Detecting Semantic Overlap and Discovering Precedents in the Biodiversity Research Literature

Position Paper

Graeme Hirst*, Nadia Talent†, and Sara Scharf‡

*Department of Computer Science, University of Toronto
†Department of Natural History, Royal Ontario Museum
‡Independent Scholar
gh@cs.toronto.edu;nadia.talent@utoronto.ca;sara.scharf@gmail.com

Abstract. Scientific literature on biodiversity is longevous, but even when legacy publications are available online, researchers often fail to search it adequately or effectively for prior publications; consequently, new research may replicate, or fail to adequately take into account, previously published research. The mechanisms of the Semantic Web and methods developed in contemporary research in natural language processing could be used, in the near-term future, as the basis for a precedent-finding system that would take the text of an author's early draft (or a submitted manuscript) and find potentially related ideas in published work. Methods would include text-similarity metrics that take different terminologies, synonymy, paraphrase, discourse relations, and structure of argumentation into account.

Keywords: Biodiversity literature, taxonomy, systematics, natural language processing, Semantic Web, paraphrase, textual entailment, text similarity, discourse relations, structure of scientific papers.

1 Introduction

Scientific progress comes from building on, and occasionally overturning, past results. It is therefore a researcher's responsibility to know the history of the topic on which they are working, and this is so for two primary reasons: (1) to do the best possible work, building upon the state of the art, and neither duplicating what has already been done nor repeating the mistakes of the past; (2) to include in any publication of the work a literature review that allows the reader to understand the work in its broader context, compare it with cognate research, and evaluate it for quality and novelty. This requires the researcher both to maintain a knowledge of current research (*current awareness*) and to perform searches for relevant work in the legacy literature when their new work necessitates it (*finding precedents*).

Nonetheless, for a variety of reasons, researchers do not always adequately achieve these tasks, and this can lead to subsequent problems both for their own work and for that of other researchers. And this is particularly so in research in biodiversity, more than perhaps most other sciences. Because of its longevous literature[1] and its need, in research on changes in biodiversity in ecosystems, to understand past conditions, finding precedents is both more important and more difficult than in the fast-moving don't-look-back-or-you'll-get-run-over sciences such as genomics.

In this position paper, we sketch the design of a proposed system that would draw on the mechanisms of the Semantic Web and methods in natural language processing to facilitate a search for precedents in the legacy biodiversity literature, especially (but not exclusively) the literature relating to systematics. It should be noted that what we are describing here is neither conventional search nor plagiarism detection (see footnote 8 below); our approach is influenced by research in the history of ideas in systematics on the detection of influence between authors and of independent re-invention (Scharf 2008).

2 What is a precedent, why do they matter, and why can they be hard to find?

We use the word *precedent* here, for want of a better term, to refer to any earlier published work or body of work that is, in an important way similar to, relevant to, or related to the current work in question. This is rather vague and subjective, but we can make it a little more concrete thus: An earlier published work is a precedent for current work if it has affected, or *should* have affected, the course of the newer work. This could include relevant methodologies, earlier attempts to solve the same problem, and earlier results and data. The most serious examples would be earlier work that is essentially the same as the present work (the new work is an independent re-invention), and, in particular, when the earlier work demonstrates that the new work is doomed to failure. Of primary interest to us in this paper are precedents in biodiversity research that, if not known and taken into account, render the current work seriously incomplete or erroneous.

Biodiversity research depends heavily on the legacy literature, which is the key source of important information about former biodiversity, and which also contains the results of massively time-consuming research that is difficult to replicate. The legacy literature of biodiversity includes a large component that is taxonomic literature. Besides the primary descriptions of new taxa, a major component of the taxonomic literature is synoptic volumes such as field guides, floras and faunas, synonymies, and 'manuals', which give varying levels of detail about the taxa present in a geographic area, including newly described taxa, summaries of opinion about previously defined taxa, and amended circumscriptions and descriptions. Modern synoptic works also include species-occurrence databases and analyses of biodiversity.

[1] "Natural history scientists work in fragmented, highly distributed and parochial communities, each with domain specific requirements and methodologies [Scoble 2008]. Their output is heterogeneous, high volume and typically of low impact, but with a citation half-life that may run into centuries" (Smith et al. 2009). "The cited half-life of publications in taxonomy is longer than in any other scientific discipline, and the decay rate is longer than in any scientific discipline" (Moritz 2005). Unfortunately, we have been unable to identify the study that is the basis for Moritz's remark.

Taxonomic nomenclature is a component of systematics that functions as a gateway to much of the taxonomic literature. It involves the application of the sets of rules that are laid down in the codes of nomenclature (ICZN 1999; McNeill et al. 2012) and periodically updated, with most provisions retroactively in force. The nomenclatural rules determine how the correct name for each species (or taxon of a higher or lower rank) must be determined. The *principle of priority* enshrined in the nomenclature rules holds as far back as the mid–eighteenth century, and literature of that vintage may be required to discover which name is correct. The definition of a taxon is anchored by the type specimen and the circumscription may be expressed either as a list of characteristics or as a list of specimens that the author considers to fit within the definition of the taxon. The specimen list may be either a list of typical specimens, or may be chosen to illustrate the range of morphological variation (or, potentially, the range of DNA sequences seen). Subsequent authors may wish to add to or subtract from the circumscription: common cases are (1) that a specimen of the other sex or a different life stage (such as a larva) is found, or (2) that a specimen originally cited is found to belong to a different taxon.

A taxonomist who wishes to create a new definitive list of the species in an geographic area or in a taxonomic group (a new "revision") must therefore search the legacy literature to find previous work that lists species in the area, or describes new species that might or might not be relevant, that amends previous descriptions, and (crucially) that works out the relationships between new or previously known species. They will need to find, evaluate, and cite prior publications that merge or split species (taxa), reclassify them into different groups, or assign new names to previously described species (taxa). All name alterations need to be re-evaluated in light of the rules of nomenclature now in force, which in practice means that previously ignored literature may resurface and lead the literature search into new areas. The precedents that were assumed for a work, and even the literature that was deliberately ignored for a work, may be listed in a way that requires a considerable sophistication in text understanding, for example in a book preface (e.g., Bentham and Hooker's *Genera Plantarum*).

Because of what has been termed the "citation gap" in the biodiversity literature (Payne et al. 2012), the taxonomic literature is massively undercited, and "such unintended omissions are likely to result in the decline of the [taxonomic] disciplines upon which the synoptic analyses depend" (Payne et al. 2012: p. 1350). This has occurred because the rules of nomenclature are now considered arcane by many researchers, and complete ignorance of the rules is common, not only among authors in ecology and biological taxonomy,[2] but lately even among the editors of major journals.[3] Large databases are being developed that already reduce the need to check the older literature, but their coverage is far from complete (Reveal 2012). Because of their ignorance and misunderstanding of the rules of nomenclature, the legacy literature becomes incompre-

[2] Systematics was traditionally a significant component of university biology courses, but the courses that provide this fundamental training have almost disappeared (Garnock-Jones 2013), replaced by courses that deal solely with molecular phylogenetic analysis, which is just one component of systematics.

[3] For an example of editorial problems, see the discussion in Taxacom at `http://mailman.nhm.ku.edu/pipermail/taxacom/2004-December/045547.html` et seq.

hensible to ecologists and inaccessible for biodiversity studies. But the consequences of mistakes, including failure to understand the older literature, can thus be very serious.[4]

Moreover, these kinds of mistakes may have a personal cost for their authors. When nomenclatural or taxonomic changes are referred to in later works, even in brief summaries, they usually carry a pointer to the authors who made the original change. Therefore, publications that err in this regard, if not ignored completely, are likely to be cited in a way that makes their transgressions apparent, an embarrassment for both the authors and the journal editors. For example, a taxonomic name may appear with an annotation such as *nomen dubium, nomen invalidum,* or *nomen illegitimum*, which indicates that the original authors erred. A correction may be published by later authors (neo- or lectotypification). When synonyms are listed, the authors commonly point to where their opinion differs from that of earlier authors, for example, *Synonyms: Leptospermum flavescens sensu W.L. Wagner et al. p.p., non Sm.* means that W.L. Wagner et al. included in the definition of *Leptospermum flavescens* some plants (*p.p. = pro parte* 'in part')) that did not match Smith's original description (*non Sm.*), and the present authors consider them to belong in another species; such a list may include implicit allegations that mistakes were made.

In the past, the principal problem had been lack of access to the required literature, but this is reducing, in large part due to the freely accessible Biodiversity Heritage Library[5] (Gwinn and Rinaldo 2009) and the (pay-walled) JSTOR collection, though much still remains inaccessible. But access helps only if researchers are willing to search this literature and can do so effectively. Non-technical barriers to doing so, in addition to the ignorance of the need and of the rules of nomenclature mentioned above, include time pressure, and the "Google effect" of just searching the Web and ignoring all but the top few results.

But even competent and well-intentioned researchers often have difficulties searching this literature. Simple Google-style keyword searches are frequently insufficient,[6] because in this literature, more so perhaps than most other fields of science, related concepts are often described or explained in different terms, or in completely different conceptual frameworks, from those of contemporary research. As a result, interesting and beneficial relations with legacy publications, or even with whole literatures, may remain hidden to term-based methods. In the case of taxonomy in particular, this implies the existence of what Nic Lughadha (2004) has called "hidden synonymies". The problem is compounded by ubiquitous Latin, non-obvious (to the modern reader) abbreviations, particularly Latin abbreviations and varied abbreviations of people's names, compact tabulations, and misspellings and multiple spellings of the same name.

[4] "International conventions and national or regional legislation concerning threatened or endangered animals specify the species or subspecies name of the animals that the law intends to protect. Thereafter, protection goes with the name rather than the endangered species itself. Any subsequent change in name could therefore affect conservation measures. The Commission often acts to protect the names of endangered species." — From the web site of the International Commission on Zoological Nomenclature (`http://iczn.org/content/conservation`)

[5] `http://www.biodiversitylibrary.org`

[6] Moreover, the quality of the OCR of many scans in the Biodiversity Heritage Library is presently so poor that keyword searches frequently result in false negatives.

Of course, none of this is to say that exact keyword matches are irrelevant or unhelpful. Term overlap can play its usual roles, and matches to names of taxa and of geographic locations are of particular importance.[7] However, our goal in the present work is to use semantic and structural relationships to discover the covert legacy literature that is not found with just a Google search or similar.

3 Foundational research

Ironically, we had great difficulty finding legacy literature on the topic of the difficulty of finding legacy literature, and on the topic of how researchers, in practice, search for and use this literature and the extent to which they do so.

The body of work that is perhaps most related to the former point is that of Swanson and colleagues (e.g., Swanson 1986; 1988; 1990) on identifying undiscovered public knowledge by analyzing the complementary but disjoint literature in two distinct fields of research and connecting knowledge in each to create new knowledge. For example, Swanson showed (1990; 1993) that studies on magnesium and studies on migraine, in two different fields, had terms in common, and the discovery that the two were related led in turn to the discovery that magnesium deficiency is connected with migraine. Superficially, the aim of this kind of analysis is the exact opposite of ours — it is looking at cases where, *a priori*, the authors are working in different research fields (rather than the same or closely related fields), and it does not operate at the level of the individual research paper. But methodologically it is similar nonetheless in that it is looking for an overlap or similarity in some aspect or aspects of the research. However, this work is limited in that the identification of related sub-fields was based simply on common terms used in both studies, and as we noted above, identical terminology cannot be assumed, even within a single research field. Moreover, the work needs, by its own background assumptions, to look at all possible pairings of topics of scholarship, and hence is prohibitively combinatorially explosive; in practice, a human must choose one topic or question as a starting point (Swanson 1993).

By contrast, in the approach that we will describe below, the search is constrained by assumption to a single, but large, field. This limits it sufficiently that it is computationally feasible with contemporary computing clusters. In the future, it will surely become computationally feasible to use our approach for Swanson's purposes.

4 Finding precedents in taxonomy and systematics

The confluence of research in natural language processing with Semantic Web technologies suggests the possibility in the near-term future of developing systems that would markedly improve researchers' ability to search and use the legacy literature in taxonomy and systematics. We assume the online availability of the literature itself — that is the continuing development of the Biodiversity Heritage Library (with improved

[7] A barrier that remains beyond the scope of this paper is the need for translation of literature written in languages not spoken by the searcher. Except for the special case of Latin, we do not address cross-lingual issues.

OCR), and access to the more-recent (still-in-copyright) twentieth-century literature in JSTOR and elsewhere. In this context, a precedent-finding system would take the text of an author's early draft (or a submitted manuscript) and find potentially related ideas in previously published work, matching not just words and phrases but ideas, regardless of how they are expressed. It would integrate current and expected near-term future research on the NLP technologies that we will describe below.[8]

We do not expect such a system to have a very high precision — many or most of its matches would be false alarms, although the design would attempt to minimize that. But the emphasis would be on high recall, bringing the potential matches to the attention of the user.

In the following subsections, we look at some of the primary elements, beyond literal keyword matching, of finding a match between new text and a potential precedent publication. We do not attempt a formal functional specification, which is the next step for this research, nor in the space available can we present examples, which would be textually large. We assume, without further comment, that a component for reasonably accurate translation of the Latin of taxonomic descriptions is available, and that the Latin is retained for keyword matching while the translation is used by other matching processes. We also assume that we have a component for recognizing taxonomic names in text, such as that of Koning, Sarkar, and Moritz (2005).

4.1 Paraphrase and similarity of meaning

The first element is the identification of sentences and phrases that are close in meaning. This has become an important research topic in computational linguistics in the last decade. It takes three forms; the first two are these:

1. *Paraphrase recognition*: identifying that two sentences or phrases are semantically equivalent or close to equivalent, even if very different in expression.
2. More generally, *recognizing textual entailment* (*RTE*): determining that the meaning of one sentence is entailed by, or is a consequence of, that of another. (Sentence-level paraphrase, then, can be thought of as mutual textual entailment.)

Dagan et al. (2013) provide a comprehensive survey of the techniques that have been developed for paraphrase recognition and RTE. Clearly, if we found this kind of a relationship between new work and a legacy publication, we would want to look further to see whether the latter might be a precedent.

The third form is this:

[8] Although there has been much research recently on *plagiarism detection* (see, for example, the evaluation lab overview by Potthast et al. (2012)), it is only peripherally relevant here, as it focuses primarily on finding matches for fragments of text that are precisely identical or differing in relatively minor ways, as when a plagiarizing student makes small changes in an attempt to evade detection. These are not the kinds of matches we are looking for. Current research in plagiarism detection has begun to take greater amounts of rewriting (including translation) into account (e.g., Barrón-Cedeño et al. 2013), making the task more like paraphrase detection (see below).

3. Measuring *semantic text similarity* (*STS*): identifying the degree to which two sentences, even if not paraphrases or entailing, are related in meaning.

Here, we are not looking for full equivalence or entailment, but rather trying to determine a degree of similarity or relatedness in meaning, and the methods that are used are rather different. Agirre et al. (2012) summarize the varied techniques and performance in a competitive evaluation of 35 STS systems. Even in the absence of equivalence or entailment, a high degree of relatedness throughout the two texts could indicate a potential precedent.

We expect that precedent-finding systems would draw on all three forms of this research. However, it should be noted that this research is presently limited to comparisons of pairs of sentences, whereas our goal inclues far broader comparisons long segments or complete texts, to find these relationships. So it will be important for this research to develop in this direction.

4.2 The low-level structure of scientific papers

The next element is the automatic analysis of the structure of scholarly discourse, especially scientific papers. Over the last decade, this has grown to become an important area of natural language processing (e.g., Ananiadou et al. 2012). This work endeavours to determine the structural purpose and discourse function of both individual sentences and of larger fragments of text in a scientific paper. Purposes or functions include such things as stating a claim, describing a gap in knowledge, criticizing or praising past work, and asserting the novelty of the present work (e.g., Teufel and Kan 2011; Angrosh et al. 2013a). This research also attempts to determine the purpose and scope of each citation in a paper (e.g., Siddharthan and Teufel 2007).

As this work becomes better and more mature, it can start to inform research on various relationships between texts (section 4.1 above), as the kind of information that it derives will be important in determining precedents. For example, if it is found that two sentences in different papers that are related in meaning are both claims, or both are statements of results, then we have a rather different situation with regards to identifying a precedent than if the sentence in the earlier paper is a result and the one in the later paper is a statement of the present state of the art.

The analysis of the structure of scientific texts will become more sophisticated in the future as it starts to incorporate more-detailed analysis of the discourse and rhetorical structures of text (e.g., Feng and Hirst 2012) — that is, the ability to find semantic discourse relationships between the clauses or sentences of a text, and then, in turn, the relationships that are built between larger fragments of text. That means not just the similarity or entailment relationships of section 4.1, but relationships such as CAUSE, CONTRAST, ELABORATION, and so on. And, in particular, it means finding them even when the author has left them only implicit in the text, which authors frequently do; in many contexts, human readers are able to recognize these relations without explicit textual cues, and authors tend to take advantage of this. Recognizing such implicit relationships is a current topic of research (Lin, Kan, and Ng 2009; Feng and Hirst 2013).

4.3 The argumentation structure of scientific papers

Our final element also relates to the structure of scientific papers, but at a higher level than the discourse relations. Ultimately, we would like to derive the structure of the overall argumentation[9] of a scientific text, and use that information too as a component of the matching process in our precedent-finding system. This is very difficult, even for people; a more realistic near-term goal based on current research (e.g., Lin, Kan, and Ng 2009; Feng and Hirst 2011) is to classify sentences as to their local role in the argumentation (e.g., premise, evidence) and use this information, and other identified discourse relations, to recognize larger components of the argumentation of the text and the kinds of argumentation scheme that it is using — for example, argument by analogy, or by induction, or by appeal to authority.

This could then allow matching of papers on the basis of the structure of the argumentation and how the content relates to this structure — or, indeed, independently of the content.[10] This kind of matching is less of an issue for the primarily fact-gathering aspects of searching the legacy literature that we described in section 2 above, but it would be of help in many other aspects of biodiversity (and other scientific) research.

4.4 Practical realization

Last, how would all this be realized in practice? Each item in the biodiversity and systematics legacy literature will need to be analyzed (including newly added items as they are published and as scanning of old literature continues) and annotated with an extensive representation for meaning and structure at all the levels of analysis. An important aspect of the representation and indexing of the legacy publications is that it must facilitate the process of checking for matches against new text, and must make this complex process as cheap as possible.

We anticipate that this representation would be based on XML and ontologies that are the topics of present-day research on mechanisms and resources for the Semantic Web. The annotation of some levels of analysis will be straightforward, such as the extraction of technical terms. Others will require further research and other design choices, as the nature of the representation will depend in part on the technical aspects of the methods chosen. For example, Dagan et al. (2013) list five distinct classes of methods for recognizing textual entailment; each implies different choices in the representation of the legacy text. One choice might involve annotating the text with details of the filled semantic roles of each sentence (Palmer, Gildea, and Xue 2010); another (not mutually exclusive) choice could be explicit annotation with contextually appropriate synonyms.

Practicality thus depends not only on our restriction of the domain (compared to the combinatorial problems of Swanson's approach, in section 3 above), but also on developing an effective representation.

[9] We refer, somewhat hyper-correctly, to *argumentation structure* to prevent the misinterpretation that we are talking about *argument structure* in the sense used in sentence-level syntax. We nonetheless refer to kinds of *argument* where there can be no terminological ambiguity.

[10] Retrieval of precedents by argumentation structure, without regard to the facts of any individual case, is also of particular concern to legal researchers (Dick 1991).

4.5 What's not included

The attentive reader will have observed that there are two things omitted from our proposal that might have been expected. The first is the use of citations and citation chains. One of our assumptions here is that our system is looking for things that are or might be completely disconnected, with respect to citations, from its starting point. Therefore, citations can play only a supporting role. Nonetheless, citations, including indirect connections, could still be a helpful factor in finding precedents; elaborating on this point is beyond the scope of this paper.

The other omission is semantic interpretation into a logical form, represented in XML, that draws on ontologies in the style of the original Berners-Lee, Hendler, and Lassila (2001) proposal for the Semantic Web. The problem with logical-form representation is that it implies a degree of precision in meaning that is not appropriate for the kind of matching we are proposing here. This is not to say that logical forms would be useless. On the contrary, they are employed by some approaches to paraphrase and textual entailment (section 4.1 above) and hence might appear in the system if only for that reason; but even so, they would form only one component of a broader and somewhat looser kind of semantic representation.

5 Conclusion

The precedent-finding system as we have sketched it here would be the culmination of a number of threads of research in computational linguistics and natural language processing and in document processing for the Semantic Web, and it can be thought of as a grand challenge for these fields. Moreover, we argue that by restricting our goals to the special case of the literature of systematic taxonomy and ecosystem biodiversity, we can achieve useful results in the near-term. But more generally, in a world in which increasingly interdisciplinary scholars must search an increasingly large legacy literature, precedent-finding systems would have great utility.

Acknowledgments. This work was financially supported by the Natural Sciences and Engineering Research Council of Canada and the Canadian Newt and Eft Foundation. We are grateful to Heike Zinsmeister for helpful discussions.

Bibliography

Angrosh, M.A.; Cranefield, Stephen; Stanger, Nigel (2013a). Context identification of sentences in research articles: Towards developing intelligent tools for the research community. *Natural Language Engineering*, to appear.

Angrosh, M.A.; Cranefield, Stephen; Stanger, Nigel (2013b). Contextual information retrieval in research articles: Semantic publishing tools for the research community. *Semantic Web Journal*, to appear. `http://iospress.metapress.com/content/q7j3606047461315`

Agirre, Eneko; Cer, Daniel; Diab, Mona; Gonzalez-Agirre, Aitor (2012). SemEval-2012 Task 6: A pilot on semantic textual similarity. *First Joint Conference on Lexical and Computational Semantics (*SEM)*, Montreal, 385–393.

Ananiadou, Sophia; van den Bosch, Antal; Sándor, Ágnes; Shatkay, Hagit; de Waard, Anita (editors) (2012). *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse, 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Korea. `http://aclweb.org/anthology-new/W/W12/W12-43.pdf`

Barrón-Cedeño, Alberto; Vila, Marta; Martí, M. Antònia; Rosso, Paolo (2013). Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, to appear.

Berners-Lee, Tim; Hendler, James; and Lassila, Ora (2001). The Semantic Web. *Scientific American*, 284(5), May 2001, 34–43.

Dagan, Ido; Roth, Dan; Sammons, Mark; Zanzotto, Fabio Massimo (2013). *Recognizing Textual Entailment: Models and Applications*. Morgan & Claypool Publishers.

Dick, Judith (1991). Representation of legal text for conceptual retrieval. *Proceedings, Third International Conference on Artificial Intelligence and Law*, Oxford, 244–252. `http://ftp.cs.toronto.edu/pub/gh/Dick-1991.pdf`

Feng, Vanessa Wei and Hirst, Graeme (2011). Classifying arguments by scheme. *Proceedings, 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon, 978–996.

Feng, Vanessa Wei and Hirst, Graeme (2012). Text-level discourse parsing with rich linguistic features. *Proceedings, 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Korea, 60–68.

Feng, Vanessa Wei and Hirst, Graeme (2013). Removing deleterious information to improve recognition of implicit discourse relations. Submitted.

Garnock-Jones, Phil (2013). The citation gap and its effects on taxonomy. In Blog: Theobrominated, 5 February 2013. `http://theobrominated.blogspot.co.uk/2013/02/the-citation-gap-and-its-effects-on.html`

Gwinn, Nancy E. and Rinaldo, Constance (2009). The Biodiversity Heritage Library: Sharing biodiversity literature with the world. *IFLA Journal*, 35(1): 25–34.

International Commission on Zoological Nomenclature 1999. *International Code of Zoological Nomenclature*, fourth edition. `http://www.nhm.ac.uk/hosted-sites/iczn/code`

Koning, Drew; Sarkar, Indra Neil; Moritz, Thomas (2005). TaxonGrab: Extracting taxonomic names from text. *Biodiversity Informatics*, 2, 79–82.

Lin, Ziheng; Kan, Min-Yen; Ng, Hwee Tou (2009). Recognizing implicit discourse relations in the Penn Discourse Treebank. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, Singapore, pages 343–351.

Moritz, Tom (2005). “Macro-economic case for open access.” Talk at Library and Laboratory: The Marriage of Research, Data and Taxonomic Literature, London, 5–6 February 2005. `http://barcoding.si.edu/LibraryAndLaboratory.htm` or `http://barcoding.si.edu/LibraryAndLaboratory/3-11_Moritz.pdf`

McNeill, J.; et 13 al. (2012). *International Code of Nomenclature for algae, fungi, and plants (Melbourne Code) adopted by the Eighteenth International Botanical Congress Melbourne, Australia, July 2011*. A.R.G. Gantner Verlag KG.

Nic Lughadha, Eimear (2004). Towards a working list of all known plant species. *Philosophical Transactions: Biological Sciences*, 359(no. 1444): Taxonomy for the Twenty-First Century (2004-04-29), 681–687. `http://www.jstor.org/stable/4142261`

Palmer, Martha; Gildea, Dan; Xue, Nianwen (2010). *Semantic Role Labeling*. Morgan & Claypool Publishers.

Payne, Jonathan L.; et 8 al. (2012). A lack of attribution: Closing the citation gap through a reform of citation and indexing practices. *Taxon* 61(6): 1349–1351. `http://www.ingentaconnect.com/content/iapt/tax/2012/00000061/00000006/art00030`

Potthast, Martin; et 11 al. (2012). Overview of the 4th International Competition on Plagiarism Detection. *Proceedings, PAN 2012 Lab: Uncovering Plagiarism, Authorship and Social Software Misuse*. In: Forner, Pamela; Karlgren, Jussi; and Womser-Hacker, Christa (editors), *CLEF 2012 Evaluation Labs and Workshop — Working Notes Papers*, Rome.

Reveal, James L. (2012). A divulgation of ignored or forgotten binomials. *Phytoneuron* 2012-28: 1–64. `http://www.phytoneuron.net/PhytoN-Divulgation.pdf`

Scharf, Sara (2008). Multiple independent inventions of a non-functional technology: Combinatorial descriptive names in botany, 1640–1830. *Spontaneous Generations*, 2(1):145–184. `http://spontaneousgenerations.library.utoronto.ca/index.php/SpontaneousGenerations/article/view/3552`

Scoble, Malcolm J. (2008). Networks and their role in e-taxonomy. In: Wheeler, Quentin D. (editor), *The New Taxonomy* New York: CRC Press, 19–31.

Siddharthan, Advaith and Teufel, Simone (2007). Whose idea was this, and why does it matter? Attributing scientific work to citations. *Proceedings, Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, NY, 316–323.

Smith, Vincent S.; et 4 al. (2009). Scratchpads: a data-publishing framework to build, share and manage information on the diversity of life. *BMC Bioinformatics*, 10(Suppl 14):S6. `http://www.biomedcentral.com/1471-2105/10/S14/S6`

Swanson, Don R. (1986). Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30, 7–18.

Swanson, Don R. (1988). Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine*, 31, 526–557.

Swanson, Don R. (1990). Somatomedin C and Arginine: Implicit connections between mutually isolated literatures. *Perspectives in Biology and Medicine*, 33, 157–186.

Swanson, Don R. (1993). Intervening in the life cycles of scientific knowledge. *Library Trends*, 41(4), 606–631.

Teufel, Simone; Kan, Min-Yen (2011). Robust argumentative zoning for sensemaking in scholarly documents. In: Bernadi, Raffaella et 4 al. (editors) *Advanced Language Technologies for Digital Libraries*, Lecture Notes in Computer Science, Volume 6699, 154–170.