9th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2016 [cit. 2015-12-5]. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings/**. ISSN 2336-5021.

WAYS OF DISSEMINATING, TRACKING USAGE AND IMPACT OF ELECTRONIC THESES AND DISSERTATIONS (ETDS)

Meinhard Kettler

meinhard.kettler@proquest.com

ProQuest Information and Learning, Germany

Abstract

The digital transformation has had a tremendous impact on graduate research workflows and output. Most theses are submitted as ETDs, although the share varies by country and by subject. Universities worldwide are running institutional repositories to showcase new graduate research, as well as recently digitized material. The presentation will highlight studies on dissertation and theses usage in repositories as well as giving insights into ProQuest's unique dissertations and theses analytics.

Keywords

ETD, Thesis, PhD Thesis, Dissertation, Usage Analytics, Impact, ProQuest

Introduction

It is evident to everyone in and outside academia that theses and dissertations represent a valuable contribution to scholarly research. With their graduate works young researchers show their ability to examine and summarize existing knowledge as well as adding new and unique findings in a detailed study. Not all dissertations and theses mark the beginning of a researcher's career, but for those graduate students that choose to continue their professional life in research the PhD thesis or dissertation will often be the first entry in their personal publication list and therefore deserves special attention. Even if the thesis submission does not open up a professional life in research, it may very well contribute to the relevant literature in a given subject.

In this presentation I will try to summarize what I have learnt about the current situation and trends of dissertations' and theses' visibility and impact, having been active in the field of dissertations dissemination for ProQuest for a bit more than a year, and also draw some personal conclusions. Hence the view comes from outside the academic community, yet from the perspective of a renowned commercial player in the field, collecting graduate research for dissemination for more than 70 years.

Availability and Discoverability

From the user's perspective there are a number of ways to retrieve content in dissertations and theses. As with other communication areas, the transformation from print to digital that began in the 1990s had a disruptive effect on dissertation and theses. Since then, policies and workflows in graduate schools have been amended, so that today the original versions of many graduate works consist of 'born-digital' PDFs and no longer bound copies that can only be found on the shelves of the institution's library.

Today a huge number of theses can easily be found by any interested user on the web. However, if unexperienced users conclude that all the world's graduate output is available online and easily discoverable by Google and other search engines, they might get frustrated when trying to retrieve specific information. Discoverability and accessibility is granted only for a part of the existing output. Furthermore, it can turn out to be extremely cumbersome, if not impossible, to find out how big the invisible part of the iceberg is. When a certain PhD thesis does not show up in the result list of a search engine, nor in the institutional repository (IR), nor at the numerous dedicated aggregator access points for theses, there might be a variety of reasons:

- Author's decision (accepted by the institution)
- Publication embargo requested by the author or due to third party copyrights
- Monograph published under a different title
- Submission in print without electronic copy
- Institutional repository or catalogue not in place
- Limited discoverability due to language
- ...and more

The above cases illustrate that it is impossible to get a full overview of global graduate works. A single international registry or authority for theses simply does not exist and most likely will not be launched in the near or mid-term future. Policies to submit and publish the works linked

to an academic degree vary not only from country to country, but also between universities in the same country and, quite often, even between faculties and departments of one university.

As to print submission, electronic full text theses and dissertations (ETDs) have only been available for the last twenty years. Before that time, submission in electronic form was not possible or at least not supported by the institutional workflows. However, when looking at today's figures, it might surprise that in some developed countries without technical barriers print submissions still play a significant role.

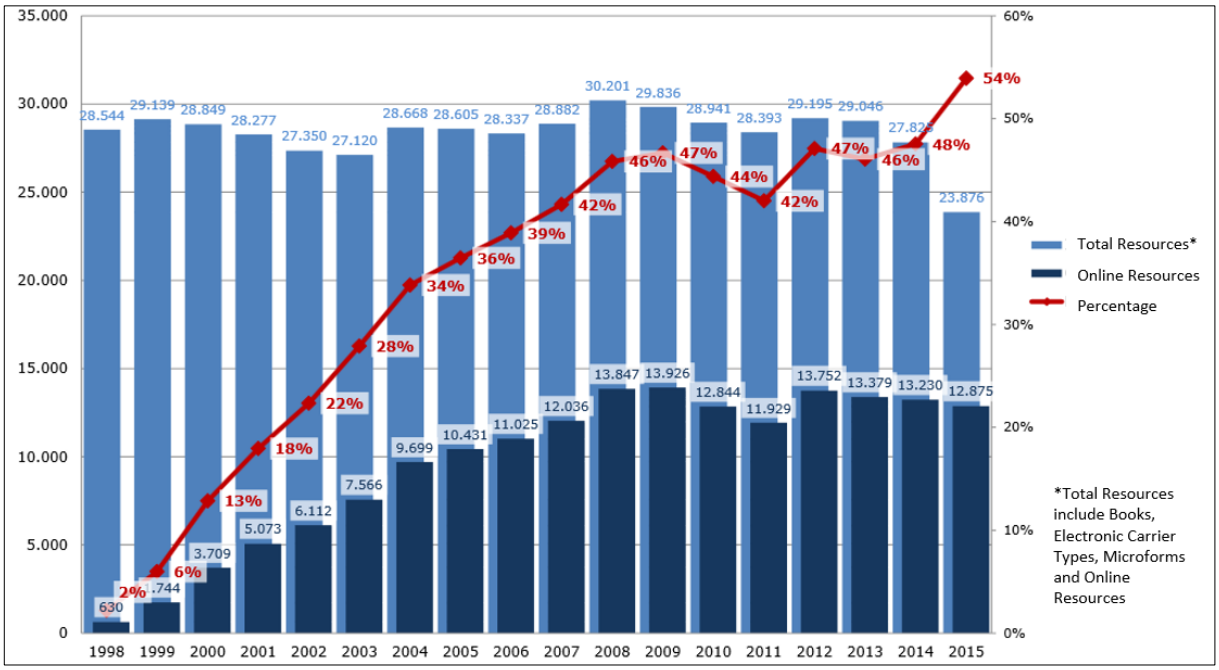

Figure 1: Online Resources Share of Total PhD Dissertations and Habilitations in German National Library Collection by Publication Year (updated 2 March, 2016).
By Deutsche Nationalbibliothek 2016,
http://www.dnb.de/DE/Wir/Kooperation/dissonline/dissonlineStatistik.html

The German National Library publishes statistics of PhD theses (Dissertationen) on its German website [Figure 1]. According to the available graph from March 2016, covering the years from 1998 to 2015, the number of PhD theses per year captured in the German National Library catalogue remains stable between 25,000 and 30,000, whereas the percentage of online available items rose from 2 to 54%. Although this looks like a remarkable increase, the remaining 46% "non-online" formats, including print, electronic carrier types and microform, in 2015 still feels very high. As explained on the website, the latest numbers may not yet be exact and complete as a result of technical changes in the harvesting interface, so that files from some universities are missing and will be ingested in the German National Library database only later. Nevertheless, many of us would have expected a much higher percentage for online availability.

For France estimates can be found on the poster *French Electronic Theses and Dissertations in Europe* presented by Hélène Prost and co-authors during the 19th International Symposium on ETDs in Lille, France (Prost, 2016). From the national platform *theses.fr* the authors retrieved the following numbers for French theses production: between 8,000 and 13,000 theses per year were deposited from 2007 to 2015 with an online share of 60% for the last two years. Not only is the percentage of online formats slightly higher than in Germany, but there

is also a clear commitment to an e-only policy. The authors mention a French decree issued in 2016 defining deposit of the digital version of a thesis as mandatory, with the aim to bring print deposit to an end in 2018.

At the same time ETD submission rates from US schools is higher. Acting as official offsite dissertation and theses repository for the U.S. Library of Congress, ProQuest received 93% of US submissions for its database *ProQuest Dissertations and Theses (PQDT)* in electronic form in 2014 (McLean, 2016).

Clearly, the trend goes in the direction of online publishing, especially in the natural sciences. Therefore, it comes as a surprise that in German universities, where students are obliged to publish their theses, doctorate students can choose how to submit and publish their theses. In the long run, the freedom regarding deposits and the lack of standards in the policy-defining faculties might turn into a competitive disadvantage for authors and institutions, negatively affecting visibility and impact of the output.

Dissertations and Theses - Meaningful Content to Showcase

How do theses submitted in digital formats reach their readers? While a global and complex discussion about the possible transition from paid to open access business models for scholarly journal articles and about new ways of funding publishing continues in many countries ETDs have been available on Open Access institutional repositories for a number of years. This offering implies benefits not only for users, but also for authors, whose graduate works have the chance to be more easily discovered, read and perhaps even cited by fellow researchers around the world. The dissemination of research increases the potential for international research collaboration as well as recognition.

Electronic theses represent a high share of content in many institutional repositories. In fact, many repositories were built for the purpose of making ETDs available, and some universities still maintain dedicated repositories or document servers only for ETDs.

In Finland, it was demonstrated that electronic theses represent valuable enrichments to repositories and enjoy surprisingly high usage by public consumers. The Finnish science community has not only managed the transition from print theses to ETDs successfully, but also extended the online availability to (master and bachelor) theses from 25 universities of applied science. More than 100,000 theses are publicly available on the portal *Theseus*, hosted by the National Library of Finland, with 15,000 new items added per year. In 2014 there were almost 18 million full text downloads - more than the total amount of downloads from all other Finnish repositories (Ilva, 2015).

Another example can be found in the repository of the prestigious and research-intensive University College London (UCL). In 2014, of the top 50 documents by downloads, 28 are doctoral theses, including 7 of the top 10.[1] Given these figures there is no reason to underestimate theses as content of "lower" relevance and meaning. For graduate schools, as

[1] **http://discovery.ucl.ac.uk/past_stats/annual-2014.html**

well as for the authors, it is important to ensure visibility and dissemination of their graduate treasures.

Usage and Impact - the Need to Measure and Analyse

In September 2016, the Directory for Open Access Repositories (OpenDOAR) counted 1,790 out of 3,220 repositories that include theses and dissertations as a content type.[2] Both numbers are still growing. Regarding the share of repositories with theses and dissertations we have to keep in mind that not all repositories in this registry are linked to degree-granting institutions and consequently some have no PhD theses to display.

On the other hand, the overwhelming and growing number of individual access points holds new challenges: how can users find their way to relevant theses content when it is distributed over hundreds or thousands of web servers with a variety of user interfaces? Google Scholar as well as aggregator platforms like the global ETD search NDLTD (Networked Digital Library of Theses and Dissertations, **http://www.ndltd.org/**), or the DART Europe E-theses portal (**http://www.dart-europe.eu/About/info.php**), respond to this need by indexing available web content or harvesting metadata from repositories, making use of available standard protocols like OAI PMH. Although not visible to the end-user, there are complex tasks to complete in the back-end, sometimes meaning portals might return incomplete results from their sources. Constantly, administrators are forced to fight data quality and transmission issues from the large variety of dynamically changing content providers.

Next to data quality, getting accurate usage statistics for Open Access repositories also provides challenges for librarians. In most cases the metrics cannot be interpreted 'as is', as usage figures are often skewed or inflated by the activity of robots crawling the open web and not always easily identified. This means the data cannot be taken as 'human' usage and compared to other analytics unless it is filtered and edited.

Some entities have undertaken the effort to create standards and services for comparable usage statistics, such as the British Jisc-funded national aggregation service IRUS-UK (**http://irus.mimas.ac.uk/**) or the initiative DINI in Germany (**https://dini.de/english/**). Most initiatives currently act on a national level.

In the context of theses collection and aggregation, ProQuest, as the first global provider of graduate works, continues to play a crucial role. The US company started microfilming PhD theses under its former name UMI in 1938, mainly for the purpose of long-term preservation. Today *ProQuest Dissertations and Theses Global* covers 99% of US output and output from thousands of universities outside the US, including nearly two million PhD and masters theses in searchable full text. Therefore, it is regarded as one of the premier one-stop-shops with a high share of unique content over all subject fields. Part of the database is openly accessible on the platform PQDT Open (**http://pqdtopen.proquest.com/about.html**).

As mentioned before, research libraries all over the world still have long shelves with enormous amounts of printed theses. How can institutions uncover the ideas in those works and preserve them for the future? In April 2016 Dimity Flanagan and Linda Bennett presented a case study

[2] **http://www.opendoar.org/find.php?format=charts**

on usage and impact of 2,000 theses recently digitized at the London School of Economics in co-operation with ProQuest (Bennett, 2016). Here is a summary of some of their observations:

- Overall download numbers nearly tripled from 2014 to 2015 with the added digitized content.
- Shortly after inclusion of 'new' content downloads per item reach the same level as before.
- Significantly increased overall traffic from social media platforms, including for some of the digitized theses.
- Seeking individual author permission too labour-intensive; a take down policy introduced instead.
- Correlation of usage and citations could not be proved.

The effect of increased usage for old material can also be seen in *ProQuest Dissertations and Theses Global*. Part of the LSE project was to make the newly digitized theses from the LSE available in PQDT in order to gain even more global visibility. As a consequence, in July 2016 an LSE PhD thesis from 1971 digitized as part of this project reached the top 25 dissertation and theses ranking published by ProQuest.[3]

Overall, there seem to be only few current studies that look at the development of usage of ETDS (and print theses) in particular. The impact of theses for future research measured by citations is even more complex to analyse. In order to help reveal connections between research topics as well as between authors, PQDT displays links to cited resources, as well as 'cited by' links. Features of this kind allow for more meaningful analysis of graduate works impact in the future.

ETDs as a Text and Data Mining Resource

The content of individual dissertations and theses can serve as an extremely useful resource for other academics or PhD students in the same field, as it often offers more detail and a more comprehensive literature review than journal articles. In recent times, ETDs as whole sets of big data have received increasing attention as objects of research with new text and data mining methods. At ProQuest our dissertation database grows by 130,000+ items per year, representing a critical mass of valuable academic output covering a wide time range, especially for the United States. This data is an interesting resource for researchers investigating the development and the volume of graduate works in certain subject fields, career paths of graduate students or the general development of language.

As an example may serve Benjamin Miller's doctoral dissertation at CUNY (City University New York) on composition and rhetoric, examining keywords and subject terms of other dissertations. The author analyzed thousands of selected dissertation records including their full texts provided to him by ProQuest (Miller, 2015).

In 2016 a study from the University of Kansas using ProQuest dissertation data as well found in the field of Psychology 80% of the dissertations (PhD theses) could not be linked to peer-

[3] **http://www.proquest.com/products-services/dissertations/ProQuest-Most-Accessed-Dissertations-and-Theses-July-2016.html**

review articles, and remain unpublished after seven years. First of all, this study demonstrates the suitability and importance of dissertation data for research of this kind. Secondly, the study's result tells us again, how important it is not to forget dissertation and theses content when researching specific scientific topics, as journal articles might not give the full picture in all subject fields.

Joachim Schöpfel and his colleagues from Lille examined "dissertations as data" in general and presented their findings at the 19th International Symposium on ETDs in Lille, France. They looked at both the dissertations' texts and the corresponding supplementary material consisting of other data types. All this can be defined as integrated dissertation data and be turned into an object of research. At the same time, the topic highlights the need of best practices regarding deposit and publishing policies for research data, not only in the area of dissertation and theses, but also in scholarly communication in general.

Regarding text and data mining of scientific content, existing grey areas and country specific inconsistencies concerning the legal basis and possible conflict with copyrights should be addressed in order to create a solid environment for better research opportunities.

Conclusion

In the examples and case studies mentioned, I have tried to highlight how relevant graduate works as part of the grey literature are for the researcher as reader and end-user. Nevertheless, with an ever growing number of online available documents, it can turn out to be a tedious task to discover the needed piece of research in the giant information 'haystack'.

For authors of dissertations or theses and for their corresponding institutions multiple channels should be explored and utilized in order to ensure maximum visibility and discoverability of the academic output. Tracking of usage and impact will probably get more attention in the future with new tools and, hopefully, standardized methods.

Peter Suber summarized in *Open Access*: "If a university requires theses and dissertations to be new and significant works of scholarship, then it ought to expect them to be made public, just as it expects new and significant scholarship by faculty to be made public. Sharing theses and dissertations that meet the school’s high standard reflects well on the institution and benefits other researchers in the field. The university mission to advance research by young scholars has two steps, not one. First, help students produce good work, and then help others find, use, and build on that good work." (Suber, 2012).

References

BENNETT, Linda and Dimity FLANAGAN, 2016. Measuring the impact of digitized theses: a case study from the London School of Economics. *Insights*. **29**(2): 111–119. DOI: **http://dx.doi.org/10.1629/uksg.300**.

CORBETT, Hillary, 2016. Out of the Archives and Into the World: ETDs and the Consequences of Openness. In: SMITH, Kevin L. and Katherine A. DICKSON, eds. *Open Access and the Future of Scholarly Communication: Implementation*. Available from **http://hdl.handle.net/2047/D20216388**.

EVANS, Spencer C. et al, 2016. *The Large Majority of Dissertation Research in Psychology Goes Unpublished*. Poster presented at the 28th Annual Convention of the Association for Psychological Science, Chicago, IL, May 2016.

ILVA, Jyrki, 2015. *Repositories and ETDs – a success story from Finland*. Presentation at Open Repositories, Indianapolis. Available from **http://urn.fi/URN:NBN:fi-fe2015061110223**.

MCLEAN, Austin, 2016. *Current Usage of Dissertations: A Global Perspective*. Presentation at the Council of Graduate Schools - Future of the Dissertation Workshop, Washington D.C., January 2016. Available from **http://cgsnet.org/sites/default/files/DissFwd_Print%20All%20Papers.pdf**.

MILLER, Benjamin M., 2015. *The Making of Knowledge-Makers in Composition: A Distant Reading of Dissertations*. CUNY Academic Works. Available from **http://academicworks.cuny.edu/gc_etds/1056**.

PROST, Hélène, Amélie BUIRETTE and Amélie HALIPRÉ, 2016. *French Electronic Theses and Dissertations in Europe – A Scientometric Approach*. Poster Presentation at the 19th International Symposium on Electronic Theses and Dissertations, Lille, July 12, 2016. Available from **https://etd2016.sciencesconf.org/98998**.

SCHOEPFEL, Joachim, Eric KERGOSIEN, Stéphane CHAUDIRON and Bernard JACQUEMIN, 2016. *Dissertations as Data*. Presentation at the 19th International Symposium on Electronic Theses and Dissertations, Lille, July 13, 2016. Abstract available from **https://etd2016.sciencesconf.org/92328/**.

SUBER, Peter, 2012. *Open Access*. Cambridge: The MIT Press. The MIT Press Essential Knowledge Series. ISBN 978-0-262-51763-8. Available from: **https://mitpress.mit.edu/sites/default/files/9780262517638_Open_Access_PDF_Version.pdf**.