Research Trends

Volume 1
Issue 37 *Special issue on alternative metrics*

Article 7

6-1-2014

The grey literature from an altmetrics perspective - opportunity and challenges

Euan Adie
Altmetric LLP

Section 6:
Value of bibliometrics

The grey literature from an altmetrics perspective - opportunity and challenges

Euan Adie
Altmetric LLP

The field of altmetrics encompasses both alternative metrics (data beyond citation counts or impact factors) and alternative research outputs (like datasets and software).

But some material falls into both camps.

Grey literature – theses, posters, preprints, patents and policy documents and similar – are created by researchers and informed by research but aren't usually viewed as first class citizens of the scholarly literature. They are not all tracked in citation indexes like Web of Science or Scopus and can be difficult to cite in academic journals, with some editors discouraging any formal citation of preprints and similar types of document. For example, the Oxford Journals author guidelines (1) states that the reference section must not include manuscripts not formally accepted for publication, e.g. preprints. There can be good reasons for this, which we'll explore further later in the article.

The term 'grey literature' comes from their position in the fuzzy grey area between academic and popular literature (2). Importantly they are resources that aren't typically controlled by academic publishers, traditionally the gatekeepers of the scholarly record. Publishers generally take this role seriously, and there is an established technical infrastructure as well as standard processes to support them doing so. It is reasonable nowadays to expect the majority of publishers to belong to an ethics program like COPE, to assign unique and persistent identifiers like DOIs and to participate in long term archiving projects like CLOCKSS (Controlled LOCKSS). This is a not-for-profit joint venture between the academic publishers and research libraries with the ambition of developing a sustainable, geographically distributed dark archive to ensure the long-term survival of Web-based scholarly publications (www.clockss.org).

No such infrastructure or processes exist for grey literature. That is part of their appeal: you can upload a preprint or present a poster without having to go through a lengthy peer review, typesetting and publishing process, or publish a report without having to go through an intermediary. It is unfortunately also a hindrance to anybody trying to mine or analyze them. Analyzing them is exactly what altmetrics initiatives should be trying to do, because policy documents and patents are potentially very interesting indicators of impact beyond scholarly impact.

The opportunity

It's not hard to imagine some use cases illustrating why altmetrics groups might want to get a handle on the subject:

- If my research is on the economic impact of river flooding then citations in other journals aren't the only thing that's important to me. I want to be kept aware of government policy that cites my work, too.
- If my work is referenced by a patent in a completely different field, I'd like to know about it.
- When looking at research outputs of my department, I don't just care about peer-reviewed research in journals, but patents, reports and policy documents too.

Being cited as evidence in a government policy report isn't impact in and of itself - perhaps the report will be locked in a filing cabinet and never acted on. It is still a valuable indicator, though, that's not easily obtainable anywhere else. It's not unusual for even the authors of a paper to not know about everywhere that their work has turned up.

The challenges

Discovering what research the grey literature material cites is just one potential opportunity to enrich impact analysis, but the challenges are fairly formidable. We've spent a lot of time and effort on building up systems to track, parse and analyze policy documents and patents and some of the more interesting challenges we've faced are:

- Identifying relevant documents
- Extracting metadata & references
- Permanence

Let's look at each one in turn.

Identifying relevant documents

The first challenge to mining grey literature is simply to find it.

It is a publisher's job (at least traditionally) to disseminate research, and there is a well-established ecosystem of discovery tools and indexing services to help individuals find and access scholarly literature that is relevant to them.

There is no such ecosystem for the grey literature, though valuable initiatives like greylit.org can give researchers a head start. Without knowing even how much grey literature material is created each year, let alone by whom, it is difficult to make assumptions about how complete any index you may build is.

Extracting metadata & references

Once relevant documents are found, you ideally want to associate basic bibliographic metadata with them – a title, some authors, a publication date.

Central databases like CrossRef or PubMed can help do this for traditional literature, returning bibliographic records originally supplied by the publisher when queried by a unique identifier.

Policy documents, to take one example, have no such canonical metadata available and they have often been published online in ways that make automatic metadata extraction impractical. A government report may be provided only as a typeset PDF, with the title and authors (if mentioned at all) in a graphic on the first page.

For the purposes of altmetrics we are interested in the research that documents cite, and common practice in scholarly articles is to keep these to a single references section. There is often no such common practice for grey literature, where references can be in figure captions, in footnotes, tables, or separate appendices to name but a few common scenarios. Furthermore, without manual curation it is hard to figure out what's a citation at all in the traditional sense of the word: we have come across medical guidelines that explicitly list out papers that may have seemed relevant but were not used in any way.

Within:	Hours	Days	Months	Years
Activity seen:	Altmetrics: First mention on social media	Altmetrics: First pickups on blogs & in news outlets	Bibliometrics: First citations in the rest of the literature	Altmetrics: First appearances in policy documents

Table 1: The attention potentially paid to research on a continuum.

Permanence

A core principle shared by most altmetrics groups is that the raw data that any numbers or assertions are based on should be available to the end user.

So if we are to report that a particular policy document links to a paper then we need to make sure that users can get to that policy document.

This leads to a couple of classic online publishing problems: firstly will we always be able to find the document again in the place we found it and secondly will the document always be available online.

There is nothing to stop an NGO or government agency from redesigning its website, shifting its online publications to a different part of the site and breaking all our links. There is also no 'dark archive' of documents to ensure that we will always have a copy even if the group that originally created it ceases to exist.

How does the grey literature fit in with other altmetrics?

One oft-mentioned advantage of altmetrics indicators is that they are usually high volume and quick to accrue, with the first data being collected within hours of publication instead of months as is usually the case with citations.

Citations to papers from policy documents buck this trend, where, anecdotally, we have seen that most of the biomedical papers referenced are five or more years older than the policy document itself – this is even slower than you might expect from traditional literature.

It is possible to imagine the attention paid to at least biomedical research on a continuum (see Table 1).

Why might citations from policy documents only appear years after a paper is published? We don't know, though it would be interesting to find out. One possibility is that it takes a long time for some types of policy document or report to actually get published, so the citations are to papers that may have actually been relatively new when the authors were still discussing whatever issue the document is addressing.

How could we improve things?

The flexibility of grey literature is a strength but also a weakness. The grey literature lacks many of the important pieces of infrastructure and best practices used by academic publishers.

Might it be possible to pull over some of the good things from academic publishing workflows, without losing too many of the benefits of occasionally being able to opt out of scholarly publishing processes?

A few key changes to the way grey literature is produced would make life much easier for anybody interested in the altmetrics that they might provide, though these must be balanced with the needs of creators who may have little interest in metrics of any kind and so lack the motivation to support change.

Use of persistent identifiers

Use of something like the Handle System (in which resources are assigned a unique identifier that can be resolved to a URL by the creator) would help ensure that groups can track documents even if they move around the internet.

Minimum standards of metadata

The best way to add basic metadata to scholarly PDFs and web pages is a problem that publishers solved long ago. PRISM (Publisher Requirements for Industry Standard Metadata) is a publisher driven initiative to agree on a standard set of metadata for academic publications (see idea alliance for more detail). Dublin Core is a broad set of standard metadata terms that can be applied to documents, videos, images and other resources. They provide standard ontologies; in PDFs these can be inserted using authoring tools or, after creation, using XMP which is a standard way of adding metadata to images and PDFs. On web pages the publishing industry has settled on <meta> tags, not least because for many journals this is a prerequisite for indexing by Google Scholar.

An index of the grey literature

An open, central index of scholarly grey literature that enforced a minimum level of metadata for each item could make searching and linking documents much easier for tool makers and help the groups authoring them with discoverability (as users would have one place to look for relevant documents) and attribution.

An alternative would be to maintain a central index of grey literature repositories – the websites of each group authoring them, perhaps – and to allow harvesting from each through a standard like OAI-PMH (Open Archives Initiatives – Protocol for Metadata Harvesting), already well adopted by institutional repositories and open access publishers.

This would allow third parties to independently provide centralized tools to search or preserve content held on each group's website, making it easier to track and discover documents.

Conclusion

The grey literature presents great opportunities for alternative metrics, providing data and indicators that cannot be found anywhere else.

Those opportunities come with great challenges, both social and technical. To work with grey literature, tools need some basic infrastructure to be put in place, but is this something that authors really want or will it compromise the advantages of publishing grey literature in the first place?

References:

1. http://www.oxfordjournals.org/our_journals/molbev/general_author_guidelines.html#References

2. Auger, C.P., Ed. (1989) Information Sources in Grey Literature (2nd ed.). London: Bowker-Saur. ISBN 0862918715.