0.154528	+
1155	16365078	0.154524	+
1156	3632216	0.154506	+
1157	8503360	0.154503	+
1158	17234732	0.154502	+
1159	16177187	0.154477	+
1160	10846237	0.154368	+

B

Rank	Pubmed Id	Similarity score	Relevant
16868	12060479	0.060701	+
16615	1309942	0.061302	+
15008	9272132	0.065686	+
15007	16060221	0.065687	+
15006	16688767	0.065687	+
15005	9173086	0.06569	+
14995	17493127	0.065714	+
14994	9720977	0.065715	-
10003	1873990	0.083392	-
10002	14692515	0.083394	+
10001	16232082	0.083396	-
10000	12671093	0.083397	-
9999	3572247	0.08342	-
9998	3107194	0.083421	+
9997	1328400	0.083421	+
9996	11978137	0.083425	+
9995	10932161	0.083435	-
9994	10080695	0.083444	-
9993	10356420	0.083453	+
9992	4424510	0.083459	-

C

Rank	Pubmed Id	Similarity score	Relevant
35926	1169117	0.011403	-
35928	10080118	0.011387	-
35204	15712987	0.015108	-
35203	7107607	0.015114	-
35202	11388477	0.015117 -	
35196	7236601	0.015143	-
35151	15564532	0.015391	-
35141	8317903	0.015441	-
35117	11204602	0.015506	-
35111	11978818	0.015533	-
35107	3932288	0.015558	+
25008	15745429	0.04189	-
25007	12110702	0.041894	-
25006	7847852	0.041894	-
25005	12068573	0.0419	+
25004	1791223	0.041903	-
25003	8536624	0.041906	+
25002	7679173	0.041907	-
25001	13679861	0.041911	-
25000	2376553	0.041911	-

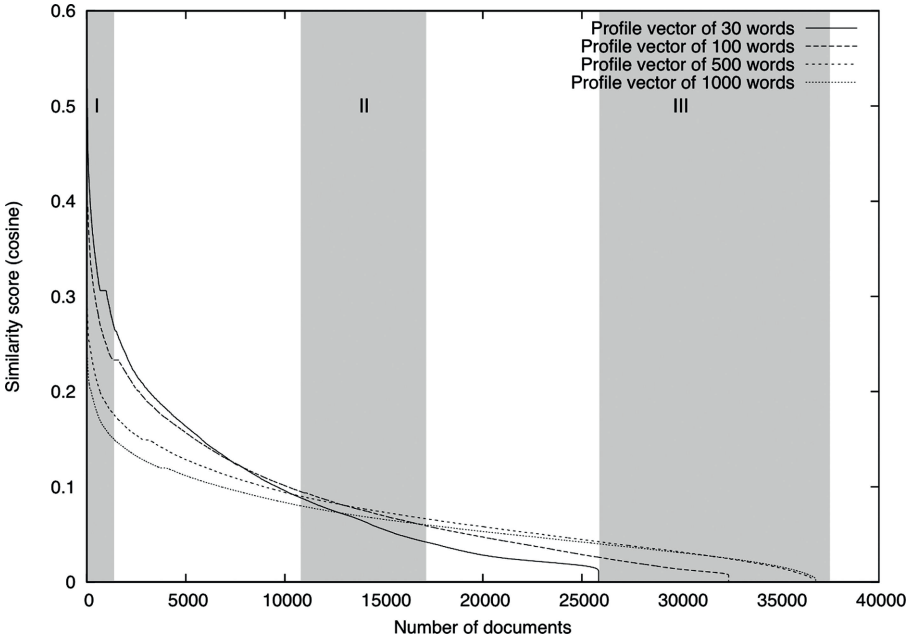


Figure 3.3: Distribution of similarity scores between a corpus containing potentially relevant publications and a seed corpus. The different graphs depict the difference in distribution for different numbers of informative terms in simdoc vectors.

This outcome confirms our expectation that the cutoff point should be selected in a range where the similarity score indicates that the term vectors mainly contain relevant terms. We decided to take the median value of the second section, which corresponds to a similarity score of 0.07.

3.3 Results

The initial query result on the PubMed query “carotenoids” resulted in a set of 51,628 references of which 34,682 contained an abstract. Of these 34,682 documents 13,579 had a similarity score >0.07 and were considered to be relevant to the construction of the vitamin A metabolism pathway.

3.3.1 Increasing journal coverage

The corpus expansion from 1,129 abstracts to 13,579 abstracts led also to changes in the distribution of journals in the corpus (Fig. 3.4). By expanding the corpus we observed an increase in the contribution of distinct journals that were not listed in the SC. The extended corpus contained 3,307 distinct journals out of which 2,586 journals were not part of the seed corpus. This clearly shows that the automated corpus expansion has led to larger journal coverage, which is practically not achievable by the manual selection of scientific articles.

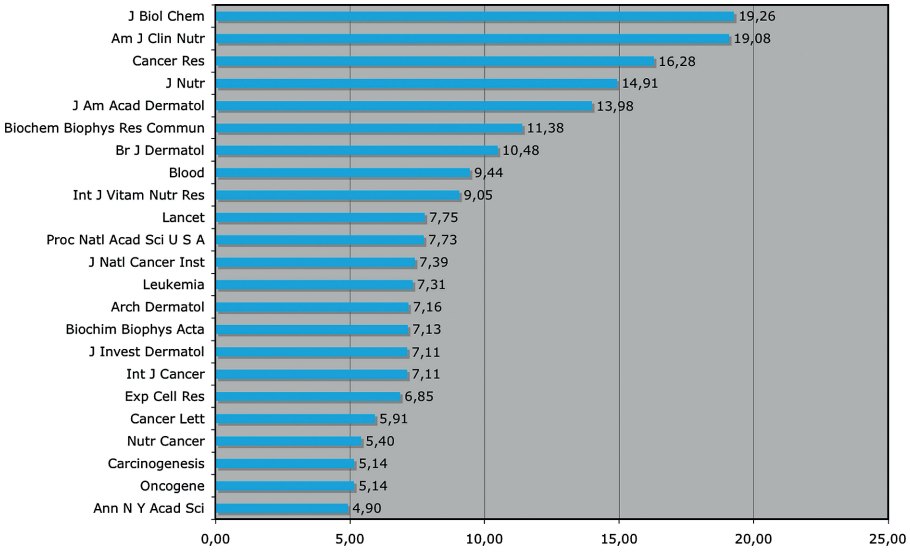


Figure 3.4: The journals that had the relatively highest increase in contribution to the extended corpus. The increase is measured in z-scores.

3.3.2 Extracting the key terminology

From the expanded corpus of 13,579 abstracts, we were able to extract 89,086 unique terms. A term may consist of multiple words. This extracted terminology contains domain-independent terms because the extraction methods

did not impose any domain-specific filter. Examples of domain-independent terms are: “Etude du Vieillissement Art,” “kg bodywt,” or “Southwest Oncology Group.” This shows that not all statistically significant terms are relevant to the analyzed pathway even though they were extracted from a corpus of relevant abstracts. From all extracted terms we identified 6,515 potential entities for the vitamin A metabolism pathway. For clarity, we distinguish single-word terms from multiple-word terms. Single-word terms are more likely to

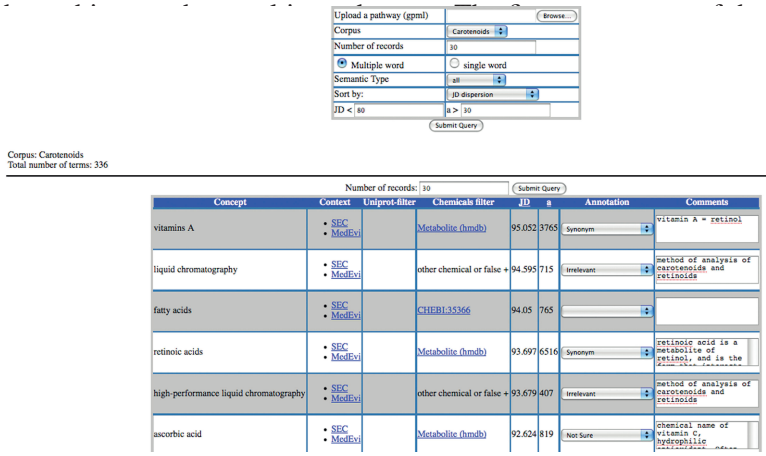


Figure 3.5: The extracted terms from the similarity extended corpus is available to pathway curators (`http://www.bigcat.unimaas.nl/public/data/textmining/carotenoids/`). Through this interface, one could browse the results by sorting on the frequency of occurrence or the Julliand Dispersion (JD). Furthermore, curators could limit the terms on specific semantic types (i.e., proteins, metabolites, chemical compounds). Curators could also consult the context either by extracting the keywords in conjunction with the sentence where they were extracted from, or by using EBI’s MedEvi (`http://www.ebi.ac.uk/Rebholz-srv/MedEvi/`).

3.3.3 Presenting potential pathway entities in context

We first presented a list of found terms that could be a potential pathway entity, in an HTML page (see *www.bigcat.unimaas.nl/public/data/textmining/carotenoids*), where a link to the abstracts from the SEC that contained this term is also presented. The list of abstracts, when opened, shows the sentences containing the concept plus a clickable PubMed identifier to the original publication. This part is primarily meant to verify the text mining results themselves. To facilitate the contextual evaluation we integrated our newly found entities as a separate list in the carotenoid pathway that is available on WikiPathways. This allows the curators to use the pathway editors functionality on WikiPathways to review the results, connect the new entities to the existing pathway, and add database information to individual concepts, leave them as a separate list of related entities, or remove them. In general, this mechanism, allows us to integrate results from automated text mining with manual curation. The final result of our text-mining workflow is a list of terms representing potential pathway entities, which are hyperlinked to their original context of occurrence in the expanded corpus. We have built an interface to browse the terminology (Fig. 3.5). In this interface, the results can be sorted by the JD values and their frequencies. It is also possible to filter for specific semantic types such as enzymes, metabolites, or other chemical compounds. The terms in Table 3.2 were selected from the generated vocabulary. They comprised the most informative term, that is, most frequent, highest JD, and high frequency but low JD. Table 3.2 contains those terms that were found with the UniProt filter (Proteins), Table 3.3 represents the metabolites found with the HMDB filter, and Table 3.4 contains terms that were identified with the chemical compound filter of Whatizit. These terms were then validated by two curators who were also involved in the creation of the initial pathway (Fig. 3.1).

We provided these terms together with an annotation tool where the curators could classify the potentially new pathway entities according to a predefined set of classes: (1) already included, (2) hyponym of a term already included, (3) hypernym of a term already included, (4) synonym of a term already in-

Table 3.2: Proteins That Are Potential Pathway Entities in Vitamin A Metabolism

Multiword Uniprot entities			
Sorted on Julliands Dispersion (JD)	Sorted on frequency of occurrence (a)	JD<70 a>10	Single-word Uniprot entities
alkaline phosphatase	protein kinases	tissue-type	dehydrogenases
cytochromes P450	alkaline phosphatase	plasminogen activator	dehydrogenase
protein kinases	cytochromes P450	Phospholipase A2	transglutaminase
Lipoprotein lipase	Lipoprotein lipase	zeta-carotene	Catalase
prostate-specific antigen	prostate-specific antigen	desaturase	peroxidase
zeta-carotene desaturase	tissue-type plasminogen activator	triglyceride lipase	GNMT
gamma-glutamyl transpeptidase	Phospholipase A2	gamma-glutamyl	thrombin
Phospholipase A2	zeta-carotene desaturase	transpeptidase	ACE
acid phosphatase	triglyceride lipase	acid phosphatase	TPA
cyclin-dependent kinase	gamma-glutamyl transpeptidase		NOS
proteinase K	acid phosphatase		AOX
DNA methyltransferase	aryl hydrocarbon hydroxylase		RetSat
tissue-type plasminogen activator	DNA methyltransferase		synthetase
Cytochrome oxidase	cyclin-dependent kinase		ADH3
triglyceride lipase	MAP kinases		PEPCK
aryl hydrocarbon hydroxylase	proteinase K		trypsin
MAP kinases	Reverse transcriptase		Lipase
tissue plasminogen activator	tissue plasminogen activator		ceruloplasmin
aldose reductase	aldose reductase		lysozyme
glutamic-oxaloacetic transaminase	Cytochrome oxidase		TGase
Reverse transcriptase	glutamic-oxaloacetic transaminas		PKA

Table 3.3: Metabolites That Are Potential Pathway Entities in Vitamin A Metabolism

Multiword metabolites			
Sorted on Julliands Dispersion (JD)	Sorted on frequency of occurence (a)	JD ¡ 70 a ¿ 10	Single-word metabolites
vitamins A	retinoic acids	25-hydroxyvitaminD	Retinol
retinoic acids	vitamins A	vitamin D2	beta-carotene
ascorbic acid	Vitamin C	phytanic acid	lycopene
Retinyl palmitate	retinyl ester	lutein esters	lutein
Vitamin C	Retinyl palmitate	nicotinic acid	alpha-tocopherol
retinoic acid	ascorbic acid	L-ascorbic acid	zeaxanthin
all-trans retinoic acids	all-trans-retinoic acids	11-cis retinaldehyde	palmitate
Folic acid	retinoic acid	pantothenic acid	alpha-carotene
all-trans retinol	retinoic acid	ferulic acids	beta-cryptoxanthin
13-cis-retinoic acid	13-cis-retinoic acid	copper ion	astaxanthin
retinyl ester	beta carotenes	alpha carotene	canthaxanthin
all-trans retinoic acid	9-cis-retinoic acids	adenosine	retinal
beta carotenes	Folic acid	monophosphate	Selenium
Linoleic acid	Vitamin D3	chlorogenic acid	Tocopherols
Vitamin D3	all-trans retinol	retinoyl glucuronide	Lycopene
beta carotenes	Linoleic acid	isopentenyl	tocopherol
9-cis-retinoic acids	retinol palmitate	diphosphate	phytoene
reduced glutathione	all-trans retinoic acid	ethanol solution	zinc
Uric acid	Uric acid	glutathione disulfide	cholesterol
retinol palmitate	Vitamin K	caffeic acids	all-trans-retinol
oleic acids	Alpha tocopherol	carnosic acid	retinaldehyde
nitric oxides	all-trans retinoic acid	retinyl ester	gamma-tocopherol
arachidonic acids	beta-carotenes	carbon dioxide	Lutein
all-trans retinoic acid	reduced glutathione	cholesterol sulfate	isotretinoin
docosahexaenoic acids	all-trans beta-carotene	nitrogen dioxide	iron
alpha-tocopherol	arachidonic acids	isopentenyl	cryptoxanthin
Superoxide anion	nitric oxides	pyrophosphate	lutein
Vitamin K	docosahexaenoic acids	alpha-linolenic acid	glutathione
all-trans beta-carotene	oleic acids	13-cis retinoic acids	violaxanthin
	benzoic acids		

https://www.overleaf.com/project/60031a7efd96da16cea3688a

Table 3.4: Chemical Compounds That Are Potential Pathway Entities in Vitamin A Metabolism

Multiword chemical			
Sorted on Julliands Dispersion (JD)	Sorted on frequency of occurrence (a)	JD ¡ 70 a ¿ 10	Single-word chemicals
liquid chromatography	fatty acids	arsenic trioxide	carotenoids
fatty acids	liquid chromatography	phenolic acid	retinoic
high-performance liquid	retinyl acetates	retinyl phosphate	carotenoid
chromatography	high-performance liquid	unesterified retinol	retinyl
polyunsaturated fatty acids	chromatography	Raman spectroscopy	retinoids
fatty acids	amino acid	all-rac-alpha-tocopheryl acetate	retinoid
retinyl acetates	hepatic vitamin	Stanol ester	acid
In spite	polyunsaturated fatty acids	lipoprotein cholesterol	antioxidant
amino acid	retinol acetate	CRBP I	carotene
electron microscopy	double bond	25-dihydroxyvitamin D	retinoic
high-pressure liquid	fatty acids	acyl coenzyme	all-trans-retinoic
chromatography	vitamins B6	heme protein	esters
thiobarbituric acid	high-pressure liquid	tannic acid	lipid
Hepatic vitamin	chromatography	retinyl methyl ether	peroxidation
hydrogen peroxide	phorbol esters	phosphatidyl choline	nutrients
vitamins B6	cholesteryl ester	petroleum ether	9-cis
free radical	bile acids	lipoprotein oxidation	provitamin
double bond	energy transfer	mitomycin C	pigment
gel electrophoresis	radical cations	cytosine arabinoside	ascorbic
Retinol acetate	electron microscopy	cyclin E	13-cis
cholesteryl ester	In spite	Eimeria acervulina	micronutrient
Subclinical vitamin	hydrogen peroxide	vitamin K1	retinoids
chain reaction	free radical	ethyl acetate	13-cis-retinoic
Carotenoids Lutein	photosystem II	redox state	pigments
Gel filtration	thiobarbituric acid	radical anion	acid
Messenger RNA	Retinol oxidation	13-cis-4-oxoretinoic acid	oxidation
vitamin B	gel filtration	Retinyl beta-glucuronide	9-cis-retinoic
gas chromatography	bile salt	11-cis-retinyl palmitate	9-cis
Actinomycin D	retinyl stearate	6-epoxyretinoic acid	Retinoids
steroid hormone	collagen synthesis	Mannosyl retinyl phosphate	xanthophylls
liquid chromatography	chain reaction		
Phorbol esters	subclinical vitamin		

cluded, (5) too generic, (6) irrelevant to the studied pathway, and (7) potential new pathway concept. Hypernyms are terms denoting more general entities than the term in question, and conversely, hyponyms are terms that are more specific than a given term. For example, lutein is the name of a carotenoid, that is, lutein is the hyponym of a carotenoid and carotenoid is the hypernym of Lutein. Our initial validation of the 266 potential pathway entities represented in Tables 2, 3, and 4 reduced the list to 36 potential pathway entities for the vitamin A metabolism pathway (Table 3.5). During this validation the curators not only read the terms themselves, but also the sentences and abstracts in which these potential pathway entities originally occurred. This validation step, based on the text from the Medline abstracts, also resulted in the identification of yet one additional pathway concept (Vitamin D3 receptor) [23–25], which had not been identified by the automatic method but has been identified by reading the most relevant abstracts. This concept has not been identified automatically because it represents a more complex multi-word concept just like the vitamin D3 receptor. Although the bigram vitamin D3 was identified by the automatic approach the trigram was not. Taking this additional concept into consideration we had a total of 37 pathway entities that have been identified as a result of our semiautomatic extraction methods. In the next step, the domain experts tried to incorporate the novel entities into the pathway. This processing step led to the result that 12 of the 36 terms and, in addition, the term vitamin D3 receptor (the 37th term) were found to be pathway entities that had to be included in the pathway for vitamin A metabolism. Fourteen of the remaining 24 entities were attributed to processes of related metabolic pathways like fatty acid metabolism, oxidative stress, coagulation cascade, and cholesterol metabolism. Because such entities form possibly yet unknown links to related pathways, we suggest that text mining not only serves to find new pathway entities but could also have a role in identifying related pathways. The 10 remaining terms could be categorized either as nutrients or biomarkers. Vitamin A, carotenoids, and other pathway concepts are frequently used as biomarkers in clinical and other health studies. The pathway containing the 13 new pathway entities was given to the expert team on carotenoid metabolism (NuGO focus team). They judged the relevance of the modifications to the pathway representation and considered the additions as a relevant improvement (Fig. 3.6).

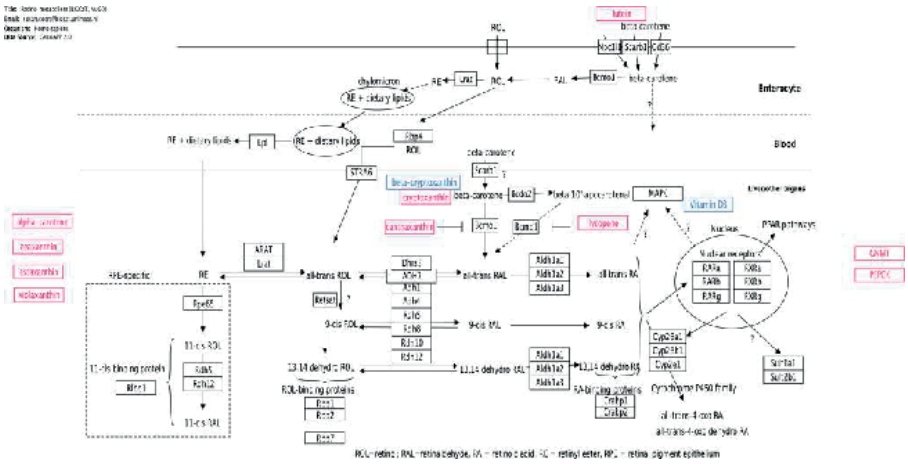


Figure 3.6: The resulting pathway of vitamin A metabolism extended with a text-mining approach. The newly included entities are indicated with colored boxes

Table 3.5: Chemical Compounds That Are Potential Pathway Entities in Vitamin A Metabolism

New pathway entities	Related processes or pathways	Biomarkers or nutrients
Vitamin D3	linoleic acid	selenium
ADH3	fatty acid	reduced glutathione gamma-tocopherol
alpha-carotene	GNMT	iron
astaxantin	MAPK	alpha-tocopherol
beta-cryptoxanthin	NOS	zinc
Canthaxanthin	nitric oxides	vitamine B
utein	oleic acids	vitamin K
ycopene	TGase	vitamine B6 prostate-specific antigen
zeaxanthin	carboxylesterases	
PEPCK	cholesterol	
cryptoxanthin	arachidonic acids	
riglyceride lipase	docosahexaenoic acids	
	tissue-type plasminogen activator	
	aldose reductase	

3.4 Discussion

Identification of pathways is ongoing research work. Extraction of pathway-relevant information from the literature, including protein–protein interactions and molecular interactions, is a challenging task. In the current state, curation of involved entities and integration of novel entities into existing pathways is necessary to deliver high-quality resources to the public. In our study, we have semi-automatically produced a list of entities that were finally included in the representation of the carotenoid pathways. Other approaches [12] have tackled the same problem but they have not been evaluated for the integration of novel entities into existing pathways. Nonetheless, both approaches demonstrate that the scientific literature is still a very rich resource to gather relevant information for the completion of pathway representations. One crucial step in our analysis is the generation of a corpus that is expanded from the SC and that can be used as input to the extraction of the key terminology. In our study, we have a review containing 1,191 references, which forms an excellent basis for an SC. Often such reviews are not available. In that case, the SC needs to be compiled by consulting experts. The resulting start set can in some cases be extended with references that were added to the studied pathway itself. KEGG and WikiPathways contain such references. This start set can in fact contain reviews like the one we used, which then can, of course, be extended. In other cases Web resources like CiteULike (http://www.citeulike.org/) and Connotea (http://www.connotea.org/), which provide means for scientists to organize and share references, can be used to find other articles that were often saved in combination with the articles from the start set. The selection of the extended corpus was based on a scoring function that discriminates between suitable documents and irrelevant documents with an empirically defined cut-off parameter. This distinction is arbitrary to a given extent but ensures that the best-ranked documents according to the judgment of a domain expert are included in the study. The main analysis is concerned with the selection of terminology and the assessment of this terminology by domain experts. The overall number of identified entities is small(only 37) in comparison to the overall number of extracted terms (89,086 terms). This reduction step is crucial to the selection of the most relevant terms and to optimize the interaction between the text-mining research group and the team of domain specialists,

that is, the amount of relevant terms has to be small enough to be processed by the domain specialists and comprehensive enough to generate sufficient benefits from their work. The identification of novel terminology from the literature for the completion of pathway representations still requires human assessment to meet the high-quality standards of curated data resources. This will not change in the near future. On the other side, our study has shown that it is beneficial to consider the whole of Medline for analysis to search for a small number of relevant terms. It is clear that the initial set of documents that have been selected from a review article did not make reference to all journals that could be relevant to the identification of entities for the selected pathway. In other words, only a text-mining approach that filters information from a large set of documents does not have to rely on assumptions that make a preselection of documents, journals, authors, or keyword queries. Considering the whole of the scientific literature gives advantages in the sense that no restrictions are applied to the primary data resource, and thus one of the many reasons for a bias to the study's results is excluded. In the future, we expect that other research teams will pick up similar approaches to filter the literature for concepts that have to be included in biomedical data resources and ontologies. To some extent, this need will be answered by online search engines that offer efficient access to the scientific literature [26, 27]. In addition, researchers will have to work together with text-mining research teams to identify the most relevant concepts. Altogether, the literature will be filtered many times for different purposes and will disclose step-by-step pieces of information that will be integrated into public biomedical data resources.

3.5 Conclusions

The increase in the volume of scientific literature requires new methods for knowledge extraction from the scientific literature. Our approach focused on the carotenoid pathway(retinol metabolism). We generated an extended corpus of Medline abstracts and based on the evaluation of sample sets, we conclude that the final extended corpora contained additional pathway-relevant documents that could only be gathered in a time-consuming manual process otherwise. With the combination of the corpus expansion and the pathway

enrichment workflow, we were able to identify 13 relevant and yet uncovered entities that were included in the vitamin A metabolism pathway, which is part of the carotenoid pathway. Our results also show that concepts can be identified that are potential links with other processes or pathways. The expanded corpus and the extracted terminology, including links to relevant Medline abstracts, are available at http://www.bigcat.unimaas.nl/public/data/textmining/carotenoids The workflow is applicable on other domains. We have also applied it to the extraction of pathways entities of the selenium metabolism pathway. Curators are currently processing the results into extending the existing pathway diagram. The evaluation of the results of our workflow still requires significant human intervention. With this workflow, we were able to select the most important information from much more abstracts than a given number of domain experts would be able to read, interpret, and understand. Essentially, we took the information from 13,579 abstracts pre-selected both by direct queries and by similarity evaluation with a seed corpus. Moreover, the application of the JD value helps in finding rare entities that are only described in a small part of the selected abstracts. The chance of missing these entities in a manual literature search is high. We believe that human intervention is decreased if our workflow is integrated into pathway repositories such as WikiPathways [28]. The community then provides the seed corpora. The final result of the workflow is a terminology of potential pathway entities. This could be provided as a list with references to the literature so that the curators (i.e., the community) can make a well-considered decision on extending the different pathways. To allow integration of this workflow in WikiPathways, we extended this wiki with an interface for automatically generated suggestions. The pathway is available on the Web page of WikiPathways [28] (`http://www.wikipathways.org`) for download. It can be used for gene expression analyses in Genmapp [8]) and EU Gene [29] and representations of interactions in Cytoscape [30].

References

[1] Marie Josèphe Amiot-Carlin, Caroline Babot-Laurent, and Franck Tourniaire. “3.1 Plant Pigments as Bioactive Substances”. *Food Colorants: Chemical and Functional Properties*. 2007. : p. 127.

[2] A. Bendich and J. A. Olson. “Biological actions of carotenoids”. *FASEB J*. 1989. 3 (8): pp. 1927–1932.

[3] N. I. Krinsky. “Actions of carotenoids in biological systems”. *Annu Rev Nutr*. 1993. 13: pp. 561–587.

[4] Kyung-Jin Yeum and Robert M. Russell. “Carotenoid bioavailability and bioconversion”. *Annu Rev Nutr*. 2002. 22: pp. 483–504.

[5] Earl H. Harrison. “Mechanisms of digestion and absorption of dietary vitamin A”. *Annu Rev Nutr*. 2005. 25: pp. 87–103.

[6] Sari Voutilainen et al. “Carotenoids and cardiovascular health”. *Am J Clin Nutr*. 2006. 83 (6): pp. 1265–1271.

[7] Michael P. Cary, Gary D. Bader, and Chris Sander. “Pathway information for systems biology”. *FEBS Lett*. 2005. 579 (8): pp. 1815–1820.

[8] Kam D. Dahlquist et al. “GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways”. *Nat Genet*. 2002. 31 (1): pp. 19–20.

[9] Gary D. Bader, Michael P. Cary, and Chris Sander. “Pathguide: a Pathway Resource List”. *Nucleic Acids Res*. 2006. 34 (Database issue): pp. D504–D506.

[10] Lars Juhl Jensen, Jasmin Saric, and Peer Bork. “Literature mining for the biologist: from information retrieval to biological discovery”. *Nat Rev Genet*. 2006. 7 (2): pp. 119–129.

[11] Hagit Shatkay and Ronen Feldman. “Mining the Biomedical Literature in the Genomic Era: An Overview”. *Journal of Computational Biology*. 2003. 10 (6): pp. 821–855.

[12] Irena Spasić et al. “Facilitating the development of controlled vocabularies for metabolomics technologies with text mining”. *BMC Bioinformatics*. 2008. 9 (5): S5.

[13] Ben van Ommen et al. “A network biology model of micronutrient related health”. *Br J Nutr*. 2008. 99 Suppl 3: S72–80.

[14] Hao Chen and Burt M. Sharp. “Content-rich biological network constructed by mining PubMed abstracts”. *BMC Bioinformatics*. 2004. 5 (1): p. 147.

[15] James E. Balmer and Rune Blomhoff. “Gene expression regulation by retinoic acid”. *J Lipid Res*. 2002. 43 (11): pp. 1773–1808.

[16] Harald Kirsch, Sylvain Gaudan, and Dietrich Rebholz-Schuhmann. “Distributed modules for text annotation and IE applied to the biomedical domain”. *Int J Med Inform*. 2006. 75 (6): pp. 496–500.

[17] Dietrich Rebholz-Schuhmann et al. “Text processing through Web services: calling Whatizit”. *Bioinformatics*. 2008. 24 (2): pp. 296–298.

[18] Amos Bairoch et al. “The Universal Protein Resource (UniProt)”. *Nucleic Acids Res*. 2005. 33 (Database Issue): pp. D154–D159.

[19] Kirill Degtyarenko et al. “ChEBI: a database and ontology for chemical entities of biological interest”. *Nucleic Acids Res*. 2008. 36 (Database issue): pp. D344–350.

[20] David S. Wishart et al. “HMDB: the Human Metabolome Database”. *Nucleic Acids Res*. 2007. 35 (Database issue): pp. D521–526.

[21] Eszter Beran. “Michael P. Oakes, 1998, Statistics for Corpus Linguistics.” *International Journal of Applied Linguistics*. 2000. 10 (2): pp. 269–274.

[22] *Text REtrieval Conference (TREC) 2007 Proceedings*. URL: `https://trec.nist.gov/pubs/trec16/t16%5C%5Fproceedings.html` (visited on 01/17/2021).

[23] David Heber and Qing-Yi Lu. “Overview of mechanisms of action of lycopene”. *Exp Biol Med (Maywood)*. 2002. 227 (10): pp. 920–923.

[24] H. Törmä et al. “Vitamin D analogs affect the uptake and metabolism of retinol by human epidermal keratinocytes in culture”. *J Investig Dermatol Symp Proc*. 1996. 1 (1): pp. 49–53.

[25] B. Vásárhelyi, A. Blázovics, and J. Fehér. "[The role of vitamin A analogues and derivatives in the regulation of cell function]". *Orv Hetil*. 1993. 134 (16): pp. 845–848.

[26] Andreas Doms and Michael Schroeder. "GoPubMed: exploring PubMed with the Gene Ontology". *Nucleic Acids Res*. 2005. 33 (Web Server issue): W783–786.

[27] Jung-jae Kim, Piotr Pezik, and Dietrich Rebholz-Schuhmann. "MedEvi: Retrieving textual evidence of relations between biomedical concepts from Medline". *Bioinformatics*. 2008. 24 (11): pp. 1410–1412.

[28] Alexander R. Pico et al. "WikiPathways: Pathway Editing for the People". *PLOS Biology*. 2008. 6 (7): e184.

[29] Duccio Cavalieri et al. "Eu.Gene Analyzer a tool for integrating gene expression data with pathway databases". *Bioinformatics*. 2007. 23 (19): pp. 2631–2632.

[30] Paul Shannon et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks". *Genome Res*. 2003. 13 (11): pp. 2498–2504.

4

Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources

Adapted from: Andra Waagmeester et al. "Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources". *PLOS Computational Biology*. 2016. 12 (6): e1004989.

Abstract

The diversity of online resources storing biological data in different formats provides a challenge for bioinformaticians to integrate and analyze their biological data. The semantic web provides a standard to facilitate knowledge integration using statements built as triples describing a relation between two objects. WikiPathways, an online collaborative pathway resource, is now available in the semantic web through a SPARQL endpoint at *sparql.wikipathways.org*. Having biological pathways in the semantic web allows rapid integration with data from other resources that contain information about elements present in pathways using SPARQL queries. In order to convert WikiPathways content into meaningful triples we developed two new vocabularies that capture the graphical representation and the pathway logic, respectively. Each gene, protein, and metabolite in a given pathway is defined with a standard set of identifiers to support linking to several other biological resources in the semantic web. WikiPathways triples were loaded into the Open PHACTS discovery platform and are available through its Web API (*dev.openphacts.org/docs*) to be used in various tools for drug development. We combined various semantic web resources with the newly converted WikiPathways content using a variety of SPARQL query types and third-party resources, such as the Open PHACTS API. The ability to use pathway information to form new links across diverse biological data highlights the utility of integrating WikiPathways in the semantic web.

4.1 Introduction

Pathway analysis and visualization of data on pathways provide insights into the underlying biology of effects found in genomics, proteomics, and metabolomics experiments [1–4]. WikiPathways is a pathway repository where content is provided by the community at large [5, 6]. In a given pathway, elements like genes, proteins, metabolites, and interactions are identified using common accession numbers from reference databases such as Entrez Gene [7], Ensembl [8], UniProt [9], HMDB [10], ChemSpider [11], PubChem [12] and ChEMBL [13]. Multiple databases can be referenced to annotate an element of the same semantic type, e.g. Ensembl and Entrez Gene to annotate gene information. Even single studies sometimes use different reference databases to annotate experimental findings. It is common for bioinformaticians to spend valuable time dealing with data mapping issues that impede the actual data analysis and interpretation. In WikiPathways we use the open-source software framework BridgeDb [14], to help resolve different identifiers representing the same (or related) entities. Capturing a semantically correct description of biological entities and their connections across datasets is the broader challenge that we have to address. The semantic web provides an approach to define entities and their relationships. By explicitly defining these entities and relationships the semantic web can provide a network of linked data [15]. The Resource Description Framework (RDF) consists of two key components: statements and universal identifiers. Each statement is captured as a triple, consisting of a subject, a predicate, and an object. For example, the following triple defines the glucose molecule as being part of the glycolysis pathway:

$$\underbrace{\text{Glycolysis}}_{\text{subject}} \quad \underbrace{\text{Has member}}_{\text{predicate}} \quad \underbrace{\text{Glucose}}_{\text{object}}$$

The notion of semantic web surfaces as you link across large sets of triples representing a vast number of objects and diverse types of concepts and predicates. The use of uniform identifiers, or URIs [16], provides consistency when specifying subjects and objects. identifiers.org [17], for example, provides a

clearinghouse for a wide variety of URIs for biological entities in the life science domain. WikiPathways provides identifiers for all its pathways and identifiers.org provides the URI scheme to make these resolvable. Standardized URIs for predicates come from efforts such as the Simple Knowledge Organization System (SKOS) [18]. For example, our example triple above can be expressed in a more universal way:

$$\underbrace{\text{<http://identifiers.org/wikipathways/WP534>}}_{\text{subject}} \quad \underbrace{\text{<http:// www.w3.org/2004/02/skos/coremember>}}_{\text{predicate}} \quad \underbrace{\text{<http://identifiers.org/chebi/CHEBI:4167>}}_{\text{object}}$$

where each element is uniquely and universally resolvable to a defined concept (glycolysis, ”has member”, and glucose respectively). Of course, the more human readable information can also be explicitly added by describing the labels in RDF. But that information is also available by resolving the URIs.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX wp: <http://identifiers.org/wikipathways/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX chebi: <http://identifiers.org/chebi/CHEBI:>

wp:WP534 skos:member chebi:4167.
wp:WP534 rdfs:label ''Glycolysis and
                    Gluconeogenesis (Homo sapiens''@en.
chebi:4167 rdfs:label ''Glucose''@en.
```

In order to contribute pathway knowledge to the semantic web, we have modeled the content of WikiPathways to form triple-based statements. The interactions and reactions curated at WikiPathways are particularly well-suited to enrich the overall connectivity of the semantic web. Pathways offer a meaningful context for relations between biological entities, such as proteins,

metabolites and diseases that are otherwise defined in disparate databases. We report on the conversion process and the development of two new vocabularies essential in capturing the semantics behind pathway diagrams. Finally, we evaluate the use of the semantically linked pathway knowledge through specialized queries and third-party resources, showing how to link WikiPathways with disease annotations (from UniProt [9] and DisGeNET [19]), with gene-expression values (from Gene Express Atlas) and with bioactive chemical compounds known to affect proteins that occur in pathways (e.g. from ChEMBL).

4.2 Results and Discussion

4.2.1 Pathway vocabularies

There are existing standards to model various aspects of pathway knowledge, such as BioPAX [20], SBGN [21], MIM [22], SBML [23] and SBO [24]. BioPAX and SBO are in fact already available in a Semantic Web-compatible language called OWL [25]. These standards provide valuable building blocks for our "WP" vocabulary that captures the biological meaning of pathways. However, not all of the graphical annotations, spatial information and other subtleties critical for the visual representation, the intuitive understanding and the usability for data visualisation of the curated content at WikiPathways are captured by these standards. Our "GPML" vocabulary directly reflects these features defined in the XML format, GPML, or Graphical Pathway Markup Language. For example, in GPML, all genes, proteins and metabolites are types of data nodes, which are rendered as a rectangular box with properties capturing among others its position, height, width, label, and external reference. For example:

```
<DataNode TextLabel="Glucose" GraphId="dba83" Type="Metabolite">
  <Graphics CenterX="279.0" CenterY="468.0" Width="112.0"
            Height="20.0" ZOrder="32768">
  <Xref Database="ChEBI" ID="CHEBI:4167" />
</DataNode>
```

In the GPML vocabulary, used for semantic representation of pathway diagrams, the markup elements and values are described as classes and properties, each with their respective URIs.

```
<http://identifiers.org/chebi/CHEBI:4167> rdf:type gpml:DataNode.
<http://identifiers.org/chebi/CHEBI:4167> rdfs:label "Glucose"@en.
<http://identifiers.org/chebi/CHEBI:4167> gpml:graphId "dba83".
<http://identifiers.org/chebi/CHEBI:4167> gpml:ZOrder 32768.
```

The GPML vocabulary, in its current form, is mainly instrumental in the representation of the spatial information captured at WikiPathways. However, as we will describe below it can also be used to convert pathway information from other semantic web resources into a format amenable to being rendered and curated at WikiPathways. Explicit mappings to external (graphical) ontologies are not added, however through plugins such as Pathvisio-MIM [26] mappings to graphical notations such as MIM or SBGN, are possible. In an analogous way, the WP vocabulary can be used to capture the biological relations from other pathways in such a way that they can be used in resources using this semantic layer of the WikiPathways RDF. We used this approach for example to make the relations from Reactome pathways available in the Open PHACTS discovery platform [27] starting from the converted pathways at WikiPathways.

The WP vocabulary, focusing on biological meaning, issues URIs for biological concepts and disregards layout and other rendering details. Using URIs from this vocabulary allows stating that something is a Pathway, or that a DataNode is a chemical compound or gene product. The vocabulary also captures descriptive elements, such as labels, shapes and lines that help annotate and contextualize the pathway reaction details. The RDF generated consist of terms from the vocabularies developed in this context. This is done to be able to reflect the semantics used in the WikiPathways community. However, to allow integration with external pathway resources—which is the primary objective of this project—we need to link to external ontologies. For the subset of concepts in common with prior vocabularies, such as BioPAX, we utilize the SKOS data model to express a range of similarities from skos:exactMatch to skos:closeMatch [18, 28].

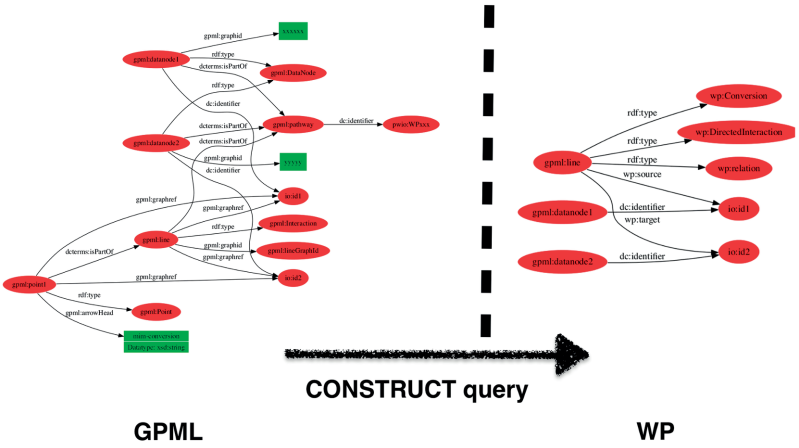


Figure 4.1: A construct query is type of SPARQL query that enables the conversion of one graph pattern to another. Here an interaction described by its spatial properties (GPML) is converted into a semantic representation reflecting its biological interpretation (WP). The SPARQL query is available in the supporting information section.

4.2.2 Pathway conversion and queries

With these vocabularies in place, the next step is the actual conversion of GPML files into triples using the GPML vocabulary. Then rules are applied to make the biological meaning explicit using the WP vocabulary. For example a directed interaction is captured in GPML as two ”DataNodes”, a line and an arrowhead. The ”DataNodes” have external references as properties. Rules are then applied to state that a line is a Directed Interaction, with a source and a target. Figure 4.1 contains an example of such a rule based reasoning query that issues triples with URIs from the WP vocabulary.

WikiPathways pathways are regularly curated by a team of volunteers that evaluate their usability for analysis and tag the pathways as ”curated”. WikiPathways contains 1000 pathways in the curated set across over a dozen species that convert to a total of 1.6 million triples. The triples are loaded in a SPARQL endpoint (*sparql.wikipathways.org*), which allows

semantic querying of the data with the SPARQL query language [29]. RDF, including new and updated pathways, is generated and tested regularly and can be delivered upon request. Updates of the RDF that is available for download and in the SPARQL endpoint are triggered by crucial events, such as Reactome or Open PHACTS data releases. This prevents discrepancies in quality control or curation, due to small differences between (frequent) releases. Example SPARQL queries and their plain language translations are given in Table 1. A broad set of approximately 50 queries is available on the help pages of WikiPathways [30].

A *federated* SPARQL query [29] enables querying over multiple SPARQL endpoints. With a variety of SPARQL endpoints available with data on disease annotations (e.g. DisGeNET and UniProt), significantly expressed genes (e.g. EBI Expression Atlas) and drug-target interactions (e.g. ChEMBL), knowledge from these remote SPARQL endpoints can be integrated. Example queries are given in Table 4.2 and on the help pages of WikiPathways [30]

4.2.3 Using linked data in common analysis platforms

Different common analysis platforms allow the integration of linked data for future analysis and visualization. One nice example of such an analysis platform is R, a widely used software environment for statistical computing and graphics. R has a SPARQL library [29], which enables the import of linked data for further processing in R. This allows for running common statistical tests or the creation of different visualization of linked data. We recently published an R library that interfaces R with PathVisio [31] and allows manipulation of pathways and data visualisation on pathways. Fig. 2 shows up and down-regulated genes in Diabetes Mellitus (efo:EFO˙0000400, efo:EFO˙0001359, and efo:EFO˙0001360) in the pathway diagram on insulin signalling in human [30]. This pathway diagram with color-coding parts indicating up- and down-regulated pathway elements was created by integrating knowledge from two geographically dispersed and independent resources, through a single SPARQL query embedded in a R script, which is available online [32].

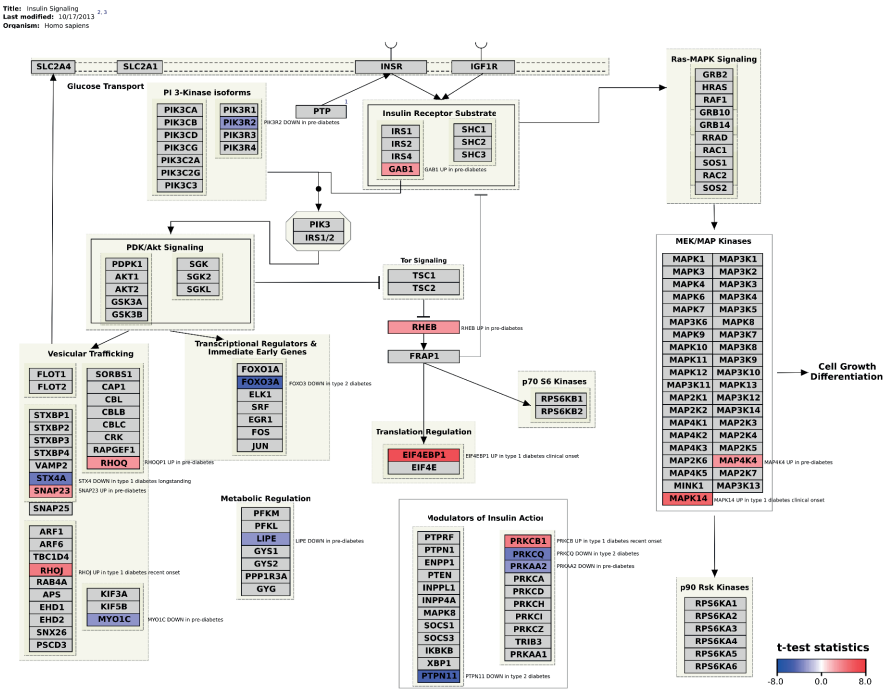


Figure 4.2: The colored boxes represent genes which are up (blue) or down (yellow) regulated in diabetes mellitus. PIK3R2, MYO1C, PRKAA2, LIPE are downregulated in pre-diabetes. STX4A is downregulated in type 1 diabetes longstanding. PRKCQ, PTPN11, FOXO3A are downregulated in type 2 diabetes. GAB1, RHEB, MAP4K4, SNAP23 are up regulated in pre-diabetes. RHOJ, PRKCB are upregulated in type 1 diabetes recent onset. MAPK14UP, EIF4EBP1 are upregulated in type 1 diabetes clinical onset. From these 17 up or down-regulated genes, 9 are being reported as being in the top 10 disease and phenotype associations for the selected gene in DisGeNET (i.e. PIK3R2, PRKAA2, LIPE, STX4A, PRKCQ, FOXO3A, MAP4K4, SNAP23, and PRKCB) (Gene-disease association data were retrieved from the DisGeNET Database, GRIB/IMIM/UPF Integrative Biomedical Informatics Group, Barcelona. (*www.disgenet.org*). 04, 2016)

Rosetta stone function

A number of resources provide content from multiple pathway databases, including Pathway Commons [33] and NCBI's BioSystems (`http://ncbi.org/biosystems`). While BioPAX in fact is RDF, the NCBI system is not. NCBI BioSystems uses NCBI's native identifiers: GeneId, ProteinId, CID. We thus have a resource with pathways from different origins that are already described in the same way. Since for WikiPathways content we know how the different entities in these resources map to the GPML and WP vocabularies we can now use that to produce RDF using these same ontologies for each of the other pathway resources present in NCBI BioSystems. In fact, we can do the same for Pathway Commons where this approach will lead to an improved version of RDF with explicit mappings to the WP vocabulary. We made a prototype script available on GitHub to be used for this type of conversion from BioSystems [32].

Use in discovery platforms

The semantically linked pathway data from WikiPathways RDF have also been integrated into the Open PHACTS discovery platform [27, 34]. Open PHACTS delivers and sustains an open pharmacological space using semantic web standards and technologies. The Open PHACTS platform currently provides 51 API methods of which thirteen deliver pathway information (`https://dev.openphacts.org/docs`). Other information collected in Open PHACTS describes other relationships like drug-target (from ChEMBL) and protein interaction (from UniProt). Having this all-in-one resource combined with a set of mapping tools allows fast analysis across the domains. By combining Open PHACTS API calls one can, for instance, find all protein targets for a drug and then all pathways that contain these targets.

List the species captured in WikiPathways and the number of pathways per species	**SELECT DISTINCT** ?organism ?label
	count(?pathway) as ?numberOfPathways
	WHERE {
	?pathway dc:title ?title .
	?pathway wp:organism ?organism .
	?pathway wp:organismName ?label .
	?pathway rdf:type wp:Pathway .
	}
	ORDER BY DESC(?numberOfPathways)

Get all gene products on a particular pathway (WP615 as an example)	**SELECT DISTINCT** ?pathway ?label
	WHERE {
	?geneProduct **a** wp:GeneProduct .
	?geneProduct rdfs:label ?label .
	?geneProduct dcterms:isPartOf
	?pathway .
	?pathway rdf:type wp:Pathway .
	FILTER regex(**str**(?pathway), "WP615").
	}

Return all PubChem compounds in WikiPathways and the pathways they are in	**SELECT DISTINCT** ?identifier ?pathway
	WHERE {
	?concept dcterms:isPartOf ?pathway .
	?concept dc:source
	"PubChem-compound"^^xsd:string .
	?concept dc:identifier ?identifier .
	?pathway rdf:type wp:Pathway
	}

Table 4.1: Example queries handled by the WikiPathways SPARQL endpoint

From DisGeNET get disease-gene pairs on asthma and get all pathways where these genes have a role

```
PREFIX identifiers: <http://identifiers.org/ensembl/>
PREFIX atlas: <http://rdf.ebi.ac.uk/resource/atlas/>
PREFIX efo: <http://www.ebi.ac.uk/efo/>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
SELECT DISTINCT ?wpId ?pwtitle
  (group_concat(distinct ?wpgene_identifier ;separator="; ")
   as ?wpgenes)
WHERE {
  SERVICE <http://rdf.disgenet.org/sparql/> {
    GRAPH <http://rdf.disgenet.org> {
      ?gda sio:SIO_000628 ?gene,?disease .
      ?gene rdf:type ncit:C16612 ;
        rdfs:label ?geneLabel .
      ?disease rdf:type ncit:C7057 ;
        rdfs:label ?diseaseLabel .
      FILTER regex(?diseaseLabel, "asthma", "i")
      ?gene sio:SIO_010078 ?protein .
    }
  }
  ?wpgene wp:bdbEntrezGene ?gene .
  ?wpgene dcterms:identifier ?wpgene_identifier .
  ?wpgene dcterms:isPartOf ?pathway .
  ?pathway a wp:Pathway .
  ?pathway dc:identifier ?wpId .
  ?pathway dc:title ?pwtitle .
}
```

For the genes differentially expressed in asthma (found in the EBI Expression Atlas), get the gene products associated to a WikiPathways pathway

```
PREFIX identifiers: <http://identifiers.org/ensembl/>
PREFIX atlas: <http://rdf.ebi.ac.uk/resource/atlas/>
PREFIX atlasterms: <http://rdf.ebi.ac.uk/terms/atlas/>
PREFIX efo: <http://www.ebi.ac.uk/efo/>
SELECT DISTINCT ?wpURL ?pwTitle ?Ensembl ?EntrezGene ?expressionValue ?pvalue WHERE {
        SERVICE <https://www.ebi.ac.uk/rdf/services/atlas/sparql> {
              ?factor rdf:type efo:EFO_0000270 .
              ?value atlasterms:hasFactorValue ?factor .
              ?value atlasterms:isMeasurementOf ?probe .
              ?value atlasterms:pValue ?pvalue .
              ?value rdfs:label ?expressionValue .
              ?probe atlasterms:dbXref ?dbXref .
        }
              ?pwElement dcterms:isPartOf ?pathway .
              ?pathway dc:title ?pwTitle .
              ?pathway dc:identifier ?wpURL .
              ?pwElement wp:bdbEnsembl ?Ensembl .
                  ?pwElement wp:bdbEntrezGene ?EntrezGene .
        }
ORDER BY ASC(?pvalue)
```

Table 4.2: Example federated queries handled by the WikiPathways SPARQL endpoint

Materials and Methods

Use of Open PHACTS RDF guidelines

In collaboration with partners in the Open PHACTS project, we proposed guidelines for presenting data as RDF [35], most of that can be considered as general guidelines to produce RDF in the biomedical domain. The guidelines consist of a prerequisite and 11 steps, covering the licensing (step 0), designing (steps 1-5), implementation (steps 6-9), and presentation (steps 10-11) of the data in the semantic web. In the work presented here, we follow these steps:

Licensing WikiPathways content is covered by the Creative Commons Attribution 3.0 Unported license (*creativecommons.org/licenses/by/3.0/*). This is stated in the VoID headers of the RDF made. These headers are automatically generated by the same script generating the WikiPathways RDF. Open PHACTS provides a template for these header files.

Implementation We used a Java RDF framework, Jena (*jena.apache.org*) [36], to generate the RDF for WikiPathways. The pathway diagrams were obtained through the web services of WikiPathways, after which they were converted into RDF with the Jena RDF framework. The code of the serializer is available on GitHub (*github.com/wikipathways/wp2lod*). The vocabularies were generated with a vocabulary framework called Deri Neologism (*neologism.deri.ie*).

Presentation The resulting RDF triples are available from (rdf.wikipathways.org) and loaded on an instance of the Virtuoso Open-Source Edition (*virtuoso.openlinksw.com/*) and available through its SPARQL endpoint at *sparql.wikipathways.org*. The triples are also loaded on the Open PHACTS discovery platform (*dev.openphacts.org/docs/1.5*) where they can be accessed through eleven API calls.

Identifier mapping

In the context of the semantic web, it is impractical to burden query writers with handling identifier mapping per resource and per query. Rather, the map-

ping results themselves need to become part of the semantic web. We applied two distinct approaches to addressing identifier mapping in our WikiPathways and Open PHACTS projects.

Query expansion

The Open PHACTS framework provides query expansion functionality through its Identifier Mappings Services. When an identifier is queried the SPARQL query is enriched with all possible identifiers to retrieve an expanded set of related entities. This approach is the most efficient in terms of the number of triples, since it requires only a single identifier per relationship, eliminating redundancy. However, it also requires a hosted identifier mapping service that it called along with every query.

Unified identifiers

In the case of WikiPathways, which does not host a mapping service, we chose a unified identifier approach, where all identifiers are mapped ahead of time to a set of common identifier systems. In this way, the database effectively contains the results of a limited number of identifier mappings in the form of partially redundant triples. For example, in the WikiPathways RDF, all identifiers have been unified to Entrez Gene [7] (`wp:bdbEntrezGene`), Ensembl [8] (`wp:bdbEnsembl`), UniProt [37] (`wp:bdbUniprot`) for gene products and HMDB [10] (`wp:bdbHmdb`), and ChemSpider [11] (`wp:bdbChemspider`) for compounds like metabolites and drugs. The original identifier provided by the pathway curator is stored as a triple, with the predicate dc:identifier, and a URI from identifiers.org, which points to both the identifier and the resource.

Summary

We present a semantic web representation of WikiPathways together with vocabularies needed to cover the graphical pathway layout and the biological

meaning and solutions to map between different identifier systems. The public availability allows rapid integration with other biological resources. The availability of two vocabularies allows to convert between different pathways resources. Different analytical tools now support the import of semantic web data, allowing integrated use of data from different resources with a single query. We demonstrate this with a federated query across multiple resources where the resulting differentially expressed genes for a disease where shown on a discovered pathway using PathVisio.

Availability

The following resources are publically available as beta releases just like WikiPathways. They are maintained as part of the open-source WikiPathways project

Vocabularies

- GPML: `http://vocabularies.wikipathways.org/gpml`
- WP: `http://vocabularies.wikipathways.org/wp`

Wikipathways on the Semantic Web

- SPARQL endpoint: `http://sparql.wikipathways.org`
- Open PHACTS: `https://dev.openphacts.org/docs/`
- RDF download: `http://rdf.wikipathways.org`

Source code

- GitHub: `https://github.com/wikipathways/wp2lod`

References

[1] Danyel G. J. Jennen et al. “Biotransformation pathway maps in WikiPathways enable direct visualization of drug metabolism related expression changes”. *Drug Discov Today*. 2010. 15 (19-20): pp. 851–858.

[2] Purvesh Khatri, Marina Sirota, and Atul J. Butte. “Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges”. *PLoS Comput Biol*. 2012. 8 (2): .

[3] Martijn P van Iersel et al. “Presenting and exploring biological pathways with PathVisio”. *BMC Bioinformatics*. 2008. 9: p. 399.

[4] Thomas Kelder et al. “Finding the Right Questions: Exploratory Pathway Analysis to Enhance Biological Discovery in Large Datasets”. *PLoS Biology*. 2010. 8 (8): .

[5] Thomas Kelder et al. “WikiPathways: building research communities on biological pathways”. *Nucleic Acids Res*. 2012. 40 (Database issue): pp. D1301–D1307.

[6] Martina Kutmon et al. “WikiPathways: capturing the full diversity of pathway knowledge”. *Nucleic Acids Res*. 2016. 44 (Database issue): pp. D488–D494.

[7] Donna Maglott et al. “Entrez Gene: gene-centered information at NCBI”. *Nucleic Acids Res*. 2011. 39 (Database issue): pp. D52–D57.

[8] Andrew Yates et al. “Ensembl 2016”. *Nucleic Acids Res*. 2016. 44 (Database issue): pp. D710–D716.

[9] The UniProt Consortium et al. “UniProt: the universal protein knowledgebase”. *Nucleic Acids Res*. 2017. 45 (D1): pp. D158–D169.

[10] David S. Wishart et al. “HMDB 3.0–The Human Metabolome Database in 2013”. *Nucleic Acids Res*. 2013. 41 (Database issue): pp. D801–D807.

[11] Harry E. Pence and Antony Williams. “ChemSpider: An Online Chemical Information Resource”. *J. Chem. Educ.* 2010. 87 (11): pp. 1123–1124.

[12] Sunghwan Kim et al. “PubChem Substance and Compound databases”. *Nucleic Acids Res*. 2016. 44 (Database issue): pp. D1202–D1213.

[13] A. Patrícia Bento et al. “The ChEMBL bioactivity database: an update”. *Nucleic Acids Res*. 2014. 42 (Database issue): pp. D1083–D1090.

[14] Martijn P. van Iersel et al. “The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services”. *BMC Bioinformatics*. 2010. 11 (1): p. 5.

[15] *The Semantic Web - Scientific American*. URL: https://www.scientificamerican.com/article/the-semantic-web/ (visited on 11/25/2020).

[16] Tim Berners-Lee, James Hendler, and Ora Lassila. “The semantic web”. *Scientific American*. 2001. 284 (5): pp. 34–43.

[17] Nick Juty, Nicolas Le Novère, and Camille Laibe. “Identifiers.org and MIRIAM Registry: community resources to provide persistent identification”. *Nucleic Acids Res*. 2012. 40 (Database issue): pp. D580–D586.

[18] *SKOS Simple Knowledge Organization System Reference*. URL: https://www.w3.org/TR/skos-reference/ (visited on 01/16/2021).

[19] Janet Piñero et al. “DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes”. *Database: The Journal of Biological Databases and Curation*. 2015. 2015: .

[20] Joanne S. Luciano. “PAX of mind for pathway researchers”. *Drug Discovery Today*. 2005. 10 (13): pp. 937–942.

[21] Nicolas Le Novère et al. “The Systems Biology Graphical Notation”. *Nat Biotechnol*. 2009. 27 (8): pp. 735–741.

[22] Kurt W. Kohn et al. “Molecular Interaction Maps of Bioregulatory Networks: A General Rubric for Systems Biology”. *Mol Biol Cell*. 2006. 17 (1): pp. 1–13.

[23] A. Finney and M. Hucka. “Systems biology markup language: Level 2 and beyond”. *Biochem Soc Trans*. 2003. 31 (Pt 6): pp. 1472–1473.

[24] N Juty et al. “BioModels: Content, Features, Functionality, and Use”. *CPT Pharmacometrics Syst Pharmacol*. 2015. 4 (2): .

[25] *OWL 2 Web Ontology Language Document Overview (Second Edition)*. URL: https://www.w3.org/TR/owl2-overview/ (visited on 11/25/2020).

[26] Augustin Luna et al. “PathVisio-MIM: PathVisio plugin for creating and editing Molecular Interaction Maps (MIMs)”. *Bioinformatics*. 2011. 27 (15): pp. 2165–2166.

[27] Joseline Ratnam et al. “The Application of the Open Pharmacological Concepts Triple Store (Open PHACTS) to Support Drug Discovery Research”. *PLoS One*. 2014. 9 (12): .

[28] Harry Halpin et al. “When owl:sameAs Isn’t the Same: An Analysis of Identity in Linked Data”. Paper presented at: *The Semantic Web – ISWC 2010*. Berlin, Heidelberg. Springer. 2010. pp. 305–320.

[29] *SPARQL 1.1 Query Language*. URL: https://www.w3.org/TR/sparql11-query/ (visited on 11/25/2020).

[30] *Help:WikiPathways Sparql queries - WikiPathways*. URL: https://www.wikipathways.org/index.php/Help:WikiPathways%5C%5FSparql%5C%5Fqueries (visited on 01/16/2021).

[31] Anwesha Bohler et al. “Automatically visualise and analyse data on pathways using PathVisioRPC from any programming environment”. *BMC Bioinformatics*. 2015. 16 (1): .

[32] *andrawaag/BioSystems2RDF*. URL: https://github.com/andrawaag/BioSystems2RDF (visited on 01/16/2021).

[33] Ethan G. Cerami et al. “Pathway Commons, a web resource for biological pathway data”. *Nucleic Acids Research*. 2011. 39 (Database issue): pp. D685–D690.

[34] Antony J. Williams et al. “Open PHACTS: semantic interoperability for drug discovery”. *Drug Discov Today*. 2012. 17 (21-22): pp. 1188–1198.

[35] Carina Haupt et al. *Guidelines for exposing data as RDF in Open PHACTS*. Oct. 2013. URL: http://www.openphacts.org/specs/2013/WD-rdfguide-20131007/.

[36] B. McBride. “Jena: a semantic Web toolkit”. *IEEE Internet Computing*. 2002. 6 (6): pp. 55–59.

[37] . “UniProt: a hub for protein information”. *Nucleic Acids Res*. 2015. 43 (Database issue): pp. D204–D212.

5

Wikidata as a knowledge graph for the life sciences

Adapted from: Andra Waagmeester et al. “Wikidata as a knowledge graph for the life sciences”. *eLife*. 2020. 9 (e52614): e52614.

Abstract

Wikidata is a community-maintained knowledge base that has been assembled from repositories in the fields of genomics, proteomics, genetic variants, pathways, chemical compounds, and diseases, and that adheres to the FAIR principles of findability, accessibility, interoperability and reusability. Here we describe the breadth and depth of the biomedical knowledge contained within Wikidata, and discuss the open-source tools we have built to add information to Wikidata and to synchronize it with source databases. We also demonstrate several use cases for Wikidata, including the crowdsourced curation of biomedical ontologies, phenotype-based diagnosis of disease, and drug repurposing.

5.1 Introduction

Integrating data and knowledge is a formidable challenge in biomedical research. Although new scientific findings are being discovered at a rapid pace, a large proportion of that knowledge is either locked in data silos (where integration is hindered by differing nomenclature, data models, and licensing terms; or locked away in free-text. The lack of an integrated and structured version of biomedical knowledge hinders efficient querying or mining of that information, thus preventing the full utilization of our accumulated scientific knowledge.

Recently, there has been a growing emphasis within the scientific community to ensure all scientific data are FAIR – FINDABLE, ACCESSIBLE, INTEROPERABLE, and REUSABLE – and there is a growing consensus around a concrete set of principles to ensure FAIRness [1, 2]. Widespread implementation of these principles would greatly advance efforts by the open-data community to build a rich and heterogeneous network of scientific knowledge. That knowledge network could, in turn, be the foundation for many computational tools, applications and analyses.

Most data- and knowledge-integration initiatives fall on either end of a spectrum. At one end, centralized efforts seek to bring multiple knowledge sources into a single database (see, for example, Mungall et al., 2017 [3]): this approach has the advantage of data alignment according to a common data model and of enabling high-performance queries. However, centralized resources are difficult and expensive to maintain and expand [4, 5], at least in part because of bottlenecks that are inherent in a centralized design.

At the other end of the spectrum, distributed approaches to data integration result in a broad landscape of individual resources, focusing on technical infrastructure to query and integrate across them for each query. These approaches lower the barriers to adding new data by enabling anyone to publish data by following community standards. However, performance is often an issue when each query must be sent to many individual databases, and the performance of the system as a whole is highly dependent on the stability and performance of each individual component. In addition, data integration

requires harmonizing the differences in the data models and data formats between resources, a process that can often require significant skill and effort. Moreover, harmonizing differences in data licensing can sometimes be impossible.

Here we explore the use of Wikidata (*www.wikidata.org* [6, 7] as a platform for knowledge integration in the life sciences. Wikidata is an openly-accessible knowledge base that is editable by anyone. Like its sister project Wikipedia, the scope of Wikidata is nearly boundless, with items on topics as diverse as books, actors, historical events, and galaxies. Unlike Wikipedia, Wikidata focuses on representing knowledge in a structured format instead of primarily free text. As of September 2019, Wikidataś knowledge graph included over 750 million statements on 61 million items (*tools.wmflabs.org/wikidata-todo/stats.php*). Wikidata was also the first project run by the Wikimedia Foundation (which also runs Wikipedia) to have surpassed one billion edits, achieved by a community of 12,000 active users, including 100 active computational 'bots' (Figure 1—figure supplement 1).

As a knowledge integration platform, Wikidata combines several of the key strengths of the centralized and distributed approaches. A large portion of the Wikidata knowledge graph is based on the automated imports of large structured databases via Wikidata bots, thereby breaking down the walls of existing data silos. Since Wikidata is also based on a community-editing model, it harnesses the distributed efforts of a worldwide community of contributors, including both domain experts and bot developers. Anyone is empowered to add new statements, ranging from individual facts to large-scale data imports. Finally, all knowledge in Wikidata is queryable through a SPARQL query interface (query.wikidata.org/), which also enables distributed queries across other Linked Data resources.

In previous work, we seeded Wikidata with content from public and authoritative sources of structured knowledge on genes and proteins [8] and chemical compounds. Here, we describe progress on expanding and enriching the biomedical knowledge graph within Wikidata, both by our team and by others in the community [9]. We also describe several representative biomedical use

cases on how Wikidata can enable new analyses and improve the efficiency of research. Finally, we discuss how researchers can contribute to this effort to build a continuously-updated and community-maintained knowledge graph that epitomizes the FAIR principles.

5.2 The Wikidata Biomedical Knowledge Graph

The original effort behind this work focused on creating and annotating Wikidata items for human and mouse genes and proteins [8], and was subsequently expanded to include microbial reference genomes from NCBI RefSeq [10]. Since then, the Wikidata community (including our team) has significantly expanded the depth and breadth of biological information within Wikidata, resulting in a rich, heterogeneous knowledge graph (Figure 1). Some of the key new data types and resources are described below.

5.2.1 Genes and proteins

Wikidata contains items for over 1.1 million genes and 940 thousand pro-teins from 201 unique taxa. Annotation data ongenes and proteins come from several key data-bases including NCBI Gene [11], Ensembl [12], UniProt [13], InterPro [14], and the Protein DataBank [15]. These annotations include information on protein families, gene functions, protein domains, genomic location,and orthologs, as well as links to related com-pounds, diseases, and variants.

5.2.2 Genetic variants

Annotations on genetic variants are primarily drawn from CIViC (*www.civicdb.org*), an open and community-curated database of cancer variants [16]. Variants are annotated with their relevance to disease predisposition, diagnosis, prognosis, and drug efficacy. Wikidata currently contains 1502 items corresponding to human genetic variants, focused on those with a clear clinical or therapeutic relevance.

5.2.3 Chemical compounds including drugs

Wikidata has items for over 150 thousand chemical compounds, including over 3500 items which are specifically designated as medications. Compound attributes are drawn from a diverse set of databases, including PubChem [17], RxNorm [18], the IUPHAR Guide to Pharmacology [19–21], NDF-RT (National Drug File – Reference Terminology), and LIPID MAPS [22]. These items typically contain statements describing chemical structure and key physicochemical properties, and links to databases with experimental data, such as MassBank [23, 24] and PDB Ligand [25]), and toxicological information, such as the EPA CompTox Dashboard [26]. Additionally, these items contain links to compound classes, disease indications, pharmaceutical products, and protein targets.

5.2.4 Pathways

Wikidata has items for almost three thousand human biological pathways, primarily from two established public pathway repositories: Reactome [27] and WikiPathways [28]. The full details of the different pathways remain with the respective primary sources. Our bots enter data for Wikidata properties such as pathway name, identifier, organism, and the list of component genes, proteins, and chemical compounds. Properties for contributing authors (via ORCID properties; [29]), descriptions and ontology annotations are also being added for Wikidata pathway entries.

5.2.5 Diseases

Wikidata has items for over 16 thousand diseases, the majority of which were created based on imports from the Human Disease Ontology [30], with additional disease terms added from the Monarch Disease Ontology [3]. Disease attributes include medical classifications, symptoms, relevant drugs, as well as subclass relationships to higher-level disease categories. In instances where the Human Disease Ontology specifies a related anatomic region and/or a causative organism (for infectious diseases), corresponding statements are also added.

5.2.6 References

Whenever practical, the provenance of each statement added to Wikidata was also added in a structured format. References are part of the core data model for a Wikidata statement. References can either cite the primary resource from which the statement was retrieved (including details like version number of the resource), or they can link to a Wikidata item corresponding to a publication as provided by a primary resource (as an extension of the WikiCite project; [31], or both. Wikidata contains over 20 million items corresponding to publications across many domain areas, including a heavy emphasis on biomedical journal articles.

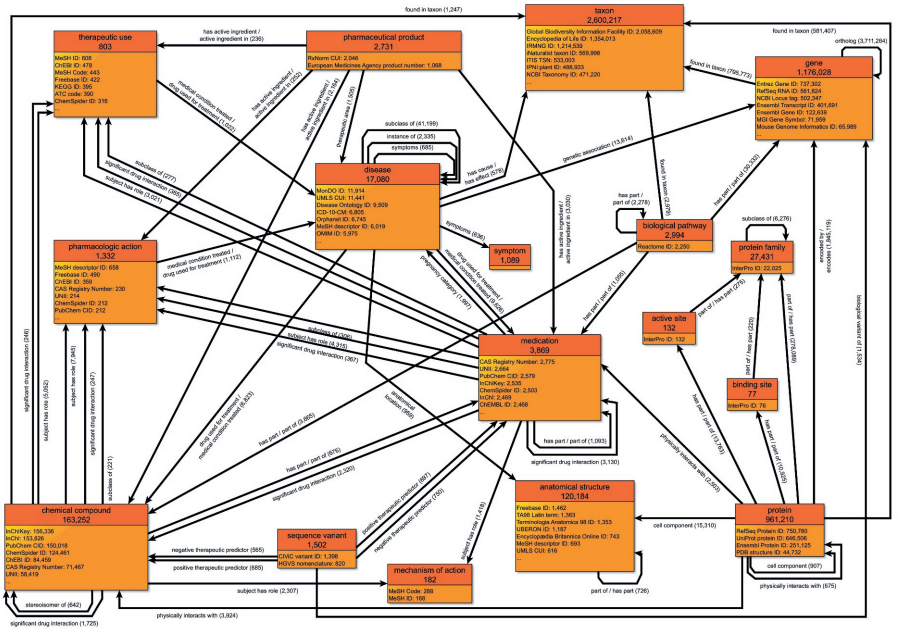


Figure 5.1: A simplified class-level diagram of the Wikidata knowledge graph for biomedical entities. Each box represents one type of biomedical entity. The header displays the name of that entity type (e.g., pharmaceutical product) and the number of Wikidata items for that entity type. The lower portion of each box displays a partial listing of attributes about each entity type and the number of Wikidata items for each attribute. Edges between boxes represent the number of Wikidata statements corresponding to each combination of subject type, predicate, and object type. For example, there are 1505 statements with 'pharmaceutical product' as the subject type, 'therapeutic area' as the predicate, and 'disease' as the object type. For clarity, edges for reciprocal relationships (e.g., 'has part' and 'part of') are combined into a single edge, and scientific articles (which are widely cited in statement references) have been omitted. All counts of Wikidata items are current as of September 2019. The most common data sources cited as references are available in Figure 5.2. Data are generated using the code in *github.com/SuLab/genewikiworld* [32]. A more complete version of this graph diagram can be found at *commons.wikimedia.org/wiki/File:Biomedical˙Knowledge˙Graph˙in˙Wikidata.svg* Most frequent data sources cited as references for the biomedical subset of the Wikidata knowledge graph shown here: *cdn.elifesciences.org/articles/52614/elife-52614-fig1-data1-v1.csv*

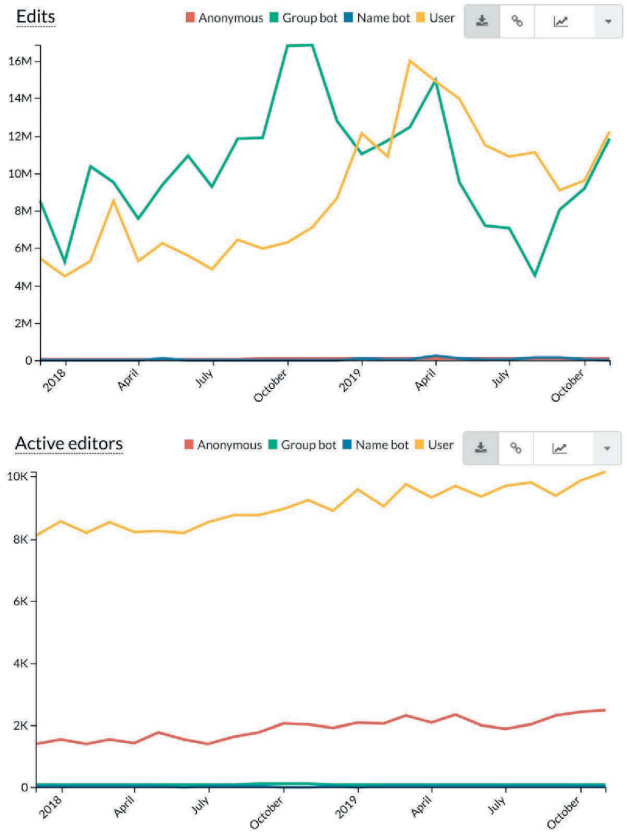


Figure 5.2: Trends in Wikidata edits. Wikidata edits are categorized into four categories: anonymous edits with no user account ('anonymous'), edits from formally registered bots ('group bot'), edits from user accounts that are presumed to be bots based on the user account name ('name bot'), and all other edits from registered, logged-in users. The top graph shows that Wikidata receives substantial contributions from both automated bots and individual users. While the overall number of edits is relatively balanced between these two groups, the lower graph shows that the number of user accounts is much higher than the number of automated bot accounts. Statistics are shown for the periods between December 2017 through December 2019. More statistics are available at *stats.wikimedia.org/v2/#/wikidata.org*

5.3 Bot automation

To programmatically upload biomedical knowledge to Wikidata, we developed a series of computer programs, or bots. Bot development began by reaching a consensus on data modeling with the Wikidata community, particularly the Molecular Biology WikiProject. We then coded each bot to retrieve, transform, normalize and upload data from a primary resource to Wikidata via the Wikidata application programming interface (API).

We generalized the common code modules into a Python library, called Wikidata Integrator (WDI), to simplify the process of creating Wikidata bots *(github.com/SuLab/WikidataIntegrator)*; archived at Burgstaller-Muehlbacher et al., 2020). Relative to accessing the API directly, WDI has convenient features that improve the bot development experience. These features include the creation of items for scientific articles as references, basic detection of data model conflicts, automated detection of items needing update, detailed logging and error handling, and detection and preservation of conflicting human edits.

Just as important as the initial data upload is the synchronization of updates between the primary sources and Wikidata. We utilized Jenkins, an open-source automation server, to automate all our Wikidata bots. This system allows for flexible scheduling, job tracking, dependency management, and automated logging and notification. Bots are either run on a predefined schedule (for continuously updated resources) or when new versions of original databases are released.

5.4 Applications of Wikidata

Translating between identifiers from different databases is one of the most common operations in bioinformatics analyses. Unfortunately, these translations are most often done by bespoke scripts and based on entity-specific mapping tables. These translation scripts are repetitively and redundantly written across our community and are rarely kept up to date, nor integrated in a reusable fashion.

An identifier translation service is a simple and straightforward application of the biomedical content in Wikidata. Based on mapping tables that have been imported, Wikidata items can be mapped to databases that are both widely- and rarely-used in the life sciences community. Because all these mappings are stored in a centralized database and use a systematic data model, generic and reusable translation scripts can easily be written (Figure 5.3). These scripts can be used as a foundation for more complex Wikidata queries, or the results can be downloaded and used as part of larger scripts or analyses.

There are a number of other tools that are also aimed at solving the identifier translation use case, including the BioThings APIs [33], BridgeDb [34], BioMart [35], UMLS [36], and NCI Thesaurus [37]. Relative to these tools,

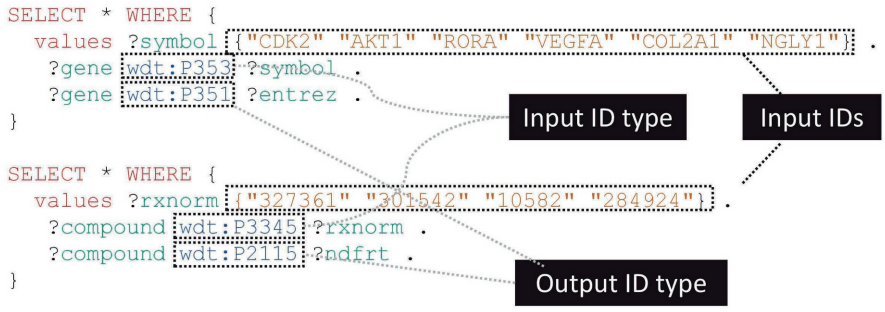


Figure 5.3: Generalizable SPARQL template for identifier translation. SPARQL is the primary query language for accessing Wikidata content. These simple SPARQL examples show how identifiers of any biological type can easily be translated using SPARQL queries. The top query demonstrates the translation of a small list of gene symbols (wdt:P353) to Entrez Gene IDs (wdt:P351), while the bottom example shows conversion of RxNorm concept IDs (wdt:P3345) to NDF-RT IDs (wdt:P2115). These queries can be submitted to the Wikidata Query Service (WDQS; https://query.wikidata.org/) to get real-time results. Translation to and from a wide variety of identifier types can be performed using slight modifications on these templates, and relatively simple extensions of these queries can filter mappings based on the statement references and/or qualifiers. A full list of Wikidata properties can be found at https://www.wikidata.org/wiki/Special:ListProperties. Note that for translating a large number of identifiers, it is often more efficient to perform a SPARQL query to retrieve all mappings and then perform additional filtering locally.

Wikidata distinguishes itself with a unique combination of the following: an almost limitless scope including all entities in biology, chemistry, and medicine; a data model that can represent exact, broader, and narrow matches between items in different identifier namespaces (beyond semantically imprecise 'cross-references'); programmatic access through web services with a track record of high performance and high availability.

Moreover, Wikidata is also unique as it is the only tool that allows real-time community editing. So while Wikidata is certainly not complete with respect to identifier mappings, it can be continually improved independent of any centralized effort or curation authority. As a database of assertions and not of absolute truth, Wikidata is able to represent conflicting information (with provenance) when, for example, different curation authorities produce different mappings between entities. (However, as with any bioinformatics integration exercise, harmonization of cross-references between resources can include relationships other than 'exact match'. These instances can lead to Wikidata statements that are not explicitly declared, but rather the result of transitive inference.)

5.5 Integrative Queries

Wikidata contains a much broader set of information than just identifier cross-references. Having biomedical data in one centralized data resource facilitates powerful integrative queries that span multiple domain areas and data sources. Performing these integrative queries through Wikidata obviates the need to perform many time-consuming and error-prone data integration steps.

As an example, consider a pulmonologist who is interested in identifying candidate chemical compounds for testing in disease models (schematically illustrated in Figure 5.4). They may start by identifying genes with a genetic association to any respiratory disease, with a particular interest in genes that encode membrane-bound proteins (for ease in cell sorting). They may then look for chemical compounds that either directly inhibit those proteins, or finding none, compounds that inhibit another protein in the same pathway.

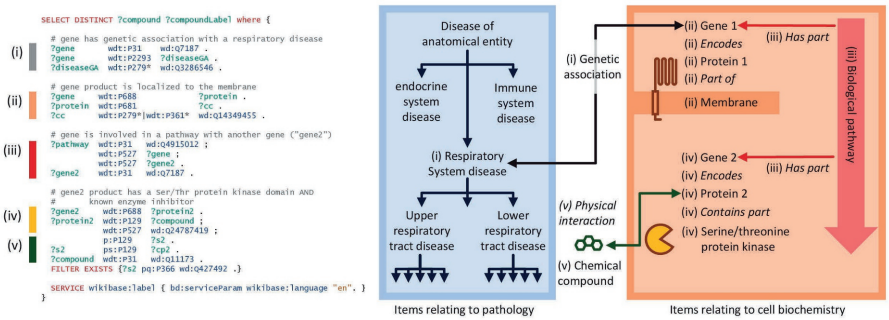


Figure 5.4: A representative SPARQL query that integrates data from multiple data resources and annotation types. This example integrative query incorporates data on genetic associations to disease, Gene Ontology annotations for cellular compartment, protein target information for compounds, pathway data, and protein domain information. Specifically, this query (depicted schematically at right) retrieves genes that are (i) associated with a respiratory system disease, (ii) that encode a membrane-bound protein, and (iii) that sit within the same biochemical pathway as (iv) a second gene encoding a protein with a serine-threonine kinase domain and (v) a known inhibitor, and reports a list of those inhibitors. Aspects related to Disease Ontology in blue; aspects related to biochemistry in red/orange; aspects related to chemistry in green. Properties are shown in italics. Real-time query results can be viewed at w.wiki/6pZ

.

Because they have collaborators with relevant expertise, they may specifically filter for proteins containing a serine-threonine kinase domain.

Almost any competent informatician can perform the query described above by integrating cell localization data from Gene Ontology annotations, genetic associations from GWAS Catalog, disease subclass relationships from the Human Disease Ontology, pathway data from WikiPathways and Reactome, compound targets from the IUPHAR Guide to Pharmacology, and protein domain information from InterPro. However, actually performing this data integration is a time-consuming and error-prone process. At the time of publication of this manuscript, this Wikidata query completed in less than 10 s and reported 31 unique compounds. Importantly, the results of that query will always be up-to-date with the latest information in Wikidata.

This query, and other example SPARQL queries that take advantage of the rich, heterogeneous knowledge network in Wikidata are available at `https://www.wikidata.org/wiki/User:ProteinBoxBot/SPARQL_Examples`. That page additionally demonstrates federated SPARQL queries that perform complex queries across other biomedical SPARQL endpoints. Federated queries are useful for accessing data that cannot be included in Wikidata directly due to limitations in size, scope, or licensing.

5.6 Crowdsourced curation

Ontologies are essential resources for structuring biomedical knowledge. However, even after the initial effort in creating an ontology is finalized, significant resources must be devoted to maintenance and further development. These tasks include cataloging cross references to other ontologies and vocabularies, and modifying the ontology as current knowledge evolves. Community curation has been explored in a variety of tasks in ontology curation and annotation (see, for example, Bunt et al., 2012; Gil et al., 2017; Putman et al., 2019; Putman et al., 2017; Wang et al., 2016 [10, 38–41]). While community curation offers the potential of distributing these responsibilities over a wider set of scientists, it also has the potential to introduce errors and inconsistencies.

Here, we examined how a crowd-based curation model through Wikidata works in practice. Specifically, we designed a hybrid system that combines the aggregated community effort of many individuals with the reliability of expert curation. First, we created a system to monitor, filter, and prioritize changes made by Wikidata contributors to items in the Human Disease Ontology. We initially seeded Wikidata with disease items from the Disease Ontology (DO) starting in late 2015. Beginning in 2018, we compared the disease data in Wikidata to the most current DO release on a monthly basis.

In our first comparison between Wikidata and the official DO release, we found that Wikidata users added a total of 2030 new cross references to GARD

[42] and MeSH (*www.nlm.nih.gov/mesh/meshhome.html*). These cross references were primarily added by a small handful of users through a web interface focused on identifier mapping (Mix'n'match, (tools.wmflabs.org/mix-n-match/#/). Each cross reference was manually reviewed by DO expert curators, and 2007 of these mappings (98.9%) were deemed correct and therefore added to the ensuing DO release. 771 of the proposed mappings could not be easily validated using simple string matching, and 754 (97.8%) of these were ultimately accepted into DO. Each subsequent monthly report included a smaller number of added cross references to GARD and MeSH, as well as ORDO [43], and OMIM [44, 45], and these entries were incorporated after expert review at a high approval rate (>90%).

Addition of identifier mappings represents the most common community contribution, and likely the most accessible crowdsourcing task. However, Wikidata users also suggested numerous refinements to the ontology structure, including changes to the subclass relationships and the addition of new disease terms. These structural changes were more nuanced and therefore rarely incorporated into DO releases with no modifications. Nevertheless, they often prompted further review and refinement by DO curators in specific subsections of the ontology.

The Wikidata crowdsourcing curation model is generalizable to any other external resource that is automatically synced to Wikidata. The code to detect changes and assemble reports is tracked online at *github.com/SuLab/scheduled-botsgithub.com/SuLab/scheduled-bots* (archived at [46]) and can easily be adapted to other domain areas. This approach offers a novel solution for integrating new knowledge into a biomedical ontology through distributed crowdsourcing while preserving control over the expert curation process. Incorporation into Wikidata also enhances exposure and visibility of the resource by engaging a broader community of users, curators, tools, and services.

5.6.1 Integrative pathway pages

In addition to its use as a repository for data, we explored the use of Wikidata as a primary access and visualization endpoint for pathway data. We used Scholia, a web app for displaying scholarly profiles for a variety of Wikidata entries, including individual researchers, research topics, chemicals, and proteins [47]. Scholia provides a more user-friendly view of Wikidata content with context and interactivity that is tailored to the entity type.

We contributed a Scholia profile template specifically for biological pathways. In addition to essential items such as title and description, these pathway pages include an interactive view of the pathway diagram collectively drawn by contributing authors. The WikiPathways identifier property in Wikidata informs the Scholia template to source a pathway-viewer widget from Toolforge (*tools.wmflabs.org/admin/tool/pathway-viewer*) that in turn retrieves the corresponding interactive pathway image. Embedded into the Scholia pathway page, the widget provides pan and zoom, plus links to gene, protein and chemical Scholia pages for every clickable molecule on the pathway diagram see, for example tools.wmflabs.org/scholia/pathway/Q29892242. Each pathway page also includes information about the pathway authors. The Scholia template also generates a participants table that shows the genes, proteins, metabolites, and chemical compounds that play a role in the pathway, as well as citation information in both tabular and chart formats.

With Scholia template views of Wikidata, we were able to generate interactive pathway pages with comparable content and functionality to that of dedicated pathway databases. Wikidata provides a powerful interface to access these biological pathway data in the context of other biomedical knowledge, and Scholia templates provide rich, dynamic views of Wikidata that are relatively simple to develop and maintain.

5.7 Phenotype based disease diagnosis

Phenomizer is a web application that suggests clinical diagnoses based on an array of patient phenotypes [48]. On the back end, the latest version of Phen-

omizer uses BOQA, an algorithm that uses ontological structure in a Bayesian network [49] . For phenotype-based disease diagnosis, BOQA takes as input a list of phenotypes (using the Human Phenotype Ontology [HPO; [50]]) and an association file between phenotypes and diseases. BOQA then suggests disease diagnoses based on semantic similarity [48]. Here, we studied whether phenotype-disease associations from Wikidata could improve BOQA's ability to make differential diagnoses for certain sets of phenotypes. We modified the BOQA codebase to accept arbitrary inputs and to be able to run from the command line (code available at *github.com/SuLab/boqa*; archived at [51]) and also wrote a script to extract and incorporate the phenotype-disease annotations in Wikidata (code available at *github.com/SuLab/Wikidata-phenomizer*; archived at [52]).

As of September 2019, there were 273 phenotype-disease associations in Wikidata that were not in the HPO's annotation file (which contained a total of 172,760 associations). Based on parallel biocuration work by our team, many of these new associations were related to the disease Congenital Disorder of Deglycosylation (CDDG; also known as NGLY-1 deficiency) based on two papers describing patient phenotypes [53, 54]. To see if the Wikidata-sourced annotations improved the ability of BOQA to diagnose CDDG, we ran our modified version using the phenotypes taken from a third publication describing two siblings with suspected cases of CDDG [55]. Using these phenotypes and the annotation file supplemented with Wikidata-derived associations, BOQA returned a much stronger semantic similarity to CDDG relative to the HPO annotation file alone (Figure 4). Analyses with the combined annotation file reported CDDG as the top result for each of the past 14 releases of the HPO annotation file, whereas CDDG was never the top result when run without the Wikidata-derived annotations.

This result demonstrated an example scenario in which Wikidata-derived annotations could be a useful complement to expert curation. This example was specifically chosen to illustrate a favorable case, and the benefit of Wikidata would likely not currently generalize to a random sampling of other diseases. Nevertheless, we believe that this proof-of-concept demonstrates the value of

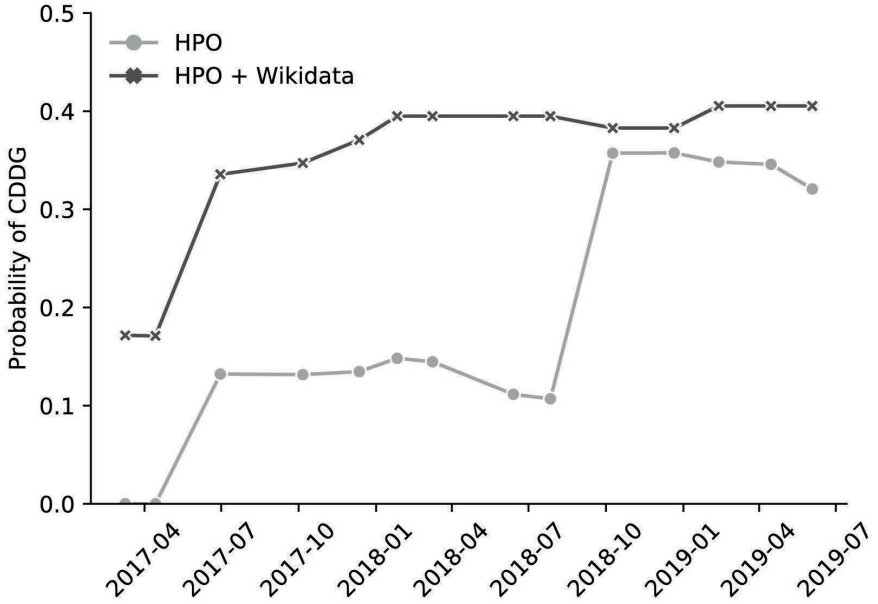


Figure 5.5: BOQA analysis of suspected cases of the disease Congenital Disorder of Deglycosylation (CDDG). We used an algorithm called BOQA to rank potential diagnoses based on clinical phenotypes. Here, clinical phenotypes from two cases of suspected CDDG patients were extracted from a published case report (Caglayan et al., 2015). These phenotypes were run through BOQA using phenotype-disease annotations from the Human Phenotype Ontology (HPO) alone, or from a combination of HPO and Wikidata. This analysis was tested using several versions of disease-phenotype annotations (shown along the x-axis). The probability score for CDDG is reported on the y-axis. These results demonstrate that the inclusion of Wikidata-based disease-phenotype annotations would have significantly improved the diagnosis predictions from BOQA at earlier time points prior to their official inclusion in the HPO annotation file. Details of this analysis can be found at https://github.com/SuLab/Wikidata-phenomizer (archived at [52]).

the crowd-based Wikidata model and may motivate further community contributions.

5.8 Drug repurposing

The mining of graphs for latent edges has been an area of interest in a variety of contexts from predicting friend relationships in social media platforms to suggesting movies based on past viewing history. A number of groups have explored the mining of knowledge graphs to reveal biomedical insights, with the open source Rephetio effort for drug repurposing as one example (Himmelstein et al., 2017) [56]. Rephetio uses logistic regression, with features based on graph metapaths, to predict drug repurposing candidates.

The knowledge graph that served as the foundation for Rephetio was manually assembled from many different resources into a heterogeneous knowledge network. Here, we explored whether the Rephetio algorithm could successfully predict drug indications on the Wikidata knowledge graph. Based on the class diagram in Figure 1, we extracted a biomedically-focused subgraph of Wikidata with 19 node types and 41 edge types. We performed five-fold cross validation on drug indications within Wikidata and found that Rephetio substantially enriched the true indications in the hold-out set. We then downloaded historical Wikidata versions from 2017 and 2018 and observed marked improvements in performance over time (Figure 5.6). We also performed this analysis using an external test set based on Drug Central, which showed a similar improvement in Rephetio results over time (Figure 5.7).

This analysis demonstrates the value of a community-maintained, centralized knowledge base to which many researchers are contributing. It suggests that scientific analyses based on Wikidata may continually improve irrespective of any changes to the underlying algorithms, but simply based on progress in curating knowledge through the distributed, and largely uncoordinated efforts of the Wikidata community.

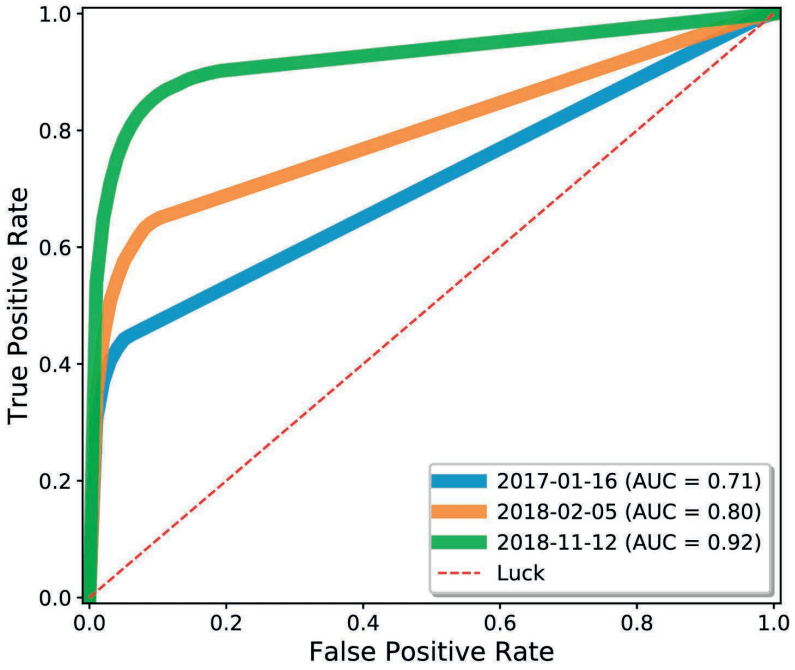


Figure 5.6: We analyzed three snapshots of Wikidata using Rephetio, a graph-based algorithm for predicting drug repurposing candidates [56]. We evaluated the performance of the Rephetio algorithm on three historical versions of the Wikidata knowledge graph, quantified based on the area under the receiver operator characteristic curve (AUC). This analysis demonstrated that the performance of Rephetio in drug repurposing improved over time based only on improvements to the underlying knowledge graph. Details of this analysis can be found at *github.com/SuLab/WD-rephetio-analysis (archived at [32]).*

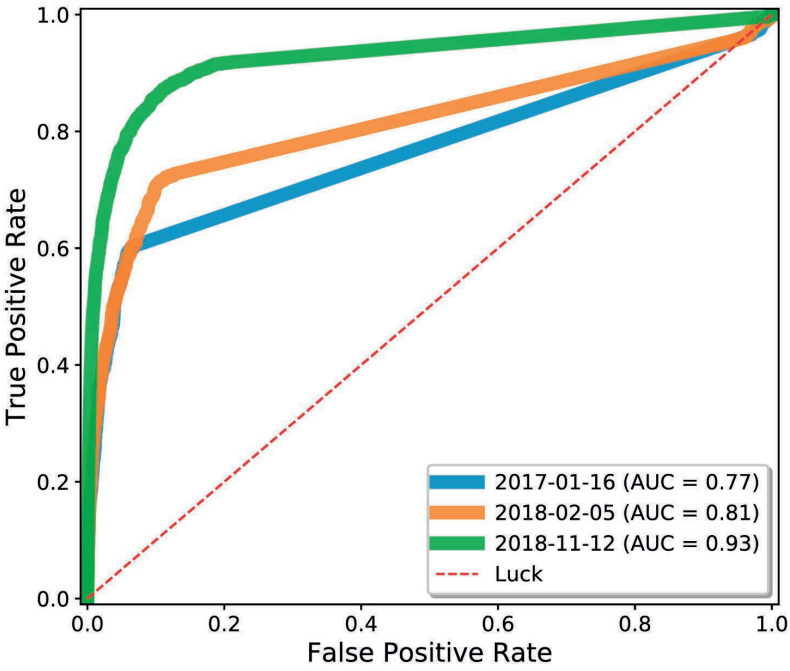


Figure 5.7: Drug repurposing using the Wikidata knowledge graph, evaluated using an external test set. The analysis in Figure 5.6 was based on a cross-validation of indications that were present in Wikidata. This time-resolved analysis was run using an external gold standard set of indications from Drug Central [57].

5.9 Outlook

We believe that the design of Wikidata is well-aligned with the FAIR data principles.

FINDABLE: Wikidata items are assigned globally unique identifiers with direct cross-links into the massive online ecosystem of Wikipedias. Wikidata also has broad visibility within the Linked Data community and is listed in the life science registries FAIRsharing (*fairsharing.org*; [58]) and *Identifiers.org* [59]. Wikidata has already attracted a robust, global community of contributors and consumers.

ACCESSIBLE" Wikidata provides access to its underlying knowledge graph via both an online graphical user interface and an API, and access includes both read- and write-privileges. Wikidata provides database dumps at least weekly (*www.wikidata.org/wiki/Wikidata:Database˙download*), ensuring the long-term accessibility of the Wikidata knowledge graph independent of the organization and web application. Finally, Wikidata is also natively multilingual.

INTEROPERABLE: Wikidata items are extensively cross-linked to other biomedical resources using Universal Resource Identifiers (URIs), which unambiguously anchor these concepts in the Linked Open Data cloud [60]. Wikidata is also available in many standard formats in computer programming and knowledge management, including JSON, XML, and RDF.

REUSABLE: Data provenance is directly tracked in the reference section of the Wikidata statement model. The Wikidata knowledge graph is released under the Creative Commons Zero (CC0) Public Domain Declaration, which explicitly declares that there are no restrictions on downstream reuse and redistribution.

The open data licensing of Wikidata is particularly notable. The use of data licenses in biomedical research has rapidly proliferated, presumably in an effort to protect intellectual property and/or justify long-term grant funding (see, for example, [61]). However, even seemingly innocuous license terms

(like requirements for attribution) still impose legal requirements and therefore expose consumers to legal liability. This liability is especially problematic for data integration efforts, in which the license terms of all resources (dozens or hundreds or more) must be independently tracked and satisfied (a phenomenon referred to as 'license stacking'). Because it is released under CC0, Wikidata can be freely and openly used in any other resource without any restriction. This freedom greatly simplifies and encourages downstream use, albeit at the cost of not being able to incorporate ontologies or datasets with more restrictive licensing.

In addition to simplifying data licensing, Wikidata offers significant advantages in centralizing the data harmonization process. Consider the use case of trying to get a comprehensive list of disease indications for the drug bupropion. The National Drug File – Reference Terminology (NDF-RT) reported that bupropion may treat nicotine dependence and attention deficit hyperactivity disorder, the Inxight database listed major depressive disorder, and the FDA Adverse Event Reporting System (FAERS) listed anxiety and bipolar disorder. While no single database listed all these indications, Wikidata provided an integrated view that enabled seamless query and access across resources. Integrating drug indication data from these individual data resources was not a trivial process. Both Inxight and NDF-RT mint their own identifiers for both drugs and diseases. FAERS uses Medical Dictionary for Regulatory Activities (MedDRA) names for diseases and free-text names for drugs [62]. By harmonizing and integrating all resources in the context of Wikidata, we ensure that those data are immediately usable by others without having to repeat the normalization process. Moreover, by harmonizing data at the time of data loading, consumers of that data do not need to perform the repetitive and redundant work at the point of querying and analysis.

As the biomedical data within Wikidata continues to grow, we believe that its unencumbered use will spur the development of many new innovative tools and analyses. These innovations will undoubtedly include the machine learning-based mining of the knowledge graph to predict new relationships (also referred to as knowledge graph reasoning [63–65].

For those who subscribe to this vision for cultivating a FAIR and open graph

of biomedical knowledge, there are two simple ways to contribute to Wikidata. First, owners of data resources can release their data using the CC0 declaration. Because Wikidata is released under CC0, it also means that all data imported in Wikidata must also use CC0-compatible terms (e.g., be in the public domain). For resources that currently use a restrictive data license primarily for the purposes of enforcing attribution or citation, we encourage the transition to CC0 (+BY), a model that "move[s] the attribution from the legal realm into the social or ethical realm by pairing a permissive license with a strong moral entreaty' [66]. For resources that must retain data license restrictions, consider releasing a subset of data or older versions of data using CC0. Many biomedical resources were created under or transitioned to CC0 (in part or in full) in recent years , including the Disease Ontology [30], Pfam [67] , Bgee [68], WikiPathways [28], Reactome [27], ECO [69], and CIViC [16].

Second, informaticians can contribute to Wikidata by adding the results of data parsing and integration efforts to Wikidata as, for example, new Wikidata items, statements, or references. Currently, the useful lifespan of data integration code typically does not extend beyond the immediate project-specific use. As a result, that same data integration process is likely performed repetitively and redundantly by other informaticians elsewhere. If every informatician contributed the output of their effort to Wikidata, the resulting knowledge graph would be far more useful than the stand-alone contribution of any single individual, and it would continually improve in both breadth and depth over time. Indeed, the growth of biomedical data in Wikidata is driven not by any centralized or coordinated process, but rather the aggregated effort and priorities of Wikidata contributors themselves.

FAIR and open access to the sum total of biomedical knowledge will improve the efficiency of biomedical research. Capturing that information in a centralized knowledge graph is useful for experimental researchers, informatics tool developers and biomedical data scientists. As a continuously-updated and collaboratively-maintained community resource, we believe that Wikidata has made significant strides toward achieving this ambitious goal.

References

[1] Mark D. Wilkinson et al. “Evaluating FAIR maturity through a scalable, automated, community-governed framework”. *Scientific Data*. 2019. 6 (1): p. 174.

[2] Mark D. Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. *Scientific Data*. 2016. 3 (1): p. 160018.

[3] Christopher J. Mungall et al. “The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species”. *Nucleic Acids Research*. 2017. 45 (D1): pp. D712–D722.

[4] Christina Chandras et al. “Models for financial sustainability of biological databases and resources”. *Database*. 2009. 2009 (bap017): .

[5] Chiara Gabella, Christine Durinx, and Ron Appel. “Funding knowledgebases: Towards a sustainable funding model for the UniProt use case”. *F1000Res*. 2018. 6: p. 2051.

[6] Denny Vrandečić and Markus Krötzsch. “Wikidata: a free collaborative knowledgebase”. *Commun. ACM*. 2014. 57 (10): pp. 78–85.

[7] Marçal Mora-Cantallops, Salvador Sánchez-Alonso, and Elena García-Barriocanal. “A systematic literature review on Wikidata”. *Data Technologies and Applications*. 2019. 53 (3): pp. 250–268.

[8] Sebastian Burgstaller-Muehlbacher et al. “Wikidata as a semantic framework for the Gene Wiki initiative”. *Database (Oxford)*. 2016. 2016: pp. 1–10.

[9] Houcemeddine Turki et al. “Wikidata: A large-scale collaborative ontological medical database”. *Journal of Biomedical Informatics*. 2019. 99: p. 103292.

[10] Tim E. Putman et al. “WikiGenomes: an open web application for community consumption and curation of gene annotation data in Wikidata”. *Database*. 2017. 2017 (bax025): .

[11] NCBI Resource Coordinators. “Database resources of the national center for biotechnology information”. *Nucleic acids research*. 2018. 46 (D1): pp. D8–D13.

[12] Daniel R Zerbino et al. “Ensembl 2018”. *Nucleic Acids Research*. 2018. 46 (D1): pp. D754–D761.

[13] The UniProt Consortium. “UniProt: a worldwide hub of protein knowledge”. *Nucleic Acids Research*. 2019. 47 (D1): pp. D506–D515.

[14] Alex L Mitchell et al. “InterPro in 2019: improving coverage, classification and access to protein sequence annotations”. *Nucleic Acids Research*. 2019. 47 (D1): pp. D351–D360.

[15] . “Protein Data Bank: the single global archive for 3D macromolecular structure data”. *Nucleic acids research*. 2019. 47 (D1): pp. D520–D528.

[16] Malachi Griffith et al. “CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer”. *Nature Genetics*. 2017. 49 (2): pp. 170–174.

[17] Yanli Wang et al. “PubChem: a public information system for analyzing bioactivities of small molecules”. *Nucleic Acids Research*. 2009. 37 (suppl_2): W623–W633.

[18] Stuart J Nelson et al. “Normalized names for clinical drugs: RxNorm at 6 years”. *Journal of the American Medical Informatics Association*. 2011. 18 (4): pp. 441–448.

[19] Simon D Harding et al. “The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY”. *Nucleic Acids Research*. 2018. 46 (D1): pp. D1091–D1106.

[20] Adam J. Pawson et al. “The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands”. *Nucleic Acids Research*. 2014. 42 (D1): pp. D1098–D1106.

[21] Christopher Southan et al. “The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands”. *Nucleic Acids Research*. 2016. 44 (D1): pp. D1054–D1068.

[22] Manish Sud et al. “LMSD: LIPID MAPS structure database”. *Nucleic Acids Research*. 2007. 35 (suppl_1): pp. D527–D532.

[23] Hisayuki Horai et al. “MassBank: a public repository for sharing mass spectral data for life sciences”. *Journal of Mass Spectrometry*. 2010. 45 (7): pp. 703–714.

[24] Gert Wohlgemuth et al. “SPLASH, a hashed identifier for mass spectra”. *Nature Biotechnology*. 2016. 34 (11): pp. 1099–1101.

[25] Jae-Min Shin and Doo-Ho Cho. “PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures”. *Nucleic Acids Research*. 2005. 33 (suppl_1): pp. D238–D241.

[26] Antony J. Williams et al. “The CompTox Chemistry Dashboard: a community data resource for environmental chemistry”. *Journal of Cheminformatics*. 2017. 9 (1): p. 61.

[27] Antonio Fabregat et al. “The Reactome Pathway Knowledgebase”. *Nucleic Acids Research*. 2018. 46 (D1): pp. D649–D655.

[28] Denise N Slenter et al. “WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research”. *Nucleic Acids Research*. 2018. 46 (D1): pp. D661–D667.

[29] Evan R. Sprague. “ORCID”. *Journal of the Medical Library Association*. 2017. 105 (2): pp. 207–208.

[30] Lynn M Schriml et al. “Human Disease Ontology 2018 update: classification, content and workflow expansion”. *Nucleic Acids Research*. 2019. 47 (D1): pp. D955–D962.

[31] Phoebe Ayers et al. “WikiCite 2018-2019: Citations for the sum of all human knowledge”. 2019. : 1993318 Bytes.

[32] *SuLab/genewikiworld: v1.1 release on 2020-01-21*. SuLab/genewikiworld. URL: https://zenodo.org/record/3620812 (visited on 02/01/2021).

[33] Jiwen Xin et al. "Cross-linking BioThings APIs through JSON-LD to facilitate knowledge exploration". *BMC Bioinformatics*. 2018. 19 (1): p. 30.

[34] Martijn P. van Iersel et al. "The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services". *BMC Bioinformatics*. 2010. 11 (1): p. 5.

[35] Damian Smedley et al. "The BioMart community portal: an innovative alternative to large, centralized data repositories". *Nucleic Acids Research*. 2015. 43 (W1): W589–W598.

[36] Olivier Bodenreider. "The Unified Medical Language System (UMLS): integrating biomedical terminology". *Nucleic Acids Research*. 2004. 32 (suppl_1): pp. D267–D270.

[37] Sherri de Coronado et al. "The NCI Thesaurus quality assurance life cycle". *Journal of Biomedical Informatics*. 2009. 42 (3): pp. 530–539.

[38] Stephanie M. Bunt et al. "Directly e-mailing authors of newly published papers encourages community curation". *Database*. 2012. 2012 (bas024): .

[39] Yolanda Gil et al. "A Controlled Crowdsourcing Approach for Practical Ontology Extensions and Metadata Annotations". Paper presented at: *The Semantic Web – ISWC 2017*. Cham. Springer International Publishing. 2017. pp. 231–246.

[40] Tim Putman et al. "ChlamBase: a curated model organism database for the Chlamydia research community". *Database*. 2019. 2019 (baz041): .

[41] Mingxun Wang et al. "Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking". *Nature Biotechnology*. 2016. 34 (8): pp. 828–837.

[42] Janine Lewis, Michelle Snyder, and Henrietta Hyatt-Knorr. “Marking 15 years of the Genetic and Rare Diseases Information Center”. *Translational Science of Rare Diseases*. 2017. 2 (1-2): pp. 77–88.

[43] Sylvie Maiella et al. “Harmonising phenomics information for a better interoperability in the rare disease field”. *European Journal of Medical Genetics*. 2018. 61 (11): pp. 706–714.

[44] Joanna S. Amberger and Ada Hamosh. “Searching Online Mendelian Inheritance in Man (OMIM): A Knowledgebase of Human Genes and Genetic Phenotypes”. *Curr Protoc Bioinformatics*. 2017. 58: pp. 1.2.1–1.2.12.

[45] Victor A. McKusick. “Mendelian Inheritance in Man and Its Online Version, OMIM”. *The American Journal of Human Genetics*. 2007. 80 (4): pp. 588–604.

[46] *SuLab/scheduled-bots: Release v1.0 2020-01-21*. SuLab/scheduled-bots. URL: `https://zenodo.org/record/3620977` (visited on 02/01/2021).

[47] Finn Årup Nielsen, Daniel Mietchen, and Egon Willighagen. “Scholia, Scientometrics and Wikidata”. Paper presented at: *The Semantic Web: ESWC 2017 Satellite Events*. Cham. Springer International Publishing. 2017. pp. 237–259.

[48] Sebastian Köhler et al. “Clinical diagnostics in human genetics with semantic similarity searches in ontologies”. *Am J Hum Genet*. 2009. 85 (4): pp. 457–464.

[49] Sebastian Bauer et al. “Bayesian ontology querying for accurate and noise-tolerant semantic searches”. *Bioinformatics*. 2012. 28 (19): pp. 2502–2508.

[50] Sebastian Köhler et al. “The Human Phenotype Ontology in 2017”. *Nucleic Acids Research*. 2017. 45 (D1): pp. D865–D876.

[51] *SuLab/boqa: Release v1.0 2020-01-21*. SuLab/boqa. URL: `https://zenodo.org/record/3620979` (visited on 02/01/2021).

[52] *SuLab/Wikidata-phenomizer: Release v1.0 on 2020-01-15*. SuLab/Wikidata-phenomizer. URL: https://zenodo.org/record/3609142 (visited on 02/01/2021).

[53] Gregory M. Enns et al. "Mutations in NGLY1 cause an inherited disorder of the endoplasmic reticulum–associated degradation pathway". *Genetics in Medicine*. 2014. 16 (10): pp. 751–758.

[54] Christina Lam et al. "Prospective phenotyping of NGLY1-CDDG, the first congenital disorder of deglycosylation". *Genetics in Medicine*. 2017. 19 (2): pp. 160–168.

[55] Ahmet Okay Caglayan et al. "NGLY1 mutation causes neuromotor impairment, intellectual disability, and neuropathy". *European Journal of Medical Genetics*. 2015. 58 (1): pp. 39–43.

[56] Daniel Scott Himmelstein et al. "Systematic integration of biomedical knowledge prioritizes drugs for repurposing". *eLife*. 2017. 6: e26726.

[57] Oleg Ursu et al. "DrugCentral: online drug compendium". *Nucleic Acids Research*. 2017. 45 (D1): pp. D932–D939.

[58] Susanna-Assunta Sansone et al. "FAIRsharing as a community approach to standards, repositories and policies". *Nature Biotechnology*. 2019. 37 (4): pp. 358–367.

[59] Sarala M. Wimalaratne et al. "Uniform resolution of compact identifiers for biomedical data". *Scientific Data*. 2018. 5 (1): p. 180029.

[60] *Wikidata as an intuitive resource towards semantic data modeling in data FAIRification*. URL: /paper/Wikidata-as-an-intuitive-resource-towards-semantic-Jacobsen-Waagmeester/cae9d1d33ed9542c6319cd889c56fa044a7b9eaf (visited on 02/01/2021).

[61] Leonore Reiser et al. "Sustainable funding for biocuration: The Arabidopsis Information Resource (TAIR) as a case study of a subscription-based funding model". *Database*. 2016. 2016 (baw018): .

[62] *Drug Indications Extracted from FAERS*. URL: https://zenodo.org/record/1436000 (visited on 02/01/2021).

[63] Rajarshi Das et al. “Go for a Walk and Arrive at the Answer: Reasoning Over Paths in Knowledge Bases using Reinforcement Learning”. *arXiv:1711.05851 [cs]*. 2018. : .

[64] Xi Victoria Lin, Richard Socher, and Caiming Xiong. “Multi-Hop Knowledge Graph Reasoning with Reward Shaping”. *arXiv:1808.10568 [cs]*. 2018. : .

[65] Wenhan Xiong, Thien Hoang, and William Yang Wang. “DeepPath: A Reinforcement Learning Method for Knowledge Graph Reasoning”. *arXiv:1707.06690 [cs]*. 2018. : .

[66] *CC0 (+BY)*. URL: `https://dancohen.org/2013/11/26/cc0-by/`.

[67] Sara El-Gebali et al. “The Pfam protein families database in 2019”. *Nucleic Acids Res*. 2019. 47 (D1): pp. D427–D432.

[68] Frederic Bastian et al. “Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species”. Paper presented at: *Data Integration in the Life Sciences*. Berlin, Heidelberg. Springer. 2008. pp. 124–131.

[69] Marcus C. Chibucos et al. “Standardized description of scientific evidence using the Evidence Ontology (ECO)”. *Database*. 2014. 2014 (bau075): .

6

A protocol for adding knowledge to Wikidata: aligning resources on human coronaviruses

Adapted from: Andra Waagmeester et al. "A protocol for adding knowledge to Wikidata: aligning resources on human coronaviruses". *BMC Biology*. 2021. 19 (1): p. 12.

Abstract

Background Pandemics, even more than other medical problems, require swift integration of knowledge. When caused by a new virus, understanding the underlying biology may help finding solutions. In a setting where there are a large number of loosely related projects and initiatives, we need common ground, also known as a "commons." Wikidata, a public knowledge graph aligned with Wikipedia, is such a commons and uses unique identifiers to link knowledge in other knowledge bases. However, Wikidata may not always have the right schema for the urgent questions. In this paper, we address this problem by showing how a data schema required for the integration can be modeled with entity schemas represented by Shape Expressions.

Results As a telling example, we describe the process of aligning resources on the genomes and proteomes of the SARS-CoV-2 virus and related viruses as well as how Shape Expressions can be defined for Wikidata to model the knowledge, helping others studying the SARS-CoV-2 pandemic. How this model can be used to make data between various resources interoperable is demonstrated by integrating data from NCBI (National Center for Biotechnology Information) Taxonomy, NCBI Genes, UniProt, and WikiPathways. Based on that model, a set of automated applications or bots were written for regular updates of these sources in Wikidata and added to a platform for automatically running these updates.

Conclusions Although this workflow is developed and applied in the context of the COVID-19 pandemic, to demonstrate its broader applicability it was also applied to other human coronaviruses (MERS, SARS, human coronavirus NL63, human coronavirus 229E, human coronavirus HKU1, human coronavirus OC4).

6.1 Background

The coronavirus disease 2019 (COVID-19) pandemic, caused by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) virus, is leading to a burst of swiftly released scientific publications on the matter [1]. In response to the pandemic, many research groups have started projects to understand the SARS-CoV-2 virus life cycle and to find solutions. Examples of the numerous projects include outbreak.info [2], Virus Outbreak Data Network (VODAN) [3], CORD-19-on-FHIR [4], KG-COVID-19 knowledge graph [5], and the COVID-19 Disease Map [6]. Many research papers and preprints get published every week and many call for more Open Science [7]. The Dutch universities went a step further and want to make any previously published research openly available, in whatever way related to COVID-19 research.

However, this swift release of research findings comes with an increased number of incorrect interpretations [8] which can be problematic when new research articles are picked up by main-stream media [9]. Rapid evaluation of these new research findings and integration with existing resources requires frictionless access to the underlying research data upon which the findings are based. This requires interoperable data and sophisticated integration of these resources. Part of this integration is reconciliation, which is the process where matching concepts in Wikidata are sought [10]. Is a particular gene or protein already described in Wikidata? Using a shared interoperability layer, like Wikidata, different resources can be more easily linked.

Wikidata is the linked-data repository of the Wikimedia Foundation. It is developed to provide Wikipedia and its sister projects with structured data. One interesting feature of Wikidata is that provenance and attribution can easily be included using the references and qualifiers which are core to the Wikidata data model.

The Gene Wiki project has been leveraging Wikidata to link different research silos by creating a brokerage system between resources on genetics, biological processes, related diseases, and associated drugs. Various use cases ranging from crowdsourced curation of biomedical ontologies, phenotype-based diagnosis of diseases, and drug repurposing can feed on this system. The project

recognizes Wikidata as a sustainable infrastructure for scientific knowledge in the life sciences.

In contrast to legacy databases, where data models follow a relational data schema of connected tables, Wikidata uses statements to store facts (see Fig. 6.1) [11–14]. This model of statements aligns well with the RDF triple model of the semantic web and the content of Wikidata is also serialized as Resource Description Framework (RDF) triples [15, 16], acting as a stepping stone for data resources to the semantic web. Through its SPARQL (SPARQL Protocol and RDF Query Language) endpoint [17], knowledge captured in Wikidata can be integrated with other nodes in the semantic web, using mappings between these resources or through federated SPARQL queries [18]. A Wikidata item, as depicted in Fig. 6.1, has properties and values. The values are editable by everyone, but the property types are restricted. Creating new properties requires a property proposal that is formally discussed online. When there is consensus on the usefulness of the proposed property, it is created by a system administrator with its attached semantics. Users cannot create new properties at will, which makes it (together with is community acceptance) highly sustainable.

The Gene Wiki project aligns novel primary data sources with Wikidata in a cycle of consecutive steps where the data schema of the primary source is aligned with the available Wikidata properties. Once the schema is in place, bots are developed to add and regularly update Wikidata with knowledge from the primary resource under scrutiny.

Automated editing of Wikidata simplifies the process; however, quality control must be monitored carefully. This requires a clear data schema that allows the various resources to be linked together with their provenance. This schema describes the key concepts required for the integrations of the resources we are interested in the NCBI Taxonomy [19], NCBI Gene [20], UniProt [21], the Protein Data Bank (PDB) [22], WikiPathways [23], and PubMed [24]. Therefore, the core elements for which we need a model include viruses, virus strains, virus genes, and virus proteins. The first two provide the link to taxonomies, and the models for genes and proteins linked to UniProt, PDB, and

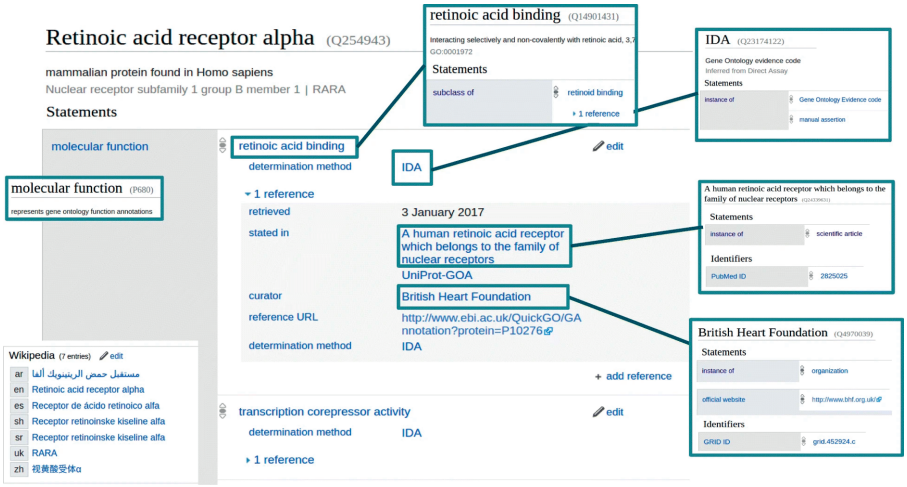


Figure 6.1: Structure of a Wikidata item, containing a set of statements which are key-value pairs, with qualifiers and references. Here the item for the angiotensin-converting enzyme 2 (ACE2) protein is given containing a statement about its molecular function. This molecular function (peptidyl-dipeptidase activity) contains a reference stating when and where this information was obtained

WikiPathways. These key concepts are also required to annotate research output such as journal articles and datasets related to these topics. Wikidata calls such keywords "main subjects." The introduction of this model and the actual SARS-CoV-2 genes and proteins in Wikidata enables the integration of these resources. The resources used were selected based on their eligibility for inclusion in Wikidata. Wikidata is available under a CC0 1.0 Universal (CC0 1.0) Public Domain Dedication which stipulates public use of the data included. Some valid resources use more restrictive licenses which prevents their inclusion in Wikidata.

This paper is a case report of a workflow/protocol for data integration and publication. The first step in this approach is to develop the data schema. Wikidata has a schema extension called EntitySchema that uses Shape Expressions (ShEx) as the structural schema language to describe and capture schemas of concepts [25, 26]. With ShEx, we describe the RDF structure by which Wikidata content is made available. These Shapes have the advantage that they are easily exchanged and describe linked-data models as a single knowledge graph. Since the Shapes describe the model, they enable discussion, reveal inconsistencies between resources, and allow for consistency checks of the content added by automated procedures. Eventually, we would like to get to a workflow where issues that can be fixed automatically are corrected, whereas biological inconsistencies will be made available for evaluation by field experts, and non-domain specific issues are acted upon by the Wikidata community at large. With the model defined, the focus can turn to the process of adding knowledge to Wikidata. In this phase, the seven human coronaviruses (HCoVs), Middle East respiratory syndrome (MERS), SARS, SARS-CoV-2 (causing COVID-19), human coronavirus NL63, human coronavirus 229E, human coronavirus HKU1, and human coronavirus OC4 [27], can be added to Wikidata. This protocol is finalized by describing how the resulting data schema and data can be applied to support other projects, particularly the WikiPathways COVID Portal.

The Semantic Web was proposed as a vision of the Web, in which information is given well-defined meaning and better-enabling computers and people to work in cooperation [28]. In order to achieve that goal, several technologies have appeared, like RDF for describing resources [16], SPARQL to query

RDF data [29], and the Web Ontology Language (OWL) to represent ontologies [30].

Linked data was later proposed as a set of best practices to share and reuse data on the web [31]. The linked data principles can be summarized in four rules that promote the use of uniform resource identifiers (URIs) to name things, which can be looked up to retrieve useful information for humans and machines using RDF, as well as having links to related resources. These principles have been adopted by several projects, enabling a web of reusable data, known as the linked data cloud [32], which has also been applied to life science [33].

One prominent project is Wikidata, which has become one of the largest collections of open data on the web [18]. Wikidata follows the linked data principles offering both HTML and RDF views of every item with their corresponding links to related items, and a SPARQL endpoint called the Wikidata Query Service.

Wikidata's RDF model offers a reification mechanism that enables the representation of information about statements like qualifiers and references [34]. For each statement in Wikidata, there is a direct property in the wdt namespace that indicates the direct value. In addition, the Wikidata data model adds other statements for reification purposes that allow enrichment of the declarations with references and qualifiers (for a topical treatise, see Ref. [35]). As an example, item Q14875321, which represents ACE2 (protein-coding gene in the Homo sapiens species) has a statement specifying that it can be found on chromosome (P1057) with value chromosome X (Q29867336). In RDF Turtle, this can be declared as:

```
wd:Q14875321 wdt:P1057 wd:Q29867336 .
```

That statement can be reified to add qualifiers and references. For example, a qualifier can state that the genomic assembly (P659) is GRCh38 (Q20966585) with a reference declaring that it was stated (P248) in Ensembl Release 99 (Q83867711). In Turtle, those declarations are represented as (see also Fig. 6.2):

```
wd:Q14875321 rdfs:label "ACE2"@en ;
             wdt:P1057 wd:Q29867336 .
wd:Q14875321 p:P1057 [
     ps:P1057 wd:Q29867336 ;
 pq:P659 wd:Q20966585 ;
 prov:wasDerivedFrom [
   pr:P248 wd:Q83867711 ;
   pr:P594 "ENSG00000130234" ;
 ]] .
wd:Q29867336 rdfs:label "human X chromosome"@en .
wd:Q20966585 rdfs:label "Genome assembly GRCh38"@en .
wd:Q83867711 rdfs:label "Ensembl Release 99"@en .
```

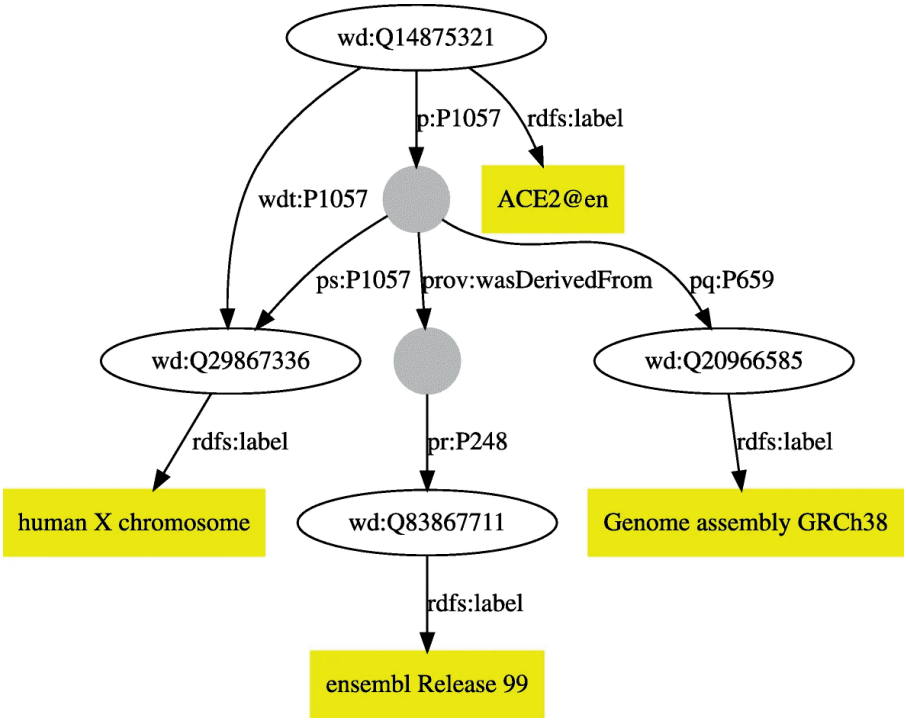


Figure 6.2: Example of an RDF data model representing ACE2, created with RDFShape [36]

6.2 Results

6.2.1 Semantic data landscape

To align the different sources in Wikidata, a common data schema is needed. We have created a collection of schemas that represent the structure of the items added to Wikidata. Input to the workflow is the NCBI taxon identifier, which is input to mygene.info (see Fig. 8b). Taxon information is obtained and added to Wikidata according to a set of linked Entity Schemas (virus, strain, disease) [37–40]. Gene annotations are obtained and added to Wikidata following the schemas virus gene and protein annotations [41, 42] are obtained and added to Wikidata following the two schemas. Pathway information is sourced according to the schema describing Wikipathways representation in Wikidata [43]. The last two schemas are an extension from more generic schemas for proteins [44] and genes [45] (Fig. 6.3).

6.2.2 ShEx validation

With the set of ShEx schemas, it is possible to check if existing data aligns with the expressed expectations. Figure 9 demonstrates two cases in which one Wikidata item (Q70644554) does not align with the tested Schema E174, while another Wikidata item on a Wikipathways Pathway (Q88292589) does conform to schema E41. The Wikidata EntitySchema extension does allow checking for conformance. There are currently five actively maintained ShEx implementations that allow checking for ShEx conformance at a larger scale [25] (Fig. 6.4).

6.2.3 Bots

The bots developed and used in this protocol are adaptations of the bots developed in the Gene Wiki project. On regular intervals, the bots run to update viral gene and protein annotations as well as pathway updates from WikiPathways. For the gene and protein annotations, we have also made a Jupyter Notebook. The bot that synchronizes virus, gene, and protein information and the Jupyter Notebook are available [46]. The bot that synchronizes the

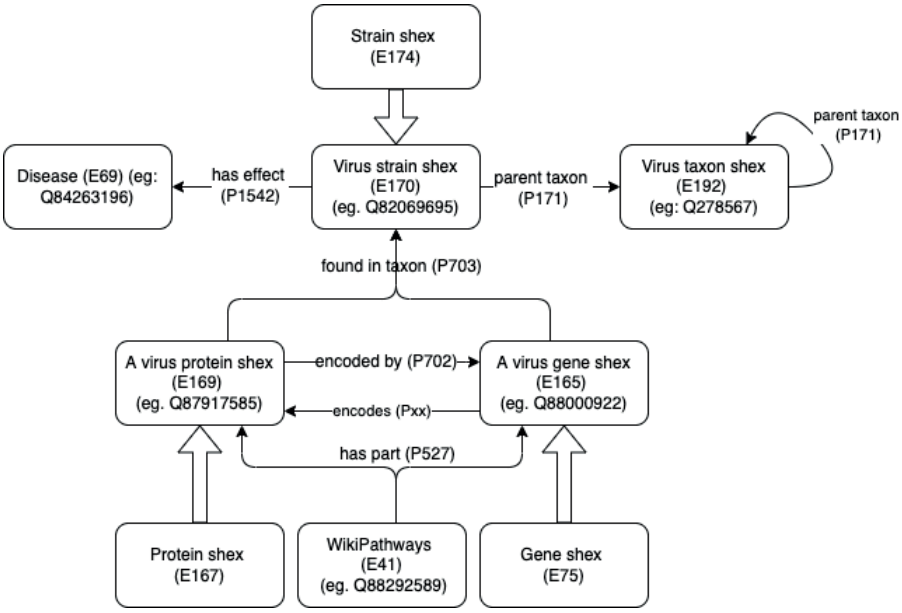


Figure 6.3: Overview of the ShEx schemas and the relations between them. All shapes, properties, and items are available from within Wikidata

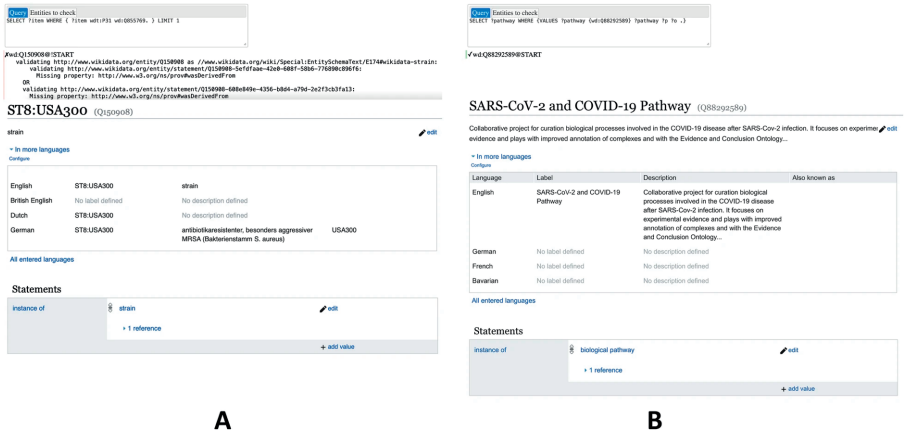


Figure 6.4: Application of the drafted ShEx schemas in the EntitySchema extension of Wikidata allows for confirmation if a set of on-topic items align with expressed expectations. In panel A, the application renders the Wikidata item invalid due to a missing reference which in turn does not conform to the expressed ShEx whereas in panel B, the item (Q88292589) conforms to the applied schema

WikiPathways pathways with Wikidata was updated from the original version to allow adding proteins annotated with Wikidata identifiers and no longer requires pathways to be part of the WikiPathways Curated Collection. The customized bot source code is available [47].

Both bots are now part of the automation server used in the Gene Wiki project. This runs on the Jenkins platform [48]. Although Jenkins is mainly aimed at software deployments, its extended scheduling capabilities the synchronization procedure at set intervals that can be changed depending on the update speed of the external resources. The Jenkins jobs are also available from [49].

6.3 Data added

Using the gene and proteins bots explained in the “Methods” section, missing genes and proteins have been added for the seven human coronaviruses. The

Table 6.1: Summary of the seven human coronaviruses, including taxon identifiers, the Wikidata items, and the number of genes and proteins. The latter two are generated by the SPARQL queries geneCount.rq and proteinCount.rq in Additional file 1

Virus strain	NCBI Taxon ID	Wikidata Qid	# Genes	# Proteins
SARS virus	694009	Q278567	14	11
MERS	1335626	Q4902157	11	9
Human coronavirus NL63	277944	Q8351095	7	6
Human coronavirus 229E	11137	Q16983356	8	8
Human coronavirus HKU1	290028	Q16983360	9	9
Human coronavirus OC43	31631	Q16991954	9	8
SARS-CoV-2	2697049	Q82069695	11	27

results are summarized in Table 1. The automatically added and updated gene and protein items were manually curated. For SARS-CoV-2, all items were already manually created, and the bot only edited gene items. Thirteen out of 27 protein entries were created by the authors. For the other species, all gene entries and most protein entries have been created by the bot. Only for MERS and SARS-CoV-2, some protein entries were added manually, including some by us. During this effort, which took 3 weeks, the bot created a number of duplicates. These have been manually corrected. It should also be noted that for SARS-CoV-2 many proteins and protein fragments do not have RefSeq or UniProt identifiers, mostly for the virus protein fragments.

6.4 Use cases

6.4.1 BridgeDb

Using the dedicated code to create a BridgeDb identifier mapping database for coronaviruses, mappings were extracted from Wikidata with a SPARQL query for the seven human coronaviruses and the SARS-related viruses. This resulted in a mapping database with 567 mappings between 380 identifiers (version 2020-11-30). This includes 171 Wikidata identifiers, 70 NCBI Gene identifiers, 71 UniProt identifiers, 58 RefSeq identifiers, and 10 Guide to Pharmacology Target identifiers. The mapping file has been released on the

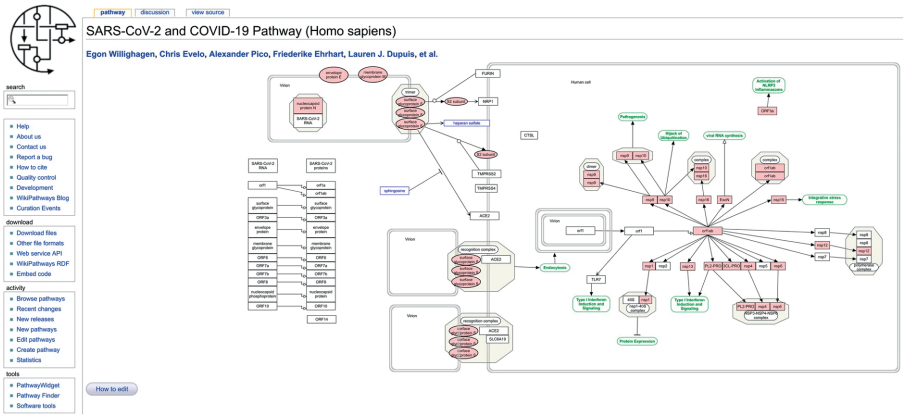


Figure 6.5: Screenshot of SARS-CoV-2 and COVID-19 Pathway in WikiPathways (wikipathways:WP4846) showing the BridgeDb popup box for the ORF3a protein, showing a link out to Scholia via the protein and gene's Wikidata identifiers

BridgeDb website and archived on Zenodo [50]. The mapping database has also been loaded on the BridgeDb webservice which means it can be used in the next use case: providing links out for WikiPathways.

6.4.2 Wikipathways

The WikiPathways project is involved in an international collaboration to curate knowledge about the biological processes around SARS-CoV-2 and COVID-19. The authors have started a pathway specifically about SARS-CoV-2 (wikipathways:WP4846). To ensure interoperability, WikiPathways uses BridgeDb and taking advantage of the enriched BridgeDb webservice, WikiPathways now links out for HCoV genes and proteins (depending on availability of mappings) to RefSef, NCBI Gene, UniProt, and Scholia (see Fig. 6.5). The latter links to the next use case and provides a link to literature about the virus. It should be noted that for each gene and protein two Wikidata identifiers with links may be given. In that case, one is for the gene and one for the protein.

6.4.3 Scholia

The WikiPathways use case shows us that literature describes our knowledge about how coronaviruses work at a rather detailed level. Indeed, many articles discuss the genetics, homology of genes and proteins across viruses, or the molecular aspects of how these proteins are created and how they interfere with the biology of the human cell. The biological pathways show these processes, but ultimately the knowledge comes from literature. Wikidata allows us to link literature to specific virus proteins and genes, depending on what the article describes. For this, it uses the "main subject" (P921) property [51]. We manually annotated literature with the Wikidata items for specific proteins and genes, particularly useful for virus concepts for which reference databases do not provide entries, such as the non-structural proteins. We developed two SPARQL queries to count the number of links between genes [17] and proteins [52] and the articles that discuss them. Scholia takes advantage of the "main subject" annotation, allowing the creation of "topic" pages for each protein. For example, Fig. 6.6 shows the topic page of the SARS-CoV-2 spike protein. As such, Scholia provides a simple user interface summarizing literature about a specific feature of the SARS-CoV-2 virus. An RSS feed is even available to get alerted about new literature about each topic (also visible in Fig. 6.6).

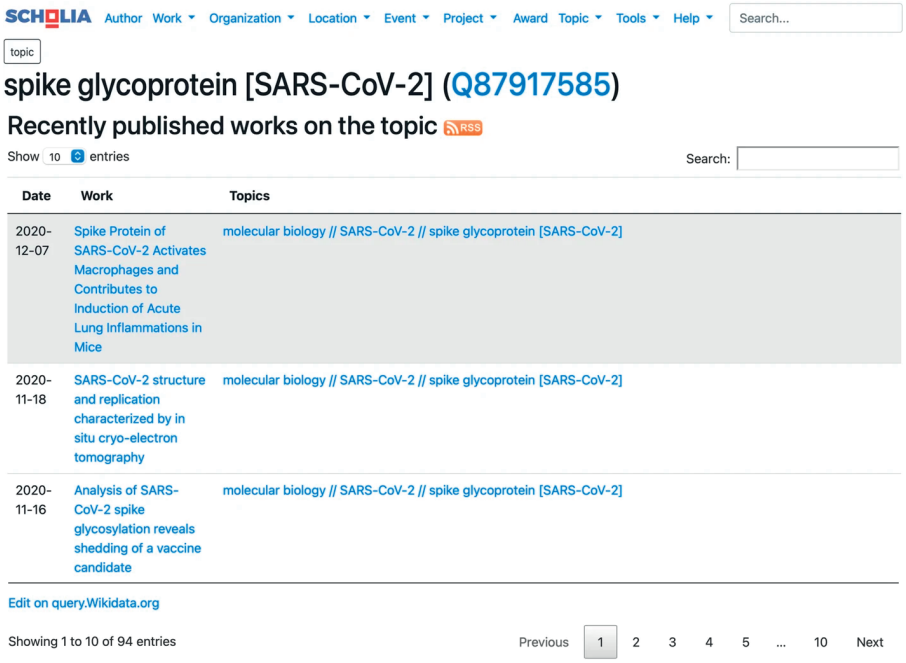


Figure 6.6: Screenshot of the Scholia page for the SARS-CoV-2 spike glycoprotein, it shows four articles that specifically discuss this protein

6.5 Discussion

This paper describes a protocol we developed to align genetic annotations from reference resources with Wikidata. Numerous annotations are scattered across different sources without any overall integration, thereby reducing the reusability of knowledge from different sources. Integration of the annotations from these resources is a complex and time-consuming task. Each resource uses different ways to access the data from a user and machine perspective. Making use of these protocols programmatically to access and retrieve data of interest requires the knowledge of various technologies and procedures to extract the information of interest.

Wikidata provides a solution. It is part of the semantic web, taking advantage of its reification of the Wikidata items as RDF. Data in Wikidata itself is frequently, often almost instantaneously, synchronized with the RDF resource and available through its SPARQL endpoint [53]. The modelling process turns out to be an important aspect of this protocol. Wikidata contains numerous entity classes as entities and more than 7000 properties that are ready for (re-)use. However, that also means that this is a confusing landscape to navigate. The ShEx Schema has helped us develop a clear model. This is a social contract between the authors of this paper, as well as documentation for future users.

Using these schemas, it was simpler to validate the correctness of the updated bots to enter data in Wikidata. The bots have been transferred to the Gene Wiki Jenkins platform. This allows the bots to be kept running regularly, pending the ongoing efforts of the coronavirus and COVID-19 research communities. While the work of the bots will continue to need human oversight, potentially to correct errors, it provides a level of scalability and generally alleviates the authors from a lot of repetitive work.

One of the risks of using bots is the possible generation of duplicate items. Though this is also a risk in the manual addition of items, humans can apply a wider range of academic knowledge to resolve these issues. Indeed, in running the bots, duplicate Wikidata items were created, for which an example is shown in Fig. 6.7. The Wikidataintegrator library does have functionality

to prevent the creation of duplicates by comparing properties, based on used database identifiers. However, if two items have been created using different identifiers, these cannot be easily identified.

Close inspection of examples, such as the one in Fig. 6.7, showed that the duplicates were created because there was a lack of overlap between the data to be added and the existing item. The UniProt identifier did not yet resolve, because it was manually extracted from information in the March 27 pre-release (but now part of the regular releases). In this example, the Pfam protein families database [54] identifier was the only identifier upon which reconciliation could happen. However, that identifier pointed to a webpage that did not contain mappings to other identifiers. In addition, the lack of references to the primary source hampers the curator's ability to merge duplicate items and expert knowledge was essential to identify the duplication. Fortunately, names used for these RNA viruses only refer to one protein as the membrane protein. Generally, the curator would have to revert to the primary literature to identify the overlap. Statements about "encoded by" to the protein-coding genes were found to be helpful as well. Reconciliation might be possible through sequence alignment, which means substantial expert knowledge and skills are required.

This makes reconciliation in Wikidata based on matching labels, descriptions and synonyms, matching statements, and captured provenance (qualifiers and references) hazardous, due to different meanings to the same label. A targeted curation query (geneAndProteinCurationQuery.rq, see Additional file 1) was developed to highlight such duplications and manually curated seven duplicate protein entries for SARS-CoV-2 alone. This duplication is common and to be expected, particularly in rapidly evolving situations like a pandemic, when many groups contribute to the same effort. In this case, this paper only represents one group contributing to the Wikidata:WikiProject COVID-19 [55].

We also discovered that virus taxonomy is different from zoological taxonomy. For example, there is no clear NCBI taxon identifier for SARS-CoV-1 and after consultation with other projects, we defaulted to using the taxon identifier for the SARS-related CoVs, something that NCBI and UniProt seem to have done as well.

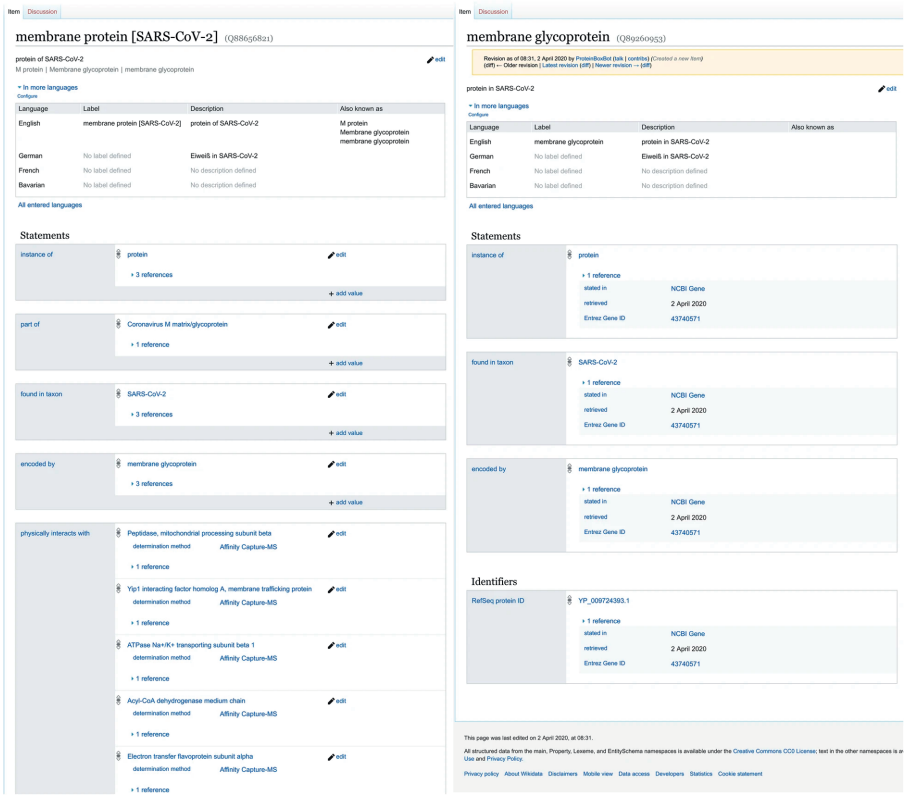


Figure 6.7: Comparison of two Wikidata entries for the SARS-CoV-2 membrane protein. An overlap between a Wikidata item and a concept from a primary source needs to have some overlap to allow automatic reconciliation. If there is no overlap, duplicates will be created and left for human inspection. Since this screenshot was made, the entries have been merged in a manually curation process

Finally, we note that during the 2 weeks, this effort took place, several other resources introduced overviews, including dedicated COVID-19 portals from UniProt [56] and the Protein DataBank in Europe [57].

6.6 Conclusion

This manuscript presents a protocol to link information from disparate resources, including NCBI Taxonomy, NCBI Gene, UniProt, PubMed, and WikiPathways. Using the existing Wikidata infrastructure, we developed semantic schemas for virus strains, genes, and proteins; bots are written in Python to add knowledge on genes and proteins of the seven human coronaviruses and linked them to biological pathways in WikiPathways and to primary literature, visualized in Scholia. We were able to do so in the period of 2 weeks, using an ad hoc team from existing collaborations, taking advantage of the open nature of the components involved.

6.7 Methods

6.7.1 Specifying data models with ShEx

Although the RDF data model is flexible, specifying an agreed structure for the data allows domain experts to identify the properties and structure of their data facilitating the integration between heterogeneous data sources. Shape Expressions were used to provide a suitable level of abstraction. Yet Another ShEx Editor (YaShE) [58], a ShEx editor implemented in JavaScript, was applied to author these Shapes [59]. This application provides the means to associate labels in the natural language of Wikidata to the corresponding identifiers. The initial entity schema was defined with YaShE as a proof of concept for virus genes and proteins. In parallel, statements already available in Wikidata were used to automatically generate an initial shape for virus strains with sheXer [60]. The statements for virus strains were retrieved with SPARQL from the Wikidata Query Service (WDQS). The generated Shape was then further improved through manual curation. The syntax of the Shape Expressions was continuously validated through YaShE and the Wikidata Entity Schema

namespace was used to share and collaboratively update the schema with new properties. Figure 6.8 gives a visual outline of these steps.

6.8 Populating Wikidata with human coronavirus data

The second step in our workflow is to add entries for all virus strains, genes, and their gene products to Wikidata. This information is spread over different resources. Here, annotations were obtained from NCBI EUtils [61], Mygene.info [62], and UniProt, as outlined below. Input to the workflow is the NCBI Taxonomy identifier of a virus under scrutiny (e.g., 2697049 for SARS-CoV-2). The taxon annotations are extracted from NCBI EUtils. The gene and gene product annotations are extracted from mygene.info and the protein annotations are extracted from UniProt using the SPARQL endpoint [56].

Genomic information from seven human coronaviruses (HCoVs) was collected, including the NCBI Taxonomy identifiers. For six virus strains, a reference genome was available and was used to populate Wikidata. For SARS-CoV-1, the NCBI Taxonomy identifier referred to various strains, though no reference strain was available. To overcome this issue, the species taxon for SARS-related coronaviruses (SARSr-CoV) was used instead, following the practices of NCBI Genes and UniProt (Fig. 6.9).

NCBI Eutils

The Entrez Programming Utilities (EUtils) [24] is the application programming interface (API) to the Entrez query and database system at the NCBI. From this set of services, the scientific name of the virus under scrutiny was extracted (e.g., “Severe acute respiratory syndrome coronavirus 2”).

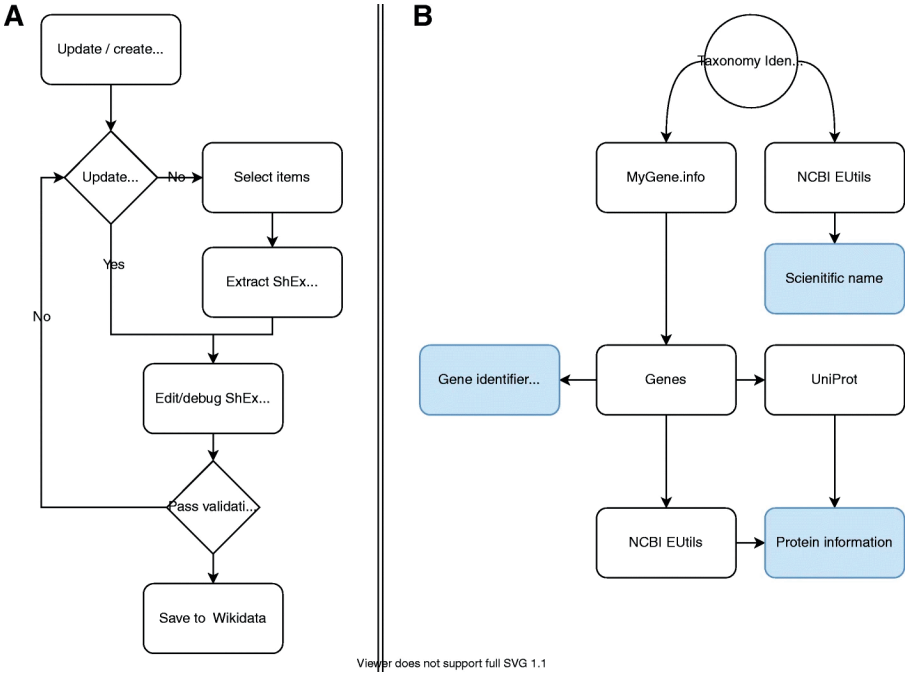


Figure 6.8: Flow diagram for entity schema development and the executable workflow for the virus gene protein bot. a The workflow of creating shape expressions. b The computational workflow of how information was used from various public resources to populate Wikidata

```
{
    "_id": "43740571",
    "_score": 15.594226,
    "accession": {
        "genomic": [
            "MN908947.3",
            "NC_045512.2"
        ],
        "protein": [
            "QHD43419.1",
            "YP_009724393.1"
        ]
    },
    "entrezgene": "43740571",
    "locus_tag": "GU280_gp05",
    "name": "membrane glycoprotein",
    "other_names": "membrane glycoprotein",
    "refseq": {
        "genomic": "NC_045512.2",
        "protein": "YP_009724393.1"
    },
    "retired": 43560233,
    "symbol": "M",
    "taxid": 2697049,
    "type_of_gene": "protein-coding"
}
```


Figure 6.9: JavaScript Object notation output of the mygene.info output for gene with NCBI gene identifier 43740571

```
PREFIX uniprotkb: <http://purl.uniprot.org/uniprot/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?id ?label ?database
WHERE {
SERVICE <https://sparql.uniprot.org/sparql> {
     VALUES ?database {
           <http://purl.uniprot.org/database/PDB>
           <http://purl.uniprot.org/database/RefSeq>
         }
     uniprotkb:uniprotID rdfs:label ?label ;
                         rdfs:seeAlso ?id .
     ?id <http://purl.uniprot.org/core/database> ?database .
   }
}
```

Figure 6.10: The UniProt SPARQL query used to obtain additional protein annotations, descriptions, and external resources

Mygene.info

Mygene.info [60] is a web service that provides a REST API that can be used to obtain up-to-date gene annotations. The first step in the process is to get a list of applicable genes for a given virus by providing the NCBI taxon id. The following step is to obtain gene annotations for the individual genes from mygene.info (e.g., [63]). This results in the name and a set of applicable identifiers (Fig. 6.10).

UniProt

The annotations retrieved from mygene.info also contain protein identifiers such as UniProt, RefSeq [64], and PDB; however, their respective names are lacking. To obtain names and mappings to other protein identifiers, RefSeq and UniProt were consulted. Refseq annotations were acquired using the earlier mentioned NCBI EUtils. UniProt identifiers are acquired using the SPARQL endpoint of UniProt, which is a rich resource for protein annotations

provided by the Swiss Bioinformatics Institute. Figure 5 shows the SPARQL query that was applied to acquire the protein annotations.

Reconciliation with Wikidata

Before the aggregated information on viruses, genes, and proteins can be added to Wikidata, reconciliation with Wikidata is necessary. If Wikidata items exist, they are updated; otherwise, new items are created. Reconciliation is driven by mapping existing identifiers in both the primary resources and Wikidata. It is possible to reconcile based on strings, but this is dangerous due to the ambiguity of the labels used [65]. When items on concepts are added to Wikidata that lack identifiers overlapping with the primary resource, reconciliation is challenging. Based on the Shape Expressions, the following properties are identified for use in reconciliation. For proteins, these are Uniprot ID (P352) and RefSeq protein ID (P637). For genes, these are NCBI Gene ID (P351) and Ensembl Gene (ID) (P594). When sourced information matches none of these properties, then a new item is created, if the concepts from the primary source reconcile with Wikidata items these are updated.

6.8.1 Wikidataintegrator

Wikidata integrator is a Python library [66] that wraps around the Wikidata API [11, 67]. From external resources such as the NCBI, gene and taxonomy statements have been compiled with provenance and assigned to the associated Wikidata items. When an item did not exist (or was not recognized), it was created. The module compiled a list of statements by parsing the primary sources under scrutiny and extracted what statements already existed on Wikidata. A JavaScript Object Notation (JSON) string was created that resembled the JSON data model used by the Wikidata API. This JSON string was then submitted to the Wikidata API for ingestion.

6.8.2 Data integration use cases/validation

WikiPathways and BridgeDb

WikiPathways is a biological pathway database and can visualize the details of interactions between genes, proteins, metabolites, and other entities participating in biological processes. It depends on BridgeDb to map identifiers of external data and knowledge to the identifiers used for the genes, proteins, and metabolites in the pathways [68]. Furthermore, mappings to Wikidata are required to establish the link of biological entities in pathways and journal articles that have those entities as main topics. Therefore, the virus genes and proteins are required to exist in Wikidata, enabling the interoperability between WikiPathways and Wikidata. Additionally, new virus mapping databases for BridgeDb are created by extracting the new virus gene and protein data, including links between Wikidata, NCBI Gene, RefSeq, UniProt, and Guide to Pharmacology Target identifiers using a SPARQL query [69]. The mapping databases will be updated regularly and will allow pathway curators to annotate virus genes and proteins in their pathways and provide link outs on the WikiPathways website.

The COVID-19-related pathways from WikiPathways COVID-19 Portal are added to Wikidata using the approach previously described. For this, a dedicated repository has been set up to hold the Graphical Pathway Markup Language (GPML) files, the internal WikiPathways file format [70]. The GPML is converted into RDF files with the WikiPathways RDF generator [23], while the files with author information are manually edited. For getting the most recent GPML files, a custom Bash script was developed (getPathways.sh in [71]). The conversion of the GPML to RDF uses the previously published tools for WikiPathways RDF [23]. Here, we adapted the code with a unit test that takes the pathways identifier as a parameter. This test is available in the SARS-CoV-2-WikiPathways branch of GPML2RDF along with a helper script (createTurtle.sh). Based on this earlier generated pathway RDF and using the Wikidataintegrator library, the WikiPathways bot was used to populate Wikidata with additional statements and items. The pathway bot was extended with the capability to link virus proteins to

the corresponding pathways, which was essential to support the Wikidata resource. These changes can be found in the sars-cov-2-wikipathways-2 branch.

Scholia

The second use case is to demonstrate how we can link virus gene and protein information to literature. Here, we used Scholia [14] as a central tool. It provides a graphical interface around data in Wikidata, for example, literature about a specific coronavirus protein (e.g., Q87917585 for the SARS-CoV-2 spike protein). Scholia uses SPARQL queries to provide information about topics. We annotated literature around the HCoVs with the specific virus strains, the virus genes, and the virus proteins as "main topic."

6.8.3 Availability of data and materials

All data and corresponding schema are available in Wikidata. All source code is available from GitHub [37–45]

References

[1] John Watkins. "Preventing a covid-19 pandemic". *BMJ*. 2020. 368: .

[2] *outbreak.info*. URL: `https://outbreak.info/` (visited on 11/24/2020).

[3] *Virus Outbreak Data Network (VODAN)*. URL: `https://www.go-fair.org/implementation-networks/overview/vodan/` (visited on 11/24/2020).

[4] *fhircat/CORD-19-on-FHIR*. URL: `https://github.com/fhircat/CORD-19-on-FHIR` (visited on 11/24/2020).

[5] Justin T. Reese et al. "KG-COVID-19: A Framework to Produce Customized Knowledge Graphs for COVID-19 Response". *Patterns*. 2020. : p. 100155.

[6] Marek Ostaszewski et al. “COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms”. *Scientific Data*. 2020. 7 (1): p. 136.

[7] *Coronavirus and Open Science: Our reads and Open use cases*. Coronavirus and Open Science. URL: https://sparceurope.org/coronaopensciencereadsandusecases/ (visited on 11/24/2020).

[8] *Speed Science*. URL: https://graphics.reuters.com/CHINA-HEALTH-RESEARCH/0100B5ES3MG/index.html (visited on 11/24/2020).

[9] Elisabeth Mahase. “Covid-19: six million doses of hydroxychloroquine donated to US despite lack of evidence”. *BMJ (Clinical research ed.)* 2020. 368: p. m1166.

[10] *Wikidata*. URL: https://www.wikidata.org/wiki/Wikidata:Main%5C%5FPage (visited on 11/24/2020).

[11] Andra Waagmeester et al. “Wikidata as a knowledge graph for the life sciences”. *eLife*. 2020. 9 (e52614): e52614.

[12] Denny Vrandečić and Markus Krötzsch. “Wikidata: a free collaborative knowledgebase”. *Commun. ACM*. 2014. 57 (10): pp. 78–85.

[13] Sebastian Burgstaller-Muehlbacher et al. “Wikidata as a semantic framework for the Gene Wiki initiative”. *Database (Oxford)*. 2016. 2016: pp. 1–10.

[14] Finn Årup Nielsen, Daniel Mietchen, and Egon Willighagen. “Scholia, Scientometrics and Wikidata”. Paper presented at: *The Semantic Web: ESWC 2017 Satellite Events*. Cham. Springer International Publishing. 2017. pp. 237–259.

[15] Fredo Erxleben et al. “Introducing Wikidata to the Linked Data Web”. Paper presented at: *The Semantic Web – ISWC 2014*. Cham. Springer International Publishing. 2014. pp. 50–65.

[16] *RDF 1.1 Concepts and Abstract Syntax*. URL: https://www.w3.org/TR/rdf11-concepts/ (visited on 11/24/2020).

[17] *Wikidata Query Service*. URL: https://w.wiki/6cAS (visited on 11/25/2020).

[18] *Getting the Most out of Wikidata: Semantic Technology Usage in Wikipedia's Knowledge Graph - International Center for Computational Logic*. URL: https://iccl.inf.tu-dresden.de/web/Inproceedings3044/en (visited on 11/25/2020).

[19] Scott Federhen. "The NCBI Taxonomy database". *Nucleic Acids Res*. 2012. 40 (Database issue): pp. D136–143.

[20] Garth R. Brown et al. "Gene: a gene-centered information resource at NCBI". *Nucleic Acids Res*. 2015. 43 (Database issue): pp. D36–42.

[21] The UniProt Consortium et al. "UniProt: the universal protein knowledgebase". *Nucleic Acids Res*. 2017. 45 (D1): pp. D158–D169.

[22] . "Protein Data Bank: the single global archive for 3D macromolecular structure data". *Nucleic acids research*. 2019. 47 (D1): pp. D520–D528.

[23] Andra Waagmeester et al. "Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources". *PLOS Computational Biology*. 2016. 12 (6): e1004989.

[24] Eric W Sayers et al. "Database resources of the National Center for Biotechnology Information". *Nucleic Acids Research*. 2020. 48 (D1): pp. D9–D16.

[25] Katherine Thornton et al. "Using Shape Expressions (ShEx) to Share RDF Data Models and to Guide Curation with Rigorous Validation". Paper presented at: *The Semantic Web*. Cham. Springer International Publishing. 2019. pp. 606–620.

[26] Eric Prud'hommeaux, Jose Emilio Labra Gayo, and Harold Solbrig. "Shape expressions: an RDF validation and transformation language". Paper presented at: *Proceedings of the 10th International Conference on Semantic Systems*. New York, NY, USA. Association for Computing Machinery. 2014. pp. 32–40.

[27] Na Zhu et al. “A Novel Coronavirus from Patients with Pneumonia in China, 2019”. *New England Journal of Medicine*. 2020. 382 (8): pp. 727–733.

[28] Tim Berners-Lee, James Hendler, and Ora Lassila. “The semantic web”. *Scientific American*. 2001. 284 (5): pp. 34–43.

[29] *SPARQL 1.1 Query Language*. URL: https://www.w3.org/TR/sparql11-query/ (visited on 11/25/2020).

[30] *OWL 2 Web Ontology Language Document Overview (Second Edition)*. URL: https://www.w3.org/TR/owl2-overview/ (visited on 11/25/2020).

[31] *Linked Data - Design Issues*. URL: https://www.w3.org/DesignIssues/LinkedData.html (visited on 11/25/2020).

[32] *The Linked Open Data Cloud*. URL: https://lod-cloud.net/ (visited on 11/25/2020).

[33] Matthias Samwald et al. “Linked open drug data for pharmaceutical research and development”. *Journal of Cheminformatics*. 2011. 3 (1): p. 19.

[34] *Help:Statements - Wikidata*. URL: https://www.wikidata.org/wiki/Help:Statements (visited on 11/25/2020).

[35] Daniel Hernandez, Aidan Hogan, and Markus Kroetzsch. “Reifying RDF: What Works Well With Wikidata?” : p. 16.

[36] *RDFSHape: Online demo implementation of ShEx and SHACL*. RDFSHape. URL: https://zenodo.org/record/1412128 (visited on 11/25/2020).

[37] *virus taxon (E192) - Wikidata*. URL: https://www.wikidata.org/wiki/EntitySchema:E192 (visited on 11/30/2020).

[38] *strain (E174) - Wikidata*. URL: https://www.wikidata.org/wiki/EntitySchema:E174 (visited on 11/27/2020).

[39] *disease (E69) - Wikidata*. URL: https://www.wikidata.org/wiki/EntitySchema:E69 (visited on 11/27/2020).

[40] *virus strain (E170) - Wikidata*. URL: https://www.wikidata.org/wiki/EntitySchema:E170 (visited on 11/27/2020).

[41] *virus gene (E165) - Wikidata*. URL: https://www.wikidata.org/wiki/EntitySchema:E165 (visited on 11/27/2020).

[42] *virus protein (E169) - Wikidata*. URL: https://www.wikidata.org/wiki/EntitySchema:E169 (visited on 11/27/2020).

[43] *biological pathway sourced from WikiPathways in Wikidata (E41) - Wikidata*. URL: https://www.wikidata.org/wiki/EntitySchema:E41 (visited on 11/30/2020).

[44] *protein (E167) - Wikidata*. URL: https://www.wikidata.org/wiki/EntitySchema:E167 (visited on 11/27/2020).

[45] *gene (E75) - Wikidata*. URL: https://www.wikidata.org/wiki/EntitySchema:E75 (visited on 11/27/2020).

[46] *SuLab/Gene_Wiki_SARS-CoV*. URL: https://github.com/SuLab/Gene%5C%5FWiki%5C%5FSARS-CoV (visited on 11/25/2020).

[47] *SuLab/scheduled-bots*. URL: https://github.com/SuLab/scheduled-bots (visited on 11/25/2020).

[48] *Jenkins*. URL: https://www.jenkins.io/index.html (visited on 11/27/2020).

[49] *SARS-COV-Wikipathways [Jenkins]*. URL: http://jenkins.sulab.org/job/SARS-COV-Wikipathways/ (visited on 11/25/2020).

[50] *BridgeDb: Human and SARS-related corona virus gene/protein mapping database derived from Wikidata*. BridgeDb. URL: https://zenodo.org/record/3860798 (visited on 11/25/2020).

[51] *main subject*. URL: https://www.wikidata.org/wiki/Property:P921 (visited on 11/27/2020).

[52] *Wikidata Query Service*. URL: https://w.wiki/6cAV (visited on 11/25/2020).

[53] *Wikidata Query Service*. URL: https://query.wikidata.org/ (visited on 11/25/2020).

[54] Sara El-Gebali et al. "The Pfam protein families database in 2019". *Nucleic Acids Res*. 2019. 47 (D1): pp. D427–D432.

[55] *Wikidata:WikiProject COVID-19 - Wikidata*. URL: https://www.wikidata.org/wiki/Wikidata:WikiProject%5C%5FCOVID-19 (visited on 11/25/2020).

[56] *UniProt*. URL: https://covid-19.uniprot.org/uniprotkb?query=%5Ctextasteriskcentered (visited on 11/25/2020).

[57] *COVID-19 ¡ EMBL-EBI*. URL: https://www.ebi.ac.uk/pdbe/covid-19 (visited on 11/25/2020).

[58] *YASHE*. URL: http://www.weso.es/YASHE/ (visited on 11/25/2020).

[59] *YaShE*. URL: https://zenodo.org/record/3739108 (visited on 11/25/2020).

[60] Daniel Fernández-Álvarez et al. "Inference of Latent Shape Expressions Associated to DBpedia Ontology." Paper presented at: *International Semantic Web Conference (P&D/Industry/BlueSky)*. 2018.

[61] Eric Sayers. *E-utilities Quick Start*. en. Publication Title: Entrez Programming Utilities Help [Internet]. National Center for Biotechnology Information (US), Oct. 2018.

[62] Chunlei Wu, Ian Macleod, and Andrew I. Su. "BioGPS and MyGene.info: organizing online, gene-centric information". *Nucleic Acids Res*. 2013. 41 (Database issue): pp. D561–565.

[63] *43740571*. URL: http://mygene.info/v3/gene/43740571 (visited on 11/25/2020).

[64] *RefSeq: NCBI Reference Sequence Database*. URL: https://www.ncbi.nlm.nih.gov/refseq/ (visited on 11/25/2020).

[65] *Never mind the logix: taming the semantic anarchy of mappings in ontologies*. Never mind the logix. URL: https://douroucouli.wordpress.com/2019/05/27/never-mind-the-logix-taming-the-semantic-anarchy-of-mappings-in-ontologie/ (visited on 11/25/2020).

[66] *SuLab/WikidataIntegrator*. URL: https://github.com/SuLab/WikidataIntegrator (visited on 11/25/2020).

[67] *MediaWiki API help - Wikidata*. URL: https://www.wikidata.org/w/api.php (visited on 11/25/2020).

[68] Martijn P. van Iersel et al. "The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services". *BMC Bioinformatics*. 2010. 11 (1): p. 5.

[69] *bridgedb/Wikidata2Bridgedb*. URL: https://github.com/bridgedb/Wikidata2Bridgedb (visited on 11/25/2020).

[70] Martina Kutmon et al. "PathVisio 3: An Extendable Pathway Analysis Toolbox". *PLOS Computational Biology*. 2015. 11 (2): e1004085.

[71] *wikipathways/SARS-CoV-2-WikiPathways*. URL: https://github.com/wikipathways/SARS-CoV-2-WikiPathways (visited on 11/25/2020).

7

General discussion: From illustrative pathways to machine-readable biological pathway

7.1 Discussion

The field of (bio)curation involves transforming biological knowledge into formats that are both understandable by humans and machines [1, 2]. Biological or scientific knowledge that still remains primarily captured in (peer-reviewed) literature. While scientific literature is the primary source of scientific knowledge, its vast size and limited accessibility pose significant challenges to the accessibility of the literature. There are several reasons why it can be difficult to access scientific literature. For example, some older documents may not have been digitized or added to databases like PubMed that researchers use to find articles. In other cases, journal publishers may put up paywalls that prevent people without a subscription from accessing articles. Additionally, some articles may be stored in file formats, making it hard for computers to read and analyze their contents. Examples are PDF and ReadCube, which are optimized for online and human reading. These factors can limit automatic access to the full text of scientific articles. The issue of hindered access to existing knowledge is exemplified in a recent review paper which describes how acute promyelocytic leukaemia (APL), a once fatal form of leukaemia, became one of the more treatable forms of the disease [3]. The paper highlights the scientific process in action and the decades-long gap between discovering a key chemical compound to APL therapy and its eventual application in modern medicine. This delay occurred primarily because the original papers reporting the early findings were published in a Chinese-language scientific journal, which was even obscure to most Chinese readers. The novel findings in this research paper are hidden in two ways. First, the paper is written in a language that limits its usefulness outside the region where Chinese is spoken. But in this case, the paper was even unknown to many people within that language community because it was published in a journal that was not widely known even to Chinese speakers. This case is similar to where Mendelian inheritance was hidden in ”plain sight” for almost four decades. Gregor Mendel published his theory on inheritance [4] but was initially ignored since his work was not recognised as being groundbreaking [5]. It took decades before the paper was accurately judged on its merit when researchers could put the theory into context [6].

Just publishing and having the paper findable is insufficient; more is needed to allow progress.

We see similar examples where connections to facts are lost by being disconnected. One prime example where integration of facts is hindered is the often-reported longstanding example of the unintended introduction of synonyms for gene labels through auto-correction rules applied using common spreadsheets programs such as Microsoft Excel [7, 8]. Gene names are incorrectly renamed through autocorrection rules used in data processing software. Fixing those issues can take a while before curators identify them [9].

Addressing this issue is of utmost importance, and it is essential to develop innovative solutions that enable the open and accessible sharing of scientific knowledge, first by making the scientific literature more accessible and primarily findable but also to make the contents of the individual articles and their metadata more machine-readable. Especially since the body of scientific literature is expanding unprecedentedly. Numerous articles have described the growth as ”nearly exponential.” Interestingly, getting the actual size of the total sum of the scientific literature is difficult to estimate. However, estimates on the size of citation databases suggest, with some level of uncertainty, confirming the sheer size. Google Scholar, for example, being the biggest, contains 389 million records as reported in 2018. [10].

But there is a nuance to be made here. In chapter 3, where a text-mining pipeline is described, we faced limitations in full access to the needed scientific literature. The main obstacle was paywalls, but automatic processing was also not always permitted, even when universities granted access through subscriptions. Automatic processing of large quantities of scientific articles was not always permitted by the licenses granted. Additionally, extracting full-text literature files automatically was often difficult due to the many file formats used to present scientific content to the public. Sometimes access was made more complicated by embedding additional code (e.g. JavaScript) to get the full content of a given article. Just as the biological databases discussed in this thesis require an interoperability layer, there is also a need for scientific literature to adopt a more machine-readable format to facilitate easier access and build an interoperability layer. Because of these limitations imposed by

(some) publishers on full access, our text-mining pipelines had to rely on abstracts from PubMed only. These abstracts come with limitations. Not all citation records contain abstracts; those abstracts are often limited to a few lines. That is, the information value of an abstract is limited: a 2018 study on 16.5 million articles "show[ed] that text mining of full-text articles consistently outperforms using abstracts only" [11]. Therefore, even though the quantity of academic publications is increasing exponentially, it is arguable whether the total amount of available knowledge is also increasing at a similar pace. In most cases, the abstract is readily available, while the underlying full text not always.

In fact, some have argued that the nearly exponential growth of scientific output could have the opposite effect of causing knowledge loss, or at least a decline in the skills needed to process and acquire it [12]. Whether the total sum of knowledge is truly becoming increasingly inaccessible due to this growth remains a subject for further exploration

In the context of pathway curation, incomplete access to knowledge on related concepts can lead to incomplete information. Pathways are an abstraction of cellular processes and metabolic interactions and thus need to be as comprehensive as possible. This does not mean that all relevant knowledge needs to be embedded into the pathway boundaries but that sufficient links to that complete knowledge should exist, as used by bioinformatics tools such as CyTargetLinker [13]. To get an accurate pathway, a pathway curator should ideally be able to find most, if not all, relevant knowledge on the process being scrutinized in the curated pathway with some level of confidence. While the limitations mentioned above exist, we identified novel pathway concepts using text-mining, showing that text-mining can be a valuable curation resource for pathway curation.

However, the limited access to existing knowledge, as described above, due to obscure file formats or poorly described access protocols, still hinders the ability of pathway curators to access potentially relevant information. To facilitate pathway curation, future work in improving access to the (relevant) literature remains relevant. A separate thread of future work is assessing the

quality by having machine-readable access to curator annotations on the quality of the individual papers. Innovations like RetractionWatch and PubPeer are slowly gaining traction [14, 15].

Although, as said, future work is needed to improve access to the relevant literature, it is also worth noting that curation also heavily relies on biological databases, as discussed in chapter 2. Pathway curators benefit from a skill set that includes knowledge extraction from the literature and the available biological databases, which are generally more structured than scientific literature and thus should be more findable and accessible. However, similar to the point made by Eve Marder et al. [12], skills in accessing the various database access methods may also be diminishing. This is also driven by the rapid developments of novel technologies where interoperability with former technologies is not always maintained. The interoperability of biological databases is an ongoing debate [12, 16, 17]. Full access to the sum of all knowledge from scientific literature and biological databases remains problematic, and more research is needed.

Full access to both the literature and biological databases is particularly relevant in the context of pathway curation and analysis [18]. As highlighted in both the introduction and Chapter 2, pathways have become essential tools for research, enabling data to be placed within a wider context [19, 20]. To achieve this, unrestricted access to the relevant knowledge stored in scientific literature and biological databases remains essential. This means direct and unambiguous access to the total sum of gained knowledge both on a conceptual and infrastructural level.

7.2 Structured Representations of Biological Knowledge through Pathways

As pathway curators abstract existing knowledge into biological pathways, these pathways become structured representations of the biological knowledge they contain. This provides researchers with a scaffold or template to align their study results against the available knowledge. The COVID-19 Disease Map project is a prime example of this role, creating a computational

repository of SARS-CoV-2 virus-host interaction mechanisms using machine-readable pathway formats that contain biological database identifiers for linking study results [21]. With technological advancements and computational power, genomics has moved beyond single-gene studies to a more comprehensive approach. Researchers can now analyze and interpret vast amounts of genomic data to gain insights into complex biological systems, including interactions between multiple cells and genes [22]. Integrating data from various file formats and access protocols is essential for pathway curators to navigate the research landscape and draft accurate pathway descriptions. Additionally, the advancement of single-cell analysis and long-read sequencing and large-scale variant evaluations and related technologies has made an ever-growing amount of data available [23–25]. The same knowledge captured in biological pathways provides the means to organize and oversee the available sequencing data and associated expression studies.

As discussed in Chapter 2, the vast amount of research data generated in biological studies can be overwhelming, and mathematical methods such as clustering, machine learning, and statistics are often employed to organize and analyze the data [26]. These methods do not always lead to increased understanding, but rather to different representations of the same data. In contrast, biological pathways have evolved from simple illustrations of processes to a valuable research tool that can increase understanding [27]. By presenting findings in visually appealing diagrams, these pathways make it easier to identify areas where changes or patterns are apparent. While pathway diagrams remain visually appealing, recent advancements allow pathways to be expressed in machine-readable formats, which can be visualized using rendering tools [28]. Moreover, with these machine-readable descriptions, research results can be projected onto pathways, provided that the pathway parts and concepts use external identifiers from commonly known biological databases [29] (see Figure 7.1). For example, expression values or changes in expression can be visualized on a pathway diagram using various colour gradients. Complementing these mathematical methods with pathway visualization can lead to a better understanding of the biological process or system being studied [22].

The main argument of this thesis is that comprehensive pathway curation re-

quires unrestricted access to all relevant biological knowledge, which can be found in the scientific literature or biological databases. It must be in machine-readable format to handle the overwhelming amount of data to enable automated processing. Although this knowledge's ideal level of completeness remains uncertain, this thesis aims to investigate techniques to support the retrieval, normalization, and standardization of biological knowledge.

Chapter 4 is crucial in normalising and standardising biological knowledge, as it focuses on transforming pathway content into linked data (i.e., RDF). In this chapter, I propose an RDF implementation for WikiPathways. By merging the native file format (GPML) with explicit semantics (including identifiers, ontologies), WikiPathways became seamlessly integrated with other data resources on the semantic web. This approach has enabled the improved integration of pathway models with biological databases and scientific literature through linked-data principles, which allow for the connection and linking of relevant information. By rendering pathway content as linked data, we have facilitated more efficient access to the available knowledge and resources in the field.

Creating an RDF version of WikiPathways was a time-consuming task requiring software development to convert the various formats for storing pathway knowledge into the RDF model. Additionally, it required the selection of controlled vocabularies and ontologies to ensure that the meaning behind concepts was unambiguously conveyed or the creation of a new ontology where none existed. Two controlled vocabularies were specifically developed for WikiPathways: one to capture biological concepts and another to describe the graphical features of a rendered pathway diagram. At the time of the first release of the RDF of WikiPathways, we had to also deviate from the linked-data principles that require that URIs point to linked data. In some cases, we had to rely on URLs instead. These URLs would then point to online documents instead of linked-data instances. One example is using identifiers.org, where `https://identifiers.org/wikipathways/WP716` resolves to the pathway Vitamin A and carotenoid metabolism (wikipathways:WP716). Minting native URIs for all pathways was administratively impossible since the hosting institute had specific requirements for pointers to resources

hosted by the institute. Using identifiers from identifiers.org instead solved this constraint by providing a URI that redirects to the pathways hosted on a web server.

After creating the RDF for WikiPathways, it is necessary to continuously curate and update the controlled vocabularies to ensure they reflect novel insights and are better integrated with other ontologies. However, it is often preferable to reuse an existing ontology that requires less maintenance or where the curation effort is shared with a broader audience. While there were existing ontologies for representing pathway knowledge, such as BioPAX [30], we chose to create novel ontologies for WikiPathways, which required, as said, significantly more effort than simply reusing. The decision to develop novel controlled vocabularies - that are however mapped to existing ontologies - for WikiPathways was motivated by the fact that the existing models did not provide the necessary level of detail for the specific use case. For example, in WikiPathways, a concept can exist multiple times within the same pathway to capture different features (e.g. an intra-cellular and extra-cellular role). Additionally, existing models did not provide the capability to capture the graphical features of a pathway diagram. However, to maintain the promise of interoperability, mappings were provided to align with those models. With current insights, if I would develop the RDF for WikiPathways from scratch, I would have developed a WikiPathways ontology instead of a controlled vocabulary, and I would be more strict on following the linked-data principles. Identifiers.org does not support linked data, and it is a lot easier to set up and maintain one's own URI space.

Often, the decision to create a new ontology or engage with existing communities to improve or detail existing models depends on various factors, such as the project's specific requirements, available resources, time constraints, and the willingness of existing communities to collaborate. As said earlier, the reuse of existing ontologies can be preferred. However, creating a specialized ontology may be necessary to address specific requirements and capture the details for a particular use case. Ultimately, it is essential to strike a balance between creating new ontologies and engaging with existing communities to ensure interoperability and sustainability in the long run. When the RDF for WikiPathways work was started, there were not many ontologies or controlled

vocabularies available that had the correct semantic meaning, which justified the decision to create a new set of controlled vocabularies. However

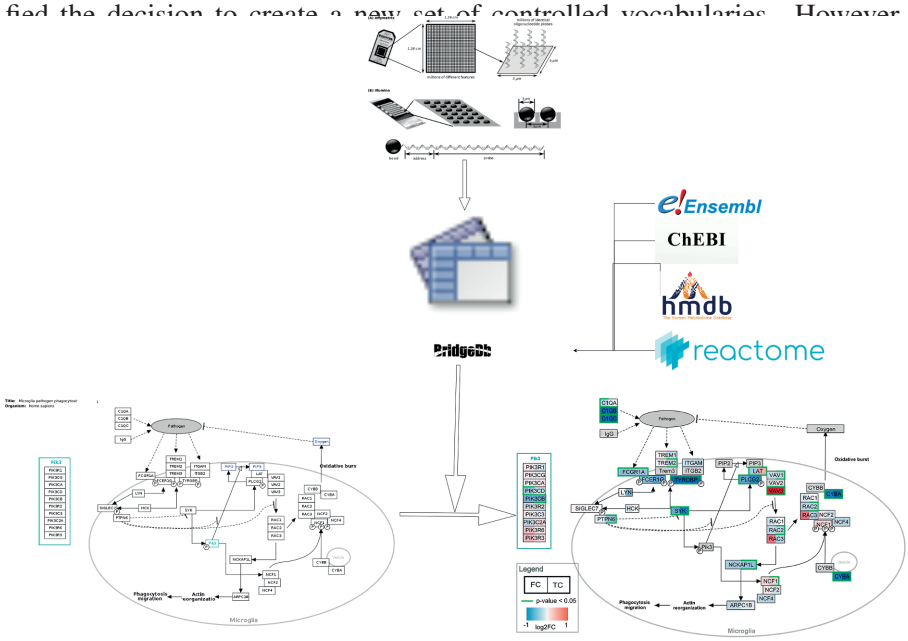


Figure 7.1: Having a pathway diagram available in a machine-readable format allows its application in the visualization of gene expression, which is nicely illustrated in the two diagrams, which visualize the frontal and temporal cortex gene expression on the Microglia pathogen phagocytosis pathway (wikipathways:WP3626) [38]. This visualization is possible even though the expression values use identifiers from different identifier schemes. These are aligned using identifier mapping in BridgeDb to use identifiers in the pathway model. The expression value is then rendered using colour gradients to show the magnitude of the expression

7.3 Text mining

With the developed RDF model for WikiPathways, we can streamline the accessibility and interoperability of pathway knowledge for pathway curators. Linking related structured data requires applying similar workflows for those related biological databases. However, while much knowledge is already captured in the many biological databases, much remains hidden in the scientific literature. In Chapter 3, a text-mining pipeline was explored to extract pathway-related concepts from the literature [39, 40]. We identified novel pathway parts by applying named entity recognition and profile extraction on a large body of text extracted from many articles.

The method described in Chapter 3 is a form of text mining, a technique used to extract useful information from large bodies of text. This study identified novel pathway extensions on a carotenoid metabolic pathway (wikipathways:WP716) by counting terms in a corpus of topical scholarly articles. This leads to topical profiles, vectors that can be used to identify similar articles. After retrieval of those articles from the full corpus - seed plus extended corpus - existing terms were extracted through named-entity recognition. A team of biocurators then assessed those terms in their context and assessed their relevance for that pathway. This way, we successfully identified novel pathway concepts that contributed to new additions to WikiPathways. However, we acknowledge the need to assess the effectiveness of our approach. Specifically, comparing the results obtained through text mining with those generated by a focus group of experts reading the same seed corpus would be valuable. Extracting knowledge automatically is a valuable tool for pathway curators and nicely illustrates the potential of text mining for biocuration, providing a solution for the extensive and ongoing influx of new literature.

We demonstrated the feasibility of extracting novel terms for the studied pathway using a text-mining pipeline. However, there are several important caveats to consider. Firstly, the seed corpus used in this study was extensive and meticulously curated, which may not be representative of other corpora. Additionally, referring to this workflow as semi-automatic is an understatement, as it required extensive input and collaboration from the curators and bioinformaticians involved. The commitment of the biocurators

and the incentive of a peer-reviewed journal publication served as motivation to see the project through to completion, as peer-reviewed publications continue to hold significant value in the scientific community [41]. It is making the possible reward of a paper, part of the workflow. For the application of this text-mining pipeline in the daily practice of biocuration, there is no incentive for routine text-mining, because it will not lead to a peer-reviewed publication, each time the workflow would be applied. Yet having to go through all the papers and extract the same terms manually would require that same team to put in a lot of time, if at all possible. To scale this workflow, the availability of seed corpora needs to be increased. In this study, we utilized a supplement to a paper to generate a seed corpus. However, exploring alternative approaches for creating topical reference lists is essential. One promising avenue is to consider reference lists of existing pathways or from community efforts such as Gene Ontology as potential seed corpora. One method could be to rely on citations attached to a pathway or a group of related pathways. WikiPathways now recognizes topical portals that might be instrumental in providing citations to publications that can be used as seed corpus.

Overall, this approach demonstrates the feasibility of text-mining techniques in extracting useful information from large volumes of scientific literature and highlights the potential for identifying novel research directions and targets for further investigation. Our method started with a seed corpus of approximately 2000 scholarly articles selected as being on topic. This was substantially extended to a difficult-to-build corpus due to its sheer number of papers that needed to be selected on their relevance.

In our approach, we focused on text mining and did not consider the potential value of bibliometrics in either composing the seed corpora or extending them. However, several alternative approaches are available for extracting knowledge or discovering relevant literature. These include utilizing citations in databases or other structured resources and conducting direct analyses of pathway figures [19, 42]. By exploring these alternative approaches, we may gain a more comprehensive understanding of relevant literature and improve the effectiveness of our research.

Finally, and as a segue into the next section, it is with noting that I have focused primarily on string similarity, which could lead to inaccuracies in the results and could use further refinement. Relying solely on string similarity for profile matching can be limited because single words can have multiple meanings or be expressed using different terminology in natural language. This can result in inaccurate identification of relevant concepts or detection of incorrect identifiers. Disambiguation of terms and synonym expansion needs to be addressed.

Chapter 4 addresses the disambiguation of terms by exploring the usefulness of the semantic web, where, instead of words, concepts are described using normalized identifiers expressed as internationalized resource identifiers (IRI).

7.4 Aligning pathways with the Semantic Web

Chapter 4 introduces the semantic web to align heterogeneous biological data sets and knowledge bases. When curating the text mining results described in Chapter 3, we had to address the issue of spelling variations and the mapping of identifiers. The latter is particularly prevalent in pathway curation since pathways as vehicles of aggregated biological knowledge can contain identical concepts identified by identifiers from different identifier providers. For example, a specific gene can be identified using gene identifiers from the NCBI gene database [43] or gene identifiers from the Ensembl gene database [44]. Sometimes the same issue with spelling variations applies to identifiers since some resources prefer using prefixes to identify a concept. e.g. a gene identified by an NCBI gene id can be identified using ”ncbi:1234567”, ”entrez:1234567” or ”1234567”. Identifiers.org and the new Bioregistry.io address this issue [45, 46].

Chapter 4 explored the value of the semantic web as an interoperability platform that deals with harmonising concepts across multiple biological resource issues. The semantic web is an extension to the World Wide Web (WWW) that became a W3C recommendation in 2008 [47, 48]. With WWW, each document requires human assessment to grasp its full context. The semantic web

uses this same protocol to link data points. However, where in the WWW the identifier points to a document on a server in the WWW, in the semantic web, the identifiers point to a concept. Another nuance is that with the WWW, that identifier, also known as Uniform Resource Locator (URL), points to a document on a physical computer giving access to that document that it hosts. On the semantic web, the identifier points to a concept, which can be hosted on a server, in which case it is also a URL, however on the semantic web, the identifier can also exist as a virtual identifier which is why it is called a Universal Resource Identifier (IRI). Its value stems from the fact that if two (independent) resources use the same URI (sequence of characters) they are addressing the same concept. While for a document pointed to by its URL, human assessment is needed to put an IRI into context because the IRI does not locate anything. If a concept is thoroughly described on the semantic web, however, the context is explicit by the collection of triples [1] of which the URI is a part of.

To give an example, on the WWW, a pathway can be depicted in its graphical representation (Figure 7.2) or a machine-readable format (Figure 7.3)

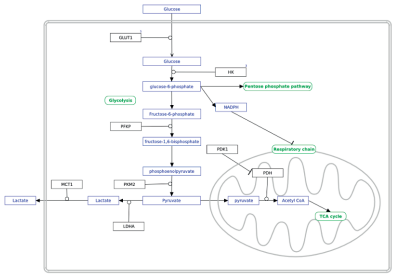


Figure 7.2: **A.** A pathway in WikiPathways where the diagram is drawn by a community of (online) curators

Figure 7.3: **B.** The same pathway but now rendered in a machine-actionable format (GPML)

The WikiPathways RDF representation differs from the earlier-described syntactical GPML representation of a pathway from which a pathway diagram is

[1] see Chapter 4 for a detailed description on URIs and triples

rendered. Often these machine-readable representations are specific to a given resource. E.g. WikiPathways utilizes the GPML [49] format, which is specific for WikiPathways (and its associated pathway analysis tool PathVisio [50]). The semantic web solves the issue of having to deal with multiple file formats, by creating a unified representation of data, where all data is represented using a statement-driven file format that reflects the subject-verb-object sentence structure used in English and many other languages. In the semantic web, these statements are called triples, of which a detailed description is given in Chapter 4. By depicting knowledge as a set of triples using normalized identifiers, such as identifiers.org, the problem of varying file formats is solved. So expressing knowledge on the semantic web requires the different existing formats to be transformed and rendered as triples. The idea is that both resources become interoperable by structuring knowledge this way and doing the same for similar resources. After all, knowledge from both resources is represented in the same format. However, reciprocal reuse also requires resources to use the same collection of unified identifiers or use explicit mapping. As is described in Chapter 6, rendering knowledge on the semantic web is a two step process—i.e. the transformation in triples and identifying the same identifier patterns across different resources [33].

Converting any file format or data structure to RDF is a straightforward task. The main challenge is to determine appropriate and pertinent identifiers. This contradicts the perception that RDF is complicated and suitable only for highly skilled computer scientists. Transforming from one format, such as CSV, to RDF can be easily accomplished using automated tools available on various platforms [2]. However, adding semantics by selecting and creating relevant controlled vocabularies and ontologies requires more domain expertise than technical skills. This is the domain of a biocurator with profound knowledge of the field.

Creating RDF pipelines is predominantly handled by individuals with technical expertise. There is a need for further research on how to involve domain experts in the overall process without requiring them to have a strong understanding of technologies such as SPARQL, RDF, ShEx, and others. One

[2] https://www.w3.org/wiki/ConverterToRdf

should not have to get a post-graduate computer science degree to be able to design and apply linked-data frameworks. Some technical skills or understanding by life scientists are needed to effectively capture knowledge as linked data. Future work should focus on developing procedures that effectively capture the knowledge of biomedical domain experts. This point is nicely reiterated in the following quote [51]:

> *"People think RDF is a pain because it is complicated. The truth is even worse. RDF is painfully simplistic, but it allows you to work with real-world data and problems that are complicated. While you can avoid RDF, it is harder to avoid complex data and computer problems. RDF brings together data across application boundaries and imposes no discipline on mandatory or expected structures. This can make working with RDF data frustrating."*

7.5 Wikidata: a linked-data proxy for the life sciences

Like with any data format, hosting RDF requires a significant amount of maintenance and updating that goes beyond just maintaining the servers. In addition to ensuring the servers run smoothly, organizations must keep their URIs' persistence by establishing an infrastructure and community to curate the ontologies and controlled vocabularies used. The ongoing cost of sustaining the infrastructure beyond funding cycles can be daunting, requiring dedicated maintenance staff and funding.

In 2012 Wikidata emerged as a sister project of Wikipedia [52]. It follows the same system principles as Wikipedia. Where Wikipedia stores its articles in a relational database, Wikidata stores its data on the same MediaWiki software as Wikipedia. The parallel can be extended to WikiPathways, where the GPML representation of the pathways was originally stored in the same platform. It is safe to say that what articles are to Wikipedia and what XML representation of pathways is to WikiPathways, is what semantically structured data is to Wikidata. The data in Wikidata follows the same structure as RDF: a concept is represented as an item consisting of a set of statements (for a detailed description of a Wikidata item, I defer to Chapter 4). The statements

resemble the RDF triples, except that a statement also consists of a set of qualifiers and references which provide the immediate context and provenance of the individual statements. Initially, Wikidata was only accessible through the same API as Wikipedia. However, early on, Wikidata was extended with a SPARQL endpoint allowing integration of Wikidata content to the semantic web. Currently, data on Wikidata exists in two (redundant) forms: first, as a collection of JSON blobs in the above-mentioned relational database, and second, a copy as RDF in a triple store with an associated SPARQL endpoint [53]. Adding the RDF layer made Wikidata a strong linked-data proxy. Without that RDF layer, Wikibase (the Wikidata backend) can only be queried with a (rich) set of predefined API queries and, thus, yet another conventional database. Adding RDF (and SPARQL) to Wikidata combines the combinatorial expression of RDF with the Wikis. The difference between Wikibase and any RDF triple store is that Wikibase allows editing a single statement without understanding the underlying RDF principles.

Wikidata has items on most Wikipedia articles (from 300 language versions). The Wikipedia narrative defines the meaning of the linked Wikidata item, which then acts as a robust, controlled vocabulary. This is subsequently enriched by many contributors on Wikidata with links to existing ontologies. Because the core backend of Wikidata is not an RDF store, Wikidata needed to invent its metadata model and cannot easily extend on existing ontologies offered by resources like the OBO foundry [54, 55], schema.org [37], BioPortal [34], etc. However, Wikidata does provide a reified redundant RDF layer and through the stored mappings to external ontologies and controlled vocabularies, it makes Wikidata part of the semantic web.

Chapter 5 describes the progress of the Gene Wiki project [56, 57] which has been enriching Wikidata with life science data on genes, proteins, diseases and chemical compounds by collecting public data on those concepts and align that with semantic models of those concepts on Wikidata. Started as a project to increase the coverage of knowledge on genes and proteins, the Gene Wiki project moved to the Wikidata project, adding gene and protein annotations and related diseases and chemical compounds. We did this by parsing public data sources on the matter and aligning them with Wikidata. By doing so,

we also aligned the sourced public resources with the semantic web, making Wikidata a gateway for life-science data to the Semantic Web.

7.6 Shape Expressions

The last chapter describes shape expressions and how these support curation. As formal language to describe RDF, Shape Expressions allow for describing use-case expectations and data descriptions. This is to communicate existing data schemas; we can detect inconsistencies automatically by having those machine-readable. The importance of this is exemplified in the following example.

Syntactic issues such as spelling variations and differences in table formats can be resolved by representing information on the semantic web as triples using URIs. As described earlier, RDF allows representing that knowledge in any preferred format. I.e. the semantic web solves any syntactical issue. However, variations in how data is being represented remain. This time in, what I would call semantic variations. An example is how gene-disease associations are expressed as RDF triples on the semantic web. It can be expressed by a single triple where the subject is a gene directly linked to the relevant predicate with its associated disease.

The same association can, however, also be expressed through intermediate steps, which means more details. To stick with the gene-disease example, the knowledge of the actual association can also be described through the involved proteins encoded by that gene and even possible protein-protein interactions, where that knowledge might even be scattered across different (RDF) data stores.

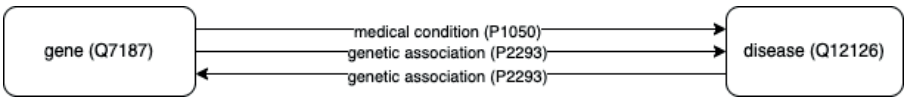


Figure 7.4: In Wikidata, at least two properties allow capturing direct gene-disease associations.

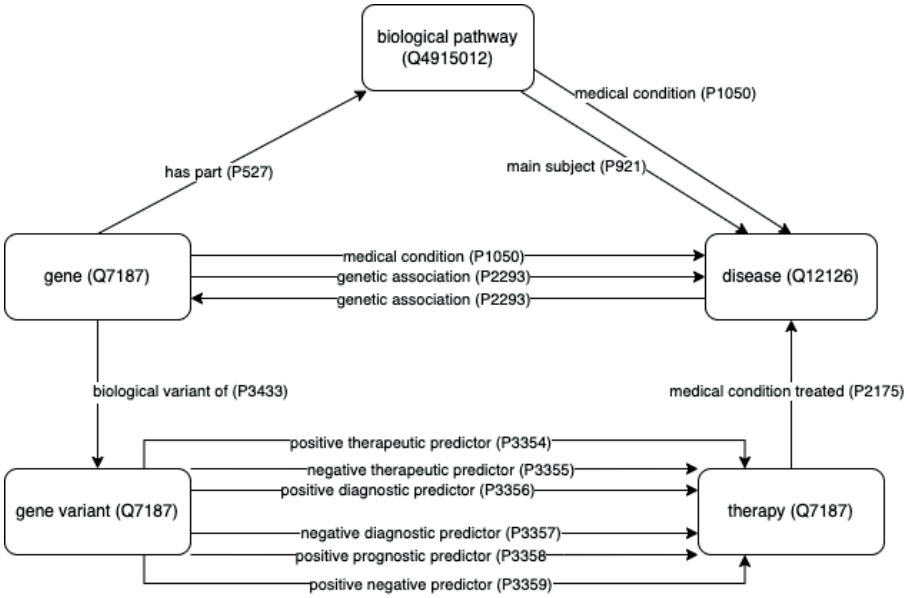


Figure 7.5: In Wikidata, at least two indirect links exist between genes and diseases. Knowing all these indirect connections is necessary to find all gene-disease associations in Wikidata

So, even while rendering knowledge on the semantic web removes the need to deal with spelling variations and different formats, substantial curation efforts are still needed to parse and process the available expertise. The knowledge can use different paths using other controlled vocabularies and ontologies, sometimes called namespaces, to capture the same expertise. Figures 7.4 and 7.5 demonstrate two examples of variations within a single namespace - i.e. Wikidata - to capture knowledge using linked concepts. A curator does need to know how the knowledge available in the source data resources exists. An intuitive way to do so is by following the different paths starting from general concepts. However, this is tedious and existing patterns are easily missed. In Chapter 6, we use Shape Expressions [58, 59] to describe graph patterns found in Wikidata to capture existing knowledge on COVID in the onset of the global pandemic. These Shape Expressions are a way to either provide documentation by the data owners or allow users to express expectations. E.g. it documents that for each virus protein we expect a statement about which virus it is encoded by.

7.7 Conclusion

Chapter 6 on Shape Expressions concludes the thesis where a full path is provided on pathway curation with the means to efficiently explore available knowledge needed to complete biological pathways with existing knowledge.

In conclusion, this thesis has provided a comprehensive pathway (pun intended) towards exposing existing biological knowledge with the aim to support more efficient pathway reuse, including curation, analysis, and integration. Steps involved are: 1. Extraction of knowledge from large bodies of text; 2. expressing this and related data on the semantic web. Biological pathways presented in machine-readable language help present documentation or user expectations to the user, making biological literature and databases more accessible. In this thesis, I have shown that this approach gives us new tools to explore, visualize, and understand the underlying biology in more detail, by making semantics explicit forcing us to think about the detailed differences

and when to generalise knowledge. With this approach, scientists can focus more on the real meaning of the research topics and less on setting up and maintaining infrastructure. While substantial efforts remain, this thesis has taken a significant step toward making biological literature easier to use with machine-readable language.

The title of this thesis is ”Biological Pathway Abstractions: from two-dimensional Drawings to Multidimensional linked-data.”. Two-dimensional panes are essential in research to abstract the available facts. These are in the form of drawings or two-dimensional data frames. While valuable, there remains a risk of knowledge loss by fitting the multidimensional reality into a limited set of two-dimensional panes. We are trying to understand our surroundings by compacting the observations into two dimensions to reconstruct a total image. Storing observations in multidimensional linked data will allow more granularity in the reconstruction of a model from the observed facts, simply by the sheer number of facets that can be stored as machine-readable research data. With the linked-data principles, we have this opportunity already. The research tools and user interface to digest and process linked data still have quite some love from our research community.

References

[1] D. Howe et al. “Big data: The future of biocuration”. *Nature*. 2008. 455 (7209): pp. 47–50.

[2] Sarah Burge et al. “Biocurators and biocuration: surveying the 21st century challenges”. *Database (Oxford)*. 2012. 2012 (0): bar059.

[3] Y. Rao, R. Li, and D. Zhang. “A drug from poison: how the therapeutic effect of arsenic trioxide on acute promyelocytic leukemia was discovered”. *Sci China Life Sci*. 2013. 56 (6): pp. 495–502.

[4] Gregor Mendel. “Versuche über Pflanzen-Hybriden”. *Verhandlungen des naturforschenden Vereines in Brünn*. (1865-1866). Bd.4 (1865-1866): pp. 3–47.

[5] Elizabeth B. Gasking. “Why was Mendel’s Work Ignored?” *Journal of the History of Ideas*. 1959. 20 (1): pp. 60–84.

[6] William Bateson, Gregor Mendel, and Arthur G. Leighton. *Mendel’s principles of heredity, by W. Bateson*. https://www.biodiversitylibrary.org/bibliography/1057. Cambridge [Eng.], University Press, 1909, 1909, p. 448.

[7] Barry R Zeeberg et al. “Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics”. *BMC bioinformatics*. 2004. 5 (1): pp. 1–6.

[8] Dyani Lewis. “Autocorrect errors in Excel still creating genomics headache.” *Nature*. 2021. : .

[9] *Can we use the citation graph to measure the quality of a taxonomic database?*

[10] Michael Gusenbauer. “Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases”. *Scientometrics*. 2018. (118): pp. 177–214.

[11] D. Westergaard et al. “A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts”. *PLoS Comput Biol*. 2018. 14 (2): e1005962.

[12] Eve Marder and Shimon Marom. “Lost knowledge”. *Current Biology*. 2022. 32 (4): R144–R145.

[13] M. Kutmon et al. “CyTargetLinker app update: A flexible solution for network extension in Cytoscape”. *F1000Res*. 2018. 7: .

[14] Robert M Kwee and Thomas C Kwee. “Retracted Publications in Medical Imaging Literature: an Analysis Using the Retraction Watch Database”. *Academic Radiology*. 2023. 30 (6): pp. 1148–1152.

[15] José Luis Ortega. “Classification and analysis of PubPeer comments: How a web journal club is used”. *Journal of the Association for Information Science and Technology*. 2022. 73 (5): pp. 655–670.

[16] Pieter BT Neerincx and Jack AM Leunissen. “Evolution of web services in bioinformatics”. *Briefings in bioinformatics*. 2005. 6 (2): pp. 178–188.

[17] Martín Pérez-Pérez et al. “Next generation community assessment of biomedical entity recognition web servers: metrics, performance, interoperability aspects of BeCalm”. *Journal of Cheminformatics*. 2019. 11 (1): p. 42.

[18] Martijn P van Iersel et al. “Presenting and exploring biological pathways with PathVisio”. *BMC Bioinformatics*. 2008. 9: p. 399.

[19] K. Hanspers et al. “Pathway information extracted from 25 years of pathway figures”. *Genome Biol*. 2020. 21 (1): p. 273.

[20] K. Hanspers et al. “Ten simple rules for creating reusable pathway models for computational analysis and visualization”. *PLoS Comput Biol*. 2021. 17 (8): e1009226.

[21] M. Ostaszewski et al. “Author Correction: COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms”. *Sci Data*. 2020. 7 (1): p. 247.

[22] Y. Hasin, M. Seldin, and A. Lusis. “Multi-omics approaches to disease”. *Genome Biol*. 2017. 18 (1): p. 83.

[23] Chandra Shekhar Pareek, Rafal Smoczynski, and Andrzej Tretyn. “Sequencing technologies and genome sequencing”. *Journal of applied genetics*. 2011. 52 (4): pp. 413–435.

[24] Tuuli Lappalainen et al. “Genomic analysis in the age of human genome sequencing”. *Cell*. 2019. 177 (1): pp. 70–84.

[25] Antoine-Emmanuel Saliba et al. “Single-cell RNA-seq: advances and future challenges”. *Nucleic acids research*. 2014. 42 (14): pp. 8845–8860.

[26] David B. Allison et al. “Microarray data analysis: from disarray to consolidation and consensus”. *Nat Rev Genet*. 2006. 7 (1): pp. 55–65.

[27] Michael P. Cary, Gary D. Bader, and Chris Sander. “Pathway information for systems biology”. *FEBS Lett*. 2005. 579 (8): pp. 1815–1820.

[28] Martina Kutmon et al. “PathVisio 3: An Extendable Pathway Analysis Toolbox”. *PLOS Computational Biology*. 2015. 11 (2): e1004085.

[29] Martijn P. van Iersel et al. “The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services”. *BMC Bioinformatics*. 2010. 11 (1): p. 5.

[30] Lena Strömbäck and Patrick Lambrix. “Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX”. *Bioinformatics*. 2005. 21 (24): pp. 4401–4407.

[31] P. L. Whetzel et al. “BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications”. *Nucleic Acids Res*. 2011. 39 (Web Server issue): W541–545.

[32] Pierre-Yves Vandenbussche et al. “Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web”. *Semantic Web*. 2017. 8 (3): pp. 437–452.

[33] J. Malone et al. “Ten Simple Rules for Selecting a Bio-ontology”. *PLoS Comput Biol*. 2016. 12 (2): e1004743.

[34] Clement Jonquet et al. “AgroPortal: A vocabulary and ontology repository for agronomy”. *Computers and Electronics in Agriculture*. 2018. 144: pp. 126–143.

[35] Simon Jupp et al. “A new Ontology Lookup Service at EMBL-EBI.” *SWAT4LS*. 2015. 2: pp. 118–119.

[36] Alasdair Gray, Carole Goble, and Rafael Jimenez. “Bioschemas: From Potato Salad to Protein Annotation”. Paper presented at: *The 16th International Semantic Web Conference 2017*. Vienna. CEUR Workshop Proceedings. 2017. pp. 1–4.

[37] Ramanathan V Guha, Dan Brickley, and Steve Macbeth. “Schema. org: evolution of structured data on the web”. *Communications of the ACM*. 2016. 59 (2): pp. 44–51.

[38] R. A. Miller et al. “Beyond Pathway Analysis: Identification of Active Subnetworks in Rett Syndrome”. *Front Genet*. 2019. 10: p. 59.

[39] T. Clark, P. N. Ciccarese, and C. A. Goble. “Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications”. *J Biomed Semantics*. 2014. 5: p. 28.

[40] A. González-Beltrán et al. “From Peer-Reviewed to Peer-Reproduced in Scholarly Publishing: The Complementary Roles of Data Models and Workflows in Bioinformatics”. *PLoS One*. 2015. 10 (7): e0127612.

[41] No authors listed. “Publish or perish”. *Nature*. 2010. 467 (7313): p. 252.

[42] *Measuring impact in online resources with the CI-number (the CitedIn Number for online impact)*. URL: https://doi.org/10.1038/npre.2011.6037.1.

[43] NCBI Resource Coordinators. “Database resources of the national center for biotechnology information”. *Nucleic acids research*. 2018. 46 (D1): pp. D8–D13.

[44] Andrew Yates et al. “Ensembl 2016”. *Nucleic Acids Res*. 2016. 44 (Database issue): pp. D710–D716.

[45] Sarala M. Wimalaratne et al. “Uniform resolution of compact identifiers for biomedical data”. *Scientific Data*. 2018. 5 (1): p. 180029.

[46] Charles Tapley Hoyt et al. “Unifying the Identification of Biomedical Entities with the Bioregistry”. *bioRxiv*. 2022. : .

[47] *The Semantic Web - Scientific American*. URL: https://www.scientificamerican.com/article/the-semantic-web/ (visited on 11/25/2020).

[48] *Semantic Web - W3C*. URL: https://www.w3.org/standards/semanticweb/ (visited on 01/16/2021).

[49] Alexander R. Pico et al. “WikiPathways: Pathway Editing for the People”. *PLOS Biology*. 2008. 6 (7): e184.

[50] Augustin Luna et al. “PathVisio-MIM: PathVisio plugin for creating and editing Molecular Interaction Maps (MIMs)”. *Bioinformatics*. 2011. 27 (15): pp. 2165–2166.

[51] *RDFSHape: Online demo implementation of ShEx and SHACL*. RDFSHape. URL: https://zenodo.org/record/1412128 (visited on 11/25/2020).

[52] Denny Vrandečić. "Wikidata: a new platform for collaborative data collection". Paper presented at: *Proceedings of the 21st International Conference on World Wide Web*. New York, NY, USA. Association for Computing Machinery. 2012. pp. 1063–1064.

[53] Stanislav Malyshev et al. "Getting the most out of Wikidata: semantic technology usage in Wikipedia's knowledge graph". Paper presented at: *International Semantic Web Conference*. Asilomar conference grounds. Monterey, California, USA. Springer Nature Switzerland AG 2018. 2018. pp. 376–394.

[54] Barry Smith et al. "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration". *Nature biotechnology*. 2007. 25 (11): pp. 1251–1255.

[55] Rebecca Jackson et al. "OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies". *Database*. 2021. 2021: pp. 1–9.

[56] III Huss Jon W. et al. "The Gene Wiki: community intelligence applied to human gene annotation". *Nucleic Acids Research*. 2009. 38 (suppl_1): pp. D633–D639.

[57] Sebastian Burgstaller-Muehlbacher et al. "Wikidata as a semantic framework for the Gene Wiki initiative". *Database (Oxford)*. 2016. 2016: pp. 1–10.

[58] Eric Prud'hommeaux, Jose Emilio Labra Gayo, and Harold Solbrig. "Shape expressions: an RDF validation and transformation language". Paper presented at: *Proceedings of the 10th International Conference on Semantic Systems*. New York, NY, USA. Association for Computing Machinery. 2014. pp. 32–40.

[59] Katherine Thornton et al. "Using shape expressions (ShEx) to share RDF data models and to guide curation with rigorous validation". Paper presented at: *European Semantic Web Conference*. Slovenia. springer. 2019. pp. 606–620.

Impact Paragraph

The main objective of the thesis is to explore how to analyze effectively and, where possible, transform biological knowledge into a structured, machine-readable and semantically enriched data format. This is to facilitate large-scale integration of knowledge into pathway diagrams and models. We could extract relevant concepts from the literature and extend this knowledge by integrating them with other formally described knowledgebases. This thesis led to better-integrating pathway knowledge with the literature and other (linked-)data sources.

Results from this thesis proved to be relevant to the broader research community. Bibliographic studies on citations of the published work stemming from some of the chapters in this thesis show subsequent and downstream usage of the pathway data with other linked-data resources. This has been demonstrated by integrating the RDF generated by the workflow developed in work from this thesis in various projects, of which Open PHACTS is the most prominent. WikiPathways RDF consistently ranks high in ranking frameworks such as YummyData [1].

Wikidata, a community-curated knowledge base, is described in this thesis as a potential source of knowledge for pathway curation. The wikidata workflows described in this thesis addressed structuring pathway knowledge in the public knowledge base but did so for other biomedical and chemical public data sources. Judging from the different citations to the work on wikidata described in this thesis has been inspirational to others. It shows at least some impact of this work beyond the narrative of pathway curation. As the author, I was able to set up a scientific research startup called Micelio BV; since its inception in 2014, it has been involved in various projects, not limited to pathway curation alone. Micelio BV is a company that provides services to different actors, primarily in the life sciences domain. Micelio is an economically viable company that became a partner in various initiatives and projects in

[1] http://yummydata.org/

healthcare, biomedical research, agriculture and cultural heritage. The company is built on the reputation gained and the results from the output in this thesis.

The results in this thesis are primarily geared towards pathway curators. However, since pathways act as an abstraction of cellular processes and interactions, they can also be seen as a hub of biological knowledge in the broader sense. Making pathway knowledge machine-readable can be instrumental in making biological knowledge more impactful, making the results relevant for a wider audience.

This thesis proved the value of methods developed in the field of bioinformatics. A more comprehensive application of these methods and research results requires further investment in capacity training for biocurators. This field needs formal training in linked data and its affiliated methods to fully appreciate its potential for biocuration.

Summary

This thesis explored methods to extract biological knowledge for pathway curation efficiently. While many platforms host scientific knowledge, scientific literature remains the central pillar of existing scientific knowledge. Scientific literature, for quite some time now, has been growing at an ever-increasing pace. With existing access constraints, this growth poses challenges for researchers to obtain knowledge efficiently.

The use case in this thesis is pathway curation. Pathways are like biological maps that show how different parts of a living organism work together. In this thesis, I explore other methods to facilitate and structure knowledge so that a pathway curator has unrestricted access to existing knowledge next to access to scientific knowledge.

In Chapter 2, we explore biological pathways and how the field of bioinformatics helps in pathway curation. Bioinformatics plays a role in organizing biological knowledge in formats that serve automatic pipelines. We also mention that pathways evolved from mere illustrations in discussions among peers towards full-fledged models that can be used in research pipelines to process and assess findings. Chapter 3 describes a text-mining pipeline that semi-automatically identifies potential pathway parts. These parts were presented to pathway curators, who validated their relevance and extended the scrutinized pathway with novel features. Text mining leads to structured data from loosely structured. Other approaches from the field of Bioinformatics lead to a myriad of structured data formats. Chapter 4 presents a representation of pathway knowledge on a framework that, next to the data, also formally captures its semantics. This framework makes rapid integration with other structured data sources possible. Chapter 5 follows up on the linked-data concept introduced in Chapter 4. In that chapter, we describe how an online linked-data platform, called Wikidata, is used as a hub in linking similarly structured linked data resources, as the one described in chapter 4. Finally, Chapter 6 describes a protocol to document the linked data. This is because while the knowledge is structured and formalized as linked data, the underlying schema of linked data

remains implicit. By using formal language describing linked-data schemas, curators and data owners alike can formally describe expectations from the data and what is offered by the data. This allows for rapid validation of expectations.

The thesis started its exploration from loosely structured scientific knowledge towards a structured representation of data and expectations.

Samenvatting

Dit proefschrift gaat over het efficiënt blootleggen van biologische en medische kennis die verstopt zit in de literatuur en biologische databanken. Er ligt een specifieke focus op de biologische pathways. Dit zijn schematische weergaven van reacties en interacties tussen cellen, hun metabolieten en hun omgeving. Hoewel er verschillende platformen en databanken bestaan die wetenschappelijke kennis kunnen opslaan en representeren, blijft de wetenschappelijke literatuur de centrale bron van wetenschappelijke kennis. Zonder evaluatie en validatie door medewetenschappers, het zogenaamde ”peer review”, bestaat kennis in de wetenschappelijke traditie niet. Echter, de wetenschappelijke literatuur groeit al geruime tijd in een steeds sneller tempo. Toegang tot die kennis wordt steeds meer een uitdaging, enerzijds door die toenemende groei, maar anderzijds ook door bestaande toegangsbeperkingen die op grote schaal toegepast worden. Er zijn vaak dure abonnementen vereist met als gevolg dat de beschikbare informatie niet altijd eenvoudig te vinden is.

Heel specifiek gaat het hier over ”pathway curatie”. Dit is een vorm van biocuratie gericht op het verzamelen, beoordelen, organiseren en annoteren van biologische gegevens voor representatie en weergave in pathways. Biocuratie omvat ook het beheren van biologische databases en het waarborgen van de kwaliteit, consistentie en nauwkeurigheid van de verzamelde informatie. Pathways zijn als biologische kaarten die laten zien hoe verschillende delen van een levend organisme samenwerken. In dit proefschrift onderzoek ik verschillende methoden om kennis te faciliteren en te structureren, zodat een bioloog of andere wetenschapper onbeperkt toegang heeft tot bestaande kennis.

In hoofdstuk 2 onderzoeken en beschrijven we verschillende vormen van deze pathways en hoe de bioinformatica helpt bij het onderhouden en actueel houden van deze biologische kaarten. Over het algemeen speelt de bioinformatica een rol bij het organiseren van biologische kennis in computer structuren, van waaruit verschillende automatische pijplijnen kunnen

worden bediend. We vermelden ook dat pathways zijn geëvolueerd van louter illustraties in discussies tussen collega's naar volwaardige modellen die kunnen worden gebruikt in onderzoekspijplijnen om bevindingen te verwerken en te beoordelen.

Hoofdstuk 3 beschrijft een automatische tekstanalyse- en extractie pijplijn die semi-automatisch potentiële nieuwe elementen identificeert, vertrekkende vanuit de wetenschappelijke literatuur. Deze potentieel nieuwe elementen werden gepresenteerd aan domeindeskundigen, die de potentiele relevantie van die elementen bevestigden en het onderzochte pathway uitbreidden met nieuwe kennis. Deze automatische tekstontginningsmethodieken leiden tot gestructureerde gegevens vanuit tekst, die vervolgens geïntegreerd kunnen worden in pathways. Vanuit de bioinformatica zijn verschillende bestandsformaten gekomen, die opslag van kennis op een gestructureerde kennis mogelijk maken. Dit biedt de mogelijkheid om te structureren, maar niet altijd op inhoud. Het verbinden van verschillende bestandsformaten kan nog lastig zijn. Daarvoor is ook een gestructureerde representatie van de inhoud van de bestanden nodig. Hoofdstuk 4 presenteert een representatie van pathway kennis op een raamwerk dat, naast de gegevens, ook formeel de semantiek vastlegt. Dit framework maakt een snelle integratie met andere gestructureerde databronnen mogelijk. Hoofdstuk 5 gaat voort op de integratie van kennis op zowel structuur als inhoudelijk niveau, zoals dat beschreven werd in hoofdstuk4. In dat hoofdstuk beschrijven we hoe een online linked-data platform, genaamd Wikidata, wordt gebruikt als een hub voor het koppelen van gelijkaardig gestructureerde linked data-bronnen Ten slotte beschrijft hoofdstuk 6 een protocol om de gekoppelde gegevens te documenteren. Dit is nodig omdat hoewel de kennis is gestructureerd en geformaliseerd is als gekoppelde gegevens, het onderliggende schema van gekoppelde gegevens impliciet blijft. Computer kunnen daar mee overweg, maar voor menselijke beoordeling blijft het nodig om die kennis schematisch weer te geven. Door gebruik te maken van een formele taal die schema's met gekoppelde gegevens beschrijft, kunnen zowel beheerders als gegevenseigenaren formeel de verwachtingen van de gegevens beschrijven en ook wat de gegevens bieden. Dit zorgt voor een snelle bevestiging van verwachtingen.

Het proefschrift begon zijn verkenning van losjes gestructureerde wetenschappelijke kennis naar een gestructureerde weergave van gegevens en verwachtingen. Dit allemaal om ten goede van volledige pathway beschrijvingen volledige toegang tot de biomedische kennis te hebben.

Acknowledgments

The thesis is the final delivery of a long and extensive journey with many Dit proefschrift is de apotheose of lang en landurig traject met veel omwegen. Ik draag dit proefschrift op aan mijn vader Nico, die dit helaas niet meer mee kan maken. De inspiratie om te promoveren is eigenlijk door hem met de paplepel ingegegoten.

Much gratitude also goes to Persephone Doupi, a colleague of mine from the time when I was still seeking a career in medical informatics. Persanophene was the one that pointed me in the direction of Bioinformatics.

The next person in the chain of influential persons is Chris Evelo. Chris, I still vividly recall your bold rejection of my application for a PhD position in your lab. The funding failed, so you could not fulfil that position. Much to my surprise, you followed up with the bold message that even if the funding would have gone through, you wouldn't hire me for that post, because I lacked wet lab experience. It is a boldness I appreciate in many to this day. Shortly after I got the opportunity to secure a Marie Curie fellowship at the Pasteur Institute in Paris, we negotiated that this fellowship would be the extended job application for a possible next PhD opportunity.

That opportunity surfaced, and I started on the Phaser project in the Jaap van den Herik group, not your group, but still at Maastricht University. When that group moved to Tilburg University, I finally ended up in BiGCaT. I sometimes wonder what would have happened with my career if I had chosen the path of most resistance. However, being content with where I am today and looking back, it is clear that this path provided me with the necessary steps to have a satisfying career in Open Science and Linked data.

Chris and Jaap, I am grateful to the way you both have shaped my career and thinking in science in general and life science specific.

I also owe much gratitude to the co-promotors Susan Coort and Egon Willighagen.

Susan, it has been a great pleasure to have worked with you during our work on the carotenoid pathways, which is now a core chapter in this thesis. Also, I would argue that our curation session on the carotenoids is probably my first series of ”Zoom” meetings. I was at the EBI, you in Maastricht. You just started there after I left for a 6-month training at the EBI. Today, it is so common to meet colleagues first on Zoom before meeting in person, but that was a first for me. I vividly remember discussing going through the lengthy Excel sheets resulting from the different text-mining pipelines. I also reflect on the fun and professional time we shared an office in Maastricht. It is a great honour that you are part of the supervising committee. Egon, we met as Biostars. You joined the group as a stellar Biostar, and your Biostar is still shining. You have been inspirational ever since, and I appreciate our lengthy discussions over the years. Sometimes, if not often, those discussions preach to our parish, but sometimes, it helps to find some resonance. Thanks for your patience, especially in the final steps towards this thesis. Susan and Egon, I thank you both for believing in me where I often had already given up.

Tina and Alex, WikiPathways RDF would not have happened without our extended coffee session in Heidelberg. That half-day breakfast, morphing into lunch, was where we did the largest part of the WPRDF work. You both have showed that most work can be done in a very short time. It is working out the details that is often underestimated.

Gratitude should also be expressed to Martijn, Stan, Thomas, Arie, Magali, Gontran, Jahn, Guido, Igor, Guillaume, Joyca, Sander, Steven, Evgemi for being my contempories. I have learned a lot from you and fondly reflect on the slacking and procrastinating that pushed us through.

This journey couldn’t have been done without social support as well. I am eternally grateful to my parents, Marijke and Nico. You both have shown me the path and the conviction that one’s destiny should be sought beyond individual borders. It is an excellent tradition that each generation switches country, at least in the Verkuijl Branch of the family. Special posthumous gratitude also goes to my grandfather Joachim. I lack his almost natural feeling for meticulous organisation, at least physically. On an organisational level, I have adopted some of his organisational traits. I think it is safe to say that

my grandfather and father have been very influential in my life and deserve much gratitude. If I have to believe Stef Bos, we will never meet again. After all, you both believed in god, and I believe in nothing. So, hopefully, by submitting this thesis, there is some eternity in my gratitude.

I also owe much gratitude to co-authors, all collaborators, colleagues and friends who have been (willingly or unwillingly) part of this project. I notably acknowledge all involved in the Rhebolz text mining group at the EBI and the WikiPathways, Open PHACTS and Gene Wiki projects.

In addition to those mentioned, I extend my heartfelt gratitude to all who have supported me in any capacity during my Ph.D. journey. Your influence has been a vital part of this endeavour, whether big or small. If I have omitted anyone, it is not due to a lack of appreciation but an oversight in the whirlwind of this extensive journey. I offer my deepest thanks to each of you who has been a part of this process, seen or unseen, spoken or silent. Your contributions have been invaluable, and this achievement is as much yours as it is mine.

Finally, special thanks should go to Anton de Vries, who has been a really good friend and constant in this journey. What started at the editorial offices of ”Verband” at the AMC, where many of our ventures started. Thanks for being such a good friend.

Finally, An, Emeline en Maïté, my biggest appreciation should maybe go to you three. I am proud and honoured that Emeline and Maïte can be part of the defence as the paranymphs. While Chris was pivotal in my pursuit of a PhD in Bioinformatics in Maastricht. An, if you hadn’t come into my life, that career would probably not have happened in Maastricht.

Andra Waagmeester
Maastricht
2024-01-16

Curriculum Vitae & list of publications

Experience

2016–current

Founder, Micelio BV

2014–2016

Freelance consultancy under the Micelio brand

2006

Visitor, Rebholz Group, European Bioinformatics Institute, Hinxton, United Kingdom

2005–2014

PhD Candidate, Maastricht University, the Netherlands

2003–2005

Marie Curie Fellowship, Pasteur Institute, Paris, France

2000–2003

Junior Researcher, Department of Medical Informatics, Erasmus MC, Rotterdam, the Netherlands

1998–2003

Medical Informatician, Department of Hematology, VUMC Medical Center, Amsterdam, the Netherlands

Publications

- Shafee, T. , Mietchen, D. , Lubiana, T. , Jemielniak, D. , **Waagmeester, A.** . ”Ten quick tips for editing Wikidata”. In: *PLoS Comput Biol*. URL: https://pubmed.ncbi.nlm.nih.gov/pmid37471307/

- Meldal, B. H. M. , Perfetto, L. , Combe, C. , Lubiana, T. , Ferreira Cavalcante, J. V. , Bye-A-Jee, H. , **Waagmeester, A.** , Del-Toro, N. , Shrivastava, A. , Barrera, E. , Wong, E. , Mlecnik, B. , Bindea, G. , Panneerselvam, K. , Willighagen, E. , Rappsilber, J. , Porras, P. , Hermjakob, H. , Orchard, S. . ”Complex Portal 2022: new curation frontiers”. In: *Nucleic Acids Res*. URL: https://pubmed.ncbi.nlm.nih.gov/pmid34718729/

- Hanspers, K. , Kutmon, M. , Coort, S. L. , Digles, D. , Dupuis, L. J. , Ehrhart, F. , Hu, F. , Lopes, E. N. , Martens, M. , Pham, N. , Shin, W. , Slenter, D. N. , **Waagmeester, A.** , Willighagen, E. L. , Winckers, L. A. , Evelo, C. T. , Pico, A. R. . ”Ten simple rules for creating reusable pathway models for computational analysis and visualization”. In: *PLoS Comput Biol*. URL: https://pubmed.ncbi.nlm.nih.gov/pmid34411100/

- Martens, M. , Ammar, A. , Riutta, A. , **Waagmeester, A.** , Slenter, D. N. , Hanspers, K. , A Miller, R. , Digles, D. , Lopes, E. N. , Ehrhart, F. , Dupuis, L. J. , Winckers, L. A. , Coort, S. L. , Willighagen, E. L. , Evelo, C. T. , Pico, A. R. , Kutmon, M. . ”WikiPathways: connecting communities”. In: *Nucleic Acids Res*. URL: https://pubmed.ncbi.nlm.nih.gov/pmid33211851/

- **Waagmeester, A.** , Stupp, G. , Burgstaller-Muehlbacher, S. , Good, B. M. , Griffith, M. , Griffith, O. L. , Hanspers, K. , Hermjakob, H. , Hudson, T. S. , Hybiske, K. , Keating, S. M. , Manske, M. , Mayers, M. , Mietchen, D. , Mitraka, E. , Pico, A. R. , Putman, T. , Riutta, A. , Queralt-Rosinach, N. , Schriml, L. M. , Shafee, T. , Slenter, D. , Stephan, R. , Thornton, K. , Tsueng, G. , Tu, R. , Ul-Hasan, S. , Willighagen, E. , Wu, C. , Su, A. I. . ”Wikidata as a knowledge graph for the life sciences”. In: *Elife*. URL: https://pubmed.ncbi.nlm.nih.gov/pmid32180547/

- Putman, T. , Hybiske, K. , Jow, D. , Afrasiabi, C. , Lelong, S. , Cano, M. A. , Stupp, G. S. , **Waagmeester, A.** , Good, B. M. , Wu, C. , Su, A. I. . ”ChlamBase: a curated model organism database for the Chlamydia research community”. In: *Database (Oxford)*. URL: https://pubmed.ncbi.nlm.nih.gov/pmid31211397/

- Slenter, D. N. , Kutmon, M. , Hanspers, K. , Riutta, A. , Windsor, J. , Nunes, N. , lius, J. , Cirillo, E. , Coort, S. L. , Digles, D. , Ehrhart, F. , Giesbertz, P. , Kalafati, M. , Martens, M. , Miller, R. , Nishida, K. , Rieswijk, L. , **Waagmeester, A.** , Eijssen, L. M. T. , Evelo, C. T. , Pico, A. R. , Willighagen, E. L. . ”WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research”. In: *Nucleic Acids Res*. URL: https://pubmed.ncbi.nlm.nih.gov/pmid29136241/

- Putman, T. E. , Lelong, S. , Burgstaller-Muehlbacher, S. , **Waagmeester, A.** , Diesh, C. , Dunn, N. , Munoz-Torres, M. , Stupp, G. S. , Wu, C. , Su, A. I. , Good, B. M. . ”WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research”. In: *Database (Oxford)*. URL: https://pubmed.ncbi.nlm.nih.gov/pmid28365742/

- **Waagmeester, A.** , Kutmon, M. , Riutta, A. , Miller, R. , Willighagen, E. L. , Evelo, C. T. , Pico, A. R. . ”Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources”. In: *PLoS Comput Biol*. URL: https://pubmed.ncbi.nlm.nih.gov/pmid27336457/

- Putman, T. E. , Burgstaller-Muehlbacher, S. , **Waagmeester, A.** , Wu, C. , Su, A. I. , Good, B. M. . ”Centralizing content and distributing labor: a community model for curating the very long tail of microbial genomes”. In: *Database (Oxford)*. URL: https://pubmed.ncbi.nlm.nih.gov/pmid27022157/

- Burgstaller-Muehlbacher, S. , **Waagmeester, A.** , Mitraka, E. , Turner, J. , Putman, T. , Leong, J. , Naik, C. , Pavlidis, P. , Schriml, L. , Good, B. M. , Su, A. I. . ”Wikidata as a semantic framework

for the Gene Wiki initiative". In: *Database (Oxford)*. URL: https://pubmed.ncbi.nlm.nih.gov/pmid26989148/

- Wilkinson, M. D. , Dumontier, M. , Aalbersberg, I. J. , Appleton, G. , Axton, M. , Baak, A. , Blomberg, N. , Boiten, J. W. , da Silva Santos, L. B. , Bourne, P. E. , Bouwman, J. , Brookes, A. J. , Clark, T. , Crosas, M. , Dillo, I. , Dumon, O. , Edmunds, S. , Evelo, C. T. , Finkers, R. , Gonzalez-Beltran, A. , Gray, A. J. , Groth, P. , Goble, C. , Grethe, J. S. , Heringa, J. , 't Hoen, P. A. , Hooft, R. , Kuhn, T. , Kok, R. , Kok, J. , Lusher, S. J. , Martone, M. E. , Mons, A. , Packer, A. L. , Persson, B. , Rocca-Serra, P. , Roos, M. , van Schaik, R. , Sansone, S. A. , Schultes, E. , Sengstag, T. , Slater, T. , Strawn, G. , Swertz, M. A. , Thompson, M. , van der Lei, J. , van Mulligen, E. , Velterop, J. , **Waagmeester, A.** , Wittenburg, P. , Wolstencroft, K. , Zhao, J. , Mons, B. . "The FAIR Guiding Principles for scientific data management and stewardship". In: *Sci Data*. URL: https://pubmed.ncbi.nlm.nih.gov/pmid26978244/

- Kutmon, M. , Riutta, A. , Nunes, N. , Hanspers, K. , Willighagen, E. L. , Bohler, A. , lius, J. , **Waagmeester, A.** , Sinha, S. R. , Miller, R. , Coort, S. L. , Cirillo, E. , Smeets, B. , Evelo, C. T. , Pico, A. R. . "WikiPathways: capturing the full diversity of pathway knowledge". In: *Nucleic Acids Res*. URL: https://pubmed.ncbi.nlm.nih.gov/pmid26481357/

- Ratnam, J. , Zdrazil, B. , Digles, D. , Cuadrado-Rodriguez, E. , Neefs, J. M. , Tipney, H. , Siebes, R. , **Waagmeester, A.** , Bradley, G. , Chau, C. H. , Richter, L. , Brea, J. , Evelo, C. T. , Jacoby, E. , Senger, S. , Loza, M. I. , Ecker, G. F. , Chichester, C. . "The application of the open pharmacological concepts triple store (open PHACTS) to support drug discovery research". In: *PLoS One*. URL: https://pubmed.ncbi.nlm.nih.gov/pmid25522365/

- Willighagen, E. L. , **Waagmeester, A.** , Spjuth, O. , Ansell, P. , Williams, A. J. , Tkachenko, V. , Hastings, J. , Chen, B. , Wild, D. J. . "The ChEMBL database as linked open data". In: *J Cheminform*. URL: https://pubmed.ncbi.nlm.nih.gov/pmid23657106/

- Evelo, C. T. , **Waagmeester, A.** . ”Nature Europe site should highlight most productive countries”. In: *Nature*. URL: https://pubmed.ncbi.nlm.nih.gov/pmid20535178/
- Adriaens, M. E. , Jaillard, M. , **Waagmeester, A.** , Coort, S. L. , Pico, A. R. , Evelo, C. T. . ”The public road to high-quality curated biological pathways”. In: *Drug Discov Today*. URL: https://pubmed.ncbi.nlm.nih.gov/pmid18652912/
- **Waagmeester, A.** , Thompson, J. , Reyrat, J. M. . ”Identifying sigma factors in Mycobacterium smegmatis by comparative genomic analysis”. In: *Trends Microbiol*. URL: https://pubmed.ncbi.nlm.nih.gov/pmid16140533/
- Ceusters, W., Buekens, F.,De Moor G., **Waagmeester, A.**. ”The distinction between linguistic and conceptual semantics in medical terminology and its implication for NLP-based knowledge acquisition”. In: *Methods Inf Med.* URL: https://pubmed.ncbi.nlm.nih.gov/pmid19865030/

About the author

Andra Waagmeester's story is one of continuous learning and adaptation in bioinformatics and medical informatics. Born in Utrecht and raised in the diverse environment of Suriname, Andra developed an early interest in the life sciences.

He pursued this interest at the University of Amsterdam, where he completed a master's degree in Medical Information Science. His career began at the VUMC Medical Center in Amsterdam and continued at the Department of Medical Informatics at Erasmus MC in Rotterdam. He contributed to projects focused on enhancing electronic patient records and clinical guidelines there.

Seeking a new challenge, Andra transitioned into the field of bioinformatics. This shift led him to the Pasteur Institute in Paris, where he engaged in comparative genome research, gaining valuable insights into the complexities of gene annotation.

Andra's journey then took him to Maastricht University, where he further expanded his expertise in bioinformatics. Specifically, focus on pathway annotations, text mining and the semantic web. Andra was also involved in teaching about database theory, programming and various other topics in various curricula in Maastricht. Next, His collaboration with the Gene Wiki project, led by Andrew Su of Scripps in San Diego, marked a significant phase in his career. Balancing personal commitments with professional aspirations, Andra engaged in remote collaboration on the Gene Wiki project, demonstrating the possibilities of modern work arrangements.

In 2016, Andra founded Micelio, a consultancy and research firm. Through Micelio, he has contributed to various projects in healthcare, biology, and biodiversity, as well as in the field of cultural heritage data. This venture represents his commitment to applying knowledge in practical and diverse contexts.

Andra's career path showcases his adaptability and dedication to the evolving world of informatics.

Retinal pigment epithelium specific
Liver/other organs
Nucleus
RE + dietary lipids
Retinyl ester
all-trans Retinol
Vitamin A
all-trans Retinal
11-cis-Retinol
11-cis-Retinal
Rhodopsin formation
9-cis-Retinol
9-cis-Retinal
9-cis-Retinoic acid
all-trans Retinoic acid
Vitamin A acid
4-oxo-Retinoic acid
13,14-dihydroretinal
ROL-binding proteins
RA-binding proteins
Cytochrome P450 family
Betacarotene
beta 10'-apocarotenal
Lycopene
Lutein
Canthaxanthin
Beta-Cryptoxanthin
Retinal
Vitamin A aldehyde
Retinol dehydrogenases
MAPK
RARa
RARb
RARg
RXRa
RXRb
RXRg
Cyp26a1
Cyp26b1
Cyp2c1
Aldh1a1
Aldh1a2
Aldh1a3
Rbp1
Rbp2
Rbp4
Rbp7
Crabp1
Crabp2
Sult1a1
Sult2b1
gap junction communication