**SLIDE 0**

Can Download stats tell us where to from here? This presentation focuses on download stats for the UNSW institutional repository, UNSWorks – for the period Jan-Jun 2019.

**SLIDE 1**

What can download stats tell us about our content and users? Can the stats give us any evidence to answer the ‘where to from here?’ question?

We found some evidence by answering these questions: What is being downloaded? Who is downloading content? And how do users get to the content?

I would be very interested to hear if our answers to these questions are similar for your repositories or not – so please approach me later to compare.

**SLIDE 2**

We also wanted to know which of the various measures we implemented during the last two years to increase discoverability of UNSWorks had the most impact, because as you can see, there has been an almost 80% increase in downloads from Jan-Jun 2018 to Jan-Jun 2019.

**SLIDE 3**

Download stats by resource type assist in answering what is being downloaded.

This chart shows the number of downloads relative to full-text content – by resource type. The left axis shows the number of downloads – and the bars are blue. The right axis shows the number of full-text records – and the bars are orange.

Almost three quarters of downloads are theseEs, even though theseEs account for only half of the open access or full-text publications in UNSWorks. Less than 15% of downloads are journal articles, even though almost 30% of the repository’s full-text content is journal articles. Institutional repositories are ideal publishing platforms for non-traditional research outputs – or NTROs - like conference materials, reports, creative works etc - but deposit and downloads of these research outputs in UNSWorks are quite low.

Since theseEs are primarily published in institutional repositories, high download numbers are expected. Journal articles are primarily published in journals, and metadata for 80% of UNSWorks journal articles contains the publisher’s DOI. Therefore, if a user is affiliated with a university holding a journal subscription, the publisher version will be the preferred copy to access.

What does this mean for us?

For one, we need to better facilitate the deposit of NTROs by improving functionality specific to these outputs, e.g. re-use licenses, streaming capabilities, NRTO-specific metadata and provision for the deposit of large files of various formats.

Secondly, although we noted that both the deposit and download of journal articles doubled in the last year, there are still opportunities to increase both by focusing on 1) the Open Access aspect of the repository and 2) the value-added functionality of the repository.

As for Open Access, it means integrating with more OA tools and services like UnPayWall and the Open Access Button; it also means encouraging researchers to view the repository as an OA platform that may assist them in fulfilling their OA mandates, especially for ‘APC-poor’ authors.

As for value-added functionality, we could do more to promote what the repository is already adding in value (like its integration with metrics and ORCID), but we could also be looking at how we can increase value-added functionality which will make researchers *want* to deposit their research in UNSWorks.

**SLIDE 4**

Looking at downloads by faculty could also help in answering *what* is being downloaded.

This chart shows the number of downloads relative to full-text content – by faculty. The left axis shows the number of downloads – and the bars are blue. The right axis shows the number of full-text records – and the bars are orange.

Almost two thirds of full-text content in UNSWorks are from the Faculties of Medicine, Engineering and Science and their download numbers for the period are highest among the faculties. Less than 9% of full-text content in UNSWorks are from the Faculties of Art & Design, Built Environment and Law and their download numbers for the period are lowest among the faculties.

This suggests a correlation between total OA or full-text content in the repository and number of downloads, i.e. the more content, the more downloads.

However, Arts & Social Sciences bucks this trend by having a high download number even though content from the faculty accounts for only 14% of total full-text content. This means the faculty’s total downloads constitute a smaller number of files which are downloaded multiple times.

**SLIDE 5**

We also did an analysis of highly downloaded publications, i.e. publications that have been downloaded 20 or more times in any one month for the 6-month period.

Almost 40% of the highly downloaded publications are from Arts & Social Sciences. A word map of the keywords of these highly downloaded titles indicates the popularity of Australian and Indigenous content.

So what does this mean for us?

Providing download figures and analysis to faculties on a regular basis may promote the repository as a publishing platform and source of research material. In particular, we could communicate to Arts & Social Sciences the popularity of its content in the repository.

**SLIDE 6**

Who is downloading content? This is the most difficult question to answer, since no sign-in is required to use an institutional repository, which means there are no user data. Instead, statistical tools rely on the user’s Internet Protocol (IP) address – which identifies, among other things, the network and location of the *device* used to access the internet – but not the *user*.

One option considered was paying for a bulk IP lookup via a third party to retrieve information on user type (e.g. educational or commercial user), but we decided not to do this due to patron privacy reasons.

**SLIDE 7**

The most that could be determined from our stored IP addresses is geographical locations of users - or rather devices - and whether files are being downloaded from within or outside the UNSW network. The geolocation of IP addresses points to almost 150 countries, the top countries being Australia, United States, United Kingdom, Germany and China.

However, geographical location may not be a reliable indicator of who is downloading content. During the period Jan-Jun 2019, the number of downloads from Australian IP addresses decreased substantially from previous periods and the number of downloads from US IP addresses increased and surpassed the Australian numbers. This does not mean that our users are physically in the United States; it is probably more related to increasing numbers of downloads that come via data centres like Google and Amazon which are based in the United States.

**SLIDE 8**

More reliable than geographical location is the fact that only 5% of downloads occurred via the UNSW network for the period Jan-Jun 2019. This percentage has dropped considerably from a year ago.

It appears that UNSWorks publications are more broadly discoverable than they were a year ago, which has increased the percentage of non-UNSW network downloads.

**SLIDE 9**

Web spiders, robots and/or search engine crawlers are also “downloading” UNSWorks publication files. As much as possible, these download numbers are removed from download counts, but even during this study we found that 5% of “counted” downloads for the period were Google bots.

What we have learnt is that it is not possible to accurately answer “who is downloading content?“ Visitor stats, rather than download stats, may provide a more reliable picture of the repository’s users – bearing in mind that visitor stats do not include users who bypass the repository pages and download files from other sites.

Google Analytics stats show that 70% of UNSWorks visitors have an Australian location and more than a quarter are using the UNSW network – which indicates that UNSW staff and/or students *are* using the repository.

**SLIDE 10**

Referrer data show how users get to content: whether they are finding content from within the repository itself, via other websites, content aggregators or via social media or search engines.

But before we look at referrers, I want to explain that our study also looked at the source of downloads. By source we mean the distinction between ‘direct file’ downloads and ‘indirect via landing page’ downloads.

We were somewhat surprised to find that 93% of UNSWorks content downloads for Jan-Jun 2019 were ‘direct file’ downloads. This means users bypassed UNSWorks landing pages and downloaded files from outside the repository. Only 7% of downloads were made via the landing pages of UNSWorks records.

Direct file downloads increased substantially since Aug 2018 as you can see. This increase coincides with our creation of a [local deep links site](http://unsworks.unsw.edu.au/publications_by_year/) in Aug 2018 which contains only direct links to UNSWorks publication files. (<http://unsworks.unsw.edu.au/publications_by_year/> )

**SLIDE 11**

Here on the left is a screenshot of some pages of this local site. It was created because our repository front-end software technology is not conducive to Google Scholar indexing. Testing indicated that Google Scholar may be able to parse and index publication files directly – without having to read record meta-tags.

Even though this site did increase Google Scholar indexing of UNSWorks records, the more notable consequence was that Google started indexing this local site – note the examples on the right: the same record is indexed twice in Google: one that resolves to UNSWorks landing page, and one that resolves to the PDF.

So why do we care whether a file is downloaded from within UNSWorks or from outside UNSWorks?

If users are not accessing files via landing pages, we need to ensure that publication files or attachments contain comprehensive citation metadata and information about rights of use – information usually contained in the landing page record. It also reminds us that the “look and feel” of the repository is secondary to other concerns.

**SLIDE 12**

Now let’s take a closer look at referrers. I already mentioned that direct file downloads increased substantially, and that Google is the main referrer for these downloads – see the pie chart on the right.

As for indirect via landing page downloads – see the pie chart on the left - almost half of users searched for and downloaded files from UNSWorks itself, otherwise users were mostly referred from Google, [Trove](https://trove.nla.gov.au/) and Open Access Theses and Dissertations. This has not changed from previous periods.

What did we learn from this?

Despite concerted efforts during 2018 to 2019 in registering UNSWorks as an OAI-PMH provider in various registries and working with various aggregators like BASE, CORE, Primo Central Index, WorldCat, OAIster and EBSCO Open Dissertations to harvest UNSWorks records, these aggregators are *not* major referrers to UNSWorks records. It seems that discoverability is increasingly dependent on interoperability with the web, especially search engines and social media – and we need to ensure that our repository platforms are geared towards this.

**SLIDE 13**

So here is a recap on what our download stats investigation indicates for the future.

Essentially, we want to promote the open access value of the repository; we want to increase the repository’ interoperability - with the web, with Google Scholar, with more tools and services – internal and external to the university. We want to improve the functionality of the repository and we want to engage more with faculties and researchers.

UNSW Library recently completed a tender process for a new repository platform. Interoperability, discoverability and researcher engagement were important considerations in our choice of a new system, and we can’t wait to see where the new system will take us.

So back to downloads. In parting, with so many variables present when counting downloads - including the differences between statistical tools and the near-impossible task of excluding non-human downloads from download counts – we found the best method of analysis is to look at trends and patterns in the stats rather than the actual numbers.

**SLIDE 14**

Please feel free to contact me if you would like further information on this presentation.

Thank you.