The LSE, the Blogs and the Metadata

Helen Williams, Metadata Manager, LSE Library

You might be somewhat wary of a paper where the title has to be explained, but *The LSE, the Blogs and the Metadata* is an allusion to the second chronicle of Narnia, *The Lion, The Witch and The Wardrobe*. Inspired by the wardrobe's power to let the children into another world, this paper looks at the power of metadata in accessing LSE's blog content. Just as the children tumbling out of the wardrobe, when they returned home at the end of C. S. Lewis' second chronicle, was only the beginning of their adventures in Narnia in the remaining chronicles, the project this paper describes is only the beginning of our blog adventures at LSE. It considers the key role of the Library, and specifically metadata, in supporting institutional goals and the wider work and outreach of the university.

In May 2016 the Library had the opportunity to bid for some Higher Education Innovation Funding (HEIF). This funding is given to support knowledge exchange between universities and the wider world which result in economic and social benefit to the UK.[1] Such a focus on social benefit is an integral part of LSE, which was established in 1895 for '*the betterment of society*'. The phrase we use 123 years later for sharing research in order to make a difference in how problems are understood and addressed around the world is '*Knowledge Exchange and Impact*'. Just one of the many ways academics are encouraged to do that is by blogging about their work to bring it into a sphere for wider academic communication.[2] A 2017 study of 45,000 academics and scientists carried out by the School of Informatics and Computing at Indiana University found that content from blogs.lse.ac.uk made the top 15 list of academic content that such groups share on social media. For political scientists, content from LSE blogs was the number one source of academic content they shared on Twitter. For sociologists it was number two and for economists it was number five.[3]

We know that LSE-generated blog content is significant for the institution, and yet by its nature it is entirely web -based content, which OCLC's 2018 report, *Descriptive Metadata for Web Archiving* describes as 'volatile', stating that '*if not preserved on a timely basis a significant percentage of web content simply ceases to exist*' which means it is '*imperative that we preserve web content on a timely basis if we are to maintain the integrity and continuity of the historical, cultural and scholarly records*.'[4] This awareness had already informed Library Strategy and our 2015-2020 strategy includes a specific action to '*secure the collection and preservation of the complete intellectual output of the School.*'[5]

1.Higher Education Funding Council, *Knowledge Exchange Funding*, http://www.hefce.ac.uk/ke/heif/ accessed 27/09/18

2. London School of Economics and Political Science, *An Introduction to Knowledge Exchange and Impact at LSE*, https://info.lse.ac.uk/staff/services/knowledge-exchange-and-impact/Assets/Documents/PDF/Toolkit/1-Introduction-to-KEI-at-LSE.pdf accessed 27/09/18

3. Ke, Qing, Ahn, Yong-Yeol, and Sugimoto, Cassidy R., *A systematic identification and analysis of scientists on Twitter*, PLoS ONE 12 (4): e0175368, 2017, https://doi.org/10.1371/journal.pone.0175368 accessed 27/09/18

4. Dooley, Jackie, and Bowers, Kate *Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*, Dublin, Ohio, 2018, pg 5 https://www.oclc.org/content/dam/research/publications/2018/oclcresearch-wam-recommendations-a4.pdf accessed 27/09/18

5. LSE Library, *LSE Library Strategy*, http://www.lse.ac.uk/library/about/lse-library-strategy accessed 27/09/2018

Furthermore the School's 2015 Knowledge Exchange and Impact Strategy includes a specific action for the Library to archive all official blog output. Having an efficient, systematic and comprehensive approach will protect this content from the risk of loss in the future. By archiving the content in LSERO we ensure that its metadata feeds through to our Library Search platform, so that it is discoverable and accessible alongside library content.

The volume of blog content at LSE, however, makes all this no mean feat. We produce 61 blogs, as well as continuing to provide access to various closed blogs, and also contribute to a number of partner blogs.[6] On closer inspection 56 of these were deemed to be 'official' output, with the internally focused community blogs, which do not include research, being outside the scope of the project. There were already some links with LSE Research Online (LSERO), the School's Institutional Repository, in cases where an author had specifically requested that their blog content was added to the repository. In 2016 this self-selecting content was an average of 63 blog posts a month, which is about a fifth of the School's monthly blog output. Our HEIF bid focused on extending and developing existing activity to retrospectively archive all official blog outputs and establishing automated procedures to make complete archiving feasible and sustainable in the future. If time allowed we also hoped to review the potential for adding DOIs to blog posts to improve discoverability and to offer the ability to measure impact through tools such as Altmetric.

We had initially anticipated hiring a project manager to bring niche technical skills to the project, but it became apparent that for a short term project the use of internal staff, with an existing knowledge of institutional practices and procedures, would be more beneficial, even where that meant factoring in some development time. However there were time restrictions attached to the funding, so it was not possible to upskill in every area of the project and we determined that the investigation of DOIs would need to be part of future plans. Our focus for this project was on retrospectively archiving the content, which meant creating PDFs, saving them in backed-up Library server space, loading them to LSERO and creating associated metadata, and establishing automated procedures to make the process sustainable. This ensured that we would have practical deliverables. Considerable initial investigative and planning work was required, including a workflows audit to check that the amount of time we thought it took to create blog records on LSERO was still correct. This allowed me to 'number-crunch' on the basis of one member of staff being able to add 200 records to LSERO in a week. However, even armed with this information, estimating project workflows proved much more difficult than I had hoped. The top ten blogs, in particular, frequently re-post each other's content, so some posts were duplicates which did not need adding but, as cross posting is only indicated at the bottom of a post, each one required manual checking. In the end we worked with two estimates; the maximum being based on the total number of posts on each of the blogs (though, for various reasons, even this was not always easy to determine) and the minimum being based on an educated guess at the number of unique posts using a sample of two months from each blog. Adding all this up, it was apparent that we had a minimum of 5800 posts to add, and a maximum of just over 20,000, which is not the kind of mathematical discrepancy you want, particularly when you are responsible for the project! We recruited three temporary staff to work on creating PDFs and the associated metadata, so if they all worked at the anticipated rate (an unknown at this stage) we would create 600 records a week, and adding all content would take us somewhere between ten weeks and 33 weeks – but our master finance spreadsheet told us we could fund three temps for a maximum of 16 weeks. This was a particularly important time for managing expectations, and we focused our aim and communications on archiving official content from the top ten blogs. Stakeholder engagement is, of course, an essential part of project management and we liaised with the School's social media manager who invited us to meet the top ten blog editors and outline the project. They were encouragingly enthusiastic, and it quickly became evident that there was a real appetite in the School for this work to take place.

6. London School of Political and Economic Science, *LSE Blogs*, http://blogs.lse.ac.uk/ accessed 27/09/2018

In an ideal world, we would have worked on establishing the automated processes for adding content before we employed the temps so that more could be achieved within the short time available, but due to the funding and timing constraints of the project we needed to work on both aspects simultaneously. The first step was to automate the creation of PDFs of each blog post. Our colleagues in IT were able to do this for us and they now put these into a shared file which we can pick up every month. For the metadata we have been able to semi-automate its creation through the use of BibTex files which can be imported along with the PDFs. This automates the process as far as is currently possible with the blog templates being used by the School. Some metadata is consistent across all blog posts, such as publisher, copyright or item type, and can be put into a template. Metadata automated from the live posts, such as date, blog name, post title, base and blog URLs can then be added to this. Despite creator, subject and description being three of the five most commonly used Dublin Core elements for archiving web content,[7] we have been unable to automate the metadata for those fields. In our blog templates the 'author' is the person who published the post, who tends to be the editor of the blog, rather than the author of the post content. The author of the content is only identified in the body of the text rather than as a separate field. We receive the blog abstract as part of the automated metadata feed, but due to a lack of consistency in the way it is formatted across multiple blogs, it often truncates part way through, and so requires manual checking. The subject tags on the blogs are not the same as the scheme we use in LSERO but, even if they were, we understand that we would not be able to auto-populate the subject trees in Eprints. So we are limited to partial templates, but this semi-automation is still better than the entirely manual process we would otherwise have, and it has enabled us to shave three minutes off the process for each record (from ten to seven minutes).

After 16 weeks of dedicated work by our temps, our project expectations were exceeded with the addition of 11,665 blog posts to LSERO, the total unique content from all of LSE's 56 official blogs. Having also completed the automation aspect of the project there was now some further 'number-crunching' to do in order to investigate how sustainable it was to archive blogs in the future.

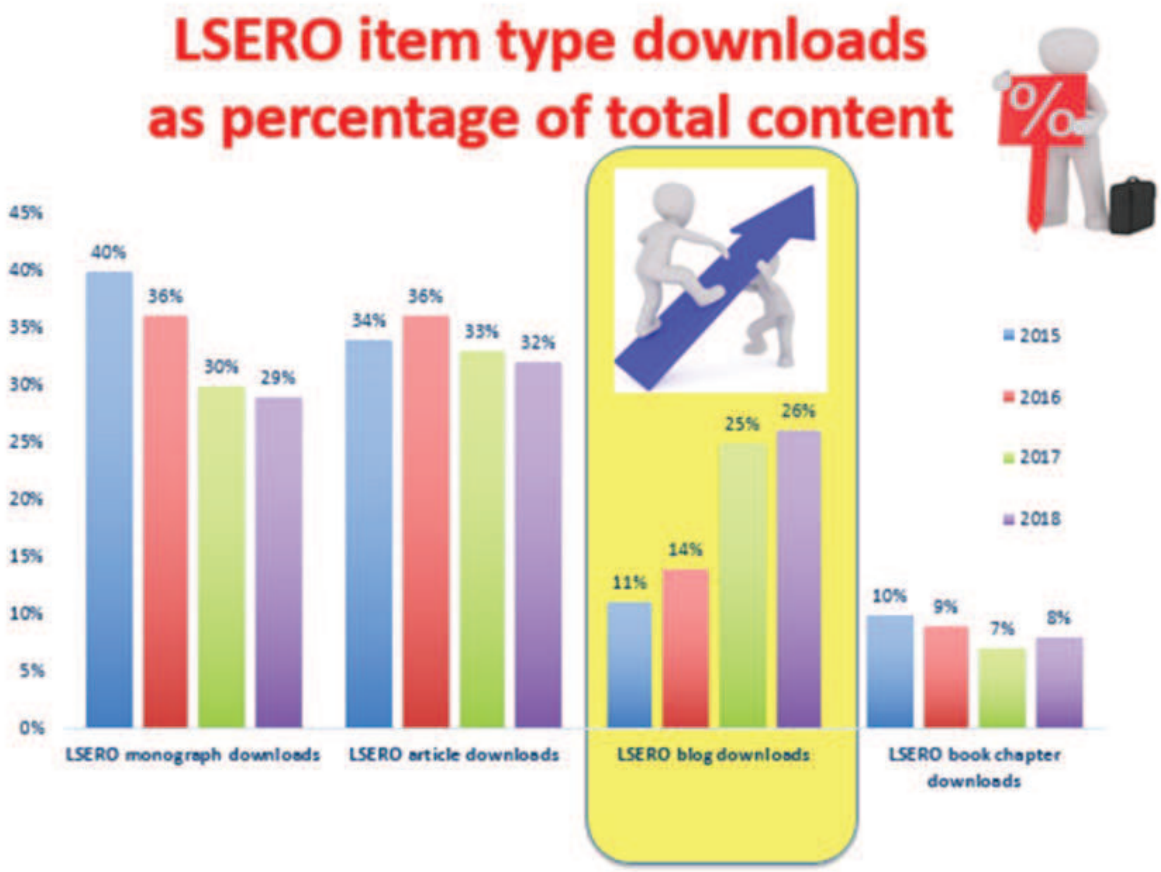
Average number of blog posts for each blog per month
x
New time to create a record on LSERO
+
Average number of re-posts per blog per month and time taken to eliminate them from the process
=
12 hours work a week to keep up with archiving all official blog output

I do not imagine anyone reading this article is operating in a team which could simply absorb that amount of extra work, nor that any reader could magically produce the finances to increase resourcing in their team by nearly 0.4 FTE. It was time to return to the 'managing expectations' aspect of the project, which had made clear that the top ten blogs were the priority. Digging into the figures a bit more it became clear that in theory archiving content from just those top ten blogs would take an average of six hours a week, and that all the other blogs put together made up the remaining six hours a week. Our team does not, of course, have six hours a week to spare. Nevertheless, six hours is significantly less than 12, and we wanted to be offering something ongoing from the project, so we took the risk of saying that we would trial absorbing those six hours into team workflows, but that beyond that we would require additional resourcing. In practice, looking at the statistics, this work has taken six and a half hours a week, but the self-selecting blog content that we had been dealing with for authors before this project began took two and a half hours a week, so we have actually only an additional four hours to add to usual workloads.

7. Dooley and Bowers, *Descriptive Metadata for Web Archiving*, pg 12 https://www.oclc.org/content/dam/research/publications/2018/oclcresearch-wam-recommendations-a4.pdf accessed 27/09/18

We have four members of staff to absorb this, so effectively each of them had to find an extra hour a week for the blog content, which suddenly sounded more hopeful. Each of those four staff takes responsibility for one or more named blogs, based on the average output for each blog, so that they are each receiving a similar amount of work. This worked excellently for four months, but then a prolonged staffing vacancy coinciding with a further new project significantly reduced our capacity and we had to put the blog work on hold for a period of months, returning to it about six months later when resources allowed.

As we closed the project, we celebrated meeting a specific goal given to the Library by the School's Knowledge Exchange and Impact strategy to archive all official blog output up to July 2017, ensuring that this key content, which falls outside the more traditional publishing channels, is protected from risk of future loss. We were also able contribute to the specific Library Strategy action to secure the collection and preservation of the complete intellectual output of the School by adding all official blog output to date. New workflows mean that we can keep up with work to secure the content of the top ten blogs each month. Content outside the top ten blogs has not been collected since the end of the project, but we are now in a much better position to do this retrospectively, should resourcing become available. Given that the blog content is still live on the LSE blogs themselves, it has been important to consider whether our work has had an impact on the visibility and discovery of this content. As might be expected, accessing blog content directly from the blogs themselves is currently its primary access point. The blogs have more than 70,000 unique users each week, and the most popular post has had 470,000 views. We cannot rival that in LSERO, but we can point to an increase of activity in accessing blog content on LSERO as illustrated in the graphic below. (Accounts for 95% of downloads with the others being conference papers, AV etc.).



Blogs are still third place in terms of most viewed content type, but they have had a significant percentage increase compared to other content types.

The success of the project led to a number of positive responses from blog editors and academics, raising the profile of the Library as a partner in disseminating LSE's research outputs.

I mentioned earlier that there had been timing constraints on our project, restricting some of the work we had hoped to do, with the result that when resourcing allows there are areas we would like to work on as a second phase of this project. OCLC's *Descriptive Metadata for Web Archiving* report addresses various issues around the capture and creation of metadata for web-based content, so we may wish to review the metadata applied to our live blog posts and consider how that could be purposed to improve discoverability. This would require liaison work outside the Library, discussing the School's blog template, and agreeing what level of consistency could be established across the blogs to facilitate more efficient application of, and subsequent capture of, descriptive metadata. We would benefit from including LSE's Website Improvement Team in this review as we think about how best to surface blogs alongside other LSE content. Our Digital Library team have developed a DOI minting service, so there are opportunities to investigate adding DOIs to blog content, both to give increased stability to content in terms of permanence and discovery, and to consider how we might then be able to measure the impact of this content via tools such as Altmetric. Finally, we could consider auto-classification for this content. *Descriptive Metadata for Web Archiving* indicates that we are not alone in struggling with the need for '*scalable descriptive metadata practices that take into account the extremely limited human resources available*'.[8] The subject classification we use for blog content, coupled with the fact that lack of staff time is our highest barrier to metadata creation for that content, means that blog content could benefit from our experimenting in this area. It would require more investigation to determine feasibility, but we do need to consider options that allow us to use metadata expertise in scaling up the blog archiving operation with our limited resources… so do watch this space, because I hope that what I have written above is just the beginning of our adventures!

Biography

Helen is the Metadata Manager at LSE Library where her team have responsibility for print, electronic and institutional repository metadata, with a strategic focus on metadata management to support discovery. Helen has previously worked for the Institute of Directors, and The London Library, and was on the CIG committee from 2009-2016.
H.K.Williams@lse.ac.uk
@HelsKRW

8. Ibid, pg 11