September 23, 2016

Harvesting ETD Metadata from Institutional Repositories to OCLC: Approaches and Barriers to Implementation

Marielle Veve, *University of North Florida*

This is the Author's Original Manuscript of an article published by Taylor & Francis. Citation: Marielle Veve (2016) Harvesting ETD Metadata from Institutional Repositories to OCLC: Approaches and Barriers to Implementation, Journal of Library Metadata, 16:2, 69-79. Available online: https://doi.org/10.1080/1051712X.2016.1215730

Harvesting ETD Metadata from Institutional Repositories to OCLC: Approaches and Barriers to Implementation

Author: Marielle Veve (m.veve@unf.edu)
Metadata Librarian, University of North Florida

Keywords:

ETD metadata, OCLC, institutional repositories, semi-automated harvesting approaches, automated harvesting approaches, OCLC Digital Gateway, WorldCat Digital Collection Gateway, ProQuest, MarcEdit OAI Harvester

Abstract:

This article discusses some of the most popular automated and semi-automated approaches currently used in libraries to harvest electronic theses and dissertations' (ETD) metadata from institutional repositories (IR) to the Online Computer Library Center (OCLC). The approaches presented are divided into four main sections: (1) completely automated approaches, (2) semi-automated approaches that rely on ProQuest Services, (3) semi-automated approaches that rely on highly technical tools, and (4) semi-automated approaches that rely on the MarcEdit OAI Harvester. An analysis, including limitations and barriers to implementation, is provided and solely based on case studies presented throughout the library literature, on presentations, and on the author's and other institutions' experiences. Recommendations for future research and suggestions for improvements are provided.

Introduction:

In order to avoid duplication of efforts and to increase accessibility, many academic libraries have used automated or semi-automated methods to harvest electronic theses and dissertations' (ETD) metadata from their institutional repository (IR) to the Online Computer Library Center (OCLC). Many of the strengths and some of the weaknesses of these approaches have been mentioned throughout the library literature. However, little has been documented on the barriers and obstacles most likely to be encountered in their implementation. One of these completely automated approaches is the WorldCat Digital Collection Gateway. The approach does not seem to work properly with some of the metadata feeds that originate from IR software that are not OCLC proprietary. By the same token, the semi-automated approaches presented either require a high level of programming expertise to be properly implemented, seem to only cater to the particular needs of the DSpace IR community, or completely rely on the ProQuest Services to generate the initial ETD metadata, a service not every library subscribes to.

The following paper discusses these automatic and semi-automatic approaches in more detail, including their limitations and barriers to implementation. Recommendations for future research and suggestions for possible improvements are provided.

Literature Review: Approaches to Harvest ETD metadata from Institutional Repositories to OCLC

The library literature contains many examples of semi-automated and some completely automated approaches to harvest institutional repository (IR) ETD metadata that is exposed to the Open Archives Initiative Protocol for Meta-data Harvesting (OAI-PMH), and transform it into Machine-Readable Cata-loging (MARC) records that can be later sent to OCLC. The following were found throughout the library literature as of April 2016.

Completely-Automated Approaches: WorldCat Digital Collection Gateway

The library literature presents only one completely automated approach to harvest metadata from an IR to OCLC that is user-friendly and can be used without deep knowledge of programming. This approach is the WorldCat Digital Collection Gateway (OCLC Digital Gateway). The tool claims in its website that it can automatically harvest metadata from "all OAI-PMH com-pliant repositories" (OCLC, 2015) to OCLC in any of the following three schemas: Dublin Core (Unqualified DC), Qualified Dublin Core (Qualified DC), and European Semantic Elements (OCLC, 2012).

Nevertheless, when trying to be implemented with IR software that is not OCLC proprietary, such as the BePress's Digital Commons or DSpace, the OCLC Digital Gateway proved it could not harvest Qualified DC feeds as claimed, but only Unqualified DC. That has been the experience of this author's institution as well as that of three other Digital Commons institutional repository users: Trinity College, University of Iowa, and Nova Southeastern University (Gibney, 2016; Harrell, 2016; Robertson, 2016). Likewise Virginia Tech, who uses another non-OCLC proprietary software, DSpace IR, reported encountering the same situation when attempting to map Qualified DC with the OCLC Digital Gateway. In Virginia Tech's situation it reported getting an error that said: "The first level metadata cannot be validated and nothing went in" (Finn, 2015). The reason for OCLC Digital Gateway's inability to harvest in Qualified DC from these IR software systems is partially caused by the fact that the Qualified DC metadata format used by both Digital Commons and DSpace is different from the metadata schema specified by the Dublin Core Metadata Initiative (DCMI). Both use their own proprietary schema, which format the usual Qualified DC elements differently with their own new elements and refinements, and this certainly interferes with services based on internationally accepted schemata and protocols, such as the OCLC's Digital Gateway.

Even though harvesting from an Unqualified DC feed is an option if the OCLC Digital Gateway service is still used, it is not recommended for ETD metadata. MARC records generated from these feeds are of low quality, do not seem to be reliant on the current Resource Description and Access (RDA) standards, and display a limited number of fields that can barely be customized with the OCLC Digital Gateway metadata map.

As of today, the only documented case in the library literature wherein the OCLC Digital Gateway was able to successfully harvest from an Unquali-fied DC feed is the case of an institution that used another OCLC proprietary IR software: CONTENTdm (Mower, Walters, & McIntyre, 2010). Otherwise, nothing has been found in the literature that can provide evidence that OCLC Digital Gateway is successful in harvesting Qualified DC feeds that originate from IR software that is not OCLC proprietary. The only positive side found to this approach when trying to harvest ETD metadata is that it is free and saves time and automatic synchronizations can be scheduled to OCLC.

Semi-Automated Approaches that Rely on ProQuest Services

The library literature contains many examples of semi-automated approaches to harvest ETD metadata from IRs to OCLC that rely on the ProQuest Ser-vices to generate part of the initial ETD metadata, which is later sent and deposited to the institutions' repositories. Most of these approaches seem to derive in some way from the original workflow presented by Averkamp and Lee (2009) but with local customizations integrated to meet individ-ual institutional needs. Examples of these workflows are the University of Massachusetts Amherst (UMASS) case (Banach & Smith, 2010), that is a Dig-ital Commons IR user and the Florida State University (FSU) case (Glerum, 2014), a DSpace IR user. In both cases, students fill in a metadata template while submitting their ETDs in ProQuest, then additional metadata is added by ProQuest and sent back to the libraries to be deposited in their IRs.

This workflow presents a couple of issues and concerns. First, it is not only extremely long and tedious but also outdated and currently unnecessary given the rise of ETD self-archiving in IRs. Second, in this approach ProQuest holds complete control of the design of the metadata template students fill, giving libraries no say over which metadata fields are chosen to describe these important unique materials. This point is also reiterated by Robertson and Routh (2010) when they mentioned: “If we didn’t rely on ProQuest, we would hopefully have more control over the form and students could even submit directly to the repository.”

Third, the practice of relying on ProQuest as metadata mediator for ETDs when students and library staff can provide the same, if not better, quality metadata is a practice some institutions and scholars have begun to question (Clement, 2013; McMillan, Halbert, & Stark, 2013) while others, such as Wichita State University (WSU), already dropped the service by mak-ing the decision to “discontinue the ProQuest service after WSU’s DSpace was implemented successfully” (Deng & Reese, 2009). The claim that Pro-Quest metadata has “not been up to the standard desired” is also brought by Middleton, Dean, & Gilbertson (2015) when they say: “One of the data elements provided by ProQuest that were less than ideal were the subject headings, as the terms were generated from a proprietary controlled vocab-ulary created and maintained by ProQuest. This controlled vocabulary did not match the Library of Congress subject headings used in the local catalog. Therefore, the library continues to provide their own subject analysis using the Library of Congress' subject headings."

Lastly, providing full text access to ETDs in both the ProQuest database and the local IR can divert potential traffic that could be directed toward the IR, thus affecting local IR statistics. A study performed by Auburn University Libraries compared the amount of views performed to the same ETD material that originated from ProQuest versus the ones that originated from the IR (Coates, 2014). The study compared views performed during 2 consecutive years, 2012 and 2013. Results showed that the amount of views generated, for exactly the same material, from the IR were significantly higher than the views generated from the ProQuest database during the 2 consecutive years studied. Anyhow, despite these findings, overall generalizations on the effect of hosting ETDs in both places cannot be made at this point, as Coates' example is only a case study and data provided is limited. If broader generalizations want to be made, then further research on the subject needs to be performed at a larger scale.

Semi-Automated Approaches that Rely on Highly Technical Tools

Other semi-automated approaches mentioned throughout the library litera-ture rely on highly technical tools to generate the initial ETD metadata or to extract them from other information sources. Some of these documented approaches rely on a Perl program to extract metadata directly from an IR or OAI feeds and later transform them into MARC records (Deng & Reese, 2009). Others rely on metadata-generating tools that apply techniques such as automatic indexing, text and data mining, or extrinsic data auto-generation (Park & Brenza, 2015) to generate initial metadata.

The problem with these tools is that they require a high level of technical or programming expertise to be properly implemented, skills the average cataloger does not possess and relying on support from a library information technology team (IT) is something many small libraries cannot afford. This problem is also mentioned by Park and Brenza (2015, p. 39) when they pointed out: "The high degree of technical knowledge needed to implement these tools means that many libraries and other institutions may not have resources to begin implementing them, let alone incorporating them into the daily workflows of the metadata creation process."

In addition to the lack of technical support, Park and Brenza (2015, p. 39) mentioned other significant barriers to the implementation of these tools in libraries. For example, she questions these tools' sustainable development in the long term as their application still "remains relatively untested in real-world scenarios." Also she states that "most of these tools only address part of the problem [by] providing solutions to the metadata generation of one or a few bibliographic elements but not the full range of elements. This means that for libraries to truly have a comprehensive tool set for the semi-automatic generation of metadata records, significant local efforts will be required to integrate the various tools into a working whole" (Park & Brenza, 2015).

Semi-Automated Approaches that Rely on MarcEdit OAI Harvester

The other semi-automated approaches mentioned in the library literature start the generation of ETD metadata within the DSpace IR software and expose them to the OAI in Qualified DC schema, where the metadata is harvested and transformed into MARC records with the assistance of an XSLT (Exten-sible Stylesheet Language Transformations) stylesheet and the MarcEdit OAI Harvester. At the end, the generated MARC records are imported to OCLC. According to Veve (2016), this is the most successful approach so far to har-vest ETD metadata as the MarcEdit OAI Harvester tool "has proved over time it can harvest metadata in any proprietary schema and from any IR software without major complications. Designed with the library community in mind, this tool is user-friendly and can be implemented by anyone with or without a programming expertise; plus its customer service is excellent. Emails are answered within the same day by the tool's creator or by other members in the discussion list. In addition, the tool has a huge community of users and followers who support its development in the long-term."

An example of a successful implementation of this workflow to harvest ETD metadata is presented by Deng and Reese (2009, p. 249), where they ap-plied it to two different institutions that use the DSpace IR software: Oregon State University (OSU) and Wichita State University (WSU). Anyhow, even though the approach proved to work smoothly and without major customiza-tions needed, it was only tested with the DSpace IR. The same happens with the other cases documented in the library literature—they address only this workflow's implementation to the DSpace IR software.

When Veve (2016) attempted to implement Deng and Reese's work-flow to another non-DSpace IR software, such as the BePress's Digital Commons, some obstacles were encountered that prevented its full imple-mentation. These obstacles were the result of differences in software capabil-ities, schema used, and element display that exist between the two software systems, DSpace and Digital Commons, and had nothing to do with the capabilities of the MarcEdit OAI Harvester tool. Examples of the obstacles encountered by Veve (2016) in her case study when trying to implement Deng and Reese's workflow to the Digital Commons software were:

1. The Qualified DC metadata exposed to the OAI by DSpace is different from the Qualified DC metadata exposed by Digital Commons. This is because, first, each institution makes its own internal decisions on which metadata elements to include in its IR metadata template and which ones to expose to the OAI and, second, because the capability of each software to display some of the metadata elements is different. An example of how each institution can make different internal decisions on which metadata elements to include in their metadata upload form or display to the OAI is illustrated in WSU and OSU cases, wherein even though they both use the same IR software (DSpace), each one made different decisions concerning how to handle the controlled and uncontrolled subject terms in their OAI feeds. On the one hand, WSU decided to display only a keyword field in its OAI feeds (Deng & Reese, 2009), while OSU decided to display both controlled and uncontrolled fields in its OIA feeds in separate fields (Reese, 2009). An example of how software capabilities to display metadata can differ between two different IR software systems is reflected in how Digital Commons and DSpace display the advisors fields differently. On one hand DSpace has the ability to display the advisors' names in inverse order (Deng, Matveyeva, & Wang, 2008) while Digital Commons cannot (Digital Commons representative, personal communication with author, October 13, 2015).

2. The DSpace IR uses a set of Qualified DC elements different from the ones used by Digital Commons IR to map their ETDs to the OAI. DSpace uses elements from the default Qualified DC in the DSpace metadata reg-istry (DSpace, 2015) while Digital Commons uses the BePress proprietary schema for Qualified DC (BePress, 2015).
3. The XSLT stylesheet used in the Deng and Reese's case study uses elements from the DSpace proprietary schema and follows the older cataloging content rule standards, the Anglo-American Cataloging Rules, second edition (AACR2). In order to be implemented to the Digital Com-mons software, this XSLT stylesheet would need to be adjusted to the Be-Press proprietary schema. Additionally, it needs to get adjusted to reflect the current Resource Description and Access (RDA) content standards.
4. Given that the initial metadata generated by DSpace will not be the same as the metadata generated by Digital Commons, it is expected that a different set of final edits will need to be made to the MARC records.

Even with these major obstacles, Deng and Reese's approach still presents a good set of practical tools and ideas that can be extended to other workflows that do not use the DSpace IR, but only if they are cus-tomized to address these differences. The idea of using a customized ETD XSLT stylesheet and the MarcEdit OAI Harvester are examples of this. The applicability of this workflow to other IR systems if customized was pointed out by the authors (Deng & Reese, 2009) when they said that even though "their experience can be applied to metadata mapping and transformation between different systems in general, the two institutions' experience demon-strates that one single crosswalk and transformation stylesheet will not meet all needs and application-specific mapping is needed."

Conclusion and Future Recommendations

With the increase in the amount of ETDs in IRs, many academic libraries have looked for ways to automatically or semi-automatically harvest metadata for these resources from the IR to OCLC. Most of the approaches presented in the library literature seem to offer a viable solution to the harvesting process, but their successful implementation to IRs will depend on various factors, first, the IR software used (some approaches can work with any IR while others cannot, and some can work better with some systems than others). The second factor is the technical expertise required for implementation (some approaches require a high level of technological skills to be success-fully implemented while others do not). A third factor is the flexibility of these approaches to integrate customizations (some approaches hold more control over the metadata generated while others provide more flexibility). A fourth factor is the quality of the metadata and the speed with which it can be produced (some approaches can produce better quality metadata than others, while others are faster). Given that each institution will have its own set of priorities when deciding on which harvesting approach to implement, these factors will have to be taken into consideration.

The OCLC Digital Gateway approach proved it can only harvest from IRs that adhere to the unaltered DCMI schema, so IRs that do not adhere to these standards and use proprietary schema will have trouble using this harvesting tool. Examples of these systems are DSpace and Digital Commons. The case studies displayed in this paper showed that even though IR systems that use proprietary schema will not be able to use the Gateway to harvest in Qualified DC, Unqualified DC may still be an option. However, harvesting ETD metadata from an Unqualified DC feed is not recommended as ETD metadata needs to be very detailed and requires a high level of granularity, something an Unqualified DC feed cannot provide. Even though harvesting metadata from Unqualified DC feeds is not recommended for ETD metadata, it can still be somehow useful for other harvesting purposes such as for collections that do not require a high level of granularity in their metadata records.

In the future, the only way IR systems that use proprietary schema would be able to use OCLC Digital Gateway to harvest Qualified DC is if these noncompliant IR systems decided to adopt the unaltered DCMI schema in addition to the proprietary schema they use. Another option would be if, vice versa, OCLC adopted the proprietary schema used by these IRs in addition to the unaltered DCMI schema they already use. Nonetheless, the probability of any of these solutions becoming a reality is very low. Only systems that adhere to the unaltered DCMI schema, such as OCLC's CONTENTdm, will be able to harvest Qualified DC from the OCLC Digital Gateway.

Given to the growth of self-archiving of ETDs in IRs, the ProQuest ap-proach seems to be redundant and unnecessary these days. Authors already provide their own metadata and library staff is better prepared than ProQuest Services to provide high quality metadata, such as controlled vocabularies. Still, the ProQuest approach may be an option for libraries that do not have enough technical services staff to catalog these items or that wish to de-crease the time staff spends in these resources. Although looking from the long and tedious ProQuest workflows followed in the case studies presented in this paper, saving time may be hardly the case. In addition, more large-scale studies are needed to assess the value, if any, of hosting ETDs in both ProQuest and the IR.

Applying some of the highly technical tools mentioned in the library literature to extract and harvest metadata from an IR may be an option for those libraries that wish to save time in the metadata-generation process. However, some factors will need to be considered before integrating any of these tools into a whole harvesting workflow. One of these factors is the level of technological skill required to implement a tool properly. This is very important as many catalogers do not have a high level of programming expertise and relying on an IT team for implementation may not be an option for some libraries. Even in cases in which it may be an option, relying completely on an IT team to implement and support a harvesting tool may not be the best decision as many libraries have a high staff turnover. Another factor to take into consideration before integrating any of these tools into a harvesting workflow is their sustainable development in the long term. As some of these relatively new tools may be there for a while, they can disappear at any moment, leaving implementers with an unfinished tool. A final factor that needs to be considered before adopting any of these tools is the tool's applicability to library settings. For tools that have been relatively untested in real-world scenarios, they would need to be tested in a library setting to assess their effectiveness with managing ETD metadata.

Lastly, from all the approaches presented here, Deng and Reese's ap-proach seems to be the most viable solution for those IR that use proprietary schema without compromising on the quality of the ETD metadata pro-duced. It is the middle ground between automation and flexibility because it still provides some level of automation (MarcEdit OAI Harvester) while still providing customization power to its users (with XSLT stylesheets).

The only time major obstacles will be encountered when implementing this approach is when another non-DSpace IR is used. These obstacles stem from differences in software capabilities, schemata used, and element display that exist between different IR systems. But as mentioned before, these ob-stacles can be overcome if additional customizations are performed at each stage of the workflow to address them. This though may be an option not everybody will want to pursue as it may be time consuming and requires some level of familiarity with XSLT stylesheets and interoperability functions that some staff may not have.

In conclusion, this paper shows that no harvesting approach, automated or semi-automated, could be applied in its entirety without encountering some type of barrier in the implementation process. For that reason further customizations on behalf of the implementer will be required.

References

Averkamp, S., & Lee, J. (2009). Repurposing ProQuest metadata for batch ingesting ETDs into an institutional repository. *Code4Lib Journal, 7.* Retrieved from http://journal.code4lib.org/articles/1647

Banach, M., & Smith, C. (March 2010). *Managing ETDs with Digital Commons: A case study at UMass Amherst.* Paper presented at the Digital Commons Webinar. Retrieved from http://digitalcommons.bepress.com/repository-research/18/

BePress. (2015). *BePress Proprietary Schema for Qualified Dublin Core*. Retrieved from http://www.bepress.com/assets/xsd/oai_qualified_dc.xsd

Clement, G. (2013, April 6). Graduate students re-FUSE! Free
US ETDs [Blog]. Retrieved from https://sites.tdl.org/fuse/?page_id=128

Coates, M. (September 2014). *Comparing apples to apples? Examining user behavior for an open-access ETDs collection vs. ProQuest*. Poster presented at the annual meeting for the United States Electronic Thesis and Dissertation Association, Orlando, Florida. Retrieved from https://conferences.tdl.org/usetda/index.php/USETDA/USETDA2014/paper/view/722/355

Deng, S., Matveyeva, S. & Wang, T.M. (October 2008). *Customized mapping and metadata transfer from DSpace/SOAR to OCLC and Voyager.* Paper presented at the Ex-Libris Southcentral Users Group (ELSUG) Meeting, Wichita, KS. Retrieved from http://soar.wichita.edu/handle/10057/1573

Deng, S., & Reese, T. (2009). Customized mapping and metadata transfer from DSpace to OCLC to improve ETD work flow. *New Library World, 110* (5/6), 255. doi: 10.1108/03074800910954271

DSpace. (2015). *Default Dublin Core Metadata Registry (DC).* Retrieved from https://wiki.duraspace.org/display/DSDOC4x/Metadata+and+Bitstream+Format+Registries

Finn, M., McIntyre, S., & Wynne, S. C. (2015, September 23). Any OCLC Digital Gateway experiences? Message posted to http://lists.monarchos.com/listinfo.cgi/metadatalibrarians-monarchos.com

Gibney, M. (2016, April 29). Experience harvesting DC records into Worldcat Digital Collections Gateway? Message posted at https://groups.google.com/forum/#!search/digital$20commons$20google$20group

Glerum, M.A. (September 2014). *Efficiencies for quality control of repurposed ETD metadata.* Poster presented at the annual meeting for the United States Electronic Thesis and Dissertation Association, Orlando, Florida. Retrieved from https://conferences.tdl.org/usetda/index.php/USETDA/USETDA2014/paper/view/743

Harrell, A. (2016, April 27). Experience harvesting DC records into Worldcat Digital Collections Gateway? Message posted at https://groups.google.com/forum/#!search/digital$20commons$20google$20group

McMillan, G., Halbert, M. & Stark, S. (July 2013). *Comprehensive study of National ETD practices*. Paper presented at the annual meeting for the United States Electronic Thesis and Dissertation Association, Claremont, California. Retrieved from https://conferences.tdl.org/usetda/index.php/USETDA/USETDA2013/paper/view/666/318

Middleton, C. C., Dean, J. W. & Gilbertson, M. A. (2015). A process for the original cataloging of theses and dissertations. *Cataloging & Classification Quarterly, 53,* (2), 234-246. doi: 10.1080/01639374.2014.971997

Mower, A., Walters, C. & McIntyre, S. (May 2010). *Dublin Core application profiles and the OCLC Digital Gateway: New tools for improving discoverability of digital collections*. Paper presented at the 2010 Utah Library Association Annual Conference, St. George, Utah. Retrieved from http://digitalcommons.usu.edu/lib_present/24/

OCLC. (August 2012). *The WorldCat Digital Collection Gateway tutorial*. Retrieved from https://www.oclc.org/content/dam/oclc/gateway/gettingstarted/tutorial.pdf

OCLC. (2015). *WorldCat Digital Collection Gateway website*. Retrieved from https://www.oclc.org/digital-gateway.en.html

Park, J.R., & Brenza, A. (2015). Evaluation of semi-automatic metadata generation tools: A survey of the current state of the art. *Information Technology and Libraries, 34*, (3), 24. doi: 10.6017/ital.v34i3.5889

Reese, T. (2009). Automated metadata harvesting: Low-barrier MARC record generation from OAI-PMH repository stores using MarcEdit. *Library Resources and Technical Services, 53*, (2), 129. doi: http://dx.doi.org/10.5860/lrts.53n2.121

Robertson, W. (2016, April 28). Experience harvesting DC records into Worldcat Digital Collections Gateway? Message posted at https://groups.google.com/forum/#!search/digital$20commons$20google$20group

Robertson, W. & Routh, R. (April 2010). *Light on ETD's: Out from the shadows.* Paper presented at the annual meeting for the ILA/ACRL Spring Conference, Cedar Rapids, Iowa. Retrieved from http://ir.uiowa.edu/lib_pubs/52/

Veve, M. (2016). From Digital Commons to OCLC: A tailored approach for harvesting and transforming ETD metadata into high-quality records. *Code4Lib Journal, 33*. In Press, coming in July 2016.