Wait! What? There's stuff missing from the scholarly record?

Why more needs to be done to include grey literature

Toby Green
Co-Founder, Coherent Digital, Paris, France
0000-0002-9601-9130

doi: 10.56012/ajel9043

Correspondence to:
Toby Green
Toby.green@coherentdigital.net

Abstract

The scholarly record is an ever-evolving network – or graph – of truth assertions on which each discipline bases its discussions, and against which each discipline measures its progress. But what if the scholarly record is missing a significant number of truths? In this article, you will learn about the scholarly record, what it comprises, and what's missing. You'll discover that the volume – and value – of what's missing, called grey literature, has been. Grey literature is a costly asset that's going to waste.

"If I have seen further it is by standing on the shoulders (sic) of Giants."

So wrote Isaac Newton to his rival, Hooke, in 1675. Newton wasn't the first to notice the importance of building on the works of others. Five hundred years earlier, Bernard of Chartres is quoted as saying "We see more and farther than our predecessors, not because we have keener vison of greater height, but because we are lifted up and borne aloft on their gigantic stature". It seems to me that the collective noun for giants' shoulders should be a "scholarly record" since this comprises "an ever-evolving network – or graph – of truth assertions"[1] "upon which each discipline bases its discussions, and against which each discipline measures its progress".[2] Some go further to suggest the scholarly record can even frame the identity of an institution.[3]

Grey literature is a costly asset that's going to waste.

If we agree that a giant's shoulder is worth standing upon, how do we ensure that their truth assertions are collected and preserved to ensure that we can indeed see further than our predecessors and measure progress? In short, how do we ensure that all giants are included in the scholarly record?

Criteria for what should be included in the scholarly record needs to be updated…

To qualify for inclusion in the scholarly record, Dougherty proposed that an item must advance or summarise knowledge, have an identifiable author, be issued through an academic publisher, be catalogued by a university library, appear in curated research databases, and belong to a recognised discipline.[4] OCLC, a global library organisation, defines the scholarly record as "published outcomes of scholarly enquiry" such as "journal articles and monographs",[5] even though others recognise that it has, of late, become much more diverse, encompassing protocols, code, and data.[6]

Let's put these definitions to the test. Let's start with the first published output on Covid-19.

On December 31, 2019, the Wuhan Municipal Health Commission published a briefing on "a pneumonia epidemic situation" and informed the WHO China Country Office. On January 5, 2020, WHO published a briefing on its website.[7] These publications marked the start of the Covid-19 pandemic and future generations of scholars and students might well want to study them. Yet, according to the definitions above, they fail to qualify for the scholarly record because they were not issued by a scholarly publisher nor did they appear in the form of a journal article or monograph. Today, the link to the Wuhan briefing returns a message of "404 – page not found". I don't know if this content is simply offline or whether is it now lost. Either way, it is no longer easily accessible and, if it's lost, shows why maintaining the scholarly record matters.

A similar story can be told about the beginning of HIV-AIDS. On June 5, 1981, the US CDC published an article in its *Morbidity and Mortality Weekly Report* (MMWR) describing rare lung infections in five young men in Los Angeles.[8] The CDC isn't an academic publisher and MMWR isn't a journal – so, again, in theory, this report doesn't qualify for the scholarly record.

Let's tack away from medicine for a moment. In 2019, two professors from University College London and King's College London published a podcast that discussed two working papers authored by economists from the Bank of England, University College London, Cambridge University, London School of Economics, and University of Warwick. The papers had made headlines in the UK press, including the *Financial Times*, and were cited in a blog run by a professor from University of London's Royal Holloway. This blog has a larger following than most journals. None of these items, including these high-impact papers, passed through the hands of an academic publisher. Hunting for them in journals, subject databases, and library catalogues will be in vain because, as with the previous examples, this is content that was released into the wild without any thought as to how it might be captured for the scholarly record.[9]

These examples lead me to conclude that Dougherty and OCLC's criteria for what should

Bernard of Chartres on the shoulders of a giant

Wikimedia: Library of Congress, Rosenwald 4, Bl. 5r. Unattributed

be included in the scholarly record needs to be updated, not least to take into account how digital and Web 2.0 tools are changing the ways in which knowledge is being published, as the 2019 example illustrates.

(Grey literature) is hard to source and is missing from secondary discovery services.

In the analogue era, authors had little choice but to find a publisher for their works: the cost of self-publishing and dissemination in print was beyond the means of most. Equally, the cost of organising and maintaining archives meant that only institutional libraries could offer readers meaningful and useful collections of previously published materials. It's no surprise that publishers and libraries were central to the creation and maintenance of the scholarly record.

Behind the scenes, publishers worked with booksellers and agents to develop an efficient, near-global, supply chain that carried their publications to libraries around the world. To reduce administration costs and speed delivery, publishers, booksellers agents, and librarians co-developed processes (e.g. ICEDIS) and metadata standards with unique identifiers (ISBNs in 1969, ISSNs in 1975). In parallel, secondary services and catalogue systems emerged to tackle the challenge of discoverability.

Since 2000, and with the transition to a digital era, both the supply chain and discovery services have been totally re-engineered. Today's standards include new persistent identifiers (PIDs) for content (DOIs) as well as for authors (ORCiDs) and their institutions (Ringgold).

This is not to say that all scholarly publications exist inside this publisher-secondary services-library "complex". They don't. Some institutions choose to self-publish because doing so has advantages, such as control over branding, timing, and pricing. Whilst some institutions, such as OECD and Brookings Institution, mimic mainstream publishers, using the same metadata standards and supply chains to channel their publications to libraries, others, especially smaller organisations, don't. In eschewing publishing norms and supply chains, their content is hard to source and is missing from secondary discovery services – frustrating for librarians and readers alike. Their content is known as "grey literature".

Grey literature

In 1984, Wood coined the term grey literature to describe material "which is not available through normal bookselling channels … leading to problems for the producers of secondary services, for librarians who wish to collect it, and for end users." Whilst noting that grey literature had a number of other distinguishing

Box 1: Prague definition of grey literature[12]

Grey literature stands for manifold document types produced on all levels of government, academics, business, and industry in print and electronic formats, that are protected by intellectual property rights, of sufficient quality to be collected and preserved by library holdings or institutional repositories, but not controlled by commercial publishers, i.e., where publishing is not the primary activity of the producing body.

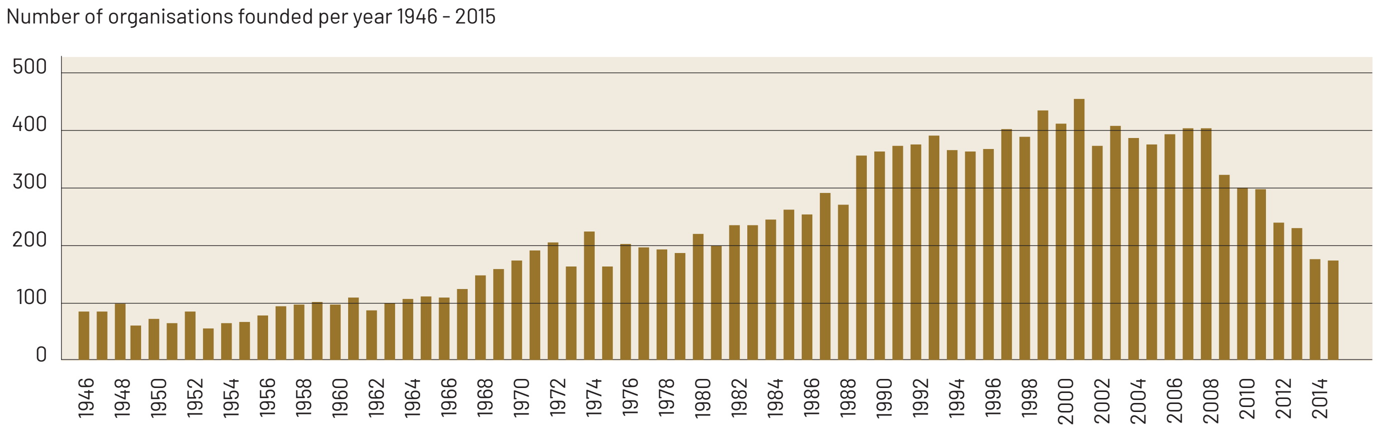


Figure 1. Number of NGOs and think tanks founded per year 1946–2015. Source: Policy Commons

characteristics – "variable standards of editing and production, poor publicity, poor bibliographic control, and poor availability in libraries", Wood rejected as "mistaken" the belief that grey literature was "essentially ephemeral and of local interest only" because "it contains information likely to be of use to a considerable number of people".[10] It is often thought that, while useful, grey literature hasn't been peer-reviewed. This is a big misunderstanding because, at least in policy, more than 60% is reviewed by experts prior to release.[11] So, no wonder that Wood reckoned grey literature "a costly public asset going largely to waste". How costly? One estimate puts it at $33BN a year.[11]

Wood's definition captured the essence of the challenge grey literature poses information professionals and readers: that this content is hard to find, capture, and use.

In 2010, the Prague definition (see Box 1) attempted to build on Wood – but the additions, to my mind, simply muddy the waters. Prague lists some producers ("government, academics, business, and industry") but excludes others (e.g., third-sector organisations and NGOs). It adds the qualifier "commercial" to publishers, which fails to understand that any publisher, for profit or not, must behave in a commercial manner if it is – as Dickens' Mr Micawber elegantly put it – to avoid financial misery. Moreover, the Prague definition is wrong to suggest that producing bodies where publishing "is not the primary activity" necessarily produce grey literature. Many universities and, as noted above, some IGOs and NGOs, run professional publishing "presses" that publish in a manner identical to houses like Elsevier and Springer and their publications are as easily found in secondary services and obtained by libraries and users alike via standard supply chains.

Wood was right in 1984 and, as I will show, his definition is just as valid in today's digital era. However, he would probably be shocked by the scale of today's public asset going to waste. That's because there is a growing amount of scholarly and professional content being published outside mainstream supply chains, which – as Wood would recognise – leads to problems for the producers of secondary services, for librarians, and for end users. The scholarly record is missing many shoulders from today's *bone fide* giants. One example is the Intergovernmental Panel on Climate Change (IPCC) who actually switched from working with publishers to self-publishing on their websites: what was formally published is now grey literature.

The core problem is the same one as Wood identified in 1984: poor bibliographic control. What compounds it is the significant increase in the supply of grey literature over the past decade. Let's look at these two issues, starting with supply.

Supply

The supply of scholarly content strongly correlates with the number of researchers.[13] So, has the supply of researchers been growing? In the 1980s, just under a third of those emerging from education systems in OECD countries did so with first degrees. Of these, roughly a quarter went on to do a masters or doctorate, so around 8% of this cohort emerged as "research capable".

In the 2000s, the proportion leaving education systems with degrees in OECD countries was up to half, of whom half went on to get postgraduate qualifications. So, 25% of the 2000s cohort emerged "research capable", a sizable increase over the 8% seen in the 1980s.[14,15] Yet, the number of jobs in academia barely changed. In the 1980s, around 15% of freshly-minted PhDs in the UK could expect to work in academia. By the 2000s, this had fallen to around 3%. So, if not into academia, where did this growing number of highly-trained, research capable people go? Some went into industry and government, but some must have joined the booming services and third sectors. (The third sector is that part of an economy or society comprising non-governmental and non-profit-making organisations or associations, including charities, voluntary and community groups, cooperatives, etc.) As the graph above shows, there has been strong growth in the number of new third sector organisations since the end of WWII, with many employing researchers to support their mission. But here's the kicker. Unlike their cousins in academia, researchers in government, industry, and third and service sectors don't have to publish in books and journals to further their careers. They are free to work with their employers to self-publish their research as reports, working papers, and other digital-first formats – and they are increasingly doing so. An analysis of the Policy Commons database, which indexes grey literature from over

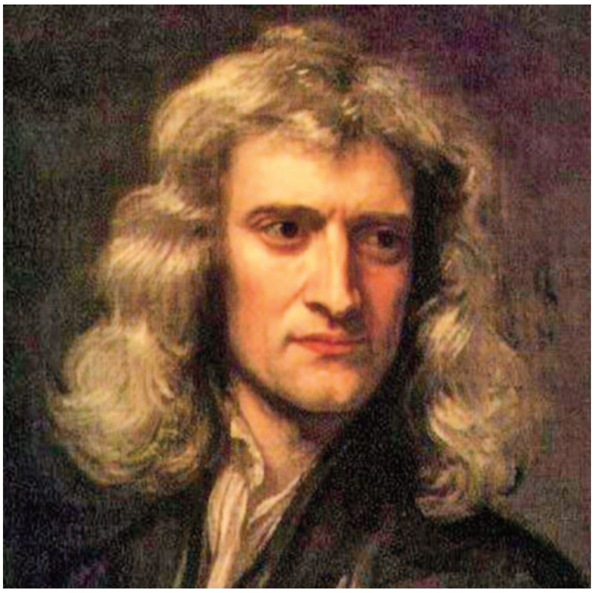
8,500 IGOs, NGOs, think tanks, and research centres from around the world, shows 55% more grey literature was released in 2020 compared with 2010 (287,545 items and 184,514, respectively). In the field of policy alone, I estimate that each year sees around 400,000 newly published items of grey literature – that's 10% of the world's journal output.

Poor bibliographic control

Today, desktop publishing, web 2.0 tools, and websites make is easy for anyone to self-publish. As Clay Shirky, an early internet "guru" and Professor at the Interactive Telecommunications Program at New York University said in a 2012 interview: "Publishing is not evolving. Publishing is going away. Because the word 'publishing' means a cadre of professionals who are taking on the incredible difficulty and complexity and expense of making something public. That's not a job anymore. That's a button. There's a button that says 'publish', and when you press it, it's done."[16] Shirky was half right. It is indeed easy to press a button and publish something online. The problem is that most people who press that button are not from that cadre of professionals who understand the incredible complexity of preparing content so it's discoverable and useful for its readers. They don't know how to wrap it in the metadata that's needed to make it discoverable and easily and reliably citable. They don't know how to ensure it is included in specialist discovery services. Nor do they understand, and more than Shirky's interviewer did, that it isn't "done" until the work has been safely preserved in the scholarly record. It's ironic that links to Shirky's interview, published in the blog *Findings*, returned a "404-page not found" within months of its publication when the blog closed and went offline.

Worse, like *Findings'* publisher, most organisations have no strategy to prevent link rot[11] and it's hardly a surprise that 75% of links in scholarly journals to "web at large" items lead to the wrong content.[17]

Plainly, it is still incredibly difficult and complex to prepare content and metadata to the standards needed to ensure that it's discoverable by users and easily available to librarians for their collections. Despite the advances in digital publishing, gathering a scholarly record of giants' shoulders is still as challenging as herding cats.



If I have seen further it is by standing on the sholders (sic) of Giants."

SIR ISAAC NEWTON

Now, you might imagine that what's missing can be quickly found via public search engines than scan open websites, like Google. The trouble with public search engines is that they deprecate content with poor metadata on low-traffic websites – most grey literature will be crowded out by content from "optimised" websites run by digital marketers.[18] Besides, public search engines seek to tailor results to each users' "bubble" of preferences, attitudes, and even location and results can change from day-to-day as algorithms evolve.[19,20] This is why most scholars and students still turn to the specialist search engines where, of course, grey literature is largely absent.[21]

Over the past two decades, publishers and librarians have been focussed on capturing research findings from the academy – mainly in books and journals – to create a digital scholarly record that's overlaid with sophisticated discovery systems for use by the academy. At the same time, they are attempting to pivot a $25BN industry to open access so the scholarly record becomes an asset not just for the academy but also for society at large.[22]

In parallel, and largely ignored, a growing number of researchers at non-academic institutions and organisations have been using digital publishing tools to post their research findings – as reports and papers – openly, via their websites. This is also a $25BN information industry, but, as I've shown, this grey literature is missing both from specialist discovery systems and library collections and is still woefully under-used. Grey literature is still a costly asset that's going to waste.

Conclusion

In 1990, I met a professor who ran a laboratory in France. He told me that the door to the library was open 24/7 but the key to the lab was given only to those who had first used the library to complete a thorough literature review of the topic they wished to investigate. At that time, when practical and financial hurdles meant there was little grey literature, the policy made sense. The professor could be confident that the library's access to the scholarly record was such that valuable lab time would only be spent looking further than was possible from the shoulders of giants who had gone before.

Today, in a world where "a button" has removed the practical and financial barriers to posting research findings on employers' websites, that policy would be increasingly undermined. Valuable lab time might be wasted because an increasing volume of giants' truth assertions are missing from the library's collection.

Researchers in government, industry, and third and service sectors don't have to publish in books and journals to further their careers.

Disclosures and conflicts of interest

The author is a co-founder of Coherent Digital LLC., whose mission is to tame wild content.

Some of the references and data used in this article come from services provided by Coherent Digital.

References

1. Eve MP. What is the scholarly record? Available from: https://eve.gd/2022/07/26/what-is-the-scholarly-record/
2. Atkinson R. Text mutability and collection administration. Library Acquisitions: Practice & Theory. 1990;14(4):355–8. doi:10.1016/0364-6408(90)90006-G
3. Proffitt M. The scholarly record: A view from the campus. Available from: https://hangingtogether.org/the-scholarly-record-a-view-from-the-campus/
4. Dougherty MV. Defining the scholarly record. In: Correcting the Scholarly Record for Research Integrity. Research Ethics Forum, vol 6. Springer, Cham; 2018. doi:10.1007/978-3-319-99435-2_2
5. Lavoie B, Childress E, Elway R, et al. The evolving scholarly record. OCLC Research. Dublin, Ohio. 2014. doi:10.25333/C3763V.
6. Tay A. Aaron Tay's musings about librarianship. Available from: http://musingsaboutlibrarianship.blogspot.com/2022/06/diversity-of-scholarly-record-push-to.html
7. WHO. Covid-19 China. Available from: https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON229
8. CDC MMWR. First report of AIDS. Available from: https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5021a1.htm
9. Coherent Digital. Policy Commons. Available from: https://policycommons.net/lists/228/wait-what-theres-lots-of-vital-stuff-missing-from-the-scholarly-record-uksg-lightning-presentation/
10. Wood DN. Management of grey literature. K.G. Saur. 1990. doi:10.1515/9783111514598.61
11. Lawrence A. Influence seekers: The production of grey literature for policy and practice. Inf Serv Use. 2017;37(4):389–403. doi:10.3233/ISU-170857
12. Schopfel J. Towards a Prague definition of grey literature. The Grey Journal. 2011;(7):1. Available from: https://www.greynet.org/images/Contents_TGJV7N1.pdf
13. Mabe M, Amin M. Growth dynamics of scholarly and scientific journals. Scientometrics. 2001;51:147–162. doi:10.1023/A:1010520913124
14. OECD. Population with tertiary education. Available from: https://data.oecd.org/eduatt/population-with-tertiary-education.htm
15. Bolton P. Education in United Kingdom: Historical statistics 1900-2010. (2012) [cited 2022 Sep 30]. Available from: https://policycommons.net/artifacts/2459320/untitled/3481117/
16. Findings Interview. How we will read: Clay Shirky. 2012 [cited 2022 Sept 30]. Available from: https://policycommons.net/artifacts/2676517/how-we-will-read_-clay-shirky/3699667/
17. Jones SM, Van de Sompel H, Shankar H, et al. Scholarly context adrift: Three out of four URI references lead to changed content. PLoS One. 2016;11(12):e0167475. doi:10.1371/journal.pone.0167475
18. Bala M, Verma D. SSRN. A critical review of digital marketing. International Journal of Management, IT & Engineering. 2018;8(10):321–39. Available from: https://ssrn.com/abstract=3545505
19. Ćurković, M, Košec A. Bubble effect: including internet search engines in systematic reviews introduces selection bias and impedes scientific reproducibility. BMC Med Res Methodol. 2018;18;130. doi:10.1186/s12874-018-0599-2
20. Urman A, Makhortykh M, Ulloa R. The matter of chance: Auditing web search results related to the 2020 U.S. presidential primary elections across six search engines. Soc Sci Comput Rev. 2002;40(5):1323–39. doi:10.1177/08944393211006863
21. Gusenbauer M. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. Scientometrics. 2019;118:177–214. doi:10.1007/s11192-018-2958-5
21. Johnson R, Watkinson A, Mabe M. The STM Report, 5th edition: An overview of scientific and scholarly publishing, STM. Netherlands 2018 [cited 2022 Sep 30]. Available from https://policycommons.net/artifacts/1575771/2018_10_04_stm_report_2018/2265545/

Author information

Toby Green is a co-founder of Coherent Digital LLC and was previously head of publishing for the Organisation for Economic Cooperation and Development (OECD). He has also held positions with Elsevier Science, Pergamon Press, and the Association of Learned and Professional Society Publishers (ALPSP). ORCID: 0000-0002-9601-9130