Jisc Open Metrics Lab

Open Metrics for Monographs Experiment: Final Report

Professor Martin Paul Eve

Birkbeck, University of London

This final report incorporates and extends Eve, Martin Paul, 'Jisc Open Metrics Lab - Open Metrics for Monographs: Background Contexts and Literature Review', 2019 <**http://repository.jisc.ac.uk/7427/**>

Table of Contents

Executive Summary

Metrics for open-access monographs inspire a range of anxieties among HSS researchers. This experiment attempted to build a project that would serve researchers using the citation graph from these texts, rather than using it to measure researcher productivity or success. We built a prototype tool that would intersect references between different open-access books but that also exposed more challenging aspects of such work.

This work was useful and important for researchers, who often need to get their head around a field as quickly as possible and wish to know which works are cited by most books in a discipline. Doing this manually, though, is an intensely time-consuming process. Our tool demonstrated that such a helpful tool could be possible, but that it – and other tools like it – require modifications to the machine-readability of open-access books.

The results of the experiment were:

- Aligning the linguistic expression of a citation with its underlying canonical citation object is a difficult computational task. The fact that references in books do not use Digital Object Identifiers (DOI)s or International Standard Book Numbers (ISBNs) to identify their target references hampered our work. Implementation of Functional Requirements for Bibliographic Records could help in this respect.
- The diversity of formats available in the Directory of Open Access Books (DOAB) means that we need to write custom parsers to extract texts in ways that make it harder to build a corpus for this type of work.
- The DOAB metadata could better signal the type of file endpoint that it makes available.
- Translations of works that are cited pose additional barriers to identifying the object that is cited.

We nonetheless managed to create a functional tool that is available at

https://github.com/BirkbeckCTP/jisc-doab.

Background Contexts

The Jisc Open Metrics Lab's Monograph Experiment takes place at a time of transition for the academic monograph in the United Kingdom. Amid debate over whether a "crisis of the academic monograph even exists", in the past few years there have nonetheless been several signals that open access for academic monographs is becoming a reality.[1] The 2015 Crossick report to HEFCE, for instance, noted that this was a time for experiment for these important media in the Humanities and Social Scientific disciplines,[2] a time in which "[o]utside the framework of any policies, funders should play a role in facilitating through pilots and the formulation of standards those developments that will help digital open access realise its potential for innovation in research communication, collaboration and practice".[3]

The true impetus, though, for a move towards open access for academic monographs was given in the 2018 announcement by the global coalition of funders, known as cOAlition S, that their uncompromising mandate for open access was to extend to monographs at some point in the near future, although the precise timeframe was left unspecified.[4] Given that Crossick's calls for experiment have

---

1 On the debate about the crisis in conventional monograph production, see Ronald Snijder, 'Measuring Monographs: A Quantitative Method to Assess Scientific Impact and Societal Relevance', *First Monday*, 18.5 (2013) <https://firstmonday.org/ojs/index.php/fm/article/view/4250> [accessed 12 May 2019]; Marilyn Deegan, 'Academic Book of the Future Project Report', 2017 <https://academicbookfuture.files.wordpress.com/2017/06/project-report_academic-book-of-the-future_deegan3.pdf>; Michael Jubb, 'Academic Books and Their Future', 2017 <https://academicbookfuture.files.wordpress.com/2017/06/academic-books-and-their-futures_jubb1.pdf>; For more on the definitions of open access, see Peter Suber, *Open Access*, Essential Knowledge Series (Cambridge, MA: MIT Press, 2012) <http://bit.ly/oa-book>; Martin Paul Eve, *Open Access and the Humanities: Contexts, Controversies and the Future* (Cambridge: Cambridge University Press, 2014) <https://doi.org/10.1017/CBO9781316161012>.

2 While it is true that monographs exist in the natural scientific spaces, and indeed many of the most important works in the history of the natural sciences have been published in book form, such as Darwin's *The Origin of the Species*, this report confines its remit to the humanities and social sciences. Indeed, the fact that the Wellcome Trust's open-access mandate includes monographs indicates that even a scientific funder with only a small cohort of medical humanities authors takes seriously this media form. Further, there have been recent attempts to appraise the bibliometrics of health monographs. See Pamela Royle and Norman Waugh, 'Bibliometrics of NIHR HTA Monographs and Their Related Journal Articles', *BMJ Open*, 5.2 (2015), e006595 <https://doi.org/10.1136/bmjopen-2014-006595>.

3 Geoffrey Crossick, 'Monographs and Open Access: A Report for the Higher Education Funding Council for England', *Higher Education Funding Council for England*, 2015, p. 68 <http://www.hefce.ac.uk/pubs/rereports/year/2015/monographs/> [accessed 24 May 2015].

4 cOAlition S, 'Plan S', *Plan S and COAlition S*, 2018 <https://www.coalition-s.org/> [accessed 12 May 2019]; Funder mandates have often driven the uptake of open access. For more on mandates, see Ulrich Herb,

barely begun, this has come as a shock to many in the humanities and social sciences as no single business model for open access has yet been developed (or may be desirable).

That said, there have been many advances since Crossick. Knowledge Unlatched continues to be the largest and most successful open-access monograph initiative,[5] facilitating the opening of hundreds of academic monographs.[6] The transfer of Knowledge Unlatched to a for-profit structure in 2018, however, has prompted some hand-wringing among libraries and criticisms from other presses around some of its activities.[7] Other smaller initiatives – often working under the banner of the ScholarLed coalition – have also shown early success though, particularly punctum books, Open Humanities Press, and Open Book Publishers. There has also been a rise of the "new" university press, specialising in open-access monographs, among which number UCL Press, Goldsmiths Press, Luminos Press, Lever Press, Calvary

---

'Recommendations, Statements, Declarations And Activities Of Science Policy Actors On Shaping The Scholarly Communication System', *Zenodo*, 2017 <https://doi.org/10.5281/zenodo.1003229>; David Sweeney and Ben Johnson, 'Seeking a Fresh Perspective: A Research Funder's View of Open Access', *Insights: The UKSG Journal*, 27.1 (2014), 51–57 <https://doi.org/10.1629/2048-7754.114>; José Carvalho and others, 'Monitoring a National Open Access Funder Mandate', *Procedia Computer Science*, 13th International Conference on Current Research Information Systems, CRIS2016, Communicating and Measuring Research Responsibly: Profiling, Metrics, Impact,Interoperability, 106 (2017), 283–90 <https://doi.org/10.1016/j.procs.2017.03.027>; Philippe Vincent-Lamarre, Jade Boivin, Yassine Gargouri, Vincent Larivière, and others, 'The Effect of Open Access Mandate Strength on Deposit Rate and Latency', 2014 <http://eprints.soton.ac.uk/366815/> [accessed 23 July 2014]; Philippe Vincent-Lamarre, Jade Boivin, Yassine Gargouri, Vincent Lariviere, and others, 'Estimating Open Access Mandate Effectiveness: The MELIBEA Score', *ArXiv:1410.2926 [Cs]*, 2014 <http://arxiv.org/abs/1410.2926> [accessed 12 May 2019]; Jingfeng Xia and others, 'A Review of Open Access Self-Archiving Mandate Policies', *Portal: Libraries and the Academy*, 12.1 (2012), 85–102; Alma Swan, 'Open Access Policy Effectiveness: A Briefing Paper for Research Institutions' (Pasteur4OA) <http://www.pasteur4oa.eu/sites/pasteur4oa/files/resource/Policy%20effectiveness%20-%20institutions%20final.pdf>; Vincent Larivière and Cassidy R. Sugimoto, 'Do Authors Comply When Funders Enforce Open Access to Research?', *Nature*, 562.7728 (2018), 483 <https://doi.org/10.1038/d41586-018-07101-w>; it can be difficult to see, though, how mandates will translate from the journal to the monographic space Martin Paul Eve and others, 'Cost Estimates of an Open Access Mandate for Monographs in the UK's Third Research Excellence Framework', *Insights*, 30.3 (2017) <https://doi.org/10.1629/uksg.392>.

5 In this report, "open access" is hyphenated as "open-access" when used as a prepositive adjective.

6 See Frances Pinter and Christopher Kenneally, 'Publishing Pioneer Seeks Knowledge Unlatched', 2013 <http://beyondthebookcast.com/transcripts/publishing-pioneer-seeks-knowledge-unlatched/>; Higher Education Funding Council for England, 'Knowledge Unlatched Pilot given HEFCE Backing', 2013 <https://www.hefce.ac.uk/news/newsarchive/2013/news85263.html> [accessed 21 December 2013]; Lucy Montgomery, 'Knowledge Unlatched:A Global Library Consortium Model for Funding Open Access Scholarly Books', *Cultural Science*, 7.2 (2014), 1–66; Knowledge Unlatched, 'How It Works', 2013 <http://www.knowledgeunlatched.org/about/how-it-works/> [accessed 5 December 2013].

7 Springer Nature, 'Open Research Library', *Open Research*, 2019 <https://www.springernature.com/gp/open-research/journals-books/books/orl> [accessed 26 May 2019].

Press, and many many others. (Notably, these new university presses are a diverse and heterogeneous grouping, with some operating out of the library on a budget close to zero, while others receive substantial budgetary subsidy. Some, also, such as Lever Press span multiple academic institutions.[8]) On the whole, the rise of open-access monographs appears set to continue.[9]

An important part of the debate around open-access monographs, though, has been usage, situated within a broader context of developing bibliometric indicators that are sensitive towards, and can work in, the humanities and social sciences.[10] Indeed, the Crossick report stressed that a "clear articulation of the opportunities and benefits of open access for monographs will be an essential component of policymaking in this area"; an articulation that can only be made when backed by evidence.[11] Some publishers, such as Springer-Nature, have already made moves in this direction, demonstrating and publicising a seven-fold increase in general usage among their open-access monographic titles, by various measures.[12] However, in addition to the regular complexities of citation and reference analysis, undertaking such analysis in the humanities and social sciences presents specific difficulties.[13] Not least of these is the issue of coverage of these disciplines' outputs within the conventional databases that are

8 For more on this theme, see Janneke Adema, Graham Stone, and Chris Keene, 'Changing Publishing Ecologies: A Landscape Study of New University Presses and Academic-Led Publishing' (Jisc, 2017).

9 Simba Information, 'Open Access Book Publishing 2016-2020', 2016 <https://www.simbainformation.com/Open-Access-Book-10410716/> [accessed 12 May 2019]; Eelco Ferwerda, Frances Pinter, and Niels Stern, *A Landscape Study On Open Access And Monographs: Policies, Funding And Publishing In Eight European Countries* (Zenodo, 1 August 2017) <https://doi.org/10.5281/zenodo.815932>.

10 Björn Hammarfelt, 'Beyond Coverage: Toward a Bibliometrics for the Humanities', in *Research Assessment in the Humanities: Towards Criteria and Procedures*, ed. by Michael Ochsner, Sven E. Hug, and Hans-Dieter Daniel (Cham: Springer International Publishing, 2016), pp. 115–31 <https://doi.org/10.1007/978-3-319-29016-4_10>.

11 Crossick, p. 68.

12 Christina Emery and others, 'The OA Effect: How Does Open Access Affect the Usage of Scholarly Books?' (Springer-Nature, 2017) <https://media.springernature.com/full/springer-cms/rest/v1/content/15176744/data/v3> [accessed 12 May 2019].

13 Cameron Neylon, 'The Complexities of Citation: How Theory Can Support Effective Policy and Implementation', 2016 <http://repository.jisc.ac.uk/6553/> [accessed 12 May 2019].

used for bibliometric analyses.[14] This is usually attributed, threefold, to “diverse publication channels, the importance of ‘local’ languages as well as the wide-ranging audience of research”.[15]

Yet there is also a problem of disciplinary definition at work here. When it is claimed, for instance, that books are more frequently cited in the humanities than journal articles (and vice versa in the social sciences), this is a generalization too far. For, in aggregating up to “the humanities” and “the social sciences”, this elides the fact that the citation of journals plays a central role in, for instance, history and linguistics, while sociology and library information sciences hold the monograph in high citation regard (although it is always worth noting that the lack of semantic value in citation metrics means that it is impossible to bestow a positive characteristic upon what is merely attention – a citation).[16] There are also, though, serious problems of obtaining accurate, centralized data for OA monographs, often by the publishers themselves.[17] For one, the permissive distribution clauses of the Creative Commons licenses – a feature, not a bug, of open-access dissemination in that copies can end up distributed in different locations – means that statistics must be aggregated and are unlikely, even where such figures are available, to be collected in standardised ways in all instances across multiple platforms (e.g. COUNTER compliance).[18]

---

14 See, for just a selection Jordi Ardanuy, ‘Sixty Years of Citation Analysis Studies in the Humanities (1951–2010)’, *Journal of the American Society for Information Science and Technology*, 64.8 (2013), 1751–55 <https://doi.org/10.1002/asi.22835>; Anton J. Nederhof, ‘Bibliometric Monitoring of Research Performance in the Social Sciences and the Humanities: A Review’, *Scientometrics*, 66.1 (2006), 81–100 <https://doi.org/10.1007/s11192-006-0007-2>; Maria Teresa Biagetti, Antonella Iacono, and Antonella Trombone, ‘Testing Library Catalog Analysis as a Bibliometric Indicator for Research Evaluation in Social Sciences and Humanities’, in *Challenges and Opportunities for Knowledge Organization in the Digital Age*, ed. by Fernanda Ribeiro and Maria Elisa Cerveira (Berlin: Ergon Verlag, 2018), pp. 892–99 <https://doi.org/10.5771/9783956504211-892>.

15 Hammarfelt, ‘Beyond Coverage’, p. 117.

16 Björn Hammarfelt, ‘Following the Footnotes: A Bibliometric Analysis of Citation Patterns in Literary Studies’ (unpublished Doctoral, Uppsala University, 2012), p. 31.

17 Charles Watkinson, Rebecca Welzenbach, and others, ‘Mapping the Free Ebook Supply Chain: Final Report to the Andrew W. Mellon Foundation’, 2017, p. 4 <https://deepblue.lib.umich.edu/handle/2027.42/137638>.

18 ‘Project COUNTER - Consistent, Credible, Comparable’, *Project Counter* <https://www.projectcounter.org/> [accessed 12 May 2019].

Existing OA Monograph Metric Initiatives

Following in these difficult methodological footsteps, there have been several projects that have, nonetheless, tried to gauge the impact and measure the changes to usage that open access has had on academic monographs. Of particular note are the OAPEN-NL and OAPEN-UK projects, which attempted to measure usage and sales figures for matching controlled sets of monographs.[19] The aforementioned Springer-Nature report has also attempted to provide a comparative measure of usage between the company's OA and non-OA books.[20] Knowledge Unlatched Research – the non-commercial research arm that sprang out of KU – has also conducted a comparative analysis of usage within the JSTOR ecosystem of the first four publishers to begin distributing their OA books through this channel.[21] This approach, as with the Springer-Nature study, has the advantage of isolating its analysis to one particular context, thereby avoiding the above noted problems of statistical aggregation. The difficulty, of course, is that such an analysis is more likely to favour OA books, as the non-availability of a title is less likely to lead to a download.

This highlights the important interrelationship between measuring "usage" (be this citations, references, or views and downloads) and understanding discoverability (how users come to find material). However, as Neylon *et al.* note, "[t]he question of visibility is [...] a complex one".[22] The problems that they identify for monographs – a print-centric discoverability system, intermediaries rather than direct reader interactions, lack of persistent identifier redirects, unexpected audience groups, poor quality assurance on

19 See OAPEN-UK, 'The Pilot', 2013 <http://oapen-uk.jiscebooks.org/pilot/> [accessed 25 March 2014]; Eelco Ferwerda, Ronald Snijder, and Janneke Adema, 'OAPEN-NL: A Project Exploring Open Access Monograph Publishing in the Netherlands Final Report', 2013 <http://www.oapen.nl/images/attachments/article/58/OAPEN-NL-final-report.pdf> [accessed 24 March 2014]; Janneke Adema, 'Overview of Open Access Models for Ebooks in the Humanities and Social Sciences', *OAPEN*, 2010 <https://curve.coventry.ac.uk/open/file/a976330e-ed7a-4bd5-b0ed-47cab90e9a5e/1/ademaoapen2comb.pdf> [accessed 12 August 2014].

20 Emery and others.

21 Lucy Montgomery and others, 'Exploring Usage of Open Access Books via the JSTOR Platform' (Knowledge Unlatched Research, 2017) <http://kuresearch.org/PDF/jstor_report.pdf>.

22 Cameron Neylon and others, 'The Visibility of Open Access Monographs in a European Context: Full Report' (Knowledge Unlatched Research, 2018), p. 7 <https://hcommons.org/deposits/objects/hc:18270/datastreams/CONTENT/content>.

data that is collected, small presses with little capacity for data collection, and inconsistent metadata – appear pervasive and will take many years to address.

Nonetheless, and despite the statistical problems encountered in the OAPEN-UK project, a subsequent and more recent OAPEN-CH project in Switzerland has managed to find some statistically significant differences between OA and non-OA books.[23] Namely that:

- "Open access had a statistically significant positive influence on the trackability and visibility of the monographs"
- "Placing open access monographs in the OAPEN Library increased international reach"
- "Open access had a statistically significant influence on the use of monographs (number of book visits, page views and downloads). Monographs in the experimental group were used more frequently than books in the control group."
- "Statistically, open access did not have a negative influence on the sales figures for printed books. The average number of monographs sold in the experimental group was only negligibly lower than the number in the control group. In fact, more copies overall were sold in the experimental group. However, the reverse conclusion – open access has a positive impact on sales figures – does not hold statistically either since there were hardly any differences between the two groups."

These findings are clearly of interest to those piloting business models for open-access monographs. Amid existing debates over whether the monograph is sustainable, the knowledge that OA appears not to have damaged sales figures is a potentially heartening finding, although there are multiple explanations for why this may be the case (poor discoverability of OA editions, unawareness of OA editions etc.). That said,

23 Eelco Ferwerda and others, *OAPEN-CH – The Impact Of Open Access On Scientific Monographs In Switzerland. A Project Conducted By The Swiss National Science Foundation (SNSF)* (Zenodo, 23 April 2018) <https://doi.org/10.5281/zenodo.1220607>.

there are also convincing rationales for how this finding should have come about (people favour reading in print, libraries buying print to support OA etc.).

The HIRMEOS project (High Integration of Research Monographs in the European Open Science Infrastructure) also has a work package devoted to metrics and monographs. Led by Ubiquity Press, this has resulted in the development of a metrics standard that includes DOI scraping, altmetrics (attention scores), and geolocation data on readers.[24] This project also convened a workshop in Paris on metrics for open-access monographs.[25] Of note, perhaps, here and stemming from the phrasing of the HIRMEOS workshop is the ambiguity in the term "open metrics", and whether this refers to metrics that are, themselves, open, or metrics pertaining to research objects that are open access.

Thinking more around the values of metrics and the ways in which such facilities are often abused in the research evaluation process, the Stateside Humane Metrics Initiative has turned its attention to developing "an initiative for rethinking humane indicators of excellence in academia, focused particularly on the humanities and social sciences (HSS)". These centre around collegiality, quality, equity, openness, and community.[26] That said, the very idea that metrics for scholarship *can even be humane* has been contested and opened for debate.[27]

Finally, there is at least one other major project in train that aims to investigate the usage of ebooks more broadly – and understanding that many of the above problems are not just specific to OA books, but

24 HIRMEOS and Ubiquity Press, 'Deliverable D6.1: Metrics Services Specification', *High Integration of Research Monographs in the European Open Science Infrastructure*, 2019 <https://www.hirmeos.eu/wp-content/uploads/2017/11/HI61-Metrics_Service_technical_specification-final.pdf> [accessed 3 June 2019].

25 HIRMEOS, 'HIRMEOS Workshops on Annotation and Metrics for OA Monographs, 10-11 Jan 2019, Paris', *High Integration of Research Monographs in the European Open Science Infrastructure*, 2019 <https://www.hirmeos.eu/2018/11/05/hirmeos-workshops-on-annotation-and-metrics-for-oa-monographs-10-11jan-2019-paris/> [accessed 3 June 2019].

26 HuMetricsHSS, 'About HuMetricsHSS', *Humane Metrics Initiative*, 2018 <http://webcache.googleusercontent.com/search?q=cache:A8FHPfPBuC8J:humetricshss.org/about/+&cd=1&hl=en&ct=clnk&gl=uk> [accessed 3 June 2019].

27 Stacy Konkiel, 'Approaches to Creating "Humane" Research Evaluation Metrics for the Humanities', *Insights the UKSG Journal*, 31 (2018) <https://doi.org/10.1629/uksg.445>; Martina Franzen, Eileen Joy, and Chris Long, *Humane Metrics/Metrics Noir* (Coventry UK: Post Office Press / meson press, 2018) <https://hcommons.org/deposits/item/hc:19823/> [accessed 30 May 2019].

simply digital books in general remains key. The Book Industry Study Group, working with Knowledge Unlatched Research, is currently undertaking a study, funded by the Andrew W. Mellon Foundation, that aims to convene "a structured community conversation around usage tracking for OA ebooks".[28] Using both conventional (that is, quantitative citation) metrics and altmetrics as its basis, this project offers a potentially promising route to addressing the systemic problems with metrics for monographs, and thereby yielding convincing rationales for any transition to open access.

The Jisc Open Metrics Lab Monograph Experiment in Context

Bibliometrics for monographs – and open-access monographs – remain extremely difficult to do well. There are a range of basic issues and problems when profiling books that are simply not as far along as they are in the world of journals. This is, in part, due to the lateness of books to come to the digital format compared to journals. That said, of course, coming late to a field also confers concomitant advantages in that one can learn from the difficulties faced in other contexts.

One of the overarching challenges, though, remains the fact that bibliometrics are inextricably associated with research assessment. In the eyes of many humanities and social scientific researchers, the only use to which the development of accurate citation metrics for books could be put is to develop ever-more coercive evaluation procedures, which is an undesirable outcome for most academics. Coupled with the extremely long citation half life in many HSS disciplines,[29] the lack of a convincing rationale for citation metrics for monographs (beyond evaluation) is hindering the uptake of open access.

---

28 Charles Watkinson, Kevin Hawkins, and others, 'Understanding OA Ebook Usage: Toward a Common Framework' (Knowledge Unlatched Research, 2018), p. 2 <https://deepblue.lib.umich.edu/bitstream/handle/2027.42/143840/Redacted%20Grant%20Narrative%20-%20OA%20Ebook%20Usage_FINAL%20SUBMISSION_042718.pdf?sequence=1&isAllowed=y> [accessed 13 May 2019].

29 See Nigel Vincent and Chris Wickham, 'Debating Open Access: Introduction', in *Debating Open Access*, ed. by Nigel Vincent and Chris Wickham (London: British Academy, 2013), pp. 4–12 and the accompanying volume.

It is within this context that the Jisc Open Metrics Lab Open Monographs Experiment places itself. For, if those in the humanities and social sciences do not want bibliometrics to be used for assessment, they are all actually already used to using the citation graph in another type of utilitarian exercise: cross-referencing in order to gain an understanding of a field. As outlined in the launch blog post for this project, one discovery technique used at present is to travel to a national deposit library and order ten or so books that appear to have pertinent titles. The researcher can then cross-reference the bibliographies of these books in order to ascertain what they cite in common. This allows the researcher to quickly understand a new field: the most-cited items in common will be good pieces to read in order to rapidly understand a new disciplinary space.

This is a labour intensive process. It involves the move to a physical space in the first place – a physical research library – which on its own has implications for accessibility for those with mobility conditions or long-term health problems. This is then followed by a search of the catalogue, a wait for the delivery of the items, and then a laborious process of note taking, observation and cross-referencing across hundreds of permutations of bibliographic entries.

For the experiment that we are undertaking, we decided to implement a digital system to perform this task using open-access monographs. For what if, in the contemporary digital publishing landscape, there were a better way than this manual searching? The project has three components:

1. This literature review of existing material on bibliometrics for open-access monographs and bibliographic intersection tools;
2. A tool that will allow people to download a corpus from the DOAB;
3. A tool that will parse references from open-access monographs and tell the user which items are cited in common among the selected titles.

There are existing tools in this space, but none, so far as we know, for monographs. The most well-known of these is CitationGecko, which acts as a visualization aide for CrossRef's repository of

interlinked citations. This relies on the publisher having deposited semantically rich citation metadata with CrossRef, which we suspect many open-access book publishers are not doing. There are also attempts at referencing mining that are variable in their success rates. For instance CERMINE uses a visual approach to PDF parsing to attempt to identify references and to parse them into uniquely identifiable data objects in the JATS XML format. Unfortunately, as with many PDF parsing solutions, there are serious problems with generically identifying the visual styling of citations and our initial attempts indicated that line breaks between entries in a test corpus of Cambridge University Press books caused serious problems. The anycite.io parser, written in Ruby, suffered from a similar problem of distinguishing references from one another. However, this latter parser also has a mode in which, if one can pass it clean, single-line, single reference plaintext, it has a high success rate for parsing the result. This makes it a viable option if one can parse the books into, at the very least, individuated references.

There are a range of strategies that we will deploy in order to convert from the free-text referencing of OA books to semantically rich and uniquely identifiable data objects that can be cross-compared. Further, the format delivery of the identifier link in the DOAB OAI feed is not consistent. Instead, when following through the OA edition of a book, we will have to be able to detect the passed format (PDF/epub/landing page with options/Dropbox-encapsulated PDF) via MIME type and select appropriately for the targeted publishers in this iteration. Existing approaches of examining the PDF files has clearly failed on projects with far more resource than this. Instead, we will use the following approaches:

- epub versions often have semantically rich TEI or HTML markup within them. Where we are presented with an epub option, we will attempt to parse this first. Many SciELO books adopt this format.
- In cases where the OAI identifier redirects the client to a URL beginning with oapen.org/view, we have found that often we can obtain semantically rich data from this source, which uses an XSLT

transformation to render references in a standard form. This appears to be the format used by Palgrave Macmillan, thereby opening a substantial catalogue of OA books to this technique.

- In some cases, publisher platforms provide semantically rich formatted HTML versions of referencing systems. For instance, in the case of Cambridge University Press, a landing page template of **https://doi.org/10.1017/CBO**<ISBN-13> yields semantically rich formatting of references, broken down into constituent components (e.g. author names, titles etc.).

In this way, we hope to demonstrate value to humanities and social science researchers in being able to parse an open citation graph that goes beyond merely developing bibliometrics for research assessment. Certainly, the task remains difficult for all the reasons above. But we believe that this tool could truly be of use to researchers in these disciplines and can then be used as further evidence of the value of opening access to monographs.

Experimental Software Architecture

The software that we built as part of this experiment is written in Python 3 (specifically, Python 3.7 and above in order to use f-strings) and uses a number of third-party open-source libraries, namely:

- beautifulsoup4==4.7.1
- bibtexparser==1.1.0
- bs4==0.0.1
- certifi==2019.6.16
- chardet==3.0.4
- crossrefapi
- docopt
- doi2bib
- EbookLib==0.17.1
- future==0.17.1
- idna==2.8
- lxml==4.3.4
- pyparsing==2.4.0
- requests==2.22.0
- Sickle==0.6.4
- six==1.12.0
- soupsieve==1.9.2
- SQLAlchemy==1.3.5
- Unidecode==1.1.1
- urllib3==1.25.3

For backend storage, the software uses the postgres database with the pg_trgm (trigram) module enabled. A makefile is included that will build a containerised Docker environment for convenience. If using the Docker install, the command-line interface (CLI) can be invoked by using the makefile argument “doab-cli” and then providing the desired command via the CMD variable. For instance: make

doab-cli CMD="extract_texts --publisher_id=1131 -d". At the time of writing in August 2019 the software was tested and developed on Ubuntu Linux 18.04.02, the only officially supported platform.

The entrypoint and command-line parsing are handled by docopt, which allows for invocation created automatically from Python documentation strings. Current invocation options are:

```
  cli.py extract_texts [--output_path=PATH] [--publisher_id=ID] [--threads=THREAD] [options]
  cli.py import_metadata [--input_path=PATH] [--threads=THREAD] [--book_id=BOOK_IDS...] [options]
  cli.py parse_references [--input_path=PATH] [--threads=THREAD] [--book_id=BOOK_IDS...] [--dry-run]
[options]
  cli.py match_reference <reference> [--parser=PARSER] [--input_path=PATH] [options]
  cli.py list_citations [--book_id=BOOK_IDS...] [options]
  cli.py list_books [--input_path=PATH] [options]
  cli.py list_publishers [options]
  cli.py list_parsers [options]
  cli.py nuke_citations [--book_id=BOOK_IDS...] [options]
  cli.py nuke_intersections [--book_id=BOOK_IDS...] [options]
  cli.py intersect [options] [-n --dry-run] [--book_id=BOOK_IDS...]
  cli.py list_intersections [options]
  cli.py list_references <book_id> [options]
```

These commands represent the atomic operations that are involved in creating an intersection between texts. The process is as follows:

1. Extract texts. This retrieves the book files for a particular publisher, specified using the --publisher-id parameter. Extractors, which can be found in the corpus_extractors.py can specify whether they can handle a particular publisher, URL, or any other feature by returning True or False in the overridden validate_identifier function. Some generic extractors, such as the PDF extractor and the JSONMetadataExtractor attempt to operate on all publishers. See the below section on “corpus extraction” for more on this.

2. Import metadata. This parses the metadata extracted by the JSONMetadataExtractor and stores it in the database for each book. This enables us to display friendly titles instead of DOAB Book IDs.
3. Parse references. This is an intensive process that attempts to parse structured citations from extracted book sources. Different extractors adopt different approaches to this computationally difficult task. These range from visual machine-learning approaches such as Cermine through to Bibtex parsing, DOI parsing, and structured XML/XHTML/HTML parsers. The approaches are detailed further below in the “reference parsing” section.
4. Match references. This takes an input reference string, runs it through the specified parser to produce a structured representation and then attempts to return books that contain the same citation. The acceptance test suite uses this method to verify that the matcher is working; see “acceptance test suite”, below. The way in which matching is handled is set out below in “reference matching”.
5. Intersect. This matches references to one another and then returns books that contain the matching references.
6. List intersections. This shows the final results of the intersection matching.
7. Other commands are for debugging and database inspection. For instance, list_publishers will show the available publishes, while list_citations will show the parsed citations for a specified set of books.

The sections below delve further into the technical complexities of each of these tasks.

Corpus Extraction

In order to determine which works are cited within an open-access book, it is first necessary to gain a copy of the files that constitute the OA book. However, this is not a straightforward matter. Not every OA book is provided in the same format, even within the DOAB records. Formats that we found varied from epub, through PDF, to structured XHTML and XML.

Further the DOAB link to read an open-access book gives no metadata about the type of content to which it points. This means that clicking this link can result, as just a few examples, in the direct download of a PDF, a direct link to the HTML display of an article, or a landing page presenting a series of formats. It is, therefore, impossible for a single generic parser to comprehensively collect/download DOAB-listed OA books in a computationally usable fashion without publisher-specific logic. That said, even within a single publisher it was often the case that DOAB links would exhibit different logic. For instance, Bloomsbury Academic, at the time of writing, presents a PDF for Jeremiah W. Cataldo's *Biblical Terror*.[30] However, the vast majority of this publisher's titles redirect to the Bloomsbury Collections website and present HTML versions of the text. It may be that funding conditions – such as those imposed by Knowledge Unlatched – have format requirements here that has led to this discrepancy. However, this nonetheless means that one cannot make assumptions as to the content type to which a DOAB remote link will resolve.

In order to handle this discrepancy, we created several distinct extractors. Some of these are high-level extractors: the PDF extractor, for instance, looks at the MIME type of the DOAB endpoint and, if it finds "application/pdf", stores the PDF in the book folder. Likewise, the JSONMetadataExtractor is an attempt to generalise DOAB metadata and to marshal this into a standardised JSON format. In many

30 Jeremiah W. Cataldo, *Biblical Terror: Why Law and Restoration in the Bible Depend upon Fear* (London: Bloomsbury Academic, 2017). The DOAB URL that resolves to a PDF is https://www.doabooks.org/doab?func=fulltext&uiLanguage=en&rid=32455.

cases, though, we had to write custom extractors that understood how to parse specific publisher websites. A good example of this is, again, the Bloomsbury Academic extractor.

Where a Bloomsbury Academic text returns a PDF, the PDF extractor will handle this. In cases where it redirects to the Bloomsbury Collections website, however, the Bloomsbury extractor will handle this instead. The Bloomsbury extractor registers itself as capable by a match on the publisher name inside the overridden validate_identifier method. The extractor then works through all HTML anchor links on the resolving page that point to a URL starting with /books. This gives a complete list of chapter links for the title in question. Regular expressions are used here to limit the links to the current book. These chapter HTML links are then fetched and stored in the book's folder.

By contrast, consider the Cambridge University Press (CUP) parser. Again, some CUP books are returned as PDFs. In these instances the PDF extractor handles them. In others, though, the books redirect to the Cambridge Core platform. This is easier to handle than Bloomsbury's setup, though, as the page to which CUP redirects contains the full-text of the book immediately on hit. In these cases, the extractor simply stores the HTML page as "cc.html" inside the book's folder.

Nonetheless, just these two instances demonstrate the difficulties that are faced merely in extracting open-access books for computational parsing. The lack of any standard format or even any metadata description of the format that is expected at a link means that the act of downloading OA books must be customized in each instance. This is a very labour-intensive process and one that makes it difficult to see how to scale experiments such as this. Further, when one writes custom extractors, one is at the mercy of future platform changes that will cause breakage to these processes. Ideally, publishers would implement a standard API to allow for the queryability of the availability of OA books that includes format specific metadata and a retrieval mechanism. This would avoid the difficulties of implementing brittle web scraping downloaders/extractors.

Reference Parsing

Reference parsing is, as noted above, a computationally difficult task if working from unstructured text. In addition to the ambiguity of the placement of different markers within a reference string ("is this a title, an author, or a publisher name?") there is also the fundamental challenge that different forms of reference resolve to the same canonical reference object. That is, "M. Foucault, *Discipline and Punish*" refers, despite the different linguistic expression, to the same object as "Michel Foucault, *Discipline and Punish: The Birth of the Prison* (London: Penguin, 1992)", at least for the purpose of this experiment. This poses two difficulties in a computational setting: the first of natural language processing (to resolve the reference into its constituent components) and the second of distance resolution (to determine whether two objects are the same, despite linguistic differences).

There is, further to this, the additional problem of determining whether a block of text is even a citation. That is, in order to know whether to parse a block of text as a citation, it must be identified as such. To demonstrate why this is problematic, consider whether there are definable and generalizable linguistic differences between the clauses "Michel Foucault, Discipline and Punish: The Birth of the Prison (London: Penguin, 1992)" and the somewhat ungrammatical but nonetheless plausible "Some of Martin Amis's novels are bad: this is not universal (see London Fields, 1989)". One might point out the individual differences here, but they are, on the whole, not generalizable.

Certain types of document provide helpful structured contexts for this work. The XML BITS (Book Interchange Tag Suite) format, for instance, provides markup tags that denote references and the sub-components within each reference (e.g. authors and so forth). This solves two of the three problems: identifying whether text is a reference and identifying the constituent components, thereby leaving only the issue of resolving the distance to the canonical reference item. Likewise, many publishers provide XHTML markup, "under the hood" of their web-page display, that includes these semantic contexts (most likely because they are using an XSL transformation from BITS to generate the XHTML). While various natural language parsers attempt to process arbitrary free-text input to solve these problems, the results

of using the structured inputs was, we felt, likely to be far superior. We also had to contend with different reference styles, with some texts having a bibliography while others had only in-text citations (footnotes). In this case, we opted to parse only bibliographies for ease of reference.

For this reason, we attempted to write extractors, as above, that would favour the retrieval of structured content that we could then parse on a publisher-by-publisher basis. All reference parsers are expected to return a series of dictionary structures with the following keys (although any key may be missing or blank):

- author
- title
- pages
- journal
- volume
- doi
- year

Parsers can register their ability to handle a specific book by specifying the publisher name and the filetypes that they can handle. These constraints are then checked by the can_handle function in the PublisherSpecificMixin class, from which all publisher-specific parsers should derive.

The ingest method for publisher-specific parsers varies hugely according to the available data. The Cambridge University Press parser, for instance, extracts the JSON representation of the citation data from a javascript variable in their web page and then processes this. The Bloomsbury Academic parser works directly from the XHTML, which is marked up with “class” attributes on “span” tags that indicate to which part-of-citation is being referred. These methods produce reasonably reliable structured data.

The best structured results, however, are obtained when we can detect a Crossref DOI within the reference string. In this instance, we make a well-behaved (etiquette-compliant) call to the Crossref API to retrieve precise and unambiguous structured metadata about the citation. This works particularly well

for our use case as the data that we store are identical between documents (i.e. if document A and document B both cite the same DOI, we will retrieve and store identical data for them both, regardless of the citation's linguistic formatting and manifestation). For the purposes of this experiment, our DOI lookup engine is restricted to Crossref DOIs as a reasonable limitation for scholarly artefacts, but as data citation becomes more common and other registries acquire more prominence, a generalised DOI parser would be desirable.

Of note in our parsing engine is that the hierarchical object-oriented structure allows for multiple parsers to register to parse a single reference string. This results in multiple parsed citation instances for a single reference. Hence, in cases where, say Cambridge University Press, passed structure data and a DOI, we store both the result of the CUP parser output and the DOI parser output. This requires a mechanism, then, to determine the weight to assign to each of these parses. We have, as a result, built an accuracy flag into the parsing engine (used, for instance, by the __str__ representation of the ParsedReference class in models.py). At present, these values are manually assigned based on our anecdotal experience of using the parsers. A more thorough approach here would be to determine the optimal values by calculating the loss function across the corpus.

Our core observations and recommendations on the reference parsing process can be summarised as follows:

- We encountered three core problems in computational reference parsing and matching:

    1. Identification of text as a citation
    2. Identification of the parts-of-citation within each citation
    3. Identification of the canonical reference object to which a citation refers

- Current unsupervised parsing software is not good at any of these tasks. Accurate results are best achieved by writing publisher-specific parsers that use structured metadata extraction. The absolute best results are achieved when a central lookup authority, such as Crossref, returns canonical exact match data for a citation.

- A scoring system for the accuracy of classifiers could be devised that would be better than the *ad hoc* metrics that we have anecdotally derived for this experiment.
- The diverse range of file formats, presentation styles, and lack of signalling metadata about the delivery of OA monographs makes it extremely labour intensive to produce an accurate citation network of these media, despite their ready availability.
- Ideally there would be a proper implementation of Functional Requirements for Bibliographic Records: **https://en.wikipedia.org/wiki/Functional_Requirements_for_Bibliographic_Records**. This would allow us to distinguish the highest-level book records from specific editions and to treat references to them as the same.

Reference Matching and Intersecting

Once we have parsed the references into their constituent components, the next step is to begin to match the references to one another. The first and most obvious match that we attempt is to see if multiple parsed references exist that have the same DOI. In these instances, we can record this as a direct match.

In other instances, we had to think more laterally about the commonality of references. The Postgres database, however, has a useful module called pg_trgm which allows for a “distance” similarity figure to be obtained for records using trigrams. A trigram is a set of three consecutive characters that are the same. For instance, the words “them” and “theta” share a trigram, “the”. The Postgres trigram module takes two records and returns a “distance” between the two based on the number of shared trigrams, normalised to fall within the range 0→ 1 (see: trgm_op.c in the Postgres source code). Although this operation can work using indexes, it is nonetheless relatively CPU intensive.

The basic procedure to match a single reference to books is:

1. Parse the input reference (we allow this via the match_reference CLI option, which also permits the user to select a suitable parser).
2. Work sequentially through the ingested corpus looking for matches.

This process allows us to return a list of books for a single reference. It is also the basis on which the acceptance tests (see below) work.

If we want to extend this to return an *intersection* – that is, to build a list of references shared in common by a set of books – then the process is as follows:

1. For each book *n* in the corpus: iterate over the references (*r*) in *n*
2. Parse *r*
3. Work sequentially through every other book in the ingested corpus looking for matches (*m*) of *r*.

4. If a match is found, add *m* to the result set and increment its count, provided that the parent book *n* is not already in the result set.

As is clear from even this example, with more than 100 references per book, this quickly becomes a complex and lengthy network graph to traverse.

The strength of the intersection tool is heavily dependent on the quality of the input data. If there are clear errors in the input data, such as journal names being mis-parsed into titles, then the similarity between entries will be misjudged and, in this example, all citations to articles *in a journal* would be returned. The match and intersect tool, therefore, suffers from the same weaknesses as the parsing command, above.

Our core observations and recommendations on the intersection process can be summarised as follows:

1. Intersecting references is a computationally intensive task. When performing this across hundreds of books it can take upwards of 24 hours on even the most recent processors. Of course, the goal of the tool is not to run on hundreds of books, but on a user-specified subset.
2. Parallelising this task is harder than other elements in our pipeline. The choice of Python as a language means that any multi-tasking or multi-threading has to contend with the GIL (Global Interpreter Lock). Because we need to query and *write* to the database simultaneously in this process, the interdependence between threads is high. This compounds the speed problems above, although it is not an insurmountable problem, merely outside of the current experiment's remit.
3. The quality of the input data greatly affects the output quality of the intersect tool. If the problems listed above cannot be resolved, then the intersect tool is of limited final use.

A test case of the reference parser working can be seen when the Palgrave and Cambridge University Press corpuses are pulled down (as of September 2019).

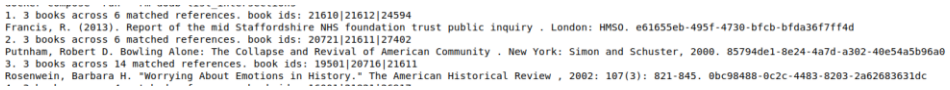

Figure 1: The first three results of list_intersectionsThe second of these cases demonstrates the flexibility that we had to adopt:

The second of these cases demonstrates the flexibility that we had to adopt:

> 2. 3 books across 6 matched references. book ids: 20721|21611|27402
>
> Putnham, Robert D. Bowling Alone: The Collapse and Revival of American Community. New York: Simon and Schuster, 2000. 85794de1-8e24-4a7d-a302-40e54a5b96a0

These three books are:

- Shepherd, Dean A., and Holger Patzelt. *Trailblazing in Entrepreneurship: Creating New Paths for Understanding the Field*. Palgrave Macmillan, 2017.
- Syvertsen, Trine. *Media Resistance: Protest, Dislike, Abstention*. Palgrave Macmillan, 2017.
- Enjolras, B., et al. *The Third Sector as a Renewable Resource for Europe: Concepts, Impacts, Challenges and Opportunities*. Palgrave Macmillan, 2019.

The citations appear, respectively, in these books as:

- Putnam, R. D. (2001). *Bowling alone: The collapse and revival of American community*. New York: Simon and Schuster
- Putnham, Robert D. *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon and Schuster, 2000.
- Putnam, R. D. (2000). *Bowling alone*. New York: Touchstone.

Note that there are two different spellings of Putnam used here, with Syvertsen erroneously giving ‘Putnham’. Also note that two of the texts cited are to the Simon and Schuster edition, but they use different dates from one another. Finally, the third citation we here detected is from a different publisher

and does not use the subtitle. These are the kinds of complexity with which we have had to deal when matching references to one another.

Acceptance Test Suite

In order to ensure that the parser was working, we developed an acceptance test suite that checks that references are returned for the correct books. These are stored in tests/acceptance_tests.py.

An example test entry might read:

```
@TestManager.register
class PalgraveAcceptanceTestA(AcceptanceTest):
    CITATION = "Foucault, M. (1991). Discipline and Punish. The Birth of the Prison St. Ives: Penguin"
    BOOK_IDS = {"24596", "20716", "27401"}
```

This specifies that the books with IDs 24596, 20716, and 27401 should match for *Discipline and Punish*.

The acceptance test suite can be run by using the makefile's "make check" command.

Outcomes

Formal Project Criteria

This section details the output fulfilment as set out in the Jisc Open Metrics Labs Open Monographs Experiment Contract.

- Output 1, a blog post, was published in May 2019: **https://openmetrics.jiscinvolve.org/wp/2019/05/open-access-monographs-and-metrics-more-than-counting-beans/**
- Output 2, a backgrounds advisory report and literature review, was published in June 2019: **http://repository.jisc.ac.uk/7427/**
- Output 3, the corpus creator, is fulfilled by the software output.
- Output 4, a computational test suite and tool for reference parsing that can handle styles from 5 books across 5 publishers is fulfilled by the software output. We have exceeded the number of parseable books and publishers by at least tenfold. For instance, the Bloomsbury Academic parser alone handles more than 150 books. The publishers handled by the tool are:
    - Accademia University Press
    - Alpara
    - Bloomsbury Academic
    - Cambridge University Press
    - Carrières Sociales Editions
    - Casa de Velázquez
    - CEDEJ
    - CEFAS
    - Central European University Press

- Centre français des études éthiopiennes
- Centre Jacques
- Centro de estudios mexicanos y centroamericanos
- Centro de Estudos Internacionais
- C.H.Beck
- CIDEHUS
- CNRS Éditions
- Collège de France
- Éditions Contrechamps
- Éditions de la Bibliothèque nationale de France
- Éditions de la Bibliothèque publique d’information
- Éditions de la Sorbonne
- Éditions de l’École des hautes études en sciences sociales
- Éditions de l’IHEAL
- Edizioni Kaplan
- ENS Éditions
- Graduate Institute Publications
- Innsbruck university press
- Institut de la gestion publique et du développement économique

- "Institut de recherches et détudes sur le monde arabe et musulman
- Institut de recherche sur le Maghreb contemporain
- Institut français de recherche en Afrique
- Institut français d’études anatoliennes
- Institut français d’études andines
- IRD Éditions
- Ledizioni
- MOM Éditions
- OpenEdition Press
- Palgrave Macmillan
- Presses de l’Ifpo
- Presses de l’Inalco
- "Presses de lUniversité de Montréal"
- Presses des Mines
- Presses Sorbonne Nouvelle
- "Presses universitaires dAix
- Presses universitaires de Caen
- Presses universitaires de la Méditerranée
- Presses universitaires de Liège

    - Presses universitaires de Louvain
    - Presses universitaires de Paris Nanterre
    - Presses universitaires de Provence
    - Presses universitaires de Rennes
    - Presses universitaires du Septentrion
    - Presses universitaires François
    - Publications du Centre Jean Bérard
    - Rosenberg & Sellier
    - Siglo del Hombre Editores
    - Société des Océanistes
    - UGA Éditions
    - Universidad Externado de Colombia
- Output 5, the intersect tool, is fulfilled by the software output.
- This report constitutes delivery of output 6.
- The presentation of this work at a community workshop is currently scheduled for the 14th October 2019.

Challenge: Consumer-Side Metadata Repositories

One of the major challenges that we faced in this project was in aligning the linguistic expression of a citation with its underlying canonical citation object. This was greatly eased when the artefact in question had a DOI. However, the fact that the current Document Object Identifier (DOI) system is a supplier-side, push mechanism means that it will never be possible or likely for all cited objects to have a DOI. Consumers of citation data, therefore, are at the mercy of content creators to register their metadata. In addition, this comes with preservation and access requirements – the PILA agreement – that are not necessarily suitable or realistic for all types of content, given that scholarly work can cite arbitrary grey literature.

One solution that could be envisaged is a crowdsourced independent metadata repository that can be linked to DOIs but that provides canonical metadata resolution functionality for cited artefacts. Good examples of such services, in another domain, are the MetaBrainz services, such as MusicBrainz. These provide “consumer”-controlled – crowdsourced – metadata records for music releases. These are queryable via an API and addressable via a canonical URL. One could imagine the same for cited objects, although an infinitely extensible open metadata schema for any kind of object (remember, data=stuff) is ambitious (impossible), to say the least. This could be linked to and defer to a DOI if one were subsequently issued.

The challenge, though, is that this mistakes the DOI architecture for a lookup service. It provides this functionality, for sure, on a producer-side push basis. But what Crossref and other registries do with DOIs are actually more social than technical. DOIs are a contract between the content producer and the registry to keep the work available and preserved and to keep the DOI resolving, even in the event of organisational failure. In the imagined environment above, there is no compact to hold the metadata to standard, no body who would have any responsibility to maintain a set of canonical identifiers that could permanently find their way into the scholarly record. In other words, it reinvents the *function* of DOIs as a technical, rather than social, matter, and neglects the *compact*.

There is also the issue of sustainability. Crossref and other DOI registries have to maintain a vast computational infrastructure that processes enormous quantities of data, on both a deposit and query basis. Breaking the system, even for a few minutes, is out of the question and can have dire consequences. This requires an onboard staff of highly competent technicians and astute critical thinkers, often converging in the same person (see: Geoff Bilder). To sustain an organisation like this requires a business model. Crossref's business model is membership based; an annual tiered membership fee based on size of organisational turnover from publishing and a (very small) fee per DOI deposit.

In the case of the envisaged repository mentioned above, who would the members be? Certainly one could envisage a coalition of libraries supporting such infrastructural provision through initiatives such as SCOSS (the global Sustainability Coalition for Open Science Services). But, there is consensus in the DOI model among publisher members that DOIs are useful and that the system should receive ongoing infrastructural support. The same cannot be said for a new consumer-controlled central metadata service.

All in all, then, the recommendation on this front from this experiment is not that such a consumer-controlled metadata service should be created – it is likely impossible to get this to work robustly in a useful way that has consensus – but to echo other recommendations from elsewhere: assign DOIs to granular scholarly objects (books, book chapters etc.) and ensure that whenever an artefact is cited, the DOI is included. This means that publisher processes for finding and inserting DOIs need to be robust; authors are incredibly unlikely to change their practices to include DOIs across the board. Given that typesetting (PDF, XML etc.) is often outsourced, this can be a tricky quality-assurance issue. Further to this, other identifier systems could be integrated into citations. ORCID IDs, for instance, could be included alongside author names to allow for robust lookup. (If it is undesirable to have such identifiers in the version for human display, then at least some form of structured representation that

includes this would be helpful.) Other databases, such as GRID, could be used to indicate institutions where these exist (University Presses, for example).

Challenge: Format Diversity

One of the other challenges that we faced was that different publishers provide their books in different formats with very little standardisation. In some ways, this contributes to a diverse and healthy book publishing infrastructure, in which the multitude of forms are part of an experimental ethos. After all, we do not want to stifle innovation through format constriction.

However, the computational parsing of references is a difficult task if one has only unstructured free text. Further, even where we could extract references themselves, often the markup contained bad tags or was simply the free text inside a tag marker that denoted a bibliography.

Without the standardisation of references and output formats, we will remain dependent on natural language processing techniques that have only a limited success rate. Where structured reference metadata is available in proprietary stylings, we become dependent on publisher-specific markups which are not generalisable/abstractable. This results in a huge level of programming labour just to implement additional publisher parsing. We would encourage publishers to provide structured XML versions of their OA books for computational parsing and to expose these for download.

Challenge: DOAB Endpoint Variance

We note that the endpoints signalled by the Directory of Open Access Books do not provide consistent access to the same format. Even within a single publisher, endpoints resolve to different file formats, with no easy way to determine in advance the file type that will be returned. As with the above recommendations on file formats, we recommend that DOAB endpoints should signal the type of resource to which they point and that they should allow multiple endpoints per title. This has economic implications for DOAB.

Challenge: Multi-Lingual Versions

One of the other challenges we encountered pertains to the resolution of canonical references across languages. For instance, should the parser treat “Michel Foucault, *Surveiller et Punir*” as a reference to “Michel Foucault, *Discipline and Punish*”? In the strictest terms, the reference object is, here, not the same. This is particularly the case in abridged or modified translations (Foucault’s *History of Madness* vs. *Madness and Civilization* is another good example here) where the versions are substantially different. However, for the purposes of this tool, it could be of use to a reader to know that a text, whether in English, French, or any other language has been commonly cited.

This would be an extremely difficult task and would require some kind of registry of canonical objects, as machine-translation alone would be insufficient to yield a solid match.

Conclusion

This experiment has demonstrated the feasibility of a full-blown monograph intersect tool that can pull down books from the Directory of Open Access books, take an arbitrary number of these books as input, and intersect them to find references cited in common.

For the experiment, we hit our target of five publishers and five books from each. However, for true scalability and practical implementation we have set out a number of structural changes that would be necessary, or, at least, helpful. Without these changes we are reliant either on advances in natural language processing technologies or on publisher-specific implementations within the intersect software. We also note, though, that these structural changes – the provision of structured XML citation data for instance – come with economic costs for publishers, which could hinder the growth of new presses operating on lower technical provision in order to save costs.

Bibliography

Adema, Janneke, ‘Overview of Open Access Models for Ebooks in the Humanities and Social Sciences’, *OAPEN*, 2010 <https://curve.coventry.ac.uk/open/file/a976330e-ed7a-4bd5-b0ed-47cab90e9a5e/1/ademaoapen2comb.pdf> [accessed 12 August 2014]

Adema, Janneke, Graham Stone, and Chris Keene, ‘Changing Publishing Ecologies: A Landscape Study of New University Presses and Academic-Led Publishing’ (Jisc, 2017)

Ardanuy, Jordi, ‘Sixty Years of Citation Analysis Studies in the Humanities (1951–2010)’, *Journal of the American Society for Information Science and Technology*, 64.8 (2013), 1751–55 <https://doi.org/10.1002/asi.22835>

Biagetti, Maria Teresa, Antonella Iacono, and Antonella Trombone, ‘Testing Library Catalog Analysis as a Bibliometric Indicator for Research Evaluation in Social Sciences and Humanities’, in *Challenges and Opportunities for Knowledge Organization in the Digital Age*, ed. by Fernanda Ribeiro and Maria Elisa Cerveira (Berlin: Ergon Verlag, 2018), pp. 892–99 <https://doi.org/10.5771/9783956504211-892>

Carvalho, José, Cátia Laranjeira, Vasco Vaz, and João Mendes Moreira, ‘Monitoring a National Open Access Funder Mandate’, *Procedia Computer Science*, 13th International Conference on Current Research Information Systems, CRIS2016, Communicating and Measuring Research Responsibly: Profiling, Metrics, Impact,Interoperability, 106 (2017), 283–90 <https://doi.org/10.1016/j.procs.2017.03.027>

Cataldo, Jeremiah W., *Biblical Terror: Why Law and Restoration in the Bible Depend upon Fear* (London: Bloomsbury Academic, 2017)

cOAlition S, ‘Plan S’, *Plan S and COAlition S*, 2018 <https://www.coalition-s.org/> [accessed 12 May 2019]

Crossick, Geoffrey, ‘Monographs and Open Access: A Report for the Higher Education Funding Council for England’, *Higher Education Funding Council for England*, 2015 <http://www.hefce.ac.uk/pubs/rereports/year/2015/monographs/> [accessed 24 May 2015]

Deegan, Marilyn, ‘Academic Book of the Future Project Report’, 2017 <https://academicbookfuture.files.wordpress.com/2017/06/project-report_academic-book-of-the-future_deegan3.pdf>

Emery, Christina, Lucraft Mithu, Agata Morka, and Ros Pyne, ‘The OA Effect: How Does Open Access Affect the Usage of Scholarly Books?’ (Springer-Nature, 2017) <https://media.springernature.com/full/springer-cms/rest/v1/content/15176744/data/v3> [accessed 12 May 2019]

Eve, Martin Paul, *Open Access and the Humanities: Contexts, Controversies and the Future* (Cambridge: Cambridge University Press, 2014) <https://doi.org/10.1017/CBO9781316161012>

Eve, Martin Paul, Kitty Inglis, David Prosser, Lara Speicher, and Graham Stone, ‘Cost Estimates of an Open Access Mandate for Monographs in the UK’s Third Research Excellence Framework’, *Insights*, 30.3 (2017) <https://doi.org/10.1629/uksg.392>

Ferwerda, Eelco, Frances Pinter, and Niels Stern, *A Landscape Study On Open Access And Monographs: Policies, Funding And Publishing In Eight European Countries* (Zenodo, 1 August 2017) <https://doi.org/10.5281/zenodo.815932>

Ferwerda, Eelco, Ronald Snijder, and Janneke Adema, ‘OAPEN-NL: A Project Exploring Open Access Monograph Publishing in the Netherlands Final Report’, 2013 <http://www.oapen.nl/images/attachments/article/58/OAPEN-NL-final-report.pdf> [accessed 24 March 2014]

Ferwerda, Eelco, Ronald Snijder, Brigitte Arpagaus, Regula Graf, Daniel Krämer, and Eva Moser, *OAPEN-CH – The Impact Of Open Access On Scientific Monographs In Switzerland. A Project Conducted By The Swiss National Science Foundation (SNSF)* (Zenodo, 23 April 2018) <https://doi.org/10.5281/zenodo.1220607>

Franzen, Martina, Eileen Joy, and Chris Long, *Humane Metrics/Metrics Noir* (Coventry UK: Post Office Press / meson press, 2018) <https://hcommons.org/deposits/item/hc:19823/> [accessed 30 May 2019]

Hammarfelt, Björn, ‘Beyond Coverage: Toward a Bibliometrics for the Humanities’, in *Research Assessment in the Humanities: Towards Criteria and Procedures*, ed. by Michael Ochsner, Sven E. Hug, and Hans-Dieter Daniel (Cham: Springer International Publishing, 2016), pp. 115–31 <https://doi.org/10.1007/978-3-319-29016-4_10>

———, ‘Following the Footnotes: A Bibliometric Analysis of Citation Patterns in Literary Studies’ (unpublished Doctoral, Uppsala University, 2012)

Herb, Ulrich, ‘Recommendations, Statements, Declarations And Activities Of Science Policy Actors On Shaping The Scholarly Communication System’, *Zenodo*, 2017 <https://doi.org/10.5281/zenodo.1003229>

Higher Education Funding Council for England, ‘Knowledge Unlatched Pilot given HEFCE Backing’, 2013 <https://www.hefce.ac.uk/news/newsarchive/2013/news85263.html> [accessed 21 December 2013]

HIRMEOS, ‘HIRMEOS Workshops on Annotation and Metrics for OA Monographs, 10-11 Jan 2019, Paris’, *High Integration of Research Monographs in the European Open Science Infrastructure*, 2019 <https://www.hirmeos.eu/2018/11/05/hirmeos-workshops-on-annotation-and-metrics-for-oa-monographs-10-11jan-2019-paris/> [accessed 3 June 2019]

HIRMEOS, and Ubiquity Press, ‘Deliverable D6.1: Metrics Services Specification’, *High Integration of Research Monographs in the European Open Science Infrastructure*, 2019 <https://www.hirmeos.eu/wp-content/uploads/2017/11/HI61-Metrics_Service_technical_specification-final.pdf> [accessed 3 June 2019]

HuMetricsHSS, ‘About HuMetricsHSS’, *Humane Metrics Initiative*, 2018 <http://webcache.googleusercontent.com/search?q=cache:A8FHPfPBuC8J:humetricshss.org/about/+&cd=1&hl=en&ct=clnk&gl=uk> [accessed 3 June 2019]

Jubb, Michael, ‘Academic Books and Their Future’, 2017 <https://academicbookfuture.files.wordpress.com/2017/06/academic-books-and-their-futures_jubb1.pdf>

Knowledge Unlatched, ‘How It Works’, 2013 <http://www.knowledgeunlatched.org/about/how-it-works/> [accessed 5 December 2013]

Konkiel, Stacy, ‘Approaches to Creating “Humane” Research Evaluation Metrics for the Humanities’, *Insights the UKSG Journal*, 31 (2018) <https://doi.org/10.1629/uksg.445>

Larivière, Vincent, and Cassidy R. Sugimoto, ‘Do Authors Comply When Funders Enforce Open Access to Research?’, *Nature*, 562.7728 (2018), 483 <https://doi.org/10.1038/d41586-018-07101-w>

Montgomery, Lucy, ‘Knowledge Unlatched:A Global Library Consortium Model for Funding Open Access Scholarly Books’, *Cultural Science*, 7.2 (2014), 1–66

Montgomery, Lucy, Neil Saunders, Frances Pinter, and Alkim Ozaygen, ‘Exploring Usage of Open Access Books via the JSTOR Platform’ (Knowledge Unlatched Research, 2017) <http://kuresearch.org/PDF/jstor_report.pdf>

Nederhof, Anton J., ‘Bibliometric Monitoring of Research Performance in the Social Sciences and the Humanities: A Review’, *Scientometrics*, 66.1 (2006), 81–100 <https://doi.org/10.1007/s11192-006-0007-2>

Neylon, Cameron, ‘The Complexities of Citation: How Theory Can Support Effective Policy and Implementation’, 2016 <http://repository.jisc.ac.uk/6553/> [accessed 12 May 2019]

Neylon, Cameron, Lucy Montgomery, Alkim Ozaygen, Neil Saunders, and Frances Pinter, ‘The Visibility of Open Access Monographs in a European Context: Full Report’ (Knowledge Unlatched Research, 2018) <https://hcommons.org/deposits/objects/hc:18270/datastreams/CONTENT/content>

OAPEN-UK, ‘The Pilot’, 2013 <http://oapen-uk.jiscebooks.org/pilot/> [accessed 25 March 2014]

Pinter, Frances, and Christopher Kenneally, ‘Publishing Pioneer Seeks Knowledge Unlatched’, 2013 <http://beyondthebookcast.com/transcripts/publishing-pioneer-seeks-knowledge-unlatched/>

‘Project COUNTER - Consistent, Credible, Comparable’, *Project Counter* <https://www.projectcounter.org/> [accessed 12 May 2019]

Royle, Pamela, and Norman Waugh, ‘Bibliometrics of NIHR HTA Monographs and Their Related Journal Articles’, *BMJ Open*, 5.2 (2015), e006595 <https://doi.org/10.1136/bmjopen-2014-006595>

Simba Information, ‘Open Access Book Publishing 2016-2020’, 2016 <https://www.simbainformation.com/Open-Access-Book-10410716/> [accessed 12 May 2019]

Snijder, Ronald, ‘Measuring Monographs: A Quantitative Method to Assess Scientific Impact and Societal Relevance’, *First Monday*, 18.5 (2013) <https://firstmonday.org/ojs/index.php/fm/article/view/4250> [accessed 12 May 2019]

Springer Nature, ‘Open Research Library’, *Open Research*, 2019 <https://www.springernature.com/gp/open-research/journals-books/books/orl> [accessed 26 May 2019]

Suber, Peter, *Open Access*, Essential Knowledge Series (Cambridge, MA: MIT Press, 2012) <http://bit.ly/oa-book>

Swan, Alma, ‘Open Access Policy Effectiveness: A Briefing Paper for Research Institutions’ (Pasteur4OA) <http://www.pasteur4oa.eu/sites/pasteur4oa/files/resource/Policy%20effectiveness%20-%20institutions%20final.pdf>

Sweeney, David, and Ben Johnson, ‘Seeking a Fresh Perspective: A Research Funder’s View of Open Access’, *Insights: The UKSG Journal*, 27.1 (2014), 51–57 <https://doi.org/10.1629/2048-7754.114>

Vincent, Nigel, and Chris Wickham, ‘Debating Open Access: Introduction’, in *Debating Open Access*, ed. by Nigel Vincent and Chris Wickham (London: British Academy, 2013), pp. 4–12

Vincent-Lamarre, Philippe, Jade Boivin, Yassine Gargouri, Vincent Lariviere, and Stevan Harnad, ‘Estimating Open Access Mandate Effectiveness: The MELIBEA Score’, *ArXiv:1410.2926 [Cs]*, 2014 <http://arxiv.org/abs/1410.2926> [accessed 12 May 2019]

Vincent-Lamarre, Philippe, Jade Boivin, Yassine Gargouri, Vincent Larivière, and Stevan Harnad, ‘The Effect of Open Access Mandate Strength on Deposit Rate and Latency’, 2014 <http://eprints.soton.ac.uk/366815/> [accessed 23 July 2014]

Watkinson, Charles, Kevin Hawkins, Lucy Montgomery, Brian O’Leary, Cameron Neylon, and Katherine Skinner, ‘Understanding OA Ebook Usage: Toward a Common Framework’ (Knowledge Unlatched Research, 2018) <https://deepblue.lib.umich.edu/bitstream/handle/2027.42/143840/Redacted%20Grant%20Narrative%20-%20OA%20Ebook%20Usage_FINAL%20SUBMISSION_042718.pdf?sequence=1&isAllowed=y> [accessed 13 May 2019]

Watkinson, Charles, Rebecca Welzenbach, Eric Hellman, Rupert Gatti, and Kristyn Sonnenberg, ‘Mapping the Free Ebook Supply Chain: Final Report to the Andrew W. Mellon Foundation’, 2017 <https://deepblue.lib.umich.edu/handle/2027.42/137638>

Xia, Jingfeng, Sarah B. Gilchrist, Nathaniel XP Smith, Justin A. Kingery, Jennifer R. Radecki, Marcia L. Wilhelm, and others, ‘A Review of Open Access Self-Archiving Mandate Policies’, *Portal: Libraries and the Academy*, 12.1 (2012), 85–102