The state of OAI-PMH repositories in Canadian Universities

Frédéric Piedboeuf [†], **Guillaume Le Berre** [†], **David Alfonso-Hermelo** [†],
Olivier Charbonneau [‡], **Philippe Langlais** [†]
[†] DIRO, University of Montreal
[‡] Concordia University
Montreal, QC, Canada
{frederic.piedboeuf, guillaume.le.berre, philippe.langlais}@umontreal.ca,
davidalfonsohermelo@gmail.com, o.charbonneau@concordia.ca

Abstract

This article presents a study of the current state of Universities Institutional Repositories (UIRs) in Canada. UIRs are vital to sharing information and documents, mainly Electronic Thesis and Dissertation (ETDs), and theoretically allow anyone, anywhere, to access the documents contained within the repository. Despite calls for consistent and shareable metadata in these repositories, our literature review shows inconsistencies in UIRs, including incorrect use of metadata fields and the omission of crucial information, rendering the systematic analysis of UIR complex. Nonetheless, we collected the data of 57 Canadian UIRs with the aim of analyzing Canadian data and to assess the quality of its UIRs. This was surprisingly difficult due to the lack of information about the UIRs, and we attempt to ease future collection efforts by organizing vital information which are difficult to find, starting from addresses of UIRs. We furthermore present and analyze the main characteristics of the UIRs we managed to collect, using this dataset to create recommendations for future practitioners.

Keywords: metadata; canada; institutional repositories; universities; OAI-PMH protocol; scholarly communication; digital scholarship; electronic thesis and dissertation; EDT

1 Introduction

The term ETD, or *Electronic Thesis or Dissertation*, refers to theses which are present in digital form, often online and in open access format. ETDs find themselves in document repositories of universities, with the associated metadata commonly entered by authors themselves, which saves considerable human and financial resources of cataloguing (Robinson et al., 2016; Reeves, 2007). Following open-access papers from journal and conferences, ETDs are estimated to be the most important document types in open archives (Schöpfel, 2013). Furthermore, the existence of centralized repositories of research publications eases the analysis of research trends over time (Gökmen et al., 2017), and has also allowed the emergence of *bibliomining*, the study of documentalist tendencies using mined data from repositories (Siguenza-Guzman et al., 2015).

Canada's presence in the international research scene is a large one, which is illustrated by its coming 7th on the Nature Index[1], despite its population being low in contrast to the other countries from the top 10. In addition, growth in Canada Institutional Repositories (IRs) has been fast; Loan and Shah (2020) report that in 2020, Canada did not even get into the ranking of countries with the most IRs, meaning it had less than 57. At the time of writing, the current number of listed repositories on OpenDOAR for Canada is 99, including IRs from both universities and other institutions. Canada also has a unique bilingual culture which, as we show in Section 7, creates additional inconsistencies in metadata for ETDs. Understanding how Canadian UIRs are organizing ETDs is therefore important to ensure that UIRs grow toward a state in which they could be easily used by aggregators, researchers, and bibliominers. In general, as noted in (Najjar et al., 2003), the study of the use of repositories allows the development of better tools for sharing information and creating better standards. The last study of UIRs in Canada was in 2010, where Park and Richard (2011) showed inconsistencies in the use of metadata on a subset of 10 Canadian universities. Amid Canada's evolving UIRs, we aim to see if the situation has changed and, if so, in which direction.

[1] https://www.nature.com/articles/d41586-021-03065-6

https://doi.org/10.23106/dcmi.953359643

In this article, we conduct an analysis of all Canadian UIRs. Out of 106 universities, we find 57 repositories that we could extract the content of (we found 62 Canadian UIRs in total, with 5 returning errors during the extraction), and present a study of their main characteristics, commonalities, and differences. We find problems in conformity, not only in the language used for describing elements (e.g.: "*thèse*" vs. "*thesis*"), but also in the use of metadata elements, such as dates in the *Description* element. In this work, we focus our analysis on the use of the OAI_DC metadata format, but we also provide a complete list of the metadata formats supported by each UIRs.

The article is structured as such. In Section 2, we provide a definition of the key concepts essential to this paper, such as IRs, UIRs, and metadata, followed by a literature review in Section 3. Sections 4 and 5 discuss the collection process of UIRs as well as the metadata formats that are used in Canadian UIRs. Then in Sections 6 and 7 we describe our collection effort and analyse the resulting metadata dataset. Finally, we use this knowledge to set forth recommendations in Section 8 and conclude in Section 9.

2 Preliminaries

We begin by providing an overview of key concepts central to this paper, namely IRs, metadata, metadata formats, and methods for accessing them. IRs, or Institutional Repositories, are essentially collections of documents hosted by institutions. Our focus in this study is on University Institutional Repositories (UIRs), which are accessible online through the OAI-PMH protocol (Open Archives Initiative Protocol for Metadata Harvesting).

As its name implies, this protocol allows the collection and sharing of metadata (data describing aspects of a document, such as the publication date, the title, or a summary) using common commands. To do so, the first thing needed is an access point to the repository. For example, Université de Montréal UIR is hosted at `https://papyrus.bib.umontreal.ca/xmlui/`, and the OAI-PMH access point is `https://papyrus.bib.umontreal.ca/oai/`. As we show in Section 4, the URL structure is not regulated or consistent and is often surprisingly difficult to find. From the access point one can, in accordance to the protocol, use additional verbs to gain access to the metadata. In this case, one could use the *Identify* command (`https://papyrus.bib.umontreal.ca/oai/request?verb=Identify`) to get information about the repository.

This paper focuses on metadata access, specifically examining three aspects: metadata formats, software utilized for OAI-PMH metadata management, and the metadata content itself.

When accessing the metadata, users must specify the desired output format. The metadata is always presented in XML, which is a markup language that employs tags to structure data. However, the chosen format determines the specific XML tags employed to convey information. In Section 5, we provide detailed descriptions of the various formats utilized.

The Dublin Core Metadata Element Set (usually shortened as Dublin Core) is a compilation and description of 15 metadata elements (formally and internationally standardized as ISO 15836). It was developed by the non-profit organization *Association for Information Science and Technology* within the Dublin Core Metadata Initiative project. Its goal is to define a minimal inventory of metadata necessary to record digital or physical resources of textual, visual, or auditory nature. It enables the documentation science community to have a common, freely-available and inter-operable specification, as well as a shared metadata vocabulary to describe resources of diverse nature.

It is distinct from the OAI_DC format (also informally called Dublin Core) which is an implementation of the metadata elements described by the Dublin Core Metadata Element Set. It formalizes the abstract metadata concepts into XML metadata tags as described in Section 5. We make in this work the distinction between the two concepts by using the term Dublin Core to refer to the Dublin Core Metadata Element Set and OAI_DC to refer to the OAI_DC format.

3 Literature Review

The management of an OAI-PMH repository is a subject of interest due to its direct impact on the sharing and accessibility of knowledge. In this section, we provide a literature survey focusing on the conformity of IRs as well as the studies of national IRs across the world. For a broader literature review, we refer to Nisa et al. (2021).

https://doi.org/10.23106/dcmi.953359643

Metadata conformity has been extensively studied in the past. Bueno-De-La-Fuente et al. (2009) take a special interest in the dissemination of *learning objects* through OAI-PMH, analyzing 47 repositories listed on OpenDOAR and reporting inconsistencies in their metadata structures, effectively rendering those IRs inefficient for sharing knowledge. Efron (2007) examines the OAI-DC format of 23 repositories from DSPACE (a software used for creating and maintaining IRs) and also conclude that there are inconsistencies in the usage of metadata, with a large portion being left empty and with some repositories returning errors when accessing them. Similarly, Ward (2003) observes the OAI-DC of 100 data providers, finding that about half the tags of OAI-DC were used less than half the time.

This conclusion is shared also by Shreeves et al. (2003), which however add that the general use of metadata attributes varies depending on the communities (museum, academic, or digital libraries). By examining several repositories, Dushay and Hillmann (2003) conclude that they can classify inconsistencies into missing data, confusing data, incorrect data, and insufficient data. Shreeves et al. (2005) establish ways to evaluate the data from repositories, noting that conformity to norms is important for aggregators. They evaluate some databases along different axes (completeness, consistency, and ambiguity), concluding that a lot of the data is not suitable for aggregation. More general reflections on the use and management of metadata have also been reported in Barton et al. (2003); Schopfel et al. (2014).

The idea to explore and analyze a country's IRs to assess their quality is not novel, as it is an important notion to assess the accessibility of the produced outputs. In fact, countries face specific cultural challenges which make individual assessment excessively important. For example, Elahi and Mezbah-ul Islam (2018) study the state of IRs and OAI in Bangladesh, a country with a fairly low literacy rate (61.5% at the time of the study, currently 74.7%[2]) and find that despite several obstacles such as lack of openness or the presence of predatory repositories, the use of IRs was steadily growing and being centralized. Ukwoma and Okafor (2017) review the state of Nigerian IR, concluding that more funding, conferences, and communications about the issue is necessary to assure a good use of IRs.

Rodríguez Bravo and Luisa Alvite Díez (2007); Abadal et al. (2010) explore the state of IR in Spain, noting that it stands behind its neighbouring countries, but also observe that there is significant growth in IRs, giving hope for the future. Similar studies have been conducted in India Singh (2016), Nigeria Christian PhD (2009), Zimbabwe (Kusekwa and Mushowani, 2014), and Bangladesh (Islam and Akter, 2013).

While most studies find things severely needing improvement, with both a lack of participation and adhesion to OAI-PMH standards from universities, in some cases findings are positive. Noticeably, Shin (2010) find the governmental repositories in Korea to be well-built, although they note that more involvement from researchers and universities is required.

As we can see, the exploration of IRs from national bodies, and most notably Universities, can reveal important trends in how effectively the country's research may be disseminated. Most studies find the metadata to be severely lacking, but this is not always the case. It is also interesting to note that the state of IRs is often linked to the geopolitical situation of the country, and Canada's, with its strong research output but also its bilingual culture, proves to be an interesting subject for metadata analysis.

4 Repositories

We first need to find the address of the repository for each university, and its OAI-PMH access point. This step is complex for three reasons: 1. There is no guarantee that a given university has such a repository, 2. OAI-PMH access points can be difficult to find from the repository URL[3], and 3. the information available online about the OAI-PMH URLs is sparse and lacking. The main resource that lists IRs with OAI-PMH enabled is OpenDOAR[4], but similarly to Islam and Akter (2013), we found that many UIRs were missing from the list or had incorrect information (out of the initial UIRs we found, 23 were not listed on OpenDOAR or had incorrect information). Another existing resource is BorealisData[5], which stores theses and documents from universities, but it does not list the URLs for the OAI-PMH access points, and we found that the data listed is severely lacking. For example, it reports only 1425 documents for the UIR of University of Montréal, in contrast with the 27040 documents we collected.

[2] `https://www.thedailystar.net/youth/education/news/bangladeshs-literacy-rate-now-7466-3080701`

[3] The OAI-PMH URL for a repository is typically found in subdirectories of the main URL. While some patterns are common (e.g., `/oai`, `/oai2`), other directory structures are unique and not easily guessable, such as `/server/oai`.

[4] `https://v2.sherpa.ac.uk/view/repository_by_country/Canada.html`

[5] `https://borealisdata.ca`

https://doi.org/10.23106/dcmi.953359643

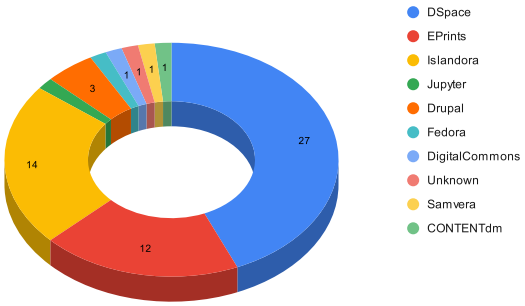

Figure 1: Distribution of the programs used for managing the 62 IRs we found the OAI-PMH access point of.

Using a combination of the resources described above, independent search, and directly contacting universities, we found 62 repository addresses, among which five were not accessible for data collection (either timed out or returned errors while trying to harvest the data). We list all collected URLs in Appendix A, which to our knowledge is the most exhaustive list of Canadian UIRs to date.

An important thing to note is the software used for the creation and management of the UIRs. This software not only impacts which metadata formats are available, but is also central to the set up of the UIRs. Most studies found that a majority of IRs used DSpace (Loan and Shah, 2020), but interestingly, we found that the proportion of DSpace repositories in Canadian UIRs is smaller. We believe that this is due to the large presence of Islandora, which is a Canadian Open Source software for setting up and managing IRs. A visualization of the softwares used in the UIRs we collected is presented in Figure 1 and the complete list can be found in Appendix A.

5 Metadata formats

In this section, we explore available metadata formats in ETDs repositories. We discovered 35 different metadata prefixes, but no standardized naming conventions. This results in multiple instances of various prefixes describing the same format (e.g. the ETDMS format is often provided under the prefixes *etdms*, *oai_etdms*, *etd-ms*, etc.). Some repositories provide multiple prefix aliases for some common metadata formats, and some providers maintain multiple competing versions of the same metadata format (e.g. *etdms10* vs *etdms11* or *marc* vs *marc21*). After removing aliases, versions, and outlier formats with minimal use, we identify 10 core metadata formats: OAI_DC, QDC, ETDMS, DIDL, MARC, METS, MODS, ORE, RDF, and UKETD_DC. We also excluded the DIM and XOAI formats as they seem to be provided only by DSpace, and we could find little to no documentation on them.

In this section, we briefly present the most commonly used metadata formats. Analysis of those have been done in the past (Burnett et al., 1999), but were not specific to UIRs, and we found wide differences between the metadata formats considered in this work and the ones provided by the UIRs we collected.

OAI_DC (Open Archives Initiative–Dublin Core): As per the OAI-PMH specifications, OAI_DC is the only required metadata format for all OAI-PMH repositories. OAI_DC is based on the Dublin Core Metadata Element[6] Set which is composed of 15 metadata elements: *Contributor*, *Coverage*, *Creator*, *Date*, *Description*, *Format*, *Identifier*, *Language*, *Publisher*, *Relation*, *Rights*, *Source*, *Subject*, *Title*, and *Type*. An example of a simplified metadata record in OAI_DC can be found in Figure 2.

QDC (Qualified Dublin Core) is a modified Dublin Core schema that adds three metadata elements: *Audience*, *Provenance*, and *Rights holder*. Furthermore, QDC specifies a set of qualifiers that can be applied to various elements to either further refine the function of an element (e.g. the qualifier *abstract* to the element *Description*) or specify the encoding scheme used for an element. When using OAI-PMH, the most frequent metadata prefixes for QDC are *qdc*, *dqc* or *oai_qdc*. Forty of the repositories provide QDC.

ETDMS (Electronic Thesis and Dissertation Metadata Standard[7]) is a modification to the OAI_DC standard explicitly made for ETDs. On top of the 15 elements already present in OAI_DC, ETDMS adds a new *Degree*

[6] https://www.dublincore.org/specifications/dublin-core/
[7] https://ndltd.org/wp-content/uploads/2021/04/etd-ms-v1.1.html

https://doi.org/10.23106/dcmi.953359643

```
<metadata>
<oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd
     http://www.w3.org/1999/02/22-rdf-syntax-ns# http://www.openarchives.org/OAI/2.0/rdf.xsd ">
<dc:title>Personality extraction through LinkedIn</dc:title>
<dc:creator>Piedboeuf, Frédéric</dc:creator>
<dc:contributor>Langlais, Philippe</dc:contributor>
<dc:contributor>Lapalme, Guy</dc:contributor>
<dc:subject>Extraction de personalité</dc:subject>
<dc:subject>MBTI</dc:subject>
<dc:subject>DiSC</dc:subject>
<dc:subject>LinkedIn</dc:subject>
<dc:subject>Réseau sociaux</dc:subject>
<dc:subject>Profilage d'auteur</dc:subject>
<dc:subject>Personality Extraction</dc:subject>
<dc:subject>Social Network</dc:subject>
<dc:subject>Author Profiling</dc:subject>
<dc:subject>
Applied Sciences - Computer Science / Sciences appliqués et technologie - Informatique (UMI : 0984)
</dc:subject>
<dc:description>
L'extraction de personnalité sur les réseaux sociaux est un domaine qui n'a que récemment commencé à capturer l'attention
     des chercheurs. La tâche consiste à, en partant d'un corpus de profils d'utilisateurs de réseaux sociaux, être capable
      de classifier leur personnalité correctement, selon un modèle de personnalité tel que défini en psychologie. Ce
     mémoire apporte trois innovations au domaine. Premièrement, la collecte d'un corpus d'utilisateurs LinkedIn.
     Deuxièmement, l'extraction sur deux modèles de personnalités, MBTI et DiSC, l'extraction sur DiSC n'ayant pas encore
     été faite dans le domaine, et finalement, la possibilité de passer d'un modèle de personnalité à l'autre est explorée,
      dans l'idée qu'il serait ainsi possible d'obtenir les résultats de multiples modèles de personnalités en partant d'un
      seul test.
</dc:description>
<dc:date>2019-11-19T19:23:16Z</dc:date>
<dc:date>NO_RESTRICTION</dc:date>
<dc:date>2019-11-19T19:23:16Z</dc:date>
<dc:date>2019-10-30</dc:date>
<dc:date>2019-05</dc:date>
<dc:type rdf:resource="http://purl.org/coar/resource_type/c_46ec" xml:lang="en">thesis</dc:type>
<dc:type rdf:resource="http://purl.org/coar/resource_type/c_46ec" xml:lang="fr">thèse</dc:type>
<dc:identifier>http://hdl.handle.net/1866/22536</dc:identifier>
<dc:language>eng</dc:language>
<dc:format>application/pdf</dc:format>
</oai_dc:dc>
</metadata>
```

Figure 2: A simplified example of an OAI-PMH record (in Dublin Core format), from the repository Papyrus, of the University of Montreal. Notably, we exclude the header which is returned along the metadata elements while querying the OAI-PMH repository.

tag containing one or more subtags among *Name*, *Level*, *Discipline*, and *Grantor*. ETDMS also provides some qualifiers for the existing elements, such as a *role* qualifier for the *Description* element, and attempts to normalize the content of some elements (e.g. only 0, 1 and 2 are supposed to be valid codes for the *Rights* element). The OAI-PMH metadata prefix for ETDMS is most often either *etdms* or *oai_etdms*. Out of the 62 UIRs used in this study, 49 provide the ETDMS format.

DIDL (Digital Item Declaration Language[8]) was established as a mean to be able to describe raw content such as music or images, and is intended as a very general mean to describe content. DIDL is supplied by 38 repositories out of the 62 queried UIRs.

MARC (MAchine Readable Cataloging[9]) is a legacy way to represent bibliographic metadata using a set of numeric codes (for example, the code 245 is used to define a title entry), which incidentally makes MARC one of the less human-readable format presented here. MARC is found in 26 repositories.

METS (Metadata Encoding and Transmission Standard[10]) is maintained by the Library of Congress and was designed to be a wrapper for other metadata formats such as MARC. In essence, while Dublin Core and MODS are intended for cataloguing and giving high level descriptions of a document, METS is intended to give a full description of the object, referencing its structure. Similarly to OAI_DC, METS is intended to structure the XML document, and recommends the use of MODS and Dublin Core for metadata elements, among others (Cantara, 2005). METS is provided by 37 of the queried repositories.

[8] https://www.xml.com/pub/a/2001/05/30/didl.html
[9] https://www.loc.gov/marc/
[10] https://www.loc.gov/standards/mets/

https://doi.org/10.23106/dcmi.953359643

Table 1: Percentage of empty entries for each element of the OAI_DC extracted documents.

Tag	Frequency of absence
Source	99.4%
Coverage	95.0%
Relation	74.4%
Contributor	60.4%
Rights	55.4%
Publisher	49.2%
Subject	32.9%
Format	29.5%
Language	23.1%
Abstract	17.7%
Creator	12.6%
Type	7.0%
Date	3.0%
Identifier	0.3%
Title	0.0005%

MODS (Metadata Object Description Schema[11]) is also maintained by the Library of Congress and is a schema intended for various purposes. It proposes 10 top-level tags and 105 sublevel tags, and is intended to be an overlay on the MARC format to convert the MARC schema into more human-readable tags. We found that 42 of our queried repositories supply this format.

ORE (Object Reuse and Exchange) facilitates aggregations of Web resources, like image collections or HTML documents with interconnected webpages. We found ORE in 26 repositories.

RDF (Resource Description Framework) is a very general structure of XML elements. In fact, all other schemas are built on RDF. By observing the RDF data of universities supporting it, we found that it was used as a wrapper for OAI_DC tags, making it not only superfluous but confusing for users. RDF was found in 37 of the repositories.

UKETD_DC is a format developed by EThOS (Electronic Thesis Online Service) for describing UK theses. Very little information can be found about it and in fact, we could find no official documentation for it. This format is provided by 37 of the queried UIRs.

6 Data collection and cleaning

We now aim to analyse how metadata is presented in the UIRs. We extract all data from all universities using the `pyoai` python package[12], which allows the extraction of multiple types of schemas. Because OAI_DC is the only format that is available through all repositories, we focus on the extraction of this format, even if it is not the most appropriate format for ETDs[13].

Metadata for a given document is organized in an XML structure, as shown in Figure 2, but using `pyoai` allows us to extract automatically the content of the XML file in a more readable format. It is important to note that for the extraction to work correctly, names and document structure need to be standardized. For instance, if a university writes the *Date* element as "*Dote*" (with a typo), the parser will not catch it. Given that the conversion of metadata to XML is generally handled by third party software, the chance of erroneous naming is quite low and in fact when viewing the data we found no such instances.

Extraction over the 57 repositories yielded 728,754 documents. Initially, we examine the percentage of empty elements in the data collected, providing insights into UIRs' prioritization and perceived usefulness of specific elements. Table 1 presents the results, revealing that some elements are largely neglected in the Canadian UIRs, particularly *Source* and *Coverage* (refer to Section 7 for element descriptions).

Shreeves et al. (2005) focus on developing efficient tools for assessing metadata and note that to be complete at least 8 tags should be present: *Title*, *Creator*, *Subject*, *Description*, *Date*, *Format*, *Identifier*, and *Rights*. We found that through our collections, only 17.5% of the extracted documents filled all these tags. This does not

[11] `https://www.loc.gov/standards/mods/`
[12] `https://github.com/infrae/pyoai`
[13] The code is available at `https://github.com/dahrs/oai-pmh_canadian_universities`

https://doi.org/10.23106/dcmi.953359643

indicate that data is absent, as the relevant information may be present in another tag. For example, keyphrases occasionally appear at the end of the abstract in the *Description* tag instead of inside the *Subject* tag.

7 Data analysis

In this section, we discuss how tags are used in the UIRs we collected. We perform a qualitative evaluation using representative samples from each university and include quantitative evaluations when needed.

Contributor: This tag is intended to specify the co-authorship, supervisor participation, affiliation to a laboratory, etc. While it is absent in more than half of the entries, those that use this tag tend to input the correct information for it. In Figure 2 the two contributors reference the co-supervisor of the master thesis.

Coverage: According to guidelines set up by OAI_DC, and followed by the ETDMS format, *Coverage* is expected to be used to represent the legal, spatial, or temporal coverage of the subject. While not used extensively, we found that in most cases, it is used for the geographical coverage. However, even within that use there is no coherence, with some ETDs using place names (e.g. "*Fidji*"), and some using coordinates (e.g. "*49.13, -122.871*"). When used for temporal coverage, dates are, as we describe in more details below, not standards. Furthermore, a proportion of the data in the *Coverage* tags represents nothing, such as "———-". Ultimately, it would be very hard for aggregators at the moment to use this tag for analysis due to not only the various information that this field could contain but also the lack of convention.

Creator: In the same vein, *Creator* is a tag that represents any entity responsible for creating the resource. Most often this matches the author(s), but in some cases it is used to divulge institutions (e.g.: "*Ontario Agricultural College*"). We found that overall this tag seems to be used appropriately, probably due to the high information value of the author names, which universities are deeply invested in cataloguing correctly. Our example in Figure 2 follows the convention of listing the author's name in the *Creator* tag.

Date: is another vital tag for data analysis. As is common practice in librarianship and information science, most dates follow ISO 8601 formats (e.g. "*2019-11-19T19:23:16Z*" or "*2019-05*"), as shown in Appendix C. However, there are exceptions such as the American anglophone standard (month-day-year), or even non-date entries such as "*NO_RESTRICTION*" (see Figure 2). Additionally, it is not uncommon to encounter multiple dates for a single document without clear indications of their specific meanings (submission date, presentation date, graduation date, etc.), as exemplified in Figure 2.

Description: Within the context of ETDs, the *Description* element is most often used for the abstract of the document. However, some universities also provide other information in that field, such as information about the origin of the document and even, in some cases, dates. When multiple abstracts are provided (e.g. when the abstract is provided in multiple languages), these abstracts are sometimes concatenated together in a single *Description* element. There are also cases of single abstracts being split between multiple *Description* elements (e.g. one line per element). *Description* is one of the most high value elements of ETDs, and the bad formatting is certainly a heavy obstacle to its employment by aggregators or analysts. Our running example presents the French abstract of the thesis as the description element.

Format: The *Format* element mostly contains references to the file type of the document the metatada references. Given that UIRs contain mostly ETDs, most records thus contain some variation of "*pdf*" (the most common being "*application/pdf*", as in Figure 2).

Identifier: This field is used to give a link to the underlying document. These can either be direct links to the IR or handles such as Handle.Net or DOI. Our example in Figure 2 gives a link to the document in the OAI-PMH IR.

Language: Since the OAI-PMH protocol only allows the extraction of the metadata and not the document itself, it is quite difficult to analyse if the *Language* tag is correctly assigned. However, as for many other elements, we notice a lack of uniformity between UIRs. We find different naming conventions across the UIRs such as fully spelled out names (e.g.: "*english*", "*french*", "*français*", etc.) as well as various ISO 639 formats ("*en*", "*fra*", "*en_us*", etc.). It is not surprising to note that across the chronology (see Figure 3), records are mostly declared to be in either English or in French. Table 2 shows the distribution of assigned languages among the extracted records. The regular expressions used for normalizing the languages are shown in Appendix 8.

Publisher: As noted in Table 1, most entries do not contain a *Publisher* tag. This is normal, since a lot of theses produced at universities do not go through the usual publisher pipeline but are simply uploaded online upon acceptance, as is reflected in our example. Still, this convention is not adopted universally. For example, we found almost 13K documents which were tagged as "*article*" but had no publishers. Some entries also had

https://doi.org/10.23106/dcmi.953359643

Table 2: Language distribution according to the provided *Language* tag. The English category regroups all records labelled "*english*", "*anglais*", "*en*", "*eng*", "*en_us*", or "*en_ca*" and the French category is composed of all the records labelled either "*french*", "*français*", "*francais*", "*fre*", "*fra*", or "*fr*".

Languages	Frequency
English	77.71%
French	18.24%
Other	4.05%

Table 3: Most frequent content of the *Type* tag as presented in the extracted metadata.

Types	Frequency
Text	20.91%
Thesis	15.82%
Dataset	13.46%
Image	12.17%
Article	9.24%
Journal contribution	9.22%
Figure	8.12%
Master thesis	5.10%
Nonpeerreviewed	3.28%
Thèse	2.86%
Journal article	2.81%
Online resource	2.31%

Table 4: Most frequent types grouped by value categories, using the regular expressions presented in Appendix D.

Types	Frequency
Thesis	30.70%
Media	16.15%
Article	13.51%
Dataset	13.46%
Book	1.28%
Presentation	1.04%
Poster	0.33%

the "*These*" type (in the *Type* tag) and the university as a publisher. This, ultimately, makes the use of the *Publisher* tag difficult for bibliomining or aggregation.

Relation: is generally of little use for ETDs, as it is supposed to represent a "related resource". In fact, this and the *Source* tags are the only two tags for which the ETDMS guidelines give no recommendations of what to input. Still, some uses of the *Relation* tag we could see were, among others, a link pointing to the specific ETD in the UIR, books or chapter names (such as "*DMS-676-IR*"), or what seems like general themes "*Education & culture*". Ultimately, this tag may be useful for searching information, but the lack of guidelines to use it hinders aggregation and data analysis. We note that in our example in Figure 2, the relation tag is also absent.

Rights: Approximately half of the theses appearing in all repositories do not contain any mention to the rights associated with the document. One of the reasons for this omission is the fact that for theses and institutional documents, the rights of property and ownership may be derived from the author(s) of the theses and set on the year of publication; two concepts that already appear in the record.

Source: among all tags in the OAI_DC format, this tag is the most absent overall, and our example in Figure 2 is no exception. We believe this is partly due to the fact that theses do not always have a clear source (for example the URL, university name, financing institution, etc. would all be valid sources) and also because the utility of the *Source* tag is higher for documents of a different nature such as maps, media files or images.

Subject: The *Subject* element is intended to contain the keyphrases for the document. However, the subject's link to the thesis sometimes seems quite distant. This is due to the policy of some universities to automatically generate (often generic) keyphrases. Furthermore, it is not rare to see keyphrases that have been provided by the users with incorrect separators, thus preventing an automatic system from correctly identifying them as distinct keyphrases. This results in several subject elements containing keyphrases of the form "*A;B;C*" for example where "*A*", "*B*", and "*C*" should have been instead identified as 3 separate entries. In our example, we can see both French and English keyphrases being present in the *Subject* tag.

Title: As noted in Table 1, entries for the *Title* tag are rarely empty. Although some other tags might be ambiguous and allow for multiple interpretations, in the case of theses this tag is straightforward, and we can see the classical use in Figure 2.

Type: One of the most present tags, *Type* is potentially also one of the most useful ones for aggregators and for analysis purposes. Here again, there is a lack of uniformity between records of different universities. Our example provides a demonstration of this, with both the "*thesis*" and "*thèse*" information provided. Table 3 presents a list of the most frequent types. In this form, it is difficult to extract any meaningful statistic. We therefore use regular expressions to group some types into more meaningful categories (presented in Appendix D), and the result is shown in Table 4.

https://doi.org/10.23106/dcmi.953359643

Table 5: Amount of theses published per decade, all repositories combined. At the moment of the writing, the 2020 decade is not over (2020-2023), making the numbers from this decade not comparable to the previous.

Year	Theses
1920	500
1930	571
1940	485
1950	882
1960	1745
1970	11476
1980	14235
1990	21008
2000	115014
2010	437199
2020	106621

Our qualitative analysis identifies three categories of metadata problems. Firstly, there are unused tags, such as the absence of a *Publisher* tag for articles or books. In our view, this is the least significant issue among the three, as missing information is easier to handle in analysis and research compared to incorrect information.

The second category is the fact that some information is stored in the wrong tag, such as the common use of the *Description* tag used for storing the date, instead of the *Date* tag. This is more problematic, even if in some cases it can be mitigated. For example, one could attempt to recognize whether the description is a date or not, by checking if it corresponds to the ISO norm. That technique, however, would not eliminate completely the problem, and would still demand considerable efforts in order to obtain clean data.

The last category of problems, which is in our opinion the biggest, is the lack of convention not only between universities, but internally as well. An example of this between universities would be the tagging of doctoral theses (in the *Type* tag) as "*Doctoral Thesis*", "*Thesis*", "*Thèse*", "*Thèse ou mémoire*", etc. An example of this internal dissonance is the fact that some departments will index documents with controlled vocabulary in the *Description* tag, such as the UMI or JEL standards, but not always. This, compounded to all tags and universities, makes it very difficult to obtain clean data either for analysis or aggregation.

8 Recommendations

From our analysis of the repositories, we offer several recommendations that could greatly improve the state of institutional repositories in Canada, namely:

Metadata schemes: Firstly, we suggest eliminating non-essential schemes, like RDF, which are merely wrappers around OAI_DC without added value. While more metadata schemes seems like a good idea, we argue that it makes little sense to have unused standards. Secondly, we advocate for a broader adoption of ETDMS, designed specifically for theses and dissertations. Furthermore, the institutional repositories that

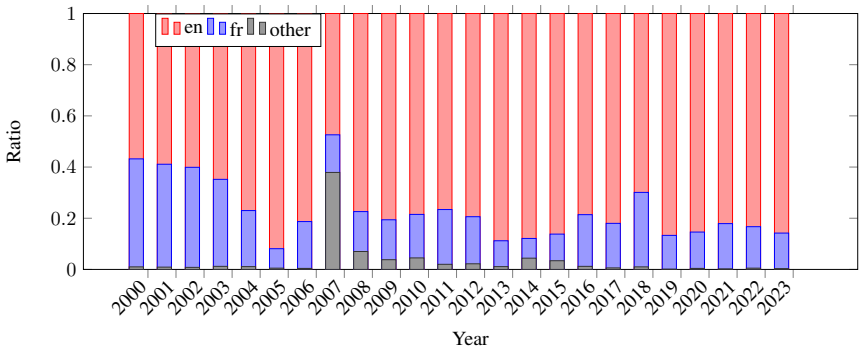

Figure 3: Ratio of theses according to the declared language of the theses, from 2000 to 2023. Please note that at the moment of analysis, the 2023 year is not over (January-March).

https://doi.org/10.23106/dcmi.953359643

provide the ETDMS format should fully familiarize themselves with the metadata schema, use it to its full capacity, follow the defined standards, and keep up to date with updated versions.

Naming convention: We propose the adoption of naming conventions for some tags such as *Type* or *Format*. For example, we recommend a finite set of possible values ("*Thesis*", "*Article*", "*Book*", etc.) for the *Type* tag, which would facilitate the aggregation and identification of documents. Such conventions could be established by independent parties, the most logical representative in this case being Theses Canada.

Language: We encourage a wider use of the *xml:lang* tag attribute to separately identify the language of each occurrence of the *Title*, *Description* (in particular for abstracts), and *Subject* tags. This language attribute should abide by the IETF's BCP 47 specification. We also encourage universities and providers to normalize the content of the *Language* describing the document language, instead of the various norms that have been used to specify the language until now. For example, while some use 2-letter (ISO 639-1) or 3-letter (ISO 639-3) descriptors, others use the language full name (e.g.: "*English*", "*French*", etc.), as shown in Section 7. In order to ease the extraction and language identification of records from multiple universities, we suggest that all institutions follow Theses Canada's recommendations and use the ISO 639-3's 3-letter code[14].

Dates: Dates should be formatted using any kind of ISO 8601 format for a more efficient parsing. The most commonly used metadata formats (including OAI_DC) don't allow specifying what type of date (submission date, release date, etc.) is provided. Therefore, we recommend that only one date should be specified (by default: the release date, when the work is made public). The providers should also prevent the users from freely inputting dates and instead prefer a selection in a calendar. This would avoid erroneous dates like "*These are the annual proceedings of the Grand Lodge A.F. & A.M. of Canada in the Province of Ontario covering (...)*".

9 Conclusion

Efficient metadata sharing is crucial for knowledge dissemination, aggregation, bibliomining, and analysis. Despite Canada's prominent position in international research and its unique cultural context, there is a lack of research on the state of metadata in Canadian UIRs.

In this paper, we address this issue by examining metadata in Canadian UIRs, highlighting their deficiencies in terms of rigour and uniformity. Our contributions include providing an up-to-date list of Canadian university UIRs with OAI-PMH access points, conducting qualitative and quantitative data analysis, and offering recommendations for improving metadata quality.

There is no easy solution to the metadata problem. As noted, user-generated metadata saves both time and money and therefore can be a valuable practice. However, even if authors have been found to generate quality metadata in experimental settings (Greenberg et al., 2002), there is always a possibility of failure from the users due to several factors, such as unclear explanations or lack of examples (Ed Barker, 2003). Additionally, evolving input mechanisms over time create inconsistencies in guidelines across repositories. Still, there is evidence that the repeated calls of researchers for better metadata standards do not fall on deaf ears, as recent years have seen some remediation and cleaning efforts in UIRs (Thompson et al., 2019; Stein et al., 2017).

We hope that this research inspires the adoption of better metadata maintenance practices, not only among Canadian universities but worldwide. By improving existing UIRs through remediation and cleaning, it becomes possible to identify significant research trends, such as thesis languages and subjects of interest. This information is vital for understanding research focus, dissemination, and for informed decision-making by funding institutions.

[14] https://library-archives.canada.ca/eng/services/services-libraries/theses/pages/information-universities.aspx

https://doi.org/10.23106/dcmi.953359643

References

Abadal, E., Anglada, L., Melero, R., Abad, F., Termens, M., and Rodríguez-Gairín, J.-M. (2010). Open access in spain. *Open access in Southern European countries. Madrid: FECYT*, pages 101–115.

Barton, J., Currier, S., and Hey, J. (2003). *Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice*. DCMI.

Bueno-De-La-Fuente, G., Hernández-Pérez, T., Rodríguez-Mateos, D., Méndez-Rodríguez, E. M., and Martín-Galán, B. (2009). Study on the use of metadata for digital learning objects in university institutional repositories (moderi). *Cataloging & classification quarterly*, 47(3-4):262–285.

Burnett, K., Ng, K. B., and Park, S. (1999). A comparison of the two traditions of metadata development. *Journal of the American Society for Information Science*, 50(13):1209–1217.

Cantara, L. (2005). Mets: The metadata encoding and transmission standard. *Cataloging & classification quarterly*, 40(3-4):237–253.

Christian PhD, G. (2009). Issues and challenges to the development of open access institutional repositories in academic and research institutions in nigeria. *Available at SSRN 1323387*.

Dushay, N. and Hillmann, D. I. (2003). Analyzing metadata for effective use and re-use. In *International Conference on Dublin Core and Metadata Applications*, pages pp–161.

Ed Barker, I. (2003). The higher level skills for industry repository.

Efron, M. (2007). Metadata use in oai-compliant institutional repositories.

Elahi, M. H. and Mezbah-ul Islam, M. (2018). Open access repositories of bangladesh: An analysis of the present status. *IFLA journal*, 44(2):132–142.

Gökmen, Ö. F., Uysal, M., Yasar, H., Kirksekiz, A., Güvendi, G. M., and Horzum, M. B. (2017). Methodological trends of the distance education theses published in turkey from 2005 to 2014: A content analysis. *Egitim ve Bilim*, 42(189).

Greenberg, J., Pattuelli, M. C., Parsia, B., and Robertson, W. D. (2002). Author-generated dublin core metadata for web resources: a baseline study in an organization.

Islam, M. A. and Akter, R. (2013). Institutional repositories and open access initiatives in bangladesh: A new paradigm of scholarly communication. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 23(1):3–24.

Kusekwa, L. and Mushowani, A. (2014). The open access landscape in zimbabwe: the case of university libraries in zulc. *Library Hi Tech*, 32(1):69–82.

Loan, F. A. and Shah, U. Y. (2020). Global electronic thesis and dissertation repositories–collection diversity and management issues. *Insights*, 33(1).

Najjar, J., Ternier, S., and Duval, E. (2003). The actual use of metadata in ariadne: an empirical analysis. In *Proceedings of the 3rd Annual ARIADNE Conference*, pages 1–6. Citeseer.

Nisa, N. T., Gulzar, F., Bashir, S., Gul, S., Khan, A., and Bashir, A. (2021). A systematic review of open access institutional repositories (oairs). *Library Philosophy and Practice*, pages 1–18.

Park, E. G. and Richard, M. (2011). Metadata assessment in e-theses and dissertations of canadian institutional repositories. *The Electronic Library*.

Reeves, S. (2007). User-generated metadata for etds: Added value for libraries. In *Proceedings of the tenth international symposium on electronic theses and dissertations*.

Robinson, K., Edmunds, J., and Mattes, S. C. (2016). Leveraging author-supplied metadata, oai-pmh, and xslt to catalog etds: A case study at a large research library. *Library Resources & Technical Services*, 60(3):191–203.

Rodríguez Bravo, B. and Luisa Alvite Díez, M. (2007). E-science and open access repositories in spain. *OCLC Systems & Services: International digital library perspectives*, 23(4):363–371.

Schöpfel, J. (2013). Adding value to electronic theses and dissertations in institutional repositories. *D-lib Magazine*, 19(3/4):n–a.

Schopfel, J., Chaudiron, S., Jacquemin, B., Prost, H., Severo, M., and Thiault, F. (2014). Open access to research data in electronic theses and dissertations: an overview. *Library Hi Tech*.

Shin, E.-J. (2010). The challenges of open access for korea's national repositories. *Interlending & document supply*, 38(4):231–236.

https://doi.org/10.23106/dcmi.953359643

Shreeves, S. L., Kaczmarek, J. S., and Cole, T. W. (2003). Harvesting cultural heritage metadata using the oai protocol. *Library hi tech*.

Shreeves, S. L., Knutson, E. M., Stvilia, B., Palmer, C. L., Twidale, M. B., and Cole, T. W. (2005). Is quality metadata shareable metadata? the implications of local metadata practices for federated collections.

Siguenza-Guzman, L., Saquicela, V., Avila-Ordóñez, E., Vandewalle, J., and Cattrysse, D. (2015). Literature review of data mining applications in academic libraries. *The Journal of Academic Librarianship*, 41(4):499–510.

Singh, P. (2016). Open access repositories in india: Characteristics and future potential. *IFLA journal*, 42(1):16–24.

Stein, A., Applegate, K. J., and Robbins, S. (2017). Achieving and maintaining metadata quality: Toward a sustainable workflow for the ideals institutional repository. *Cataloging & Classification Quarterly*, 55(7-8):644–666.

Thompson, S., Liu, X., Duran, A., and Washington, A. (2019). A case study of etd metadata remediation at the university of houston libraries.

Ukwoma, S. C. and Okafor, V. N. (2017). Institutional repository in nigerian universities: Trends and development. *Library Collections, Acquisitions, & Technical Services*, 40(1-2):46–57.

Ward, J. (2003). A quantitative analysis of unqualified dublin core metadata element set usage within data providers registered with the open archives initiative. In *2003 Joint Conference on Digital Libraries, 2003. Proceedings.*, pages 315–317. IEEE.

https://doi.org/10.23106/dcmi.953359643

A Repository URLs

Table 6: List of all OAI-PMH access points. Stars indicate repositories for which we couldn't extract the data due to errors returned when using the OAI-PMH

Universities	OAI Repositories	Software
Acadia University	https://scholar.acadiau.ca/oai2	Islandora
Athabasca University	https://dt.athabascau.ca/oai/request	DSpace
Bishop's University*	https://eprints.ubishops.ca/cgi/oai2	EPrints
British Columbia Institute of Technology	https://circuit.bcit.ca/repository/oai2	Drupal
Brock University	https://dr.library.brocku.ca/oai/request	DSpace
Cape Breton University	https://cbufaces.cairnrepo.org/oai2	Islandora
Capilano University	https://capu.arcabc.ca/oai2/request	Islandora
Carleton University	https://curve.carleton.ca/oai-pmh/all	Drupal
Collège Militaire Royal du Canada	https://espace.rmc.ca/oai/request	DSpace
Dalhousie University	https://dalspace.library.dal.ca/oai/request	DSpace
Emily Carr University of Art and Design	https://ecuad.arcabc.ca/oai2	Islandora
HEC Montréal	https://reflexion.hec.ca/in/rest/oai	Unspecified
Institut National de la Recherche Scientifique	https://espace.inrs.ca/cgi/oai2	EPrints
Kwantlen Polytechnic University	https://kora.kpu.ca/oai2	Islandora
Lakehead University	https://knowledgecommons.lakeheadu.ca/oai/request	DSpace
MacEwan University	https://roam.macewan.ca/server/oai/request	Islandora
McMaster University	https://macsphere.mcmaster.ca/oai/request	DSpace
Memorial University of Newfoundland	https://research.library.mun.ca/cgi/oai2	CONTENTdm
Mount Royal University	https://mru.arcabc.ca/oai2	Islandora
Mount Saint Vincent University	https://ec.msvu.ca/server/oai/request	DSpace
NSCAD University	https://nscad.cairnrepo.org/oai2	Islandora
National School of Public Administration	https://espace.enap.ca/cgi/oai2	EPrints
OCAD University	https://openresearch.ocadu.ca/cgi/oai2	EPrints
Ontario Tech University	https://ir.library.ontariotechu.ca/oai/request	DSpace
Queen's University	https://qspace.library.queensu.ca/oai/request	DSpace
Royal Roads University	https://viurrspace.ca/oai/request	DSpace
Saint Mary's University	https://library2.smu.ca/oai/request	DSpace
Simon Fraser University*	https://summit.sfu.ca/oai/request	Drupal
St. Francis Xavier University	https://stfxscholar.cairnrepo.org/oai2	Islandora
TELUQ University	https://r-libre.teluq.ca/cgi/oai2	EPrints
Thompsom Rivers University	https://tru.arcabc.ca/oai2/request	Islandora
Trinity Western University	https://twu.arcabc.ca/oai2/request	Islandora
University of Alberta	https://era.library.ualberta.ca/oai	Jupyter
Université Concordia d'Edmonton	https://spectrum.library.concordia.ca/cgi/oai2	EPrints
University of Calgary	https://prism.ucalgary.ca/oai/request	DSpace
University of Guelph	https://atrium.lib.uoguelph.ca/oai/request	DSpace
University of Manitoba	https://mspace.lib.umanitoba.ca/oai/request	DSpace
University of Northern British Columbia	https://unbc.arcabc.ca/oai2/request	Fedora
University of Prince Edward Island	https://islandscholar.ca/oai2	Islandora
University of Quebec, Abitibi-Temiscamingue	https://depositum.uqat.ca/cgi/oai2	EPrints
University of Quebec at Chicoutimi	https://constellation.uqac.ca/cgi/oai2	EPrints
University of Quebec in Montreal	https://archipel.uqam.ca/cgi/oai2	EPrints
University of Quebec, Trois-Rivieres	https://depot-e.uqtr.ca/cgi/oai2	EPrints
University of Regina	https://ourspace.uregina.ca/oai/request	DSpace
University of Saskatchewan	https://harvest.usask.ca/oai/request	DSpace
University of Toronto	https://tspace.library.utoronto.ca/oai/request	DSpace
University of Victoria	https://dspace.library.uvic.ca/oai/request	DSpace
University of Waterloo	https://uwspace.uwaterloo.ca/oai/request	DSpace
University of Winnipeg	https://winnspace.uwinnipeg.ca/oai/request	DSpace
University of the Fraser Valley	https://ufv.arcabc.ca/oai2/request	Islandora
Université Laurentienne	https://zone.biblio.laurentian.ca/oai/request	DSpace
Université Laval	https://corpus.ulaval.ca/oai/request	DSpace
Université McGill*	https://escholarship.mcgill.ca/catalog/oai/request	Samvera
Université de Lethbridge*	https://opus.uleth.ca/oai/request	DSpace
Université de Moncton	https://udmscholar.cairnrepo.org/en/oai2	Islandora
Université de Montréal	https://papyrus.bib.umontreal.ca/oai/request	DSpace
Université de Sherbrooke	https://savoirs.usherbrooke.ca/oai/request	DSpace
Université d'Ottawa	https://ruor.uottawa.ca/oai/request	DSpace
York University	https://yorkspace.library.yorku.ca/oai/request	DSpace
Western University*	https://ir.lib.uwo.ca/do/oai	DigitalCommons
École Polytechnique de Montréal	https://publications.polymtl.ca/cgi/oai2	EPrints
École de Technologie Supérieure	https://espace.etsmtl.ca/cgi/oai2	EPrints

https://doi.org/10.23106/dcmi.953359643

B Supported metadata prefixes

Table 7: Supported formats for all extracted universities. Stars indicate repositories for which we couldn't extract the data due to errors returned when querying the IR, and empty entries are universities for which the OAI-PMH command to list metadata formats returned an error.

Universities	oai_dc	qdc	oai_qdc	etdms	oai_etdms	others
Acadia University	X		X		X	mods
Athabasca University	X				X	
Bishop's University*	X					didl, mets, oai_bibl, rdf, uketd_dc
British Columbia Institute of Technology	X				X	mods
Brock University	X	X		X		didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
Cape Breton University	X		X		X	mods
Capilano University	X		X		X	mods
Carleton University	X					mods
Collège Militaire Royal du Canada	X	X		X		didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
Dalhousie University	X	X		X	X	didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
Emily Carr University of Art and Design	X		X		X	mods
HEC Montréal	X					inmedia
Institut National de la Recherche Scientifique	X					didl, mets, oai_bibl, rdf, uketd_dc
Kwantlen Polytechnic University	X		X		X	mods
Lakehead University	X	X		X		didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
MacEwan University	X	X		X		didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
McMaster University	X	X		X		didl, dim, lac2, lac_mac, marc, mets, mods, ore, rdf, uketd_dc, xoai
Memorial University of Newfoundland	X					didl, etd-ms, mets, oai_bibl, rdf, uketd_dc
Mount Royal University	X		X		X	mods
Mount Saint Vincent University	X	X		X		didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
NSCAD University	X		X		X	mods
National School of Public Administration	X					didl, mets, oai_bibl, rdf, uketd_dc
OCAD University	X				X	didl, mets, oai_bibl, rdf, uketd_dc
Ontario Tech University	X	X		X	X	didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
Queen's University	X	X		X		didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
Royal Roads University	X	X		X		didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
Saint Mary's University	X	X		X		didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
Simon Fraser University*						
St. Francis Xavier University	X		X		X	mods
TELUQ University	X				X	didl, mets, oai_bibl, rdf, uketd_dc
Thompsom Rivers University	X		X		X	mods
Trinity Western University	X		X		X	mods
University of Alberta	X				X	
Université Concordia d'Edmonton	X				X	didl, mets, oai_bibl, oai_openaire, oai_ore_atom, oai_ore_rdf, rdf, uketd_dc
University of Calgary	X	X		X		didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
University of Guelph	X	X		X	X	didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
University of Manitoba	X	X		X	X	didl, dim, marc, mets, mods, oai_lockss, ore, rdf, uketd_dc, xoai
University of Northern British Columbia	X		X		X	mods
University of Prince Edward Island	X		X		X	mods
University of Quebec, Abitibi-Temiscamingue	X					didl, mets, oai_bibl, rdf, uketd_dc
University of Quebec at Chicoutimi	X					didl, oai_bibl, uketd_dc
University of Quebec in Montreal	X					didl, mets, oai_bibl, rdf, uketd_dc
University of Quebec, Trois-Rivieres	X				X	didl, etd_ms_uqtr, mets, oai_bibl, oai_openaire, rdf, uketd_dc
University of Regina	X	X		X		didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
University of Saskatchewan	X	X		X		didl, marc, mets, mods, ore, rdf, uketd_dc
University of Toronto	X	X		X	X	didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
University of Victoria	X	X		X	X	didl, dim, marc, mets, mods, oai_etdms_old, oai_openaire, ore, rdf, uketd_dc, xoai
University of Waterloo	X	X		X		didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
University of Winnipeg	X	X		X	X	didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
University of the Fraser Valley	X		X		X	mods
Université Laurentienne	X	X		X	X	didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
Université Laval	X	X		X		didl, dim, etdms11, marc, mets, mods, ore, rdf, uketd_dc, xoai
Université McGill*						
Université de Lethbridge*	X	X		X	X	didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
Université de Moncton	X		X		X	mods
Université de Montréal	X	X		X	X	didl, dim, etdms10, etdms11, marc, marc21, mets, mods, oai_openaire, oai_openaire4science, ore, rdf, uketd_dc, xoai
Université de Sherbrooke	X	X			X	didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
Université d'Ottawa	X	X		X	X	didl, dim, marc, mets, mods, ore, rdf, uketd_dc, xoai
York University	X	X		X	X	didl, dim, marc, mets, mods, musicmods, ore, rdf, uketd_dc, xoai
Western University*	X	X			X	dcq, dcs, oai-dc, qualified-dublin-core, simple-dublin-core
École Polytechnique de Montréal	X				X	didl, mets, oai_bibl, oai_openaire, rdf, rem_atom
École de Technologie Supérieure	X					didl, mets, oai_bibl, rdf, uketd_dc

https://doi.org/10.23106/dcmi.953359643

C Date ISO distribution and formats

Figure 4: Proportion of dates in an ISO 8601 format in each university repository (we discard any repository that never includes the date tag).

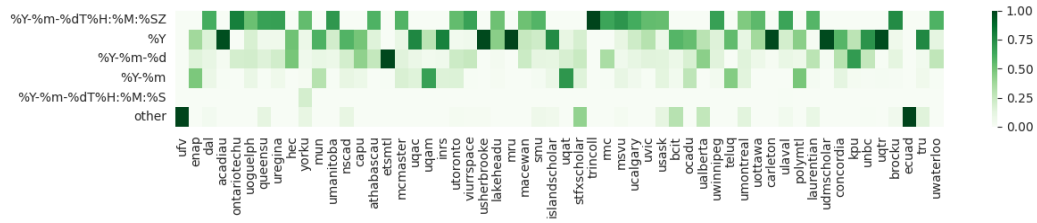

Figure 5: Distribution of most used date formats in each university repository (we discard any repository that never includes the date tag).

D Type Regexes

Type	Regex
Article	`/*article*/i`
Book	`/*(book)|(livre)*/i`
Dataset	`/*dataset*/i`
Media	`/*(photo)|(m[eé]dia)|(image)|(picture)|(vid[ée]o)|(audio)*/i`
Poster	`/*poster*/i`
Presentation	`/*pr[eé]sentation*/i`
Thesis	`/*(thesis)|(th[eè]se)|(m[eé]moire)|(dissertation)*/i`

E Language capturing Regexes

Table 8: Regular expressions used to capture different variations of language names appearing in the *Language* tag and map them to a unambiguous standard name. The Regular expressions are simple since the variety of languages mentioned in Canadian UIRs is quite limited and does not require to cover complex cases. All strings were lowercased previous to using the regular expressions. Our capturing script was written in Python and therefore, we used its native Traditional DFA (deterministic finite automaton) regular expression engine.

Targeted language	Regex	Clarification
English	`/*(^en)|(^angl)*/i`	Matches a string whose first characters are *en* or *angl*.
French	`/*(^fr)*/i`	Matches a string whose first characters are *en*.

https://doi.org/10.23106/dcmi.953359643