Enhancing the bioscience literature for people and machines

Colin Caine

May 2015

Introduction

The world of scientific literature, particularly in the biosciences, has a "big" problem: there's too much of it and it is growing too fast.

As the volume of published work increases, and, indeed, as the rate accelerates[6], it becomes ever harder for researchers to keep up with their fields[4] and may make it easier for useful ideas and inferences to be lost.

Further, the way that scientific literature is distributed, searched and accessed has changed dramatically with the advent of the internet, partially enabling this new volume of papers. The papers themselves, however, are much as they were 100 years ago: facsimiles of long-form printed articles taking little advantage of the new media and giving little consideration to the massively increased volume of papers.

We believe that there is a need for new mechanisms to guide scientific reading, perhaps by highlighting major claims in articles and providing easy access to the data, perhaps by providing more and better data on the links between articles or to other knowledge bases.

In order to build these mechanisms, we need to understand what scientists are looking for when they read articles, what types of tasks are causing them to read, and what reading strategies they deploy. Hence, we need to understand how and why scientists read papers.

While there have been a large number of observational and interview studies investigating how and why scientists read scientific papers, there have been few quantitative studies, and none in our subdomain.

We intend to perform such a quantitative study by instrumenting a PDF reader and monitoring participants' gaze with an eye tracker. We hope this will allow us to build on existing studies to construct better models of scientific reading.

With a better understanding of scientific paper reading behaviour, we will investigate how well current communication media (PDFs, primarily) support and inhibit reading activities and what we can do to enhance scientific paper reading. We will also investigate whether it is possible to infer useful data about

users' intentions, attitude and beliefs towards a paper from their monitored behaviour.

Qayyum [16] describes a related experiment, their recommendations and discoveries. They describe how changes to annotation entry tools and their display could both enhance user experience and improve the quality of data available to machines from this annotation experience.

Our desired outcomes are similar in principle, though rather more ambitious. We would like to improve user experience and productivity while improving machine comprehension of the literature as a noninvasive side effect. Indeed, we believe that building a machine-readable version of the scientific literature is an essential pre-requisite for building the most powerful tools to aid researchers.

Project Lazarus is a project lead by Steve Pettifer and is Manchester's attempt to do just this. Lazarus aims to build a machine-readable version of the bioscience scientific literature by combining state of the art automated analysis with crowdsourcing microtasks completed as a noninvasive side effect of experts' reading behaviour and tool use.

My research questions are:

1. Why do bioscientists read scientific papers?
2. How do bioscientists read scientific papers?
3. How can we optimise that process?
4. Can the power of expert crowds be harnessed to recover machine-readable knowledge as a side effect of reading behaviour and use of our tools?
    1. Is this knowledge useful?
    2. Can we use these data to create tools to improve researcher productivity and/or enhance the scientific reading process?
    3. How can scientists be encouraged to contribute?

The first two questions will be answered by literature review and an experiment, possibly a series of experiments, currently being designed by Steve Pettifer, Robert Stevens, Caroline Jay, Teresa Attwood and myself. Question three shall be answered by speculative tool design and implementation, informed by analysis of our paper-reading experiments and literature review. Question four will be answered with an attempt to use our greater understanding of reading behaviour to recover meaning from the user behaviour we observe in Utopia Documents, our specialist PDF reader.

More details on our experiment design are in the next section. An overview of Project Lazarus, Utopia Documents and our novel crowdsourcing approach are similarly presented below. The Impact section contains a number of speculative tool proposals.

How do bioscientists read scientific papers?

Steve Pettifer, Robert Stevens, Caroline Jay, Teresa Attwood and myself are collaborating to design an experiment, and perhaps a short series of experiments to investigate how bioscientists read scientific papers.

Fundamentally, we want to collect usage and gaze information from participants who are reading scientific papers. The remaining questions are: What task should participants attempt? How will participants collect the papers that they need to read? How should we select participants?

The former questions concern a trade-off between ecological validity and comparability of data: we can get participants to complete similar, and therefore comparable, tasks; or we can ask them to perform tasks that are closer to what they would normally do, which are also less likely to be comparable between participants.

Our current proposed task is a protein annotation task, much like that undertaken by the Swiss-Prot curators[9]. Participants will be presented with a protein name and asked to find out various properties of the protein (such as location and related diseases) and supporting evidence in the literature.

This task has been selected because it is somewhat realistic and because it is completable in a reasonable time (~45 minutes) by our likely participants (UoM biomedicine graduate students). We are advised that this is a realistic task because bioscientists will need to search for the properties of proteins, genes or diseases in the literature as a typical research activity; moreover, this is a genuine task for a niche but valuable population of Swiss-Prot researcher/curators.

The expected data from the experiment are a time series of interaction events and gaze events for each participant. Interaction events will include mouse movements, scrolling, button and key presses, menu interaction, etc. Gaze events will be timestamped coordinates indicating where the eye-tracker believes the participant is looking.

By analysing these data together, we hope to discover patterns of behaviour that we can compare to prior studies and theory of reading. The analysis will draw on typical quantitative human-computer interaction studies practice, such as finding n-grams.

We do not yet know how we would like participants to discover and collect the papers they will need to read. This is significant as participants are likely to filter papers by reading titles, abstracts or the html preview before downloading them. If we do not instrument a web browser as well, we will not be observing an important part of the reading/researching process. If we strictly control how participants acquire papers to read then we lose some ecological validity.

Later experiments may be performed in "the wild", that is if we believe we can recover useful data from the instrumented PDF reader alone we may release it for researchers to use for their actual work. Markel Vigo et al. are pursuing a similar strategy in their usability studies for Protégé. They have performed a controlled study[18] and are now performing a study of Protégé users in the wild.

The generally passive nature of reading may preclude this PDF-reader-only approach, though Qayyum [16] has shown that studies encouraging annotation, and, we may speculate, other interaction, can prove successful.

Project Lazarus

Scholarly communications continue to be published as facsimiles of printed documents; prose heavy and littered with tables and figures intended for human, rather than machine consumption.

There is simply too much work to read and understand as an unaided human; we need support from machines, and to do that we need to build a literature that machines can understand[4].

While we hope that the future of scholarly communication will be created open access and increasingly machine-readable as tooling and understanding improve, it is unrealistic to expect the past decades and centuries of communications to be republished similarly.

This implies that we need to retrospectively create a machine readable version of the current literature to fuel our tools. Unfortunately, we speak a very different language to our machines and extracting meaning for machines from literature intended for human consumption is hard.

In traditional literature, factual assertions are written in prose and embedded in complex narratives, data is presented in tables and figures encoded more like images than data. Relations between articles are expressed primarily with citations, but sources are rarely identified uniquely, almost always lack version data and the citation graph itself is not available.

Text mining has had some success extracting meaning from prose and image recognition research some success with the kind of data tables and some other graphical representations of data we find in PDFs, however, these tools are generally not reliable enough to run unsupervised and, frustratingly, the vast majority of literature is not available to anyone en-masse for machine processing. Articles are available on a paper-by-paper basis to humans, and while a recent government ruling provides the right to process papers that one has access to for reading, licenses and technical methods are used to prevent bulk-download of content.

We observe that although no individual or single organisation can access the whole literature, the global community of scientists does have collective access to the literature (or at least all those bits anyone thinks are interesting). Lazarus can exploit this and use the community to provide access to the literature.

Project Lazarus proposes to combine state of the art automated analysis with noninvasive crowdsourcing microtasks to extract the data we need to build a machine-readable version of the current life science literature.

Useful data will be extracted and combined into an enormous graph relating articles, authors, institutions, concepts (drugs, body parts, diseases, etc), data, and more. We hypothesise that this graph will contain many novel, valuable insights. We further hypothesise that exposing life scientists to these useful insights will create an incentive to use our crowdsourcing tools, thus enhancing our graph and generating more insight.

Utopia Documents

The public facing element of Project Lazarus will be built into Utopia Documents (Utopia). Utopia is a specialist PDF reader for life science literature developed by Steve Pettifer, David Thorne, James Marsh, Teresa Attwood and others[4]. It has been downloaded over 30,000 times with around 500 more downloads each day.

To access papers for automated analysis, Utopia will be modified to send to our servers the full text of research papers that scientists read with it. We will then perform analysis, and store new extracted data in the graph.

Further data will be recovered by providing tooling within Utopia to make various tasks that scientists routinely perform while reading articles easier: annotation, organisation/categorisation and even extracting data tables. By using these tools, scientists will automatically generate citable nano-publications that we will incorporate into the graph and share with the community.

As Lazarus develops, we will incorporate new, motivating, features into Utopia that exploit the data we are collecting.

Crowdsourcing approach

A review of Good and Su [10] suggests that our approach is unique within life sciences. In typical crowdsourcing environments, participants are incentivised to perform a task they would otherwise not do in return for some reward, often money, entertainment, fame or access to a restricted resource[12].

In Utopia, the effort required to perform our tasks is very small, zero or even negative (if the task had to be done anyway and our tools make it easier than it would otherwise have been).

For example, using Utopia is (hopefully) preferable to less specialised PDF readers and merely using it completes a useful microtask by submitting papers for analysis and represents nil or negative effort; extracting a data table or molecular structure from a PDF "by-hand" is much more difficult than with our tools; extracting data with third-party tools, where such tools exist, will require similar or more effort than with our tools.

In these cases, the user simply uses the tool and as a transparent and noninvasive side effect, the tool contributes data back to the Lazarus project. Users will be notified appropriately that this collection will take place.

In place of effort, our users are trading in something more like risk. Risk that completion of a task will indirectly benefit a rival by making their lives easier or by leading them to realisations that they might have otherwise missed.

Initially, Lazarus will focus on tasks and data collection where we believe that the convenience offered by Utopia far outweighs the risk. The data gathered by these low risk activities will in itself by invaluable.

A second innovative crowdsourcing technique is being pioneered by our colleagues in the Scripps Institute. They are developing crowdsourcing systems that combine

expert and naive users' contributions while making the most of both categories of user. Expensive, expert users perform tasks requiring their domain knowledge and cheaper naive users do the less skilled jobs that fall out as a result.

We intend to investigate what role a paid naive crowd might have in developing or cleaning up the data produced noninvasively by our expert users.

Exploring new crowdsourcing strategies is important because, despite their attraction, effective, repeatable crowd-sourcing strategies have proved elusive; as Howe [12] says, "We know crowdsourcing exists because we've observed it in the wild. However, it's proven difficult to breed in captivity.".

Related work

Reading behaviour

As we considered possible tools for development, we realised we don't really know how bioscientists read papers. An initial literature review suggests that readers use a variety of strategies dependent on their knowledge, reading objective, familiarity with the work and more[5, 8, 14, 19].

Early studies in this area are primarily interview or observation based. More recent studies often use instrumented software, sometimes combined with eye-tracking data, to monitor subjects[7, 11, 15]. This latter approach offers a potentially greater payoff: if we can develop techniques to identify user's reading strategies or intentions from their interactions with the reader we can potentially use that data to build user aids and to tell us something about the paper we can incorporate into our knowledgebase.

Hornbæk and Frøkjær's work[11] is particularly relevant to us. Not only are they able to identify some different reading strategies from scrolling data alone, they show that different reading interfaces affect reading behaviour, such as exploration and reading speed, in a statistically significant manner. This is promising for our eventual goal of providing better reading tools.

So far, we have not been able to find another quantitative study of PDF-reader interaction and gaze of biomedical scientists during a research-related reading task.

Project Lazarus data collection

I have not yet performed an extensive literature survey on this topic.

Project Lazarus and my role within it are all about the unusual and unique nature of our datasets, how we hope to build them and the inferences and tools we hope to draw and build with them.

As discussed in Crowdsourcing approach, our data acquisition approach appears to be unique.

In my research so far I have been unable to find work producing substantially similar open access crowd-sourced data. The closest I have discovered are those produced by Mendeley.

Mendeley acts as a kind of social networking and article collection and sharing site. Formerly, it allowed users to evade subscription charges on papers as users with access would effectively make publicly available articles they collected, presently Mendeley allows users to share metadata on the articles they collect.

Mendeley users are encouraged to create libraries and groupings of articles to share with other users and also to upload their own articles to share and to track readership statistics.

Mendeley has an annotation feature in its desktop software, but the annotations are not easily shared or published.

While Lazarus intends to collect similar data on user categorisation habits, we will also help scientists extract, verify and share data from tables and figures, aid the creation of other valuable nano-publications and annotations, and collect data on user reading behaviour.

Our tool for reconstructing chemical structures from tables of R-groups appears to be unique. Some of our other data extraction tools, including our PDF-to-semantic-representation tool, Hamburger to cow (H2C), and our table extraction tools have competitors[13, 3, 2, 17]. The competitors use similar computer vision strategies to our tools. The H2C competitor, pdfextract, is much less capable than H2C, indeed, CrossRef reportedly ceased development of pdfextract after they discovered H2C's predecessor, pdfx.

Impact

Our experiments and study of how and why bioscientists read scientific papers will provide us with an understanding of the tasks that bioscientists seek to accomplish and the strategies that they use to accomplish them. The impact of our work will primarily be as a result of the tools and user interface improvements we will develop and distribute to make these tasks easier and through the creation and distribution of the Project Lazarus datasets.

Lazarus will provide means for millions of restricted-access articles to be processed by automatic systems, something that would be extremely difficult otherwise. Tools embedded in Utopia will reduce the effort required of scientists to perform important research tasks, including reading, data extraction and citation tracing, whilst sharing the results for the common good.

The data from these processes will form a precious and unique open-access resource for future research and tools.

I will benefit the Lazarus project directly by working with my colleagues to develop the crowdsourcing strategy and software as informed by our studies of reading behaviour, to integrate more data into Lazarus, and to refine our automated analysis tools.

Recovering insight from our data and developing tools for readers will benefit the project by increasing developer and user interest in the project (and thus hopefully stimulating greater crowdsourcing activity), by informing our data acquisition priorities, and by gaining and sharing expertise in the practicalities of this kind of data analysis.

The tools and techniques themselves will accelerate research by enabling hypothesis generation, simplifying literature review, and encouraging and enabling greater cooperation.

We believe the following tools (illustrated with indented use cases) will be of particular value.

Finding transitive relations

> Community 1 has observed that eating squid activates gene DA42. In another corner of life science, community 2 observes that DA42 is related to poor eyesight.
>
> Community 3 is producing a review of the effects of squid eating. Searching for that concept with our tools presents the scientists with a graph of relationships, perhaps weighted by how well known the links are. And each relation is supported by direct links to the literature for investigators to follow.

This tool allows scientists to discover and investigate connections that have never been explicitly stated. It will also guide scientist's reading, showing them exactly where in papers claims come from and performing a great deal of the article searching and collation automatically.

After discovering these relations, the authors can publish their review in the traditional literature and it's claims will be incorporated into Lazarus when the first users read it.

Suggesting collaboration

> Communities 1 and 2 are both investigating HN200, a flu virus, and it's relationships with various genes but publish in different journals.
>
> If made aware of each other they are likely to be able to collaborate.

Analysis of the citation and concept graphs in Lazarus can highlight communities or literature that are well connected or clustered together by concept, and badly connected by citation.

We can perform this analysis continually and alert a Lazarus researcher, mailing list, or perhaps automatically contact a corresponding author if we are very confident in the usefulness of our results.

Alerting scientists to new potential collaborators and to areas of related work they don't seem to know about could spark valuable collaborations that might have otherwise happened later or not at all.

The results of similar clustering analysis on the graph could be compared to the stated topics of conferences or journals (or work published in same) and suggest topics that are not well served.

While performing manual searches for matching keywords or similar may be easy, Lazarus will offer the advantages of more sophisticated clustering (perhaps incorporating domain knowledge on related concepts from ontologies as well as from Lazarus) and continuous, automated monitoring.

Finding evidence

> Alice is reading a paper and comes across an assertion that is new to her. With Utopia, she can ask Lazarus when this assertion/relationship first appears, perhaps filtering by whether Lazarus believes the assertion is evidenced.
>
> Alice is presented by a number of early papers that mention, say, the relationship between ear infections and the liver that she can then investigate. Not only are the papers enumerated, she can jump straight to the assertion.
>
> Perhaps she determines the topic is worthy of further review, so she queries Utopia for every data table it knows whose headers or immediately surrounding text contain both concepts and receives a load of tables with associated provenance data in a format she can easily import into her analysis software. There will still be significant work involved to unifying and comparing them, but hours of labour have been saved.
>
> And, of course, every new table that Alice might extract or assertion she annotates will be integrated into Lazarus for other researchers.

We hope that a tool like this will help us find "Wuuzl facts": beliefs that become widespread but are in fact ultimately all cited from an unevidenced musing in an otherwise unrelated paper (more generally, assertions that are repeated and perhaps "well known" but have never actually been supported by evidence).

We hope to identify evidenced assertions by looking at where they appear in papers (primarily which section they appear in) and whether their concepts appear in or near artifacts that tend to embody supporting evidence (like data tables and figures).

Future publishing

While these tools, and the Lazarus project as a whole, obviously support the reading, discovery and analysis of present and past literature, many of these tools and techniques will be transferable to the data likely to be provided by future, more machine-readable literature.

The necessary resilience of Lazarus tools against noisy data (due to our inexact data collection tools) also fits quite well with less discriminatory publishers and archives, like arXiv, whose strategies could conceivably play a larger role in the future of publishing.

My contribution

Since February 3rd, I have been familiarising myself with the existing work on Project Lazarus and with one of our initial data sources, termite text-mining analysis of a library of full-text life science journal articles.

More recently, I have been investigating how bioscientists read scientific papers by literature review and preparing for our first experiment.

The graph

Termite is a domain specific text-mining tool developed by SciBite. At present, we use it to find expressions in each article that appear to relate life science concepts. From these expressions we build a graph that relates expressions to concepts (diseases, body parts, genes, etc), to their host sentence, and through that to a source article.

This graph is instantiated in neo4j, a popular graph database software.

I have spent about two months learning how to manipulate and query the database we have created from cypher (neo4j's built-in declarative query language) and from plugins (Java classes that are incorporated into the database server).

As a first exercise, intended mostly to explore methods of querying the graph, I have developed a simple plugin and assorted primitive tooling to identify and rank other articles that are related by concept to a given article.

In our graph, a typical article contains a number of sentences that themselves contain expressions. Each expression is further considered to "have" at least two concepts.

I consider expression A to be related to expression B if and only if A "has" every concept that B "has". Note that this is not a commutative relationship: if A "has" more concepts than B then A may be related to B, but B will not be related to A.

For each article, I find the related expressions of each of its expressions, then trace each related expression back to a source article and record the number of related expressions for each source article.

Given some article x, each other article y are considered to be related to x with a strength proportional to the number of expressions in y that are related to expressions in x.

This is a rather simplistic approach but does find papers that look related to the original by concept (to this naive Computer Scientist). While a number of possible improvements are immediately apparent, not least to the noncommutative notion of expression relatedness, this task has fulfilled its immediate purpose of familiarising myself with the graph.

This task also highlights the difficulties we are likely to encounter evaluating our tools in the future: our dataset is very large and establishing, for example, to what extent two papers are related or that this assertion is evidenced here, is likely to be an expensive and skilled job. Manually annotating a subset of our data is also problematic: a small random sample is unlikely to contain much inter-relation and a sample informed from our tools is obviously biased.

Some thought will be needed here.

I have also investigated building a citation graph in Lazarus by recognising the reference sections of papers, querying CrossRef, and then storing that knowledge appropriately in Lazarus. I've put this on hold for now as the research focus has moved towards building a better understanding of reading activity and because H2C, a tool I intend to build my work on is currently being rewritten by the Utopia team.

The experiment

The experiment is discussed in the Related work subsection, Reading behaviour, and in the subsection How do bioscientists read scientific papers? in the introduction.

My primary contribution thus far has been reviewing the available literature and participating in the group discussions of the experiment.

Plan

The broad objectives of this project are outlined in the introduction. In this section I will discuss my specific plans for the current and next phases of my research and briefly discuss what I will be doing after that.

As a reminder, my research questions are:

1. Why do bioscientists read scientific papers?
2. How do bioscientists read scientific papers?
3. How can we optimise that process?
4. Can the power of expert crowds be harnessed to recover machine-readable knowledge as a side effect of reading behaviour and use of our tools?
    1. Is this knowledge useful?

2. Can we use these data to create tools to improve researcher productivity and/or enhance the scientific reading process?
3. How can scientists be encouraged to contribute?

Phase one

This is where I am now. In this phase I intend to complete a thorough literature review of prior experiments and studies on the topic of scientific reading strategies and motivations. This literature review will support the development of the primary deliverables for this phase: a typology or taxonomy of reading goals (and possibly behaviours); and a completed design for our first experiment on how bioscientists read scientific papers.

A sensible starting point for the taxonomy of reading goals will by O'Hara's excellent 1996 work *Towards a Typology of Reading Goals*[14], which discusses both reading goals and some behaviours. The primary work will be in reviewing O'Hara's work more thoroughly and in seeking more recent scholarship.

As discussed in How do bioscientists read scientific papers?, my collaborators and I have already developed some ideas about our future experiment. The design is, however, far from complete. Effort will be required to address the issues raised in that section and to ensure that our design benefits as much as possible from the experience of prior scholars. The primary work here will be the continuation of the literature review discussed in Reading behaviour and its integration into the experimental design.

Phase two

In this phase, I will perform our first experiment and some analyses of its results. I intend to analyse the gaze and interaction data using a variety of standard human-computer interaction (HCI) techniques to establish what readers are doing. In preparation for this, I have discussed the particulars of these kind of studies and their analysis with my peers Markel Vigo and Aitor Apaolaza whose recent work[18, 1] is related.

My analysis will further be informed by Phase one's literature review and promising any non-standard or non-HCI analysis that I find there.

Later phases

Following the analysis, we will determine if we need or want to run further experiments to improve our understanding of reading behaviour, motivations or perhaps even something else.

Having, hopefully, established a working understanding of bioscientists' reading behaviour and goals, we will look for goals that are inhibited by current paper and PDF-reader design. We will then develop tools to overcome these inhibitions, perhaps including some of the tools discussed in Impact.

Our increased understanding of reading goals and behaviour will also be used to improve our understanding of our crowdsourcing data and to determine what user behaviours we should monitor in Utopia and what those user behaviours will mean.

As necessary to this work, I will improve the Lazarus infrastructure and incorporate other data sources and tools into Lazarus.

Tools developed will be evaluated for their ability to enhance reader productivity and experience and for their ability to contribute useful data back to Lazarus.

I expect that evaluation of both the tools and the utility of our crowdsourced data will be difficult and may require novel evaluation frameworks. Likewise, I expect that understanding how to exploit the unusual and unreliable data sources that Lazarus will generate will be difficult.

References

[1] Aitor Apaolaza, Simon Harper, and Caroline Jay. "Understanding Users in the Wild". In: *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*. W4A '13. New York, NY, USA: ACM, 2013, 13:1–13:4. ISBN: 978-1-4503-1844-0. DOI: 10.1145/2461121.2461133. URL: http://doi.acm.org/10.1145/2461121.2461133 (cit. on p. 12).

[2] Manuel Arisrarán and Mike Tigas. *Introducing Tabula - Features - Source: An OpenNews project*. https://source.opennews.org/en-US/articles/introducing-tabula/. Apr. 2013. URL: https://source.opennews.org/en-US/articles/introducing-tabula/ (cit. on p. 7).

[3] Manuel Arisrarán, Mike Tigas, and Jeremy B. Merrill. *Tabula: Extract Tables from PDFs*. http://tabula.technology/. July 2013. URL: http://tabula.technology/ (cit. on p. 7).

[4] Teresa K. Attwood et al. "Calling International Rescue: knowledge lost in literature and data landslide!" In: *Biochemical Journal* 424.3 (Dec. 2009), pp. 317–333. ISSN: 0264-6021, 1470-8728. DOI: 10.1042/BJ20091474. URL: http://www.biochemj.org/bj/424/bj4240317.htm (cit. on pp. 1, 4, 5).

[5] Charles Bazerman. *Shaping written knowledge: The genre and activity of the experimental article in science*. Madison: University of Wisconsin Press, 1988. ISBN: o-299-11690-5. URL: http://wac.colostate.edu/books/bazerman_shaping/ (cit. on p. 6).

[6] Lutz Bornmann and Rüdiger Mutz. "Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references: Growth Rates of Modern Science: A Bibliometric Analysis Based on the Number of Publications and Cited References". In: *Journal of the Association for Information Science and Technology* 66.11 (Nov. 2015), pp. 2215–2222. ISSN: 23301635. DOI: 10.1002/asi.23329. URL: http://doi.wiley.com/10.1002/asi.23329 (cit. on p. 1).

[7] Georg Buscher et al. “Eye Tracking Analysis of Preferred Reading Regions on the Screen”. In: *CHI ’10 Extended Abstracts on Human Factors in Computing Systems*. CHI EA ’10. New York, NY, USA: ACM, 2010, pp. 3307–3312. ISBN: 978-1-60558-930-5. DOI: 10.1145/1753846.1753976. URL: http://doi.acm.org/10.1145/1753846.1753976 (cit. on p. 6).

[8] Davida Charney. “Study in Rhetorical Reading: How Evolutionists Read ’The Spandrels of San Marco’”. In: *Understanding scientific prose* (Jan. 1993), pp. 203–231. URL: http://tc.eserver.org/36509.html (cit. on p. 6).

[9] “UniProtKB/Swiss-Prot - Springer”. In: ed. by David Edwards. Methods in Molecular Biology™ 406. Humana Press, 2007. ISBN: 978-1-58829-653-5 978-1-59745-535-0. URL: http://link.springer.com/protocol/10.1007/978-1-59745-535-0_4 (cit. on p. 3).

[10] Benjamin M. Good and Andrew I. Su. “Crowdsourcing for Bioinformatics”. In: *Bioinformatics* (June 2013). PMID: 23782614, btt333. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btt333. URL: http://bioinformatics.oxfordjournals.org/content/early/2013/06/19/bioinformatics.btt333 (cit. on p. 5).

[11] Kasper Hornbæk and Erik Frøkjær. “Reading Patterns and Usability in Visualizations of Electronic Documents”. In: *ACM Trans. Comput.-Hum. Interact.* 10.2 (June 2003), pp. 119–149. ISSN: 1073-0516. DOI: 10.1145/772047.772050. URL: http://doi.acm.org/10.1145/772047.772050 (cit. on p. 6).

[12] Jeff Howe. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. 1st ed. New York, NY, USA: Crown Publishing Group, 2008. ISBN: 0-307-39620-7 978-0-307-39620-4 (cit. on pp. 5, 6).

[13] CrossRef Labs. *pdfextract*. 2012. URL: http://labs.crossref.org/pdfextract/ (cit. on p. 7).

[14] Kenton O’Hara. *Towards a Typology of Reading Goals*. Technical Report. Xerox, 1996. URL: http://www.xrce.xerox.com/Research-Development/Publications/1996-107 (cit. on pp. 6, 12).

[15] Takehiko Ohno. “EyePrint: Using Passive Eye Trace From Reading to Enhance Document Access and Comprehension”. In: *International Journal of Human-Computer Interaction* 23.1-2 (June 2007), pp. 71–94. ISSN: 1044-7318. DOI: 10.1080/10447310701362934. URL: http://dx.doi.org/10.1080/10447310701362934 (cit. on p. 6).

[16] Muhammad Asim Qayyum. “Capturing the online academic reading process”. In: *Information Processing & Management*. Evaluating Exploratory Search Systems Digital Libraries in the Context of Users’ Broader Activities 44.2 (Mar. 2008), pp. 581–595. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2007.05.005. URL: http://www.sciencedirect.com/science/article/pii/S0306457307001112 (cit. on pp. 2, 3).

[17] ScraperWiki. *Accurately extract tables from PDFs — PDF Tables*. https://pdftables.com/. URL: https://pdftables.com/ (cit. on p. 7).

[18] Markel Vigo, Caroline Jay, and Robert Stevens. "Constructing Conceptual Knowledge Artefacts: Activity Patterns in the Ontology Authoring Process". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. New York, NY, USA: ACM, 2015, pp. 3385–3394. ISBN: 978-1-4503-3145-6. DOI: 10.1145/2702123.2702495. URL: http://doi.acm.org/10.1145/2702123.2702495 (cit. on pp. 3, 12).

[19] David Wyatt et al. "Comprehension strategies, worth and credibility monitoring, and evaluations: Cold and hot cognition when experts read professional articles that are important to them". In: *Learning and Individual Differences* 5.1 (1993), pp. 49–72. ISSN: 1041-6080. DOI: 10.1016/1041-6080(93)90026-O. URL: http://www.sciencedirect.com/science/article/pii/104160809390026O (cit. on p. 6).