Mining the History Sections of Wikipedia Articles on Science and Technology

Wolfgang Kircheis
Leipzig University and ScaDS.AI

Marion Schmidt
DZHW

Arno Simons
Technische Universität Berlin

Benno Stein
Bauhaus-Universität Weimar

Martin Potthast
Leipzig University and ScaDS.AI

ABSTRACT

Priority conflicts and the attribution of contributions to important scientific breakthroughs to individuals and groups play an important role in science, its governance, and evaluation. Debates and dynamics around these processes are analyzed by science studies. Our objective is to transform Wikipedia into an accessible, traceable primary source for analyzing such debates. In this paper, we introduce Webis-WikiSciTech-23, a new corpus consisting of science and technology Wikipedia articles, focusing on the identification of their history sections. We extract such articles from Wikipedia dumps through iterative filtering of the category network. The identification of passages covering the historical development of innovations is achieved by combining heuristics for section heading analysis and classifiers trained on a ground truth of articles with designated history sections.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; **Language resources**.

KEYWORDS

Wikipedia, Science Studies, Priority Disputes, Science, Technology, Science and Technology, Innovation

1 INTRODUCTION

In science, particularly following significant technological breakthroughs, disputes often arise over scientific priority and credit allocation among individuals and research groups. These disputes are crucial for stakeholders, science governance, awarding prizes, funding, and commercial applications. Science and innovation studies investigate these dynamics, identifying key factors and dispute resolutions. Take the example of the 2020 Nobel Prize in Chemistry awarded to Jennifer Doudna and Emmanuelle Charpentier for contributing to the CRISPR-Cas method for genome editing [13]: The UC Berkeley team surrounding Doudna and Charpentier filed a patent first, but Feng Zhang's team at the Broad Institute pursued a fast-tracked review, leading to a patent interference claim from Berkeley [9]. The situation became even more complex when the *Patent Trial and Appeal Board* determined that the Broad Institute had priority for inventions not included in Berkeley's patent [5]. This patent conflict and the underlying dispute thus joined the long history of similar disputes in the field of biotechnology.

Numerous sources exist for studying priority disputes in science, but there are few standardized approaches. Bibliometrics has developed field delineation methods using keyword queries, clustering, bibliometric coupling, co-citation, or co-author networks [27], but these approaches cannot provide insights into narratives and controversies. Editorial accounts provide interpretation but are prone to subjective bias. In contrast, collaborative accounts, such as Wikipedia, play an increasingly significant role in collaborative historiography based on multilateral communication. Wikipedia's 'freedom to edit' allows for near-real-time updates, and its open access, coverage, and topicality result in high web search rankings, constantly attracting new readers and editors. Researchers can gain insights into the editing process and contributors' decisions through Wikipedia talk pages. The platform's article revision tracking provides a timeline of scientific debates in a research field, supported by Wikipedia's policy of requiring cited sources [24]. Previous studies have specifically examined Wikipedia's coverage of the CRISPR development [2, 18, 19].

Wikipedia articles on scientific innovations and discoveries often include sections that concisely outline the discovery's history. In a case study on CRISPR-Cas [20], we have observed that this section is frequently changing and subject to debate on which researchers should be acknowledged within this section. We propose a method to access the history sections of Wikipedia articles on scientific innovations as primary sources for researchers. In a first step, we apply an iterative filtering of Wikipedia's category network to identify relevant articles. Given the lack of a standardized approach both for naming and identifying history sections in Wikipedia, we then assess the effectiveness of title-based heuristics and computational classification approaches in identifying history sections. Our evaluation demonstrates that a combination of heuristics and classification is the optimal strategy, enabling us to compile the Webis Wikipedia Science and Technology 2023 corpus (Webis-WikiSciTech-23)[1], a high-precision resource for science studies to track the evolution of priority disputes.[2]

2 RELATED WORK

Wikipedia serves as a critical tool for researchers and educators, with its quality maintained by human editors adhering to community standards [12]. It is recognized as a significant source of encyclopedic knowledge [8, 21, 26] and a resource for exploring historical development in societal controversies [3]. As a continuously growing resource, it covers past and present developments, making it a lexical semantic resource [10]. Consequently, Wikipedia has been utilized in text categorization, information extraction, information retrieval, question answering, computing semantic relatedness, and named entity recognition, as well as a collaborative knowledge base for domain-specific named entities, phrases, and terms [26].

[1]Code: https://github.com/webis-de/JCDL-23
[2]Data: https://doi.org/10.5281/zenodo.7845809

Wikipedia is used for a variety of text extraction tasks, often focused on article sections, as they “are the building blocks of Wikipedia articles” [16]. As about one quarter of all English language articles have only one or even no sections and the vast majority of headings are only ever used once, Piccardi et al. recommend (sub)sections for articles by finding sections from similar articles using topic modeling, collaborative filtering, and Wikipedia’s category system, the latter being most successful [16]. WIKITABLET (‘Wikipedia Tables to Text’) matches tabular and metadata in Wikipedia articles with their respective sections [4]. Schenkel et al. extend Wikipedia dumps with “semantically rich, self-explaining tags” by exploiting Wikipedia’s category network [17]. As many Wikipedia entries lack section subdivision and have inconsistent headings, Field et al. generate section titles for Wikipedia articles with BERT encoders and RNN decoders [7]. Liu and Iwaihara extract representative phrases for sections from external articles containing the same words as the target article. They retrieve candidate articles by calculating the TF-IDF-based cosine-similarity between related articles and each section and using frequent phrases to extract co-occurring word sets, then pipe phrases into search engines and rank them using Gradient Descent [11]. Aprosio and Tonelli record a growing interest in the task of extracting biographical information from data and name Wikipedia “the main source of information for research in this direction”. Seeing as Wikipedia’s lack of consistent templates for describing biographies has led to various page types to describe a person’s life, they employ Conditional Random Fields (CRFsuite) and compare them to Support Vector Machines (YAMCHA) but conclude that a baseline using the most frequent words appearing in the section heading is the most successful approach [1]. According to Lin et al. many Wikipedia-based studies and systems incorrectly assume that similar concepts have a one-to-one mapping across different language editions. They address this article-as-concept assumption and try to solve the sub-article matching problem to “identify all corresponding sub-articles in the same language edition”. They parse out sub-article candidates, mostly using regular expressions, then use SVMs, Random Forests, Decision Trees, Naïve Bayes, Logistic Regression, and Adaboost to identify sub-articles [10]. As a significant number of Wikidata entries has no corresponding article in any language, Ostapuk et al. map these ‘orphan entries’ to (sub)sections using graphs and token-key comparison [14].

3 METHODOLOGY

3.1 Finding Science and Technology Articles

From the more than 16 million entries in the Wikimedia dump from 1 January 2022, all articles are extracted using their Wikipedia namespace [23], yielding a total of 6,129,024 articles with an extractable section tree. A custom WikitextReader cleans and processes the Wikitext, extracting headings, text, and categories, and builds a section tree. The articles are filtered using their categories to find articles on innovative technologies, scientific concepts, theories, and procedures (‘**science and technology**’). Only articles with extractable sections are taken into consideration. The corpus is not complete but provides clean training data for classifiers to assess the most successful strategy to extract history sections. While the list of inclusive strings initially only included the terms ‘science’ and ‘technology’ and was later reduced to just ‘technolog’, the list of exclusive strings (‘**stopcategories**’) was extended over several iterations of manual list expansion and sampling (Table 1). Even though Wikipedia’s categories span a graph, this approach was the most viable as the “category network is noisy and ill-conceived [...] and notoriously incomplete” [16], and “authors often tend to overstrain the features” [17]. While categories are useful for article classification, some are simply administrative in nature, only reference the subject matter, or the article is not an instance of the category [17]. For Iteration 2, 3, 4, and 5, categories were checked for the most frequent tokens in addition to the most frequent categories. In Iteration 4, 96 of the 100 articles sampled already featured science and technology. However, as the sample still contained a large number of articles and categories which proved difficult to assess, the stopcat list was extended one last time. Iteration 5 introduced a second list of stopping strings (‘**stoptitles**’) for which articles are checked and, if matched, excluded.

Table 1: Number of stopcategories (SCs), number of articles, categories, sample size, and number of science and technology articles (S&T) in sample per iteration; iteration 3 and 4 also assessed history sections (cf. Section 3.2); bug fixing in iteration 2.

Iteration	SCs	Articles	Categories	Sample	S&T
1.1	0	104,155	168,187	50	24 (48%)
1.2	2	57,681	98,004	50	20 (40%)
1.3	29	27,819	43,612	50	33 (66%)
2	55	17,085	18,034	–	–
3	56	16,961	17,840	100	88 (88%)
4	73	15,177	14,667	100	96 (96%)
5	79	8,402	8,752	650	621 (96%)

3.2 Finding History Sections

Level. Each iteration recorded the number of articles with a section heading matching the regex ‘history’ (exact match) or ‘*histor[y|i]*’ (partial match). Figure 1 shows that the article on data compression contains two history sections, but both are subsubsections, and neither describes the historical development of data compression but the history of its applications. Table 2 gives an overview of the number of history sections in each iteration. As many partial-match history sections are not history sections, and because lower-level exact-match history (as in the article on data compression) sections occur in less than one percent of all articles, training data is sourced from articles with a designated exact-match history section at top level (‘**designated history section**’).

Heuristic. The baseline approach checks all headings and filters out sections titled ‘History’. Sampling during Iteration 3 (P = 0.98, R = 0.70, F_1 = 0.82) and 4 (P = 0.98, R = 0.72, F_1 = 0.83) had shown that, while most sections labeled ‘History’ do describe the development of the technology featured in the article, a considerable number of articles without a designated history section has a history section (‘**non-designated history section**’). Only articles with 10 or more top-level sections were taken into consideration for

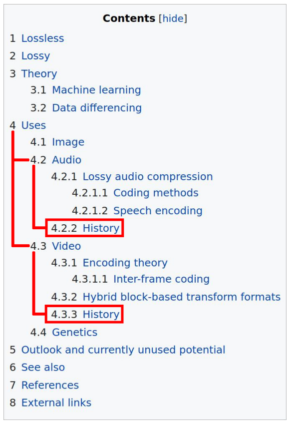

Figure 1: Two subsubsections in one article describe the history of the application of data compression [22].

the articles to be sufficiently long and have enough exploitable structure (cf. [1]). All articles with designated history sections at top level in this corpus happen to have exactly one such section.

Classification. Table 3 details the make-up of the the corpus. In order to find non-designated history sections which cannot be found by the above baseline approach of selecting sections with the heading 'History', various classifiers were trained to find history sections using their section text. Articles with designated history sections are extracted and their sections divided into the classes *HISTORY* and *OTHER* (cf. [1]), which serve as ground truth [7] for training and cross-validation. In order to be sufficiently long and have enough exploitable structure to get examples from both classes, an article now needs to have three or more sections excluding 'See also', 'References', 'Further Reading', 'External links', and 'Notes'.

Sklearn & BERT. In addition to BERT [6, 25], 26 Sklearn [15] classifiers underwent a 5-fold-cross-validation for candidate retrieval. Hyperparameters were set to default for all classifiers, with the exception of four Support Vector Machines, which were set up with

Table 2: Number of history sections in science and technology articles, both as partial (heading contains string 'history' or 'histori') and exact (heading 'history') match. From iteration 3 onwards the level of the section (any or top) was recorded.

		History Sections in S&T Articles			
		'*histor[y|i]*'		'history'	
Iteration	Articles	Any	Top	Any	Top
1.1	104,155	13,965	–	11,145	–
1.2	57,681	10,308	–	8,066	–
1.3	27,819	7,340	–	6,288	–
2	17,085	4,454	–	3,847	–
3	16,961	4,743	4,564	3,953	3,861
4	15,177	4,093	3,933	3,419	3,332
5	8,402	2,363	2,289	2,068	2,021

Table 3: Articles and sections (history and other). Boilerplate sections like 'See also', 'References', 'Further Reading', 'External links', and 'Notes' are excluded.

<table>
<tr><td colspan="3">8,402 articles</td></tr>
<tr><td colspan="3">4,409 articles > 2 sections (excluding boilerplate)</td></tr>
<tr><td>2,825 without designated history</td><td colspan="2">1,584 with designated history</td></tr>
<tr><td rowspan="2">12,520 sections
> 100 characters</td><td colspan="2">8,179 sections > 100 characters</td></tr>
<tr><td>1,574 history</td><td>6,605 other</td></tr>
</table>

two regularization parameters, resulting in 30 classifier setups. Individual term frequency dictionaries are built from both classes of sections, and the feature vector vocabulary is built from the union of the most frequent tokens in both. Oversampling, unifying years, unifying persons, vocabulary size (10, 100, 1,000, and 10,000 most frequent tokens), and term frequency (binary or relative) optimize feature selection, yielding 1,920 classifier–feature setups (Table 4). The top five setups are selected, ignoring setups where the same classifier ranks higher with differing parameters.

4 EVALUATION

Evaluation I. 650 articles from the 4,409 articles with three or more sections were sampled, split into 10 batches (34 without and 31 articles with designated history sections per batch), and distributed among 9 labelers to evaluate whether they contain a history section, with one batch being labeled by 5 evaluators. Table 5 shows the inter-labeler agreement (Cohen's Kappa). The agreement between each pair of labelers is moderate to almost perfect; Fleiss's Kappa is 0.819, indicating almost perfect overall agreement. All articles were labeled with regard to science and technology by the author. More than 95% of the sampled articles cover topics of science and technology (95.29%, or 324 out of 340 articles without a designated history section, 95.81%, or 297 out of 310 articles with a designated history section). While the vast majority of sampled articles with a

Table 4: Sklearn classifier (Random Forests, Extra-Trees, RBF Support Vector, Gradient Boosting, and Multi-Layer perceptron Classifier) performance on training data using 5-fold cross-validation; parameters oversampling (OS), mapping of years (Y), mapping of people (P), vocabulary size (V) and term frequency (T); with precision $P \geq 0.75$ and recall $R \geq 0.45$; sorted precision first, recall second.

Classifier	OS	Y	P	V	T	P	R
RF	0	1	0	1000	relative	0.866	0.481
ET	0	0	0	1000	binary	0.860	0.459
...	...	...	...	...	...	...	...
RBFSV	0	0	1	1000	binary	0.832	0.482
...	...	...	...	...	...	...	...
GB	0	0	1	1000	binary	0.809	0.538
...	...	...	...	...	...	...	...
MLP	0	0	0	10000	binary	0.763	0.613
...	...	...	...	...	...	...	...

Table 5: Inter-labeler agreement for 5 of the 9 labelers.

Labelers	labeler 02	labeler 04	labeler 06	labeler 08	author
labeler 02	-	0.849	0.715	0.908	0.816
labeler 04	0.849	-	0.699	0.939	0.851
labeler 06	0.715	0.699	-	0.753	0.752
labeler 08	0.908	0.939	0.753	-	0.909
author	0.816	0.851	0.752	0.909	-

Table 6: Precision and recall per classifier on section (S) and article (A) level, as well as during cross-validation (C).

	Precision			**Recall**			$F_{0.\bar{3}}$		
	C	**S**	**A**	**C**	**S**	**A**	**C**	**S**	**A**
RF	0.87	0.66	0.79	0.48	0.17	0.17	0.80	0.51	0.58
ET	0.86	0.55	0.70	0.46	0.14	0.13	0.79	0.43	0.49
RBFSV	0.83	0.57	0.60	0.48	0.32	0.53	0.78	0.53	0.60
GB	0.81	0.55	0.60	0.54	0.31	0.50	0.77	0.51	0.59
MLP	0.76	0.45	0.53	0.61	0.45	0.72	0.74	0.45	0.54
BERT	0.81	0.50	0.59	0.37	0.28	0.41	0.72	0.47	0.56

heading 'History' were labeled as having a section that describes the history of the technology featured in the article (99.03%, or 307 out of the 310 articles), many articles without a designated history were also labeled as containing a history section (13.24%, or 45 out of the 340 articles). Given that we can expect the corpus to contain $1584 \cdot 0.9903 = 1569$ articles with a designated history section and $2825 \cdot 0.1324 = 374$ articles with a non-designated history section, we can estimated the overall recall to be $R = \frac{1569}{1569+374} = 0.808$ ($F_1 = 0.890$).

Evaluation II. The five most promising Sklearn classifiers and BERT were trained using the 1,584 articles with designated history sections and applied against the 2,825 articles without designated history sections. The evaluation pool contains 1,013 articles, which were labeled by 8 labelers. Inter-labeler agreement is not available for Evaluation II, but 7 of the 8 labelers participated in Evaluation I. According to the labelers, 615 articles contain a history section. Precision and recall are calculated for all classifiers over the pool of all sections and on article level (Table 6). The latter, more lenient approach considers a classifier's decision correct if it (a) correctly identifies at least one history section, or (b) ignores the article if it does not contain a history section, and considers the classifier to be wrong if (a) it does not find any history sections even though the article contains one, or (b) none of the sections it identifies are history sections. This precision-oriented approach provides researchers with an indication whether an article contains a history section and only requires them to double-check the articles.

5 DISCUSSION

With more than 95% of all articles sampled describing science and technology topics, filtering articles by their assigned categories proves successful. Discarding categories iteratively results in a fine-tuned list of excluding categories. Designated history sections can reliably be identified by their heading 'History'. However, Evaluation I indicates that this heuristic alone is insufficient, as there are a considerable number of articles with a non-designated history section. Evaluation II confirmed this assessment, with the number of articles with non-designated history sections (615) as found by the classifiers and labeled considerably exceeding the estimate based on the results of Evaluation I (374). It is worth noting that labelers viewed articles in the browser, possibly biasing them towards articles with a heading 'History'. Evaluation II did not contain articles with designated history sections, and labelers were asked to name the section(s) which they considered to cover history, which could have made them more attentive to non-designated history sections. Finally, the slightly skewed history section distributions in both Evaluation I and II may have also affected the outcome.

All five Sklearn classifiers and BERT fall behind the expectations based on the cross-validation. The Random Forest classifier scores the best precision, with around a third of all sections identified being history sections, but it only manages to find less than a fifth of all history sections. The Extra-Trees classifier, the second-best model in the cross-validation, suffers the lowest overall recall at a mediocre precision. The RBF Support Vector classifier achieves the highest F-Score but only manages to identify about a third of all history sections at a low precision. Only the Multi-Layer Perceptron classifier manages to find a satisfying number of articles but scores a precision below 50%. BERT manages to find a quarter of all history sections but labels every other section incorrectly.

Given the updated number of 615 articles featuring a non-designated history section, using a section-title-based heuristic only yields a precision of $P_H = 0.990$ and a recall of $R_H = 0.718$. Using the heuristic first and the lenient, recall-focused RBF Support Vector classifier ($P_C = 0.60$, $R_C = 0.53$) as fallback would increase the overall recall to $R_{H+C} = 0.868$ and reduce precision to $P_{H+C} = 0.891$. Using the heuristic first and the lenient, precision-focused Random Forest classifier ($P_C = 0.79$, $R_C = 0.17$) as fallback would increase the overall recall to $R_{H+C} = 0.761$ but only reduce precision to $P_{H+C} = 0.975$.

6 CONCLUSION

We present Webis-WikiSciTech-23, a high-precision corpus of science and technology articles mined using Wikipedia's category system. Webis-WikiSciTech-23 is used as the basis for an in-depth analysis of various classifiers and their capability to identify non-designated history sections of Wikipedia articles on science and technology. We demonstrate that using these classifiers as a fallback option can increase recall while maintaining high precision when compared to the baseline approach of using section headings to identify text segments that cover the historical development of science and technology featured on Wikipedia. Together with insights gleaned from our evaluations, Webis-WikiSciTech-23 can help science studies researchers unlock Wikipedia's unique position as a community-driven, up-to-date, and traceable account of science priority disputes.

REFERENCES

[1] Alessio Palmero Aprosio and Sara Tonelli. 2015. Recognizing Biographical Sections in Wikipedia. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (Eds.). The Association for Computational Linguistics, 811–816. https://doi.org/10.18653/v1/d15-1095

[2] Omer Benjakob, Olha Guley, Jean-Marc Sevin, Leo Blondel, Ariane Augustoni, Matthieu Collet, Louise Jouveshomme, Roy Amit, Ariel Linder, and Rona Aviram. 2022. Wikipedia as a Tool for Contemporary History of Science: A Case Study on CRISPR. *bioRxiv* (2022). https://doi.org/10.1101/2022.11.25.517950v1

[3] Erik Borra, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, and Tommaso Venturini. 2015. Societal Controversies in Wikipedia Articles. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, Bo Begole, Jinwoo Kim, Kori Inkpen, and Woontack Woo (Eds.). ACM, 193–196. https://doi.org/10.1145/2702123.2702436

[4] Mingda Chen, Sam Wiseman, and Kevin Gimpel. 2021. WikiTableT: A Large-Scale Data-to-Text Dataset for Generating Wikipedia Article Sections. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021 (Findings of ACL, Vol. ACL/IJCNLP 2021)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 193–209. https://doi.org/10.18653/v1/2021.findings-acl.17

[5] Jon Cohen. 2020. The Latest Round in the CRISPR Patent Battle Has an Apparent Victor, but the Fight Continues. https://doi.org/10.1126/science.abe7573 Online; accessed 6 October 2022.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[7] Anjalie Field, Sascha Rothe, Simon Baumgartner, Cong Yu, and Abe Ittycheriah. 2020. A Generative Approach to Titling and Clustering Wikipedia Sections. *CoRR* abs/2005.11216 (2020). arXiv:2005.11216 https://arxiv.org/abs/2005.11216

[8] Dan Flickinger, Stephan Oepen, and Gisle Ytrestøl. 2010. WikiWoods: Syntacto-Semantic Annotation for English Wikipedia. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*, Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias (Eds.). European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2010/summaries/432.html

[9] Heidi Ledford. 2016. Titanic Clash over CRISPR Patents Turns Ugly. , 460–461 pages. https://doi.org/10.1038/537460a Online; accessed 6 October 2022.

[10] Yilun Lin, Bowen Yu, Andrew Hall, and Brent J. Hecht. 2017. Problematizing and Addressing the Article-as-Concept Assumption in Wikipedia. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017, Portland, OR, USA, February 25 - March 1, 2017*, Charlotte P. Lee, Steven E. Poltrock, Louise Barkhuus, Marcos Borges, and Wendy A. Kellogg (Eds.). ACM, 2052–2067. https://doi.org/10.1145/2998181.2998274

[11] Shan Liu and Mizuho Iwaihara. 2016. Extracting Representative Phrases from Wikipedia Article Sections. In *15th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2016, Okayama, Japan, June 26-29, 2016*. IEEE Computer Society, 1–6. https://doi.org/10.1109/ICIS.2016.7550850

[12] Elizabeth M. Nix. 2010. Wikipedia: How It Works and How It Can Work for You. *The History Teacher* 43, 2 (2010), 259–264. http://www.jstor.org/stable/40543291

[13] NobelPrize.org. 2022. The Nobel Prize in Chemistry 2020. https://www.nobelprize.org/prizes/chemistry/2020/summary Online; accessed 28 October 2022.

[14] Natalia Ostapuk, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2020. SectionLinks: Mapping Orphan Wikidata Entities onto Wikipedia Sections. In *Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference(OPub 2020), Virtual Conference, November 2-6, 2020 (CEUR Workshop Proceedings, Vol. 2773)*, Lucie-Aimée Kaffee, Oana Tifrea-Marciuska, Elena Simperl, and Denny Vrandecic (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-2773/paper-14.pdf

[15] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[16] Tiziano Piccardi, Michele Catasta, Leila Zia, and Robert West. 2018. Structuring Wikipedia Articles with Section Recommendations. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.). ACM, 665–674. https://doi.org/10.1145/3209978.3209984

[17] Ralf Schenkel, Fabian M. Suchanek, and Gjergji Kasneci. 2007. YAWN: A Semantically Annotated Wikipedia XML Corpus. In *Datenbanksysteme in Business, Technologie und Web (BTW 2007), 12. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), Proceedings, 7.-9. März 2007, Aachen, Germany (LNI, Vol. P-103)*, Alfons Kemper, Harald Schöning, Thomas Rose, Matthias Jarke, Thomas Seidl, Christoph Quix, and Christoph Brochhaus (Eds.). GI, 277–291. https://dl.gi.de/20.500.12116/31804

[18] Marion Schmidt, Wolfgang Kircheis, Arno Simons, Martin Potthast, and Benno Stein. 2021. Does Wikipedia Cover the Relevant Literature on Major Innovations Timely? An Exploratory Case Study of CRISPR/Cas9.. In *18th International Conference on Scientometrics & Informetrics*, Wolfgang Glänzel, Sarah Heeffer, Pei-Shan Chi, and Ronald Rousseau (Eds.). AAAI Press, 2021–2026.

[19] Marion Schmidt, Wolfgang Kircheis, Arno Simons, Martin Potthast, and Benno Stein. 2023. A Diachronic Perspective on Citation Latency in Wikipedia Articles on CRISPR/Cas-9: An Exploratory Case Study. *Scientometrics* (2023). https://doi.org/10.1007/s11192-023-04703-81

[20] Arno Simons, Wolfgang Kircheis, Marion Schmidt, Martin Potthast, and Benno Stein. 2023. Who are the "Heroes of CRISPR"? Tracing Negotiations of Academic Micro-Notability in Wikipedia. *(in submission)* (2023).

[21] Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. 2006. Semantic Wikipedia. In *Proceedings of the 15th International Conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23-26, 2006*, Les Carr, David De Roure, Arun Iyengar, Carole A. Goble, and Michael Dahlin (Eds.). ACM, 585–594. https://doi.org/10.1145/1135777.1135863

[22] Wikipedia contributors. 2022. Data compression — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Data_compression&oldid=1112925399 Online; accessed 6 October 2022.

[23] Wikipedia contributors. 2022. Wikipedia:Namespace — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Namespace&oldid=1112342017 Online; accessed 17 November 2022.

[24] Wikipedia contributors. 2022. Wikipedia:Reliable sources — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Reliable_sources&oldid=1114796177#Primary,_secondary,_and_tertiary_sources Online; accessed 15 November 2022.

[25] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[26] Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2008/summaries/420.html

[27] Michel Zitt, Alain Lelu, Martine Cadot, and Guillaume Cabanac. 2019. Bibliometric Delineation of Scientific Fields. In *Springer Handbook of Science and Technology Indicators*, Wolfgang Glänzel, Henk F. Moed, Ulrich Schmoch, and Mike Thelwall (Eds.). Springer, 25–68. https://doi.org/10.1007/978-3-030-02511-3_2