



Testing Google Scholar bibliographic data: Estimating error rates for Google Scholar citation parsing

by David Zeitlyn and Megan Beardmore-Herd

Abstract

We present some systematic tests of the quality of bibliographic data exports available from Google Scholar. While data quality is good for journal articles and conference proceedings, books and edited collections are often wrongly described or have incomplete data. We identify a particular problem with material from online repositories.

Contents

[Introduction](#)

[Procedures and results](#)

[Discussion and conclusions](#)

Introduction

Bibliographic databases such as *Web of Science* and Elsevier's *Scopus* were once standard tools that academics used on a regular basis to locate literature that they should either consult, read or include as a citation in their work. However, times have changed and a glut of new tools are available to use. Among these tools are search tools such as *PubMed* [1], *Microsoft Academic Search* [2], *Semantic Scholar* [3], *Meta* [4] and, most recently, *Dimensions* [5]. Another facet of the scholarly search space are pre-print servers such as *arXiv.org* [6] and now *bioRxiv.org* [7] as well as data sharing services such as *Dryad* [8] and *Figshare* [9]. Of all these tools that help researchers locate, cite and interact with primary content, *Google Scholar* [10] is probably the most used with the greatest coverage. Google introduced *Scholar* in 2004 and it quickly became a favourite with academics due to its ease of use and the fact that it searches not just title and abstract but the full-text of articles.

A core use case for search discovery services is that a researcher can quickly and easily download citation information for inclusion in their literature library. This is usually done by downloading a bibliographic data file such as a BibTeX, Reference Manager, RefWorks or EndNote file for inclusion in a local or online reference library. The standard reference file format underlying several of these export formats is .ris, a standardised bibliographic tagging format developed by Research Information Systems, the creators of Reference Manager [11].

As bibliographic data has moved forward over the last few decades, the core of the .ris file format has remained the same and has led to potential data loss in records as the bibliographic data landscape and use cases have become more complex. This article is inspired by the practical issues experienced in accomplishing day-to-day tasks that academics need to perform with downloads from *Google Scholar* (henceforth *GS*). Specifically, there is regular irritation around downloading citations from *GS* and other sources [12] and finding that they migrate incorrectly into bibliographic software because the downloaded file appears to result from poor parsing of the source. The result of this failure is that around the world many academics and students waste time correcting records downloaded from *GS* which could have been better prepared at the intermediary source, such as *GS*.

It is well known that Google prefers algorithmic or automated approaches to conglomerating metadata. This relies heavily on sources tagging and formatting their data in an appropriate "Google friendly"

manner. Thus, *Google Scholar* team is perhaps limited in the functionality that they can deliver due to the inhomogeneity of the data that they are handling to deliver the *GS* Web site. However, the issues that we raise below, we believe to be within Google's power to improve.

The problems with the data fall into three main classes: i) the completeness of data harvested from sources; ii) the representation of data harvested in Google's own data system; and iii) the inhomogeneity and poor quality of data standards used in displaying and coding bibliographic information on Web sites (for example, the Dublin Core [13] standard used as a basis for data holdings in many institutional digital repositories). Indeed, institutional repository records where full bibliographic information is typically included in a repository entry but it is not harvested by Google. Moreover, our study strongly suggests that some repository software seems not to put all the information into html metatags which *GS* harvest, so the reference generated is likely to be incomplete.

It is clear from *GS* documentation [14] that they chiefly harvest the information from html metatags and in their absence from parsing pdfs versions of the article. We think this may be part of the reason why repository records are so poorly recorded (see below). Although repository software such as *ePrints* [15], *DSpace* [16] and *Fedora* [17] do provide (at least some of) the metatags that *GS* looks for, the, the data is often "lite" due to the lack of clarity of the Dublin Core metadata standard (and variation in how it has been locally interpreted). Indeed, almost every institutional digital repository has custom data crosswalks that have been determined by the local institution based on their own preferences and interpretation of the data standard — there is no standard prescription. So, even if you know that a repository is a *DSpace* repository, or indeed that it implements the Dublin Core standard, there is no guarantee that the data will have been treated in a completely standardised manner.

In the following we define a downloaded *GS* citation reference file to be *correct* when

- a. reference type is accurate; and
- b. standard items of information for that reference type that are demonstrably there in the original are included.

It is important for the reader to understand a key point around the reproducibility of the results listed below. It is not in the remit of the current project to source a full copy of the *Google Scholar* data or to obtain a copy of Google's code base in order to allow the results below to be reproduced. The Google database and codebase is dynamically evolving, hence, our study cannot be replicated in the strictest sense of the term: the exact same searches can be repeated but they will not be searching the same dataset so the results may differ. This is an inevitable feature of research online and does not invalidate our results. For the record the data was collected over a period of months from October 2017 to February 2018 and the full dataset (which includes the data and time of the searches) is being made available for other researchers. We have saved both the lists of results received and the ris files that were downloaded and analysed. These are available as an open data appendix to this article on Figshare; DOI: <https://doi.org/10.6084/m9.figshare.5984845>.

Procedures and results

T1: Generic test.

For this the search term was chosen deliberately from a subject where monographs and chapters in edited collections are as important as journal articles and conference proceedings. (In some hard science subjects only the latter two types predominate.) Procedure: We searched on "early Roman History", a topic chosen to span many different reference types and evaluated the first 100 items. In this and all subsequent tests we saved the html page showing the first 100 results and downloaded the ris files containing the bibliographic data for those items. Our analysis consisted in the examination of these ris files to check that they correctly identified the reference type and included all the data associated with it.

Summary results	
Correct reference type	Complete data
76%	49%

Specific tests: T2: How often are journal articles correctly flagged?

Procedure: We searched for a journal title that does not contain a clue to reference type (*i.e.*, the word journal) and evaluated the first 100 items from that journal. Our chosen journal was *American Anthropologist*. This was well represented in the GS ris files. Our suspicion is that instances where journal articles get flagged as 'generic' may be a result of using references coming from repository copies, an issue we will discuss below.

Summary results	
Correct reference type	Complete data
100%	99%

T3: How often are papers in conference proceedings recognised as such?

Procedure: We searched for a conference title (Actual search: "Conference on Machine Learning") and evaluated the first 100 items. We note that this contains a clue to reference type (the word Conference. We could not find a conference without give-away words such as *conference* or *annual proceedings*).

Summary results	
Correct reference type	Complete data
98%	82%

As was noted above in some hard science subjects the most important types of publication are journal articles and conference proceedings. It is perhaps unsurprising (and somewhat reassuring) that these data types are well served by GS.

T4: How often are books recognised as such?

Procedure: We searched for a book title and looked at first 100 items. Chosen book: "Origin of Species" (the actual search was: "Origin of Species" Darwin). This should be (Type: BOOK): Darwin, Charles. *On the origin of species by means of natural selection*. 1859 (or other date for reprints). London: John Murray (or other subsequent publishers). Both place of publication and the name of publisher should be given (although many widely used publishing conventions omit one or other of these pieces of information).

Summary results	
Correct reference type	Complete data
73%	58%

T5: How often are edited books recognised as such?

Procedure: We searched for book title and evaluated the first 100 items. Chosen edited book: "The Invention of Tradition" (the actual search was: "The Invention of Tradition" Hobsbawm). This should be (Type: edited book) Hobsbawm, E., and T. Ranger, eds. 1983. *The invention of tradition*. Cambridge: Cambridge University Press.

Summary results

Correct reference type	Complete data
67%	62%

T6: How often are chapters in books recognised as such?

Procedure: we searched for a chapter title that does not contain a clue to the reference type (ie the word Introduction) and evaluated the first 100 items. Chosen chapter: "The Internal African Frontier: the making of African political culture." (Actual search: "The Internal African Frontier: the making of African political culture." Kopytoff) This should be (Type: book section) Kopytoff, Igor. 1987. "The Internal African Frontier: the making of African political culture." In *The African Frontier: the reproduction of traditional African societies*, edited by Igor Kopytoff, 3–84. Bloomington: Indiana University Press.

Summary results	
Correct reference type	Complete data
81%	64%

T7: A final issue: Repository copies as noise generators.

Procedure: We searched for "Oxford University Research Archive" and evaluated the first 100 results.

Summary results	
Correct reference type	Complete data
66%	55%

Repository records are confusing for both humans and it seems for robots. They are problematic because they are both a standalone document (*e.g.*, Paper X by Zeitlyn deposited in Oxford University Research Archive on this date at this URL) but referring to another one (Paper X by Zeitlyn published in Journal Y on another date at another URL). We think because of this GS have taken the entirely reasonable decision to say we are indexing the repository item so we will only describe the repository copy, irrespective of where else it has appeared. However, this makes it odd (to say the least) that they do not include repository URLs in some of the ris files.

Consider an example from another online repository.

Conditional random fields: Probabilistic models for segmenting and labeling sequence data
J Lafferty, A McCallum, FCN Pereira — 2001 — repository.upenn.edu
... Not for redistribution. The definitive version was published in Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001), pages 282–289.
Publisher URL: <http://portal.acm.org/citation.cfm?id=655813>

The GS citation is Lafferty, John, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001).

This does not include the URL, repository name or original source.

Consider another example from Oxford University Research Archive:


[CITATION] The development of an African working class
R Sandbrook, R Cohen — 1975 — ora.ox.ac.uk
... Citable link to this page: Title: The Development of an African working class. Subtitle: Studies in class formation and action.

<https://ora.ox.ac.uk/objects/uuid:bcb737f0-6709-479e-ad7a-c71ff23bbe25>

There are several problems with this as it has been indexed in *GS*. The work is an edited collection which the metatag and hence *GS* describe as a book. The value of the “DC.Type” metatag should use the DCMITYPE vocabulary which for an edited book should be ‘collection’ but is often confusingly given as ‘book’ reflecting a disdain for the distinction between ‘book’ and ‘edited book’ that goes back to the management issues of libraries dealing with physical volumes where the difference between a monograph and edited collection is unimportant. Although the repository entry has the book publisher it has not been picked up by *GS*. And most worryingly, in this case as in that above no URL is given in the ris file, so there is no clue from the bibliographic data download that the reference comes from a repository. Repositories contain full-text items of items published elsewhere, for example, from Green open access publishing. They also contain bibliographic information about items for which no full-text is available from the repository itself. In these cases the repository functions as a bibliographic database but this data is not being harvested by *GS*. For practical purposes this decreases the value of Google’s offering since a *GS* link does not always lead in a straightforward fashion to a full text version of a scholarly work that a researcher can read. While it would be reasonable to report full-text repository items as such (were URL and repository name given), in the absence of full-text availability it seems to us that the source data should be harvested.



Discussions and conclusions

This exercise has shown that *GS* works well for journal articles and conference papers the publishing arenas in which computer scientists, engineers and hard scientists are most comfortable. Other publication types as used by other academic disciplines are less well served by *GS*. Books, and edited collections, often are indexed with incomplete bibliographic information creating a distributed, invisibly duplicated, scatter of additional work for scholars at least some of which could have been reduced by further parsing work by *GS*. The recent moves to promote repositories driven by the laudable aims of open access publishing has introduced further noise into the system since there appears to be considerable inhomogeneity in the implementation of data standards, or possibly in clarity around how these standards should be applied. This has led to a mismatch between repository software and the harvesting protocols employed by *GS*. Our data suggest that the accuracy of *GS* ris files for books and repository records is unacceptably low (in the context of meeting academic needs) but seemingly quite easily improvable. At the very least repository software needs to report more and better quality information in html metatags and *GS* need to be better at providing the full set of data in the downloads they provide. 

About the authors

David Zeitlyn works as a social anthropologist as much as he can on Mambila spider divination in Cameroon where he has been doing research since 1985. At other moments he thinks about ICT and bibliography. Some reflections on this appeared in 2017 as “No vacation from citation” in the *Times Higher Education Supplement*, at <https://www.timeshighereducation.com/opinion/no-vacation-from-citation>.
Post: Institute of Social and Cultural Anthropology, School of Anthropology and Museum Ethnography, University of Oxford, 51 Banbury Road, Oxford, OX2 6PF, UK.
E-mail: david [dot] zeitlyn [at] anthro [dot] ox [dot] ac [dot] uk

Megan Beardmore-Herd is studying for a D.Phil. in anthropology, and has worked as a research assistant in primatology and palaeoanthropology.
Post: Institute of Cognitive and Evolutionary Anthropology, School of Anthropology and Museum Ethnography, University of Oxford, 64 Banbury Road, Oxford, OX2 6PN, UK.
E-mail: megan [dot] beardmore-herd [at] anthro [dot] ox [dot] ac [dot] uk

Acknowledgements

Daniel Hook made extremely helpful comments on an early draft of this paper for which we are very grateful. The research received no external funding and the authors are not aware of any conflicts of interest save the desire not to further waste time correcting downloaded bibliographic data files.

Notea

1. <https://www.ncbi.nlm.nih.gov/pubmed/>.
2. <http://academic.research.microsoft.com/>.

3. <https://www.semanticscholar.org/>.
4. <http://meta.com/>.
5. <https://app.dimensions.ai/discover/publication>.
6. <https://arxiv.org>.
7. <https://www.biorxiv.org>.
8. <https://datadryad.org>.
9. <https://figshare.com>.
10. <http://scholar.google.com>.
11. [https://en.wikipedia.org/wiki/RIS_\(file_format\)](https://en.wikipedia.org/wiki/RIS_(file_format)).
12. Including in some cases from publishers — the latter we note have no excuses for not providing full and correct bibliographic data.
13. <http://dublincore.org>.
14. <https://scholar.google.co.uk/intl/en/scholar/inclusion.html#indexing>.
15. <http://www.eprints.org/us/>.
16. <https://duraspace.org/dspace/>.
17. <https://duraspace.org/fedora/>.

Editorial history

Received 26 March 2018; accepted 7 August 2018.



This paper is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Testing Google Scholar bibliographic data: Estimating error rates for Google Scholar citation parsing
by David Zeitlyn and Megan Beardmore-Herd.

First Monday, Volume 23, Number 11 - 5 November 2018

<https://firstmonday.org/ojs/index.php/fm/rt/prINTERfriendly/8658/7607>

doi: <http://dx.doi.org/10.5210/fm.v23i11.8658>