Towards Knowledge Maintenance in Scientific Digital Libraries with the Keystone Framework

Yuanxi Fu
School of Information Sciences
University of Illinois at Urbana-Champaign
Champaign, IL USA
fu5@illinois.edu

Jodi Schneider
School of Information Sciences
University of Illinois at Urbana-Champaign
Champaign, IL USA
jodi@illinois.com

ABSTRACT

Scientific digital libraries speed dissemination of scientific publications, but also the propagation of invalid or unreliable knowledge. Although many papers with known validity problems are highly cited, no auditing process is currently available to determine whether a citing paper's findings fundamentally depend on invalid or unreliable knowledge. To address this, we introduce a new framework, the keystone framework, designed to identify when and how citing unreliable findings impacts a paper, using argumentation theory and citation context analysis. Through two pilot case studies, we demonstrate how the keystone framework can be applied to knowledge maintenance tasks for digital libraries, including addressing citations of a non-reproducible paper and identifying statements most needing validation in a high-impact paper. We identify roles for librarians, database maintainers, knowledgebase curators, and research software engineers in applying the framework to scientific digital libraries.

CCS CONCEPTS

• **Information systems~Digital libraries and archives** • **Computing methodologies~Discourse, dialogue and pragmatics** •Applied computing~Document management and text process~Document metadata

KEYWORDS

Knowledge maintenance; argumentation theory; citation contexts; citation; knowledge claims; retraction of research; citation of retracted papers; argument retrieval; scientific literature

ACM Reference format:

Yuanxi Fu and Jodi Schneider. 2020. Towards knowledge maintenance in scientific digital libraries with the keystone framework. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (*JCDL '20). ACM, New York, NY, USA, 10 pages.* DOI: 10.1145/3383583.3398514

JCDL '20, August 1–5, 2020, Virtual Event, China

ACM ISBN 978-1-4503-7585-6/20/08.
https://doi.org/10.1145/3383583.3398514

1 INTRODUCTION

Scientific digital libraries make the dissemination of scientific publications easier and faster. Yet this also facilitates the propagation of invalid or unreliable knowledge. Many papers with known validity problems are highly cited [3]. Citations to invalidated papers pose a threat to scientific knowledge maintenance. However, currently, these citations are not flagged for review, and no auditing process is available to determine whether a new paper's findings fundamentally depend on invalid or unreliable knowledge.

Our goal in this paper is to set an agenda for knowledge maintenance in scientific digital libraries. This work is motivated by the questions: Does it matter when citing authors make use of a paper whose findings are no longer considered valid? Are papers citing it necessarily wrong? Our work introduces a framework for addressing these questions by combining argumentation theory and citation context analysis, pilot tests our new framework in two case studies, and suggests future directions for applying the framework.

1.1 Scope and Importance of the Problem

We estimate that over 800,000 articles directly cite a retracted paper. The Retraction Watch Database[1] lists over 23,000 retracted publications as of May 2020. In biomedicine 94% of retracted papers have received at least one citation, with an average citation count of 35 [15]. Nor are all citations to these articles negative; even Wakefield's fraudulent paper linking the MMR vaccine to autism received 94 positive citations [42].

Retraction "is a mechanism for correcting the literature and alerting readers to articles that contain such seriously flawed or erroneous content or data that their findings and conclusions cannot be relied upon" [46]. Consequently, articles that substantively use the findings of retracted papers need reexamination. Fundamental errors can result from certain uses of retracted papers, including for synthesizing medical evidence [2, 22]. As of 2019, the industry Committee on Publication Ethics (COPE) warns: "Articles that relied on subsequently retracted articles in reaching their own conclusions, such as systematic

[1] http://retractiondatabase.org/

reviews or meta-analyses, may themselves need to be corrected or retracted." [46]. Such errors have made their way into heavily used documents: Avenell et al. show how citation to 12 retracted clinical trials has impacted clinical literature reviews and guidelines from the American Heart Association, the American College of Physicians, and the U.S. Agency for Healthcare Research and Quality [2].

Of course, merely citing a retracted paper does not necessarily invalidate the citing paper, because not all citations are used in the logical argument. For instance, it is customary to cite a paper when critiquing it [42], and sometimes a foundational paper is cited as an 'homage to pioneers' [20] without the intention that it support the logical argument.

While retracted papers are the easiest to enumerate (and explicitly marked), their citation underestimates the scope of the problem since retraction is a recent and field-specific practice. Another common practice is to silently *abandon* works; such abandoned papers "contain conclusions that are refuted by later studies while still remaining in the record of scientific publications" [9].

The structure of the remainder of the paper is as follows: Section 2 summarizes related work. Section 3 introduces the keystone framework. Sections 4 presents our first case study, a keystone analysis of citations to a single unreliable paper. Section 5 presents our second case study, a keystone analysis of a single mouse model paper from Alzheimer's disease research, a field which has faced challenges in translating results of animal model research into insights for human treatment. Section 6 presents our research agenda for knowledge maintenance in digital libraries based on the framework. The paper concludes in Section 7.

2 RELATED WORK

2.1 Argumentation-based Curation

Argumentation theory is an interdisciplinary field with multiple branches studying persuasion, rhetoric, dialectic, defeasible reasoning, and related topics [16]. Argument schemes describe stereotypical reasoning patterns along with critical questions used in validating the reasoning [47]. Argument maps can be used to diagram relationships between evidence and the statements they support or challenge [41]. Of particular interest is whether support comes from a single source, multiple independent sources, or multiple linked sources that must be combined [19].

Digital library applications of argumentation theory include argument-based retrieval [21, 30] and curation of data and other resources associated with a paper [12]. Three main approaches to argument-based curation are rhetoric-based, provenance-based, and argument scheme-based curation.

Rhetoric-based approaches seek to extract information based on rhetorical features. The use of general linguistic features limits the need for domain knowledge. The best known rhetoric-based approach, Argumentative Zoning [43], has been applied to digital libraries at scale using computational linguistics for about a decade [e.g., 44]. Similar approaches have also been used more recently for citation recommendation [10].

Provenance-based approaches seek to interlink scientific documents and their supporting data. They model documents based on the work process through which scientific discoveries are generated. One well-known example is the micropublication model [13], which indicates the support relationships between a paper's claims, methods, materials and data. Modeling research articles into the micropublication model is currently done by humans. For the Ph.D. level curators who maintain knowledge bases and databases, this takes 10-20 minutes per article [12].

Argument scheme-based approaches rely on deep domain analysis to generalize argument structures that capture the expert tradecraft used in a field. For example, in genetics, several argument schemes can be used to determine whether a gene variant affects human health [23]. Such schemes document the underlying logic of how a discipline justifies its findings, which can make the reasoning more explicit and more visible to non-experts. Once a document's argument has been curated into schemes, it also becomes possible to identify potential weakness through critical questions [47] which could be useful for knowledge maintenance.

2.2 Citation Context Analysis

A citation context consists of some text (generally a sentence but also potentially a clause or multiple contiguous sentences) along with a reference to one or more cited items (i.e., the support for the text). Citation context analysis [39] has been used for a variety of purposes, including to study citation motivation [18] or to classify citation function [7]. The best-known citation functions include substantiating claims, paying homage to pioneers, criticizing previous work, and providing leads to hard-to-find works [20]. Moravcsik and Murugesan influentially distinguished 4 facets: conceptual or operational, organic or perfunctory, evolutionary or juxtapositional, and conformative or negational [31]. More recent work has defined and automated 'meaningful citations' [45] and 'influential citations' [52].

3 THE KEYSTONE FRAMEWORK

We now introduce the keystone framework, for tracing the impact of citing a paper whose findings are invalidated. The term 'keystone' is inspired by masonry, where damage to the keystone can threaten the arch it supports. The keystone framework combines argumentation theory and citation context analysis. Argumentation theory forms the basis of understanding how retracted articles impact the argument of a citing paper while citation context analysis helps us distinguish citations that support the argument from those not used in the argument. We first show how argumentation theory explains an existing retraction guideline before detailing our framework.

3.1 Argumentation Theory Explains COPE Retraction Guideline

Argumentation theory can be used to explain why COPE suggests that systematic reviews be corrected or retracted when a paper they relied on is retracted [46]. The logic of a systematic review is to synthesize a number of studies [2], forming a linked argument [19], as shown in Figure 1. In a linked argument, all evidence must be combined together to support a conclusion. Thus, if any one of the synthesized studies is retracted, the overall conclusions need to be reexamined. Avenell et al. describe, for instance, how retraction of 3 of 7 studies included in a review on Vitamin K for the prevention of fractures led to a correction of the review [2].

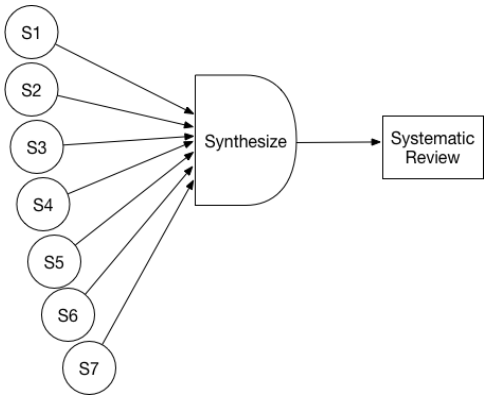

Figure 1: An argument diagram for a systematic review.

Systematic reviews make direct use of studies they synthesize, following a linked argument structure. However, other types of publications (e.g., empirical studies) often employ more intricate argument structures [23], and the impact of a single cited paper on the argument may not be immediately apparent. Furthermore, a systematic review may also cite but not synthesize a paper, whose retraction does not threaten the synthesized conclusion. To make the distinction between citations that support the argument and those not used in the argument, we must resort to the citation context. To summarize, this COPE guideline is an intuitive response to the knowledge maintenance challenge posed by retractions. A fully defined framework is needed to make further progress in knowledge maintenance in scientific digital libraries.

3.2 Defining the Keystone Framework

Under our framework, a scientific research paper puts forward at least one main finding, along with a logical argument, giving reasons and evidence to support the main finding. The main finding is accepted (or not) on the basis of the logical argument. Evidence from earlier literature may be incorporated into the argument by citing a paper and presenting it as support, using a citation context.

[2] Each "study" comprises a group of related articles; such grouping is important in clinical systematic reviews to avoid double-counting of the same patients.

We define a **keystone statement** as any statement whose unreliability threatens the argument for a main finding of a paper. We focus on **keystone citation contexts**, which we define as citation contexts supporting keystone statements. Our framework makes two further distinctions, to better understand the risks posed by citing a retracted, abandoned, or non-reproducible paper.

First, we distinguish *how many items are cited*, since removal of a unique item can be considered more risky to the argument than removal of one of several redundant items:

- A **singleton citation context** cites one item, e.g. '[2]'
- A **cluster citation context** cites multiple items, e.g., '[2, 16]' or '(DeKosky and Scheff, 1990; Scheff and Price, 2006; Terry et al., 1991)'.

Second, we delineate *whether the cited item's main findings support the citation context.* This gives an indication of the strength of support, and indicates the risk posed if the item's main findings are overturned.

- **Main-findings support**, if the citation context closely relates to a main finding of the cited item.
- **Pass-through support**, if support can be found within the cited item but only in an unsupported statement or a statement referencing one or more other work(s).
- **No clear support**, if the citation context does not clearly relate to the cited item, either its main findings, or other statements it makes.

Table 1 summarizes the three types of keystone citation contexts observed in our case studies.

Table 1: The three types of keystone citation contexts we observed

Properties of the keystone citation context	**Removing the citation context would weaken the argument supporting a main finding**	**Only one paper is cited.**	**The main findings of the cited paper(s) provide evidence to support the argument.**	**Corresponding cited article**
Singleton, main-findings support	+	+	+	Main-finding keystone citation
Cluster, main-findings support	+	-	+	Main-finding keystone citation cluster
Singleton, pass-through support	+	+	-	Pass-through keystone citation

3.3 Comparison to other Approaches to Citation

Previous approaches to analyzing papers do not consider the argument structure along with citations or citation contexts. Keystone citations are only found in the logical argument supporting the main findings, which does not include the rationale for conducting a study. In contrast, citations that provided inspiration for the study may be considered 'influential citations' [52] (e.g., for posing new ideas or research problems) and 'Argumentation-Active Support' [18] (for "calling for further research"). Keystone citations must also be distinguished from 'meaningful citations' [45]; these concepts overlap when *using* a paper in the main argument; however, *extensions* of a cited work are 'meaningful' but (unless there is a logical dependence) not keystone. Keystone citations do fit in broader categories of some citation function schemes, such as Garfield's 'substantiating claims' [20], Moravcsik and Murugesan's organic citation—"truly needed for the understanding of the referring paper" [31]. None correspond precisely with keystone citations.

3.4 Overview of the Case Studies

We next demonstrate keystone analysis through two case studies, to show how the keystone framework can help support knowledge maintenance. Our first case study investigates whether the non-reproducibility of a computational chemistry protocol [28] affects 10 recent citing papers called out by a paper pointing out the programming error [5], and demonstrates that by using the keystone framework, we can narrow down to a fraction of citing papers for follow-up actions such as alerts for validation and verification. Our second case study analyzes a high-impact paper in Alzheimer's disease research [14] in-depth and demonstrates the paper's findings depend on the validity of the findings from its keystone citations. Both case studies rely on the domain expertise of the first author, a PhD-level chemist.

4 CASE STUDY 1: CITING NON-REPRODUCIBLE CODE

4.1 Overview

In late 2019, a paper published in the journal *Organic Letters* [5] reported a glitch in a piece of widely used computational chemistry protocol for calculating Nuclear Magnetic Resonance (NMR) chemical shifts. The study found that the output of the Python scripts depends on the platform (Windows/Mac/Linux). Further investigation revealed that the glitch originated from differences in the platforms' default file sorting mechanisms; a flowchart showing the problem is given in Figure 2. The discovery also caught the mass media's attention, prompting news articles with titles such as "A Code Glitch May Have Caused Errors in More Than 100 Published Studies" [4].

To assess the real impact of the code glitch, we applied the keystone framework to analyze the ten 'affected articles' mentioned in the paper identifying the glitch [5]. We focused the origin of non-reproducibility, which is script D of the protocol, and assessed whether the non-reproducibility could impact the arguments for the main findings of the citing papers.

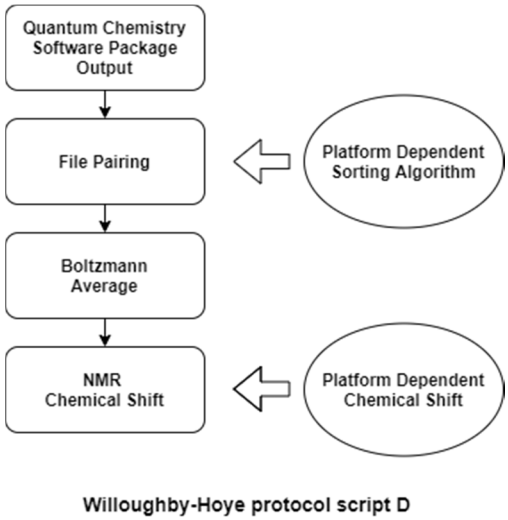

Figure 2: The NMR code glitch specifically relates to Willoughby-Hoye protocol script D, nmr-data_compilation.py.

To do this, first we developed an argument graph to illustrate the role that the Willoughby-Hoye protocol (WH protocol) could play in formulating an argument within a research article, shown in Figure 3. The computational protocol first generates a few sets of NMR chemical shifts, often called "theoretical NMR chemical shifts," based on speculated structures of a compound. Then those theoretical NMR chemical shifts are compared with experimental NMR chemical shifts to find the best match. The molecular structure of the best match is identified as the structure of the compound. Mistakes in the theoretical NMR would result in wrong identifications and thus threaten the validity of the conclusions drawn.

We then devised a list of questions to help us determine the impact of the code glitch on the citing papers. The first author manually analyzed the ten articles and recorded the answer to each of the questions shown in Table 2.

Table 2. Questions used to analyze the impact of the code glitch on the citing paper in Case Study 1

No.	Question
1	Does the citing paper use the WH protocol to calculate NMR chemical shifts?
2	What NMR chemical shifts were calculated using the protocol?
3	What does the theoretical NMR chemical shifts support?
4	Will a main finding be threatened if the numerical values of the chemical shifts are unreliable? Why?

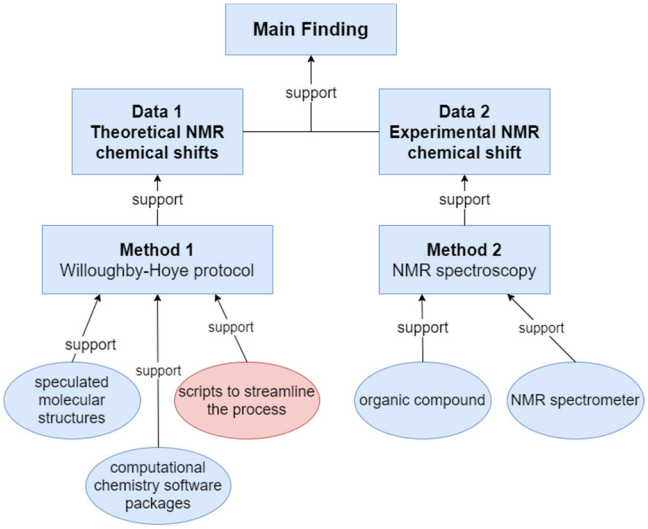

Figure 3: An argument diagram illustrating the role that the Willoughby-Hoye protocol could play in formulating an argument within a research article.

4.2 Results

We conclude that, contrary to the speculation in Organic Letters [5], six of the ten articles are not be affected by the code glitch. The reasons, shown with examples in Table 3, are as follows.

Table 3. Results of keystone analysis for Case Study 1

Group	Function	Sample citation context
Unaffected by the code glitch	Support linear regression [1, 50, 51]	*"Scaling factors (slope = -1.0522, intercept ¼ 181.2412) are applied to the 13C NMR shielding tensors (B3LYP/6-311 þ G (2d,p)//M06-2X/6-31 þ G (d,p) to calculate the 13C NMR chemical shifts [26]."*
	Support conformational analysis [9, 22]	*"Conformational analysis of 4 was performed with Schrödinger MacroModel 2016 by following the method of Willoughby et al [35]."*
	Successful example [53]	*"The successful characterization of karlotoxin 2 (KmTx2) followed by KmTx8 and KmTx9 was supported by NMR chemical shift calculation tools including gauge-including atomic orbitals (GIAO) and DP4+ probability studies in conjunction with heteronuclear single quantum coherence (HSQC) spectroscopy studies [37–40]."*
Potentially affected by the code glitch	Support NMR chemical shift calculations [17, 28, 33, 40]	*"Therefore, we turned to a protocol that relies on density functional theory-based computations of 1H and 13C NMR chemical shifts and the use of statistical tools to assign the experimental data to the correct isomer of a compound [28]."*

Three of the articles [1, 50, 51] cited the WH protocol in order to support a linear regression fitting between the theoretical and experimental NMR shifts. They are pass-through keystone citations, because the support is provided not directly by the WH protocol but rather by its bibliography (WH protocol's Ref 19). Note that this reference did not investigate different curve fitting approaches (which would provide the ideal evidence), but just showed that that for a known molecule, linear regression fitting was sufficient.

Two articles [8, 25] cited the WH protocol for adopting conformation analysis, which is up-stream in the protocol from the problematic Python script D, thus not impacted by the code glitch. One [53] cited the WH protocol in the introduction as a previous successful example of characterizing molecules by using NMR chemical shift calculations.

The remaining four papers [17, 28, 33, 40] stated that they used the WH protocol to calculate the theoretical NMR chemical shift and thus could be significantly impacted, because the theoretical NMR chemical shifts supported findings that went into abstracts or the conclusion section. We say "could be" because the non-reproducibility comes from the protocol's implementation rather than the protocol itself. More precisely, only if an article used the incorrect implementation (i.e., used the supplied Python scripts D on Linux based operating systems), were its findings no longer valid. Yet since such implementation details were not provided in the articles, clarification is needed from authors. Specifically, authors of these four papers should double-check their results and either amend their findings or document how the findings are sustained despite the code glitch.

4.3 Lessons for Digital Libraries

Digital libraries should mark the papers whose conclusions rely on unreliable findings and should alert authors to check their conclusions. Subsequently, reliability flags should remain on the papers until their conclusions are checked and either verification or correction statements can be published. A sophisticated digital library could mark only keystone citations to unreliable findings. In this case, 4/10 of the papers examined were affected, i.e., keystone citations. This is a shockingly large proportion. Yet it shows the utility of keystone analysis: identifying that the problem from the cited paper did not transfer to the other 6/10 of the papers examined helps both authors and readers.

Digital libraries should provide authors and peer reviewers with support for identifying and evaluating statements. Pass-through support citations disadvantage authors whose works are cited indirectly and not 'credited'; they make it difficult for reviewers to assess relevance of the evidence; and from the perspective of knowledge maintenance, they demonstrate that when validity problems are found, a second or further generation of citations may need assessment.

5 CASE STUDY 2: CITATIONS SUPPORTING ONE PAPER'S ARGUMENT

The second case study seeks to determine which external knowledge was most essential for supporting the argument of a single influential paper.

5.1 Overview

In this case study, we analyzed one neuroscience paper [14] in-depth using the keystone framework. Our analysis was guided by argument diagrams we previously developed [34, 35], which delineated the reasoning of the four major findings made by the article. One author manually screened each citation in the body of the text (omitting "Experimental Procedures" which described methods in a smaller font at the end of the paper) and recorded the answers to each of the questions shown in Table 4.

Table 4. Questions used for in-depth analysis in Case Study 2

No.	Question
1	What does the citation context support?
2	Does the citation context support one or more components of the argument diagrams?
3	What is the function of the citation context?
4	What type of citation context is it (i.e., keystone or non-keystone, if keystone, which type)? Why?

Ultimately, we discovered four main-finding keystone citations, one main-finding keystone citation cluster, and one pass-through keystone citation. All of them came from the results section (in this case the methods were given in an appendix) and are shown in Figures 4, 5, and 6.

5.2 Keystone Citations Support Experimental Materials

Three main-finding keystone citations support the mouse model itself, while a fourth supports a biomarker.

The authors' goal is to simulate the progression of Alzheimer's disease in a mouse. To obtain a mouse that has Alzheimer's-like traits, which is labeled "Children" in Figure 4, the authors bred two different kinds of mice that each have one Alzheimer's-related trait. One parent mouse can make human tau protein (Main-finding Keystone Citation 2/Parent 2), and the other is capable of producing proteins in a particular part of the brain (the EC region, where Alzheimer's starts from in human brains) (Main-finding Keystone Citation 1/Parent 1). These two keystone citations support the experimental materials, i.e., the mouse the authors bred and used for their experiments. Additionally, the authors performed experimental verification, using an antibody, 5A6. Main-finding Keystone Citation 3 supports the fact that the antibody 5A6 can recognize human tau protein and thus can verify that the human tau protein is only expressed in the EC region of the mouse model. All three of these keystone citations support the main finding that the authors have created a mouse model that can express human tau protein only in a particular part of the brain.

Pass-through Keystone Citation 1, shown in Figure 5, justifies a statement that supports another choice of experimental material, "PSD-95 is a good synaptic biomarker." This is a pass-through keystone citation to Zhao et al. 2006, whose main findings do not directly support this statement, because Zhao's experiment did not involve PSD-95 at all.

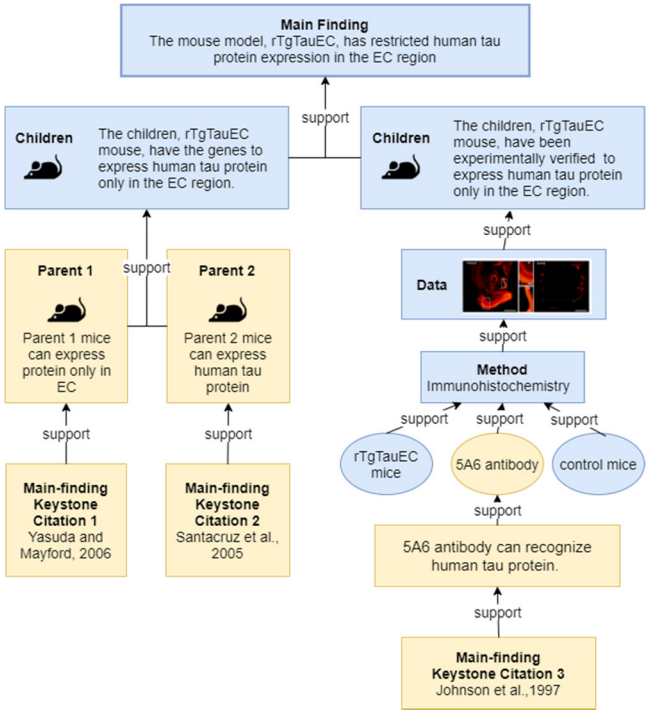

Figure 4: Three main-finding keystone citations that support experimental materials in [14].

Only its introduction summarizes a number of papers (Zhao et al. 2006's Refs 15-19) that had reported the reduction of PSD-95 early in neurodegeneration (including in Alzheimer's disease). However, unlike the pass-through keystone citations seen in the first case study, which are tangentially supported by main findings (though not of the cited paper itself), in this case, this support comes from an informal literature review, Zhao's introduction, which is cited for the statement "PSD-95 decreases early in neurodegeneration." The authors took a risk by using this ad-hoc literature review to support a key experimental decision. Citing an ad-hoc literature review is a bad practice, because it takes the risk that knowledge may have slipped into the literature without rigorous review or explicit validation.

5.3 Keystone Citations Support Experimental Methods

A main-finding keystone citation cluster of three references, reading "(DeKosky and Scheff, 1990; Scheff and Price, 2006; Terry et al., 1991)" supports a keystone statement for the choice of experimental method (use of synaptic markers to measure the degree of neurodegeneration). Two citations in the cluster provide experimental evidence from independent labs: DeKosky and Scheff, 1990 and Terry et al., 1991, are experimental demonstrations of correlation between synapse density and neurodegeneration. The third citation is a review article: Sheff and Price, 2006 reviews the Alzheimer's disease-related alterations in synaptic density.

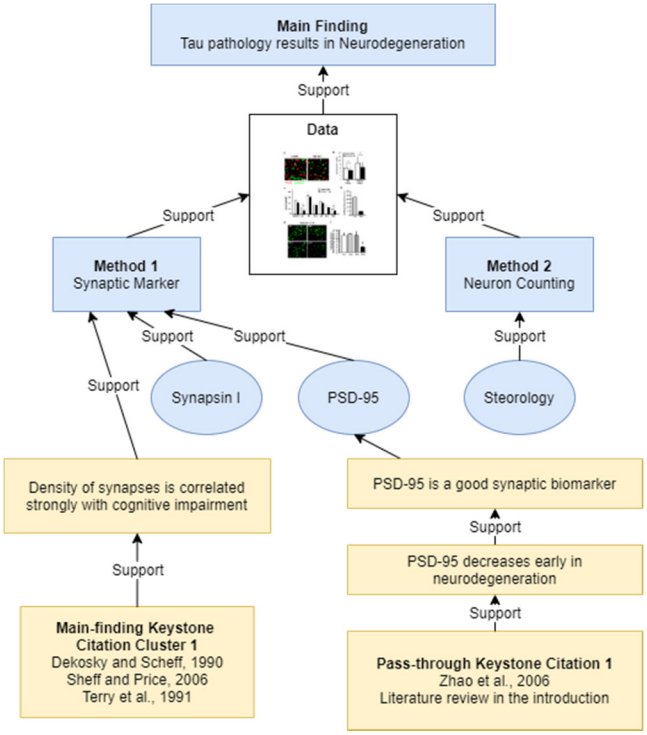

Figure 5: An ad-hoc literature review from Pass-through Keystone Citation 1 supports the choice of PSD-95 as a synaptic marker in [14].

5.4 Keystone Citations Support Interpretations

One main-finding keystone citation helps in the interpretation of data. In this case, Data 1 shows that tau protein aggregates were found in a couple of brain regions, and Data 2 rules out the possibility that the tau protein aggregates were caused by locally produced proteins. The main-finding keystone citation, Witter et al., 1988, supplies crucial information about the brain anatomy, that those regions received direct neuron input from the EC region (i.e., were connected via synapses). The resulting interpretation, Interpretation 1, was then synthesized with Interpretation 2 (the interpretation of Data 3) in order to reach the main finding, "Human tau protein propagates synaptically in mice," as shown in Figure 6. Noteworthy is that the second interpretation of data could also involve Witter et al., 1988. However, the citation appeared immediately after Interpretation 1 was introduced in the article, and therefore, in Figure 6 we add a question mark to indicate our uncertainty about whether the author intended Witter et al., 1988 to also help support Interpretation 2.

5.5 Lessons for Digital Libraries

Previous problems in translating Alzheimer's research into viable treatments make application of the keystone framework valuable. Digital libraries can take a panoptic view of the literature and identify the keystone statements over an entire body of literature. Repeated keystone statements have fundamental importance. Significant benefit can result from prioritizing the validation of keystone statements underlying important work, particularly at key phases in research translation (such as when preclinical studies move into human trials). With this approach, digital libraries can support funding agencies in identifying research priorities, and researchers in finding impactful problems.

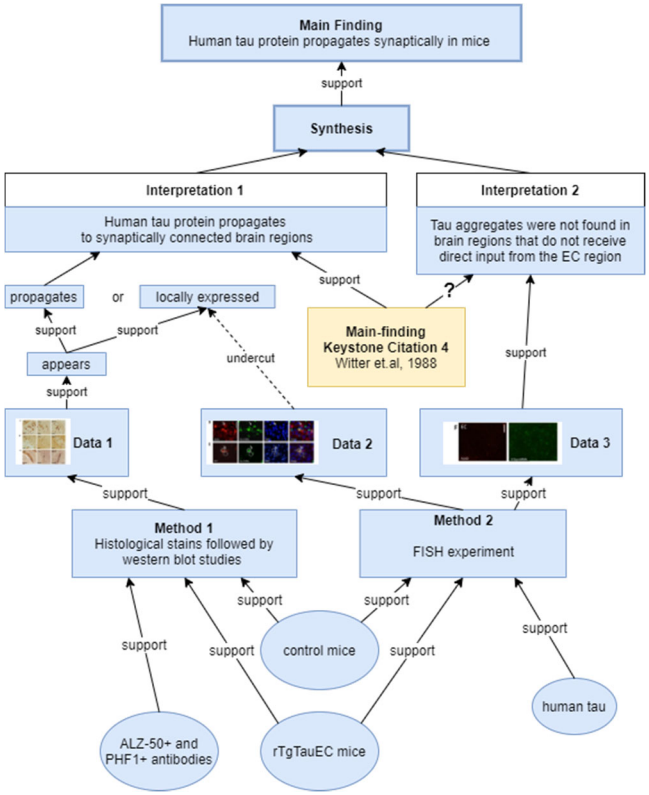

Figure 6: Main-finding Keystone Citation 4 in [14] supports the interpretation of experimental data.

6 DISCUSSION

Our case studies demonstrate how identifying keystone citations can support knowledge maintenance, in addressing citations of non-reproducible papers and identifying dependencies of a high-impact paper. Now we consider more generally the research agenda for knowledge maintenance in scientific digital libraries.

6.1 Need for Knowledge Maintenance in Scientific Digital Libraries

Scientific digital libraries need to take a knowledge maintenance perspective and create a sustainable knowledge infrastructure that will nourish the research community in the long run, because "Shared, reliable knowledge is among the human society's most precious resources" [6]. In digital libraries, the current state of knowledge maintenance is, at best [36], to mark retracted papers. However, when a part or all of the findings of a paper are invalidated, updating the state of knowledge, including the citing papers, is crucial. A top journal asked "what do we still know" after twenty-one articles by the pain researcher Scott Reuben were retracted for data fabrication in 2009 [49]. Current scientific digital libraries mainly focus on searching and sharing papers. The most advanced ones also support finding related

work, navigating citation networks, comparing the number of positive and negative citations[3] or identifying 'meaningful citations'[4] [45]. However, scientific digital libraries so far still lack a framework to guide them in maintaining a collection of reliable literature for scientists to use, and this shortcoming has motivated us to develop the keystone framework.

Librarians, database maintainers, knowledgebase curators, and research software engineers, can all play a role in applying the keystone framework to scientific digital libraries. Librarians and database maintainers should develop and incorporate metadata and indexes, to support knowledge maintenance tasks, for instance to present a panoptic view of keystone statements over an entire body of literature, or to interlink papers via keystone citations. Knowledgebase curators in fields with a tradition of focused curation [29] are well-placed to identify keystone statements. Research software engineers could develop text mining pipelines to reduce curators' workload in identifying keystone statements.

6.2 Patching Papers using Keystone Citation Alerts

The keystone framework could support selective alerting of authors whose publications are most likely to be impacted by errors in a paper they cited. When significant errors or reliability issues are identified in a paper, its citing papers should be assessed. For instance, the infamous retracted paper linking MMR vaccine to autism received 94 positive citations [42]; those papers should be assessed for reliability, to understand whether their argument depends in any way on the fraudulent paper. A greater sense of urgency in updating papers might result from this kind of selective alerting, because in general, only a fraction of citing publications are expected to be keystone. In our first case study, we narrowed down a list of 10 recent citations that experts posited [5] might be affected by a programming error. Six of the ten papers we examined were not impacted by the bug to this part of the script. This means that a smaller number of authors would need to be contacted, enabling authors and editors to follow up more aggressively on the most significant problems.

In the wake of a retraction or an erratum, hundreds or thousands of citing papers may need to be screened. We envision this task being performed by a team comprising of one domain expert and several non-experts. The domain expert develops a generalized argument model showing the roles that a particular piece of unreliable knowledge can play (e.g., Figure 3). A checklist of screening questions then can be developed based on the model (e.g., Table 2). With such a list, non-experts can quickly and consistently identify which articles are most likely to be impacted and to set up alerts for the authors.

[3] https://scite.ai
[4] https://www.semanticscholar.org

6.3 Limitations of the Keystone Framework

Our current framework has several limitations. First, it depends on explicit citation, and does not scrutinize statements that slip into literature without citation. Uncited statements are common and can be problematic: for instance, 7% (8/115) of review papers cited no supporting literature for the questionable statement that iron level increased in Alzheimer's Disease [37]. Second, the keystone framework does not scrutinize the validity of domain-specific argumentation patterns themselves. Third, it depends on explicit labeling of invalid literature, which is only standard practice for retracted papers. Fourth, application of the keystone framework beyond retracted papers needs to proceed cautiously to avoid stifling innovation, since 'abandoned' or 'non-reproducible' findings may turn out be correct. For instance, early evidence on adult neurogenesis was dismissed for nearly 30 years [24], and lack of reproducibility can sometimes be attributed to hidden variables, the study of which may lead to new insights [48].

6.4 Research Agenda

Our two case studies used different procedures in applying the keystone framework, depending on the task at hand. The first case study started from the source of the unreliability, which is most appropriate for addressing citation of papers with validity issues. The second case study started from the argument structure of the paper, which is most appropriate for focused curation in high-stake subject areas in search of breakthroughs (e.g., Alzheimer disease research). The two case studies demonstrated the flexibility and generalizability of our framework. Future discussions with stakeholders, such as journal editors retracting papers and curators maintaining knowledge bases, will help us perfect the existing procedures and discover other opportunities for applying our framework.

We argue that next generation scientific digital libraries should attend to knowledge maintenance. Current digital libraries' focus on individual papers causes challenges for tasks that require comparison and synthesis of multiple papers. With typical heuristics such as sorting papers by citation counts and recency, it is challenging to determine the state of the art in a field, identify the level of evidence supporting a statement, or find an appropriate citation. Systems prioritizing knowledge maintenance can better support these tasks.

Large-scale identification of keystone statements would be particularly valuable for several applications. A funding agency or research field could identify common keystone statements that appear repeatedly and prioritize them for further validation or verification. An editor could continuously surveil possible validity challenges in published work in their portfolio. An author could search an index of keystone citations, rather than searching by keywords, to find an appropriate citation.

Our work also highlights open questions about citation behavior. Further work on pass-through citations is needed to address why and how often authors cite main findings compared to other parts of empirical research papers. The citation of ad-hoc literature reviews needs particular examination because it takes the risk that

knowledge may have slipped into the literature without rigorous review or explicit validation. Also, citation behavior related to interpretations should be studied.

In the future we will experiment with using different approaches to argument-based curation, in place of the micropublications model. Argumentative zoning provides a coarser argumentative structure but has automated well in prior work. Adapting its existing automation might support scaling the identification of keystone citations and keystone statements. Domain-specific argumentation schemes are finer grained and hold particular promise for identifying unstated assumptions (enthymemes), critical questions, and dependencies underlying field-specific norms for argumentation. Identifying new argumentation schemes associated with fine-grain methods could be useful in combination with machine learning tools for methods prediction [26]. Analysis approaches used to identify argumentation schemes in genetics should be helpful [23].

Further exploration of the different categories of keystone citations, and their ramifications for knowledge maintenance is also a priority. In particular, certain subtypes of keystone citations may be possible to recover through existing automation, such as identification of 'method papers' [38]. Currently, distinguishing support as main findings, pass-through, or non-support is particularly challenging, and this is likely to remain a manual process for the foreseeable future. However, the Sci-Summ Shared Task[5], for instance, seeks to identify in cited text spans that most accurately reflect a given citation context [11].

Future work should develop a taxonomy of validity for indicating the confidence a reader can have in relying on or reusing the methods and the findings of a paper. This can be informed by insights from meta-research, such as that triangulation across multiple different methods increases confidence in research [32] while novel studies are frequently contradicted in subsequent research [27], which shows the urgency of integrating knowledge amongst sets of papers.

7 CONCLUSIONS

We have argued that scientific digital libraries need to take a knowledge maintenance perspective. Towards this end, we introduced the keystone framework, designed to identify when and how a citing paper is impacted by citing unreliable findings. We demonstrated how to use the framework in a digital library to trace the possible impact of error from a cited paper. Our first case study investigated whether the non-reproducibility of a computational chemistry protocol affected 10 citing papers, and demonstrated that by using the keystone framework, we can narrow down to a fraction of citing papers for follow-up actions such as alerts for validation and verification. Our second case study screened a high-impact paper in Alzheimer's disease research for keystone citations and elucidated how that paper's findings depend on the validity of the findings from its keystone citations. We presented a research agenda for knowledge maintenance in digital libraries. Librarians, database maintainers, knowledgebase curators, and research software engineers can play a role in applying the keystone framework to scientific digital libraries, and in developing infrastructures to further support knowledge maintenance.

[5] https://github.com/WING-NUS/scisumm-corpus

ACKNOWLEDGMENTS

Tim Clark, J. Stephen Downie, David Dubin, Kiel Gilleade, Michael Gryk, Daniel S. Katz, Halil Kilicoglu, Allen Renear, Karen Wickett. Alfred P. Sloan Foundation G-2020-12623. NIH R01LM010817.

REFERENCES

[1] Alvarenga, E.S., Santos, J.O., Moraes, F.C. and Carneiro, V.M.T. 2019. Quantum mechanical approach for structure elucidation of novel halogenated sesquiterpene lactones. *Journal of Molecular Structure*. 1180, (2019), 41–47. DOI:https://doi.org/10.1016/j.molstruc.2018.11.085.

[2] Avenell, A., Stewart, F., Grey, A., Gamble, G. and Bolland, M. 2019. An investigation into the impact and implications of published papers from retracted research: systematic search of affected literature. *BMJ Open*. 9, 10 (Oct. 2019), e031909. DOI:https://doi.org/10.1136/bmjopen-2019-031909.

[3] Bar-Ilan, J. and Halevi, G. 2018. Temporal characteristics of retracted articles. *Scientometrics*. 116, 3 (Jun. 2018), 1771–1783. DOI:https://doi.org/10.1007/s11192-018-2802-y.

[4] Bender, M. 2019. A code glitch may have caused errors in more than 100 published studies. *Vice*. https://www.vice.com/en_us/article/zmjwda/a-code-glitch-may-have-caused-errors-in-more-than-100-published-studies

[5] Bhandari Neupane, J., Neupane, R.P., Luo, Y., Yoshida, W.Y., Sun, R. and Williams, P.G. 2019. Characterization of Leptazolines A–D, polar oxazolines from the cyanobacterium *Leptolyngbya* sp., reveals a glitch with the "Willoughby–Hoye" scripts for calculating NMR chemical shifts. *Organic Letters*. 21, 20 (Oct. 2019), 8449–8453. DOI:https://doi.org/10.1021/acs.orglett.9b03216.

[6] Borgman, C.L., Edwards, P.N., Jackson, S.J., Chalmers, M.K. and Bowker, G.C. 2013. *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges*. https://escholarship.org/uc/item/2mt6j2mh

[7] Bornmann, L. and Daniel, H. 2008. What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*. 64, 1 (Jan. 2008), 45–80. DOI:https://doi.org/10.1108/00220410810844150.

[8] Bracegirdle, J., Robertson, L.P., Hume, P.A., Page, M.J., Sharrock, A.V., Ackerley, D.F., Carroll, A.R. and Keyzers, R.A. 2019. Lamellarin sulfates from the Pacific tunicate Didemnum ternerratum. *Journal of Natural Products*. 82, 7 (Jul. 2019), 2000–2008. DOI:https://doi.org/10.1021/acs.jnatprod.9b00493.

[9] Burnett, S., Singiser, R. and Clower, C. 2014. Teaching about ethics and the process of science using retracted publications. *Journal of College Science Teaching*. 043, 3 (2014), 24–29. DOI:https://doi.org/10.2505/4/jcst14_043_03_24.

[10] Chan, J., Chang, J.C., Hope, T., Shahaf, D. and Kittur, A. 2018. SOLVENT: A Mixed Initiative System for Finding Analogies between Research Papers. *Proceedings of the ACM on Human-Computer Interaction*. 2, CSCW (Nov. 2018), 1–21. DOI:https://doi.org/10.1145/3274300.

[11] Chandrasekaran, M.K., Yasunaga, M., Radev, D., Freitag, D. and Kan, M.-Y. 2019. Overview and results: CL-SciSumm Shared Task 2019. *Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019) co-located with the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)* (2019), 153–166.

[12] Clark, T. 2015. Argument graphs: Literature-data integration for robust and reproducible science. *First International Workshop on Capturing Scientific Knowledge Collocated with the Eighth International Conference on Knowledge Capture (K-CAP)* (Palisades, NY, Jul. 2015), 1–8.

[13] Clark, T., Ciccarese, P.N. and Goble, C.A. 2014. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. *Journal of Biomedical Semantics*. 5, 1 (2014), 28. DOI:https://doi.org/10.1186/2041-1480-5-28.

[14] De Calignon, A., Polydoro, M., Suárez-Calvet, M., William, C., Adamowicz, D.H., Kopeikina, K.J., Pitstick, R., Sahara, N., Ashe, K.H. and Carlson, G.A. 2012. Propagation of tau pathology in a model of early Alzheimer's disease. *Neuron*. 73, 4 (2012), 685–697. DOI:https://doi.org/10.1016/j.neuron.2011.11.033.

[15] Dinh, L., Sarol, J., Cheng, Y.-Y., Hsiao, T.-K., Parulian, N. and Schneider, J. 2019. Systematic examination of pre- and post-retraction citations. *Proceedings of the*

Association for Information Science and Technology (2019), 390–394. DOI:https://doi.org/10.1002/pra2.35.

[16] Eemeren, F.H. van, Garssen, B., Krabbe, E.C.W., Snoeck Henkemans, A.F., Verheij, B. and Wagemans, J.H.M. 2014. *Handbook of argumentation theory*. Springer Reference.

[17] Elkin, M., Scruse, A.C., Turlik, A. and Newhouse, T.R. 2019. Computational and synthetic investigation of cationic rearrangement in the putative biosynthesis of justicane triterpenoids. *Angewandte Chemie International Edition*. 58, 4 (2019), 1025–1029. DOI:https://doi.org/10.1002/anie.201810566.

[18] Erikson, M.G. and Erlandson, P. 2014. A taxonomy of motives to cite. *Social Studies of Science*. 44, 4 (Aug. 2014), 625–637. DOI:https://doi.org/10.1177/0306312714522871.

[19] Freeman, J.B. 2011. *Argument Structure: Representation and Theory*. Springer.

[20] Garfield, E. 1965. Can citation indexing be automated. *Statistical Association Methods for Mechanized Documentation, Symposium Proceedings* (1965), 189–192.

[21] Grabmair, M., Ashley, K.D., Chen, R., Sureshkumar, P., Wang, C., Nyberg, E. and Walker, V.R. 2015. Introducing LUIMA: An experiment in legal conceptual retrieval of vaccine injury decisions using a UIMA type system and tools. *Proceedings of the 15th International Conference on Artificial Intelligence and Law* (New York, NY, USA, 2015), 69–78.

[22] Gray, R., Al-Ghareeb, A., Davis, J., McKenna, L. and Amichai Hillel, S. 2018. Inclusion of nursing trials in systematic reviews after they have been retracted: Does it happen and what should we do? *International Journal of Nursing Studies*. 79, (Mar. 2018), 154. DOI:https://doi.org/10.1016/j.ijnurstu.2017.12.006.

[23] Green, N.L. 2018. Towards mining scientific discourse using argumentation schemes. *Argument & Computation*. 9, 2 (Jul. 2018), 121–135. DOI:https://doi.org/10.3233/AAC-180038.

[24] Gross, C.G. 2009. Three before their time: neuroscientists whose ideas were ignored by their contemporaries. *Experimental Brain Research*. 192, 3 (Jan. 2009), 321–334. DOI:https://doi.org/10.1007/s00221-008-1481-y.

[25] Guillen, P.O., Jaramillo, K.B., Jennings, L., Genta-Jouve, G., de la Cruz, M., Cautain, B., Reyes, F., Rodríguez, J. and Thomas, O.P. 2019. Halogenated tyrosine derivatives from the Tropical Eastern Pacific zoantharians Antipathozoanthus hickmani and Parazoanthus darwini. *Journal of Natural Products*. 82, 5 (May 2019), 1354–1360. DOI:https://doi.org/10.1021/acs.jnatprod.9b00173.

[26] Hoang, L., Boyce, R.D., Brochhausen, M., Utecht, J. and Schneider, J. 2019. A proposal for determining the evidence types of biomedical documents using a drug-drug interaction ontology and machine learning. *Proceedings of the AAAI 2019 Spring Symposium on Combining Machine Learning with Knowledge Engineering (AAAI-MAKE 2019)* (2019), 1–2. http://ceur-ws.org/Vol-2350/xposter3.pdf

[27] Ioannidis, J.P.A. 2005. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 294, 2 (Jul. 2005), 218–228. DOI:https://doi.org/10.1001/jama.294.2.218.

[28] Kasza, P., Trybula, M.E., Baradziej, K., Kepczynski, M., Szafrański, P.W. and Cegła, M.T. 2019. Fluorescent triazolyl spirooxazolidines: Synthesis and NMR stereochemical studies. *Journal of Molecular Structure*. 1183, (May 2019), 157–167. DOI:https://doi.org/10.1016/j.molstruc.2019.01.052.

[29] Keseler, I.M., Skrzypek, M., Weerasinghe, D., Chen, A.Y., Fulcher, C., Li, G.-W., Lemmer, K.C., Mladinich, K.M., Chow, E.D. and Sherlock, G. 2014. Curation accuracy of model organism databases. *Database*. 2014, (2014), bau058. DOI:https://doi.org/10.1093/database/bau058.

[30] Kircz, J.G. 1991. Rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval. *Journal of Documentation*. 47, 4 (Apr. 1991), 354–372. DOI:https://doi.org/10.1108/eb026884.

[31] Moravcsik, M.J. and Murugesan, P. 1975. Some results on the function and quality of citations. *Social Studies of Science*. 5, 1 (Feb. 1975), 86–92. DOI:https://doi.org/10.1177/030631277500500106.

[32] Munafò, M.R. and Davey Smith, G. 2018. Robust research needs many lines of evidence. *Nature*. 553, 7689 (Jan. 2018), 399–401. DOI:https://doi.org/10.1038/d41586-018-01023-3.

[33] Neupane, R., Parrish, S.M., Bhandari Neupane, J., Yoshida, Wesley Y., Yip, M.L.R., Turkson, J., Harper, M.K., Head, J.D. and Williams, P.G. Cytotoxic sesquiterpenoid quinones and quinols, and an 11-membered heterocycle, kauamide, from the Hawaiian marine sponge Dactylospongia elegans. *Marine Drugs*. 17, 7, 423. DOI:https://doi.org/10.3390/md17070423.

[34] Sandhu, N. and Schneider, J. 2018. Argument analysis of Alzheimer's Disease. Poster at University of Illinois Undergraduate Research Symposium. https://www.ideals.illinois.edu/handle/2142/106017

[35] Schneider, J. and Sandhu, N. 2018. Modeling Alzheimer's Disease research claims, evidence, and arguments from a biology research paper. (Jul. 2018). Presentation at the 9th International Conference on Argumentation, International Society for the Society of Argumentation, Amsterdam, Netherlands, https://www.ideals.illinois.edu/handle/2142/100340

[36] Schneider, J., Yi, D., Hill, A.M. and Whitehorn, A.S. 2020. Continued post-retraction citation of a fraudulent clinical trial report, eleven years after it was retracted for falsifying data. *Under submission to Scientometrics Special Issue on "Bibliometrics and Information Retrieval."* (2020).

[37] Schrag, M., Mueller, C., Oyoyo, U., Smith, M.A. and Kirsch, W.M. 2011. Iron, zinc and copper in the Alzheimer's disease brain: A quantitative meta-analysis. Some insight on the influence of citation bias on scientific opinion. *Progress in Neurobiology*. 94, 3 (Aug. 2011), 296–306. DOI:https://doi.org/10.1016/j.pneurobio.2011.05.001.

[38] Small, H. 2018. Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty. *Journal of Informetrics*. 12, 2 (May 2018), 461–480. DOI:https://doi.org/10.1016/j.joi.2018.03.007.

[39] Small, H. 1982. Citation context analysis. *Progress in Communication Sciences*. 3, (1982), 287–310.

[40] Spaltenstein, P., Cummins, E.J., Yokuda, K.-M., Kowalczyk, T., Clark, T.B. and O'Neil, G.W. 2019. Chemoselective carbonyl allylations with alkoxyallylsiletanes. *The Journal of Organic Chemistry*. 84, 7 (Apr. 2019), 4421–4428. DOI:https://doi.org/10.1021/acs.joc.8b03028.

[41] Stede, M., Schneider, J. and Stede, G. 2019. Modeling Arguments [Chapter 3]. *Argumentation mining*. Morgan & Claypool. 27–43.

[42] Suelzer, E.M., Deal, J., Hanus, K.L., Ruggeri, B., Sieracki, R. and Witkowski, E. 2019. Assessment of citations of the retracted article by Wakefield et al with fraudulent claims of an association between vaccination and autism. *JAMA Network Open*. 2, 11 (Nov. 2019), e1915552. DOI:https://doi.org/10.1001/jamanetworkopen.2019.15552.

[43] Teufel, S. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. University of Edinburgh. https://www.cl.cam.ac.uk/~sht25/thesis/t.pdf

[44] Teufel, S. and Kan, M.-Y. 2011. Robust argumentative zoning for sensemaking in scholarly documents. *Advanced Language Technologies for Digital Libraries* (2011), 154–170. DOI:https://doi.org/10.1007/978-3-642-23160-5_10.

[45] Valenzuela, M., Ha, V. and Etzioni, O. 2015. Identifying meaningful citations. *Scholarly Big Data: AI Perspectives, Challenges, and Ideas: Papers from the 2015 AAAI Workshop* (Apr. 2015), 21–26.

[46] Wager, E., Barbour, V., Kleinert, S. and Yentis, S. 2019. *COPE Guidelines for retracting articles [2019]*. Committee on Publication Ethics. https://publicationethics.org/node/19896

[47] Walton, D., Reed, C. and Macagno, F. 2008. *Argumentation Schemes*. Cambridge University Press.

[48] Weitz, D. 2017. *Report of the NSF workshop on Robustness, Reliability, and Reproducibility in Science Research*. http://www.mrsec.harvard.edu/2017NSFReliability/include/NSF_Workshop_Robustness.Reliability.Reproducibility.Report.pdf

[49] White, P.F., Kehlet, H. and Liu, S. 2009. Perioperative analgesia: What do we still know?: *Anesthesia & Analgesia*. 108, 5 (May 2009), 1364–1367. DOI:https://doi.org/10.1213/ane.0b013e3181a16835.

[50] Xu, H.-C., Hu, K., Sun, H.-D. and Puno, P.-T. 2019. Four 14(13→12)-abeolanostane triterpenoids with 6/6/5/6-fused ring system from the roots of Kadsura coccinea. *Natural Products and Bioprospecting*. 9, 3 (Jun. 2019), 165–173. DOI:https://doi.org/10.1007/s13659-019-0203-4.

[51] Zhu, J.S., Li, C.J., Tsui, K.Y., Kraemer, N., Son, J.-H., Haddadin, M.J., Tantillo, D.J. and Kurth, M.J. 2019. Accessing multiple classes of 2H-indazoles: Mechanistic implications for the Cadogan and Davis–Beirut reactions. *Journal of the American Chemical Society*. 141, 15 (Apr. 2019), 6247–6253. DOI:https://doi.org/10.1021/jacs.8b13481.

[52] Zhu, X., Turney, P., Lemire, D. and Vellino, A. 2015. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*. 66, 2 (2015), 408–427. DOI:https://doi.org/10.1002/asi.23179.

[53] Zou, Y., Wang, X., Sims, J., Wang, B., Pandey, P., Welsh, C.L., Stone, R.P., Avery, M.A., Doerksen, R.J., Ferreira, D., Anklin, C., Valeriote, F.A., Kelly, M. and Hamann, M.T. 2019. Computationally assisted discovery and assignment of a highly strained and PANC-1 selective alkaloid from Alaska's deep ocean. *Journal of the American Chemical Society*. 141, 10 (Mar. 2019), 4338–4344. DOI:https://doi.org/10.1021/jacs.8b11403.