Evaluating Utility and Automatic Classification of Subject Metadata from Research Data Australia†

Mingfang Wu*, Ying-Hsang Liu**, Rowan Brownlee***, Xiuzhen Zhang****

* Australian Research Data Commons, Melbourne, Australia <mingfang.wu@ardc.edu.au>

** Oslo Metropolitan University <Ying-Hsang.Liu@oslomet.no>

*** Australian Research Data Commons, Canberra, Australia <rowan.brownlee@ardc.edu.au>

**** Royal Melbourne Institute of Technology University, Australia <xiuzhen.zhang@rmit.edu.au>

Mingfang Wu is senior research data specialist at the Australian Research Data Commons. She has conducted research in the areas of natural language processing, interactive information retrieval, search log analysis, interfaces supporting exploratory search and enterprise search. Her recent research focuses on the data discovery paradigms as part of the Research Data Alliance initiative and for improving data discovery service of Australian national research data catalogue, as well as a few data management related topics such as data provenance, data versioning and data quality. Mingfang holds a PhD from RMIT University, Australia.

Ying-Hsang Liu is a Senior Researcher in Information Studies at Oslo Metropolitan University in Norway. He holds a PhD in information science from Rutgers University. His research lies at the intersections of knowledge organisation, interactive information retrieval and human information behaviour. His research program has focused on the design of interactive information technologies, with a particular emphasis on user perceptions and individual differences and the relationship between visual search and user search behaviour. He has served on the editorial boards of Information Processing & Management and Online Information Review and chaired several ASIS&T and ALISE committees.

Rowan Brownlee works with vocabulary services & projects at ARDC, and previously worked in related areas at the University of Sydney and the State Library of NSW.

Xiuzhen (Jenny) Zhang is a professor of Data Science at School of Computing Technologies, RMIT University, Australia. Her research interests include data mining, text mining and machine learning. She has published over 100 papers and her research has been supported by The Australian Research Council. Prof Zhang is an associate editor for the Information Processing and Management journal. She has chaired and served on the program committee of international and Australian conferences. Prof Zhang holds a PhD from The University of Melbourne, Australia.

Wu, Mingfang, Ying-Hsang Liu, Rowan Brownlee and Xiuzhen Zhang. 2021. "Evaluating Utility and Automatic Classification of Subject Metadata from Research Data Australia." *Knowledge Organization* 48(3): 219-230. 40 references. DOI:10.5771/0943-7444-2021-3-219.

Abstract: In this paper, we present a case study of how well subject metadata (comprising headings from an international classification scheme) has been deployed in a national data catalogue, and how often data seekers use subject metadata when searching for data. Through an analysis of user search behaviour as recorded in search logs, we find evidence that users utilise the subject metadata for data discovery. Since approximately half of the records ingested by the catalogue did not include subject metadata at the time of harvest, we experimented with automatic subject classification approaches in order to enrich these records and to provide additional support for user search and data discovery. Our results show that automatic methods work well for well represented categories of subject metadata, and these categories tend to have features that can distinguish themselves from the other categories. Our findings raise implications for data catalogue providers; they should invest more effort to enhance the quality of data records by providing an adequate description of these records for under-represented subject categories.

Received: 14 January 2021; Revised 17 September 2021; Accepted 13 October 2021

Keywords: subject metadata, user search, automatic classification

† This paper was originally presented at the NKOS Consolidated Workshop 2020, September 9-10, 2020 (Virtual), under the title "Evaluating Utility of Subject Headings in a Data Repository: A Preliminary Finding from a Data Search Log". https://nkos-eu.github.io/2020/content/MingfangWu-NKOS2020-Slides-min.pdf

1.0 Introduction

In recent years, research data has become more accessible through institutional data repositories and data catalogues. In a survey of data service providers, several important system design issues have been recognised, such as the metadata elements indexed, search results ranking models, and system evaluation methods (Khalsa et al., 2018). The importance of metadata and portal functionalities has been identified in user requirements and recommendations for data repositories (Wu et al., 2019). One of the key functions of metadata schemas is the provision of subject metadata to support users to find relevant data. Yet the outstanding research questions of how data providers describe their data records with subject metadata, and how users utilise the subject metadata for data search remain unanswered.

A controlled vocabulary is designed to address the problem of "many-one and one-many relationships between words and their referents" (Svenonius, 1986, 332) in the use of natural language in information retrieval. Controlled vocabularies have long been studied by the library science community for their role in improving resource discovery by addressing the issue of synonyms and homonyms (Gross et al. 2015; Hjørland, 2016; Liu & Wacholder, 2017). In parallel, in the data management field, the FAIR data principles (Findable, Accessible, Interoperable, Reusable) encourage datasets to be described with rich metadata, using formal, accessible, shared, and broadly applicable language for knowledge representation and discovery (Wilson 2016). We have seen much effort put into the development of vocabularies to define various data attributes such as data subject, data format, data license, etc., but less effort in investigating how those vocabularies are deployed and whether a deployment can be assisted by automation (See Khalsa, Cotroneo, & Wu, 2018, for a recent survey of data service providers).

In this study, we take an Australian research data national catalogue - Research Data Australia (RDA) as a case study to explore user data search behaviour. Specifically, we investigate the following three research questions:

1. How well do data providers describe their data records (or metadata) with subject metadata?
2. How often do data seekers utilise the subject metadata when searching for data?
3. Can the automated assignment of subject metadata reach an agreed level of accuracy?

To answer the first two research questions, we analyse the RDA catalogue content and RDA search logs respectively. To explore the third research question, we compare four state-of-the-art machine learning methods that automatically assign subject categories to data records using the Fields of Research (FoR) classification code from the Australian and New Zealand Standard Research Classification scheme (ANZSRC-FoR), one type of knowledge organization system (KOS) (Hodges, 2000). The ANZSRC-FoR code (ANZSRC 2008)[1] is widely used to measure and analyse research and experimental development (R&D) undertaken in Australia and New Zealand.

The paper is structured as follows: Section 2 discusses related work and challenges related to our research questions, Section 3 describes the catalogue Research Data Australia used for our study, Section 4 presents our findings to each research question, Section 5 concludes and discusses the implication of this study and future studies.

2.0 Related work

2.1 The role of classification in library catalogue search

The purpose of classification is to help us organise and orient ourselves within a vast information and knowledge landscape, to better understand the world around us, and to more effectively communicate with each other. That is, classification systems can shape our worldviews and social interactions (Bowker & Star, 2000). For example, libraries have long used classification systems such as the Dewey Decimal Classification and the Library of Congress Classification to organise books, learning materials and research publications. These classifications help library patrons to find a required resource (e.g., a book) in the physical environment of a library but also in a digital, online version of a library catalogue (Dai, et al, 2020). Gross (et al. 2015) looked at the importance of controlled vocabularies in library catalogues. Their study showed about 27.7 percent of relevant resources would be lost if no subject headings were presented in catalogue records. Liu and Wacholder (2017) demonstrated that domain experts are more able than less expert searchers to use subject headings to achieve high precision search results. Application of structure and controlled vocabularies is an important characteristic that distinguishes a digital library collection from the unstructured web collection, where keyword search pre-dominates.

There are two main approaches to applying a classification scheme to resources: manual and automated. In a manual approach, a resource owner or provider assigns the resource to a classification category and may also assign it to subcategories if they wish to provide more specific access points. For example, when a researcher submits a paper to a journal, they are asked to allocate a few subject terms to describe the paper. In an automated approach, a pre-trained classification model assigns a resource to a category label based on the probability that the resource is similar to other resources in that category. An automated approach can be top down, applying an existing classification scheme to resources, or bottom up, finding natural groups among resources through clustering (Wu et al. 2001). Although the manual approach may produce more precise categorisation than the automated approach, the manual approach is laborious, and may result in resources being overlooked, and not receiving category labels.

MacFarlane (2016) argued that knowledge organisation still plays its role in multimedia information retrieval. By applying controlled vocabularies to resources, online library catalogues can provide facet search and facet filter to enable users to navigate online resources, and to refine or narrow a search, especially when a user's needs are vague or when their search terms do not quite match index terms. Kemman, Kleppe and Maarseveen (2013) revealed the relationship between user search behaviour and search interface facets. Nielson and Turney (2015) demonstrated that facets and filters are extremely effective in information retrieval systems. Bogaard et al. (2019) analysed a search log from a digital library of well described historical newspaper collection, and their study shows that faceted search is more prevalent than non-faceted search in terms of the number of unique queries, time spent, clicks and downloads.

Facets or categories have been also found to provide additional context that assists users to navigate search results from an information retrieval system. For example, Pratt, Hearst et al. (1999) studied a knowledge-based method for dynamically categorising a search result into a hierarchical term structure, a method that utilises subject terms and structures provided by the National Library of Medicine and the Unified Medical Language System. Their user study suggested that the hierarchical categorisation approach helped users find answers to certain types of questions more efficiently and easily than when search results were presented in a ranked list. Wu, Fuller and Wilkinson (2001) tried to use WordNet (Fellbaum, 1998) for inferring expected answer facets from a query question, they classified and organised search results into answer facets, and compared this faceted classification interface with an interface that instead provided a ranked list for a facet recall task. Although they didn't find a significant difference between the two interfaces in terms of helping users find more answer facets to a query, the faceted classification approach was highly preferred by the test subjects. Hearst (2006) summarised that the success of applying a faceted approach in organising information resources or search results is confounded, highly dependent on and sensitive to the details of the interface design. Kules and Capra's (2012) eye tracking study of user searching in a faceted library catalogue showed that facets account for approximately 10–30% of interface uses and users do not apply the faceted interface elements quickly within a search session. These findings have been applied to the design of organisation and navigation systems in information architecture (See Ding, Lin, & Zarro, 2017; Rosenfeld, Morville, & Arango, 2015 for examples).

Overall, the role of classification schemes and controlled vocabularies and their usefulness for information access in information retrieval systems, such as library online catalogues and digital libraries, remains an open-ended research question. Their usefulness has not been fully realised since we still have a limited understanding of the optimal search interfaces for facets.

2.2 Dataset search

In past decades, there has appeared a large number of research data repositories and data catalogs. As of 6 February 2020, re3data.org (Registry of Research Data REpositories) had registered more than 2450 research data repositories from 246 countries, and this number is still increasing (Cousijn and Fenner, 2020). It is becoming a challenge for discovering data when more and more research data is made open and accessible. If data are not discoverable, the benefit and effort that is invested in making data openly available will not be fully realised.

There have been a few studies on the context within which data seekers search for data. Based on the analysis of 79 use cases collected from potential data seekers (de Waard, et al., 2017), Wu et al. (2019) summarised ten recommendations for a data repository to implement in order to support a wide range of data search needs. These recommendations included the provision of multiple access points to find data. Chapman et al (2020) conducted a literature survey of dataset search within open data portals that included data, databases, information retrieval, entity centric search and tabular search. They identified challenges for dataset search, challenges including but not limited to formal query language, providing additional information to support the query process, and facilitating user exploration and interaction with a result set. A user study of data sensemaking behaviour by Koesten et al. (2021) has suggested that the provision of contextual information is needed for data reuse based on the findings of user activity patterns and attributes.

In a survey of where data repositories are at concerning their data retrieval systems (Khalsa el al 2018), about 73% of

participating repositories had deployed out-of-the-box search systems that usually adopt the document retrieval models, and about 77% of repositories had not conducted any kind of evaluation on data retrieval or discovery performance. Carevic et al. (2020) argued that due to the diverse nature of datasets, document retrieval models often do not work as efficiently for retrieving datasets (compared to retrieval of publications), and their study of user search logs indicates that queries for dataset search are different from those for publication search. The data search track from the NII Testbeds and Community for Information Access Research Project (NTCIR) is an initiative to test and evaluate dataset search, using a common collection and search tasks, with the aim of advancing dataset search engine performance, and the Text Retrieval Conference (https://trec.nist.gov/overview.html) has also played a significant role in advancing document search and web search (Kato et al., 2020). A test collection for dataset search in the domain of biodiversity research by Löffler et al. (2021) is another effort to advance the dataset search technology; however, the question corpus could be expanded to enable a more robust testing of the effectiveness of search algorithms, and the relevance judgments by human assessors could further consider the partial relevance and involve more than one judge for relevance judgement.

In summary, among those studies on dataset search, there are few which explore the role of controlled vocabularies in dataset discovery. This topic has however been widely studied in the context of library catalogue search and other contexts such as commercial product search etc. This exploratory study aims to bridge the gap by presenting a case study that analyses the implementation of a specific controlled vocabulary in a data catalogue.

3.0 Case study: the Research Data Australia catalogue collection

Research Data Australia (RDA)[2] is an Australian national research data catalogue. RDA is developed and maintained by the Australian Research Data Commons (ARDC)[3], which is an Australian Government funded initiative supporting research through promotion of FAIR principles and development of high-quality research data assets. RDA has two major components: a catalogue registry at the backend and a data discovery portal at the frontend.

3.1 RDA registry

The registry harvests dataset metadata from approximately 100 Australian universities, research organisations, cultural heritage institutes and public sector agencies. The registry contains about 150,000 metadata records of datasets (as of June 2020). These records are encoded in the Registry Interchange Format - Collections and Services (RIF-CS) schema.

RDA, as an aggregator, harvests metadata from different schemas (e.g. ISO19115, Dublin Core, Data Documentation Initiative). Metadata in those schemas are automatically mapped to the RIF-CS schema during the harvesting process through pre-defined crosswalks.

ARDC encourages metadata contributors to adopt standard or community endorsed vocabularies to make metadata interoperable. For subject headings, RDA accommodates a number of vocabularies[4] including the three component classifications of the Australian and New Zealand Standard Research Classification (ANZSRC), and others such as the Global Change Master Directory (GCMD)[5], Powerhouse Museum Object Name Thesaurus (PONT)[6], Library of Congress Subject Headings (LCSH), Australian Pictorial Thesaurus (APT), Thesaurus of Psychological Index Terms (PSYCHIT), and ISO639-3. The RIF-CS schema includes properties enabling metadata contributors to identify the source of vocabulary terms.[4] We hope that by accommodating both discipline-agnostic and discipline-specific vocabularies, more search scenarios can be supported.

The ANZSRC comprises three classifications. The Fields of Research (FoR) is disciplinary agnostic and widely used for classifying research undertaken in Australia and New Zealand. The ANZSRC-FoR is reviewed periodically; the latest version was released in June 2020. This study uses the 2008 release in alignment with the timelines when data records were generated for inclusion within RDA. The 2008 release of the ANZSRC-FoR has 1417 category/terms, arranged in 3 hierarchical levels; the top level has 22, the second level 157 and the third level 1238 terms respectively, representing broad to narrow research areas. Each term is assigned a code - terms from the top level a two digit code, then terms from the second and the third levels four and six digit codes respectively; for example, 04 Earth Sciences → 0403 Geology → 040310 Sedimentology.

3.2 RDA data discovery portal

The RDA portal is for human users to discover data; it provides a standard search box for keyword search and advanced search, as well as collection browse and search result filtering. Subject metadata are used in five ways:

1. All subject metadata are included in the collection index and weighted the same way as other indexed terms, thus subject metadata are searchable.
2. Subject metadata are used for browsing the catalogue (Figure 1).
3. Subject metadata are used in facet search (as part of advanced search) (Figure 2).

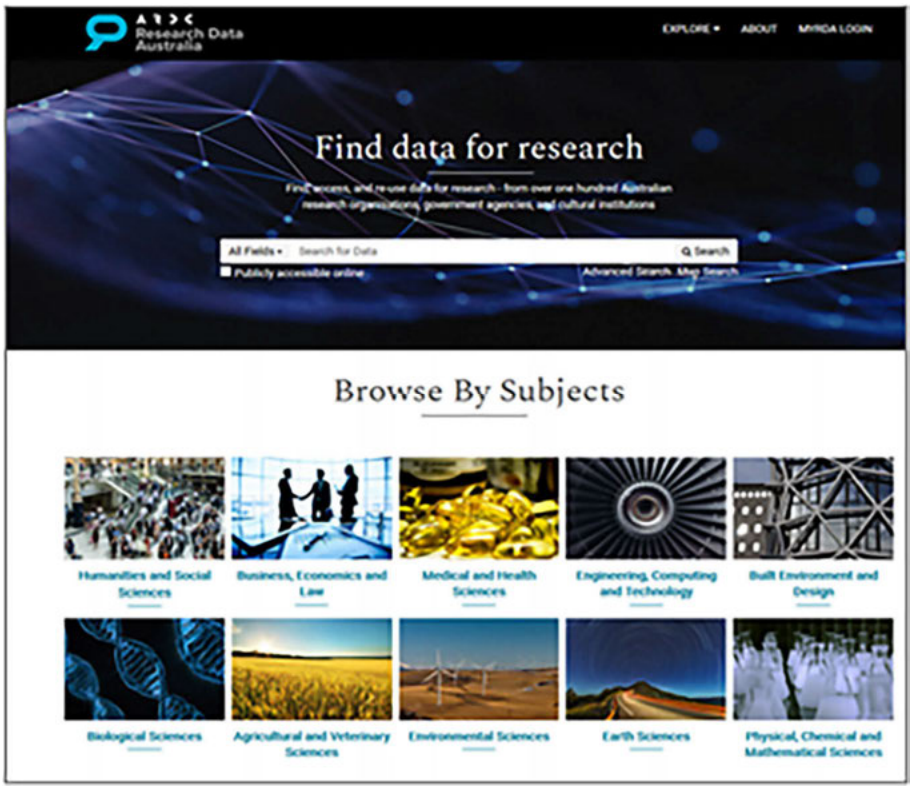

Figure 1. Browse By Subjects from the RDA homepage

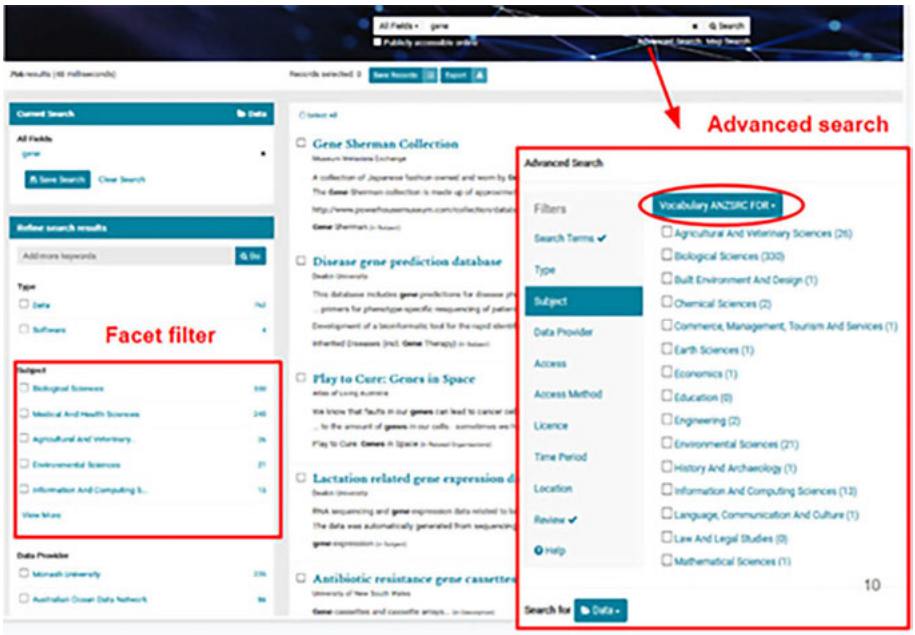

Figure 2. Subject metadata used in advanced search and facet filter of a search result

4. ANZSRC-FoR subject metadata are used as facet filters of a search result (Figure 2).
5. Within a metadata record, all subject metadata are displayed as a hyperlink.

A click on a link will result in retrieving all records with that subject metadata (Figure 3). Note that in method 2 and 4, only ANZSRC-FoR subject metadata are included. This means the records that don't have a subject heading from

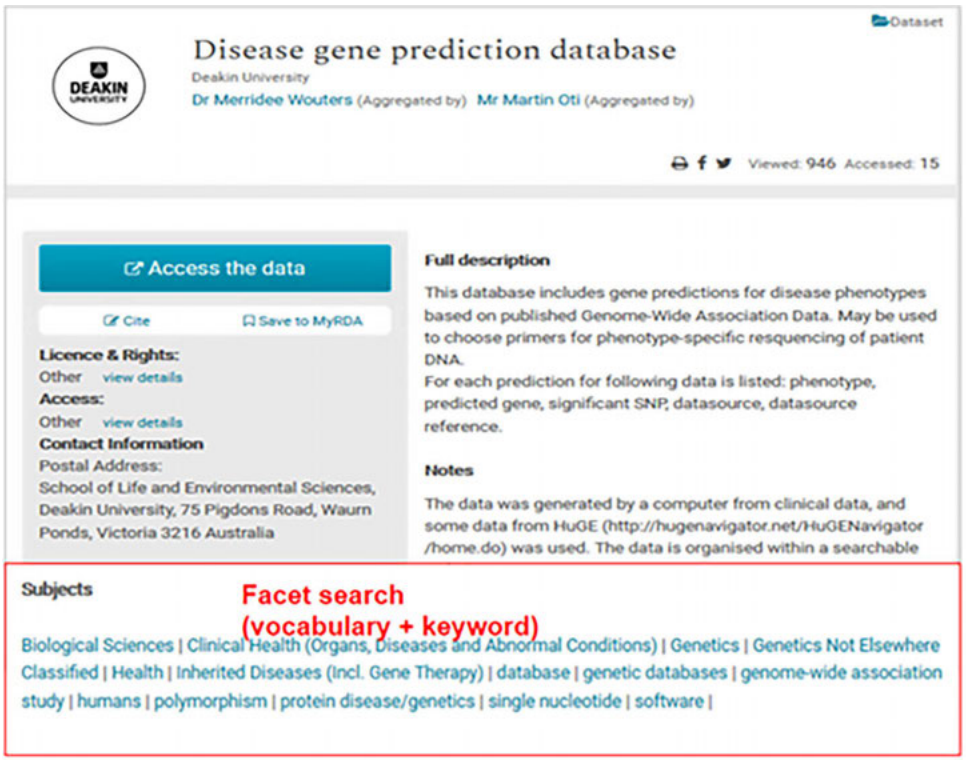

Figure 3. Subject heading is used in facet search

the ANZSRC-FoR terms are not discoverable via collection browse or by facet filtering.

4.0 Analysis

4.1 Analysis of the distribution of subject metadata

To answer the first research question about how well data records are described with subject metadata, we analysed the number of RDA data records that have at least a subject heading regardless of which classification code or vocabulary is used. The analysis includes 142,792 data records (as registered in June 2020). Figure 4 shows the number of records per type of subject vocabulary. The ANZSRC-FoR code is used much more often than other subject vocabularies; about 55% of records have at least an ANZSRC-FoR term, while few records have been assigned with terms from disciplinary vocabulary such as PONT and PSYCHIT. The blue bars in Figure 5 show the usage of the ANZSRC-FoR vocabulary: about 45% of records have 0 ANZSRC-FoR terms, 29% 1 to 2 terms, 9% 3 to 5 terms, 16% 6 to 10, and the other 1% have more than 10 terms.

These numbers show some issues, including:

1. There is inconsistent distribution of subject metadata across metadata records, due to different approaches taken to metadata creation, by upstreaming data repositories that contribute metadata to RDA. For example, some repositories mandate that the subject metadata contains at least one heading, while other repositories leave the subject metadata as optional.
2. The uneven distribution of RDA records towards a few research areas indicates ARDC may need to proactively seek content from underrepresented research areas to enrich RDA holdings and to demonstrate provision of service to the whole research community.
3. As discussed in the previous section, only ANZSRC-FoR terms are included in browsing and facet filtering, so about 45% of records (those without ANZSRC-FoR terms) are excluded from browsing and from facet filtering, which results in fewer discoverability options for these records.

4.2 Log analysis of the use of subject metadata in data discovery

Log analysis has been widely used to discover what users search for and how users interact with a search system (e.g Jansen 2009, Kacprzak et al. 2018, Schultheiß et al., 2020, Walsh et al, 2019). Here we also undertook log analysis to explore our second research question: how often data seekers utilise subject metadata in their data search. We analysed a 2019 user interaction log from RDA. The yearly log has 1,013,193 entries from 321,695 unique IPs. We first seg-

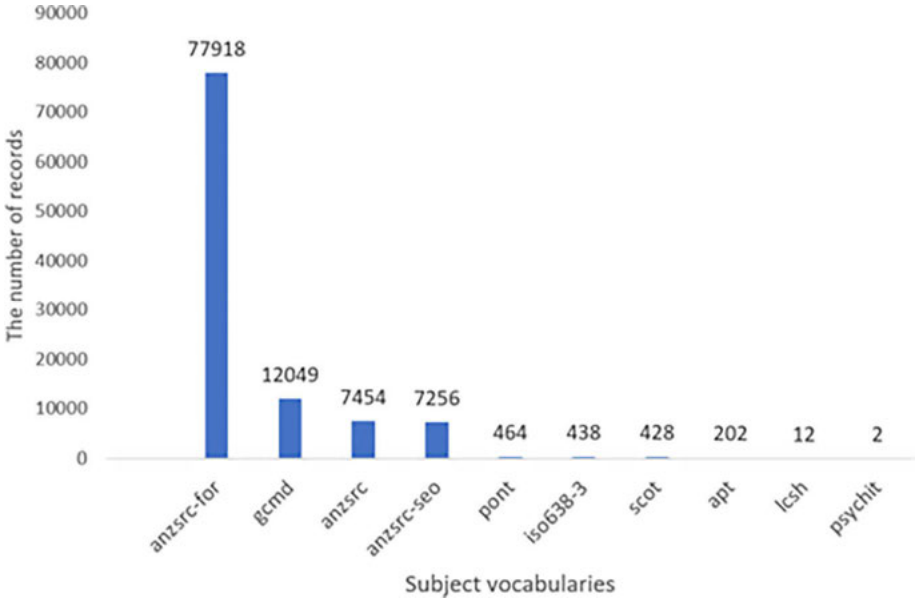

Figure 4. The number of records with subject terms from each controlled vocabulary

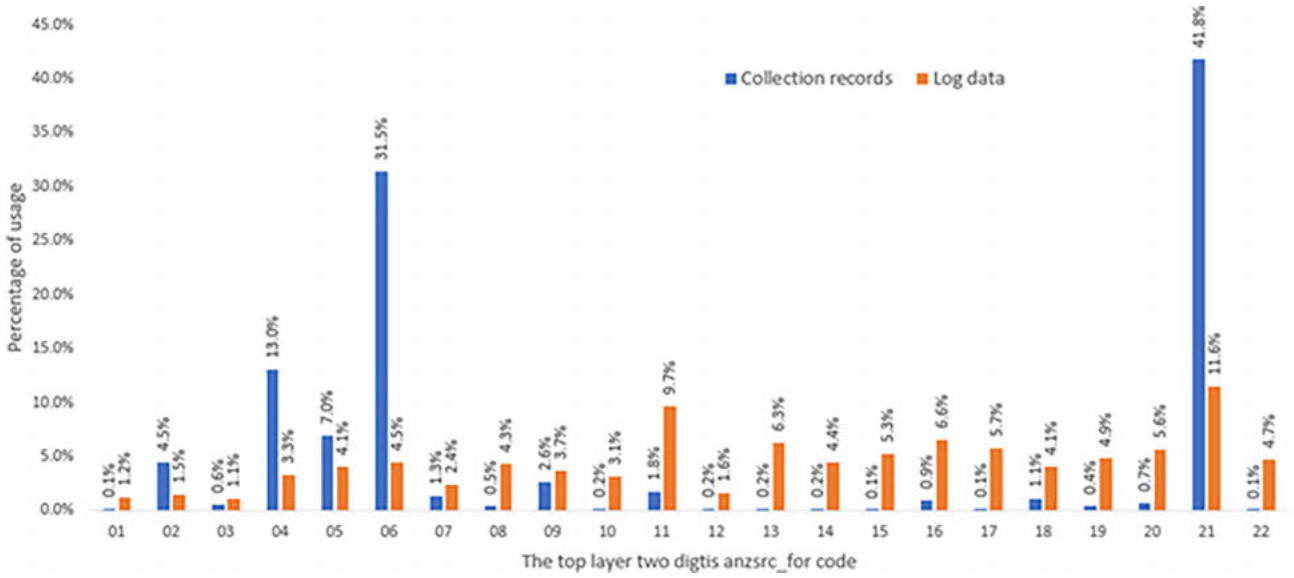

Figure 5. Comparison of the usage and the distribution of the ANZSRC-FoR subject metadata

mented the entries into sessions; a session starts from either a new IP address or from the same IP address following 30 minutes of inactivity. This results in 460,168 sessions in total.

There are in total 42,185 sessions with at least a search event. A search event can be a keyword search, an advanced search, browse by subject or a search updated with subject filters. We observed that a generic subject vocabulary (ANZSRC-FoR) is used much more often than a discipline-specific vocabulary: 7010 (16.6%) of these search sessions involved facet searches or filters with the ANZSRC-FoR subject filters, while only 59 (0.2%) filtered with the GCMD. This may be due to:

1. constraint of the search interface, only ANZSRC-FoRis included in both facet search and facet filter, all other vocabularies are in facet search only.
2. RDA is a generalist data catalogue targeting users from all research fields, and so users as recorded may not be familiar with terms that are specific to a discipline, thus less likely to select disciplinary specific terms (e.g. GCMD) in facet search.

The red bars in Figure 5 show the breakdown of those sessions utilising each ANZSRC-FoR code. We can see that the level of use of each ANZSRC-FoR code is quite similar, although the codes with a higher number of assigned records tend to attract more usage; the pearson correlation coefficient (0.46) indicates this correlation is weak, i.e. the utilising of subject metadata in facilitating data discovery is not dominated by a few research fields.

4.3 Machine learning approach for automatic tagging the RDA data records

As discussed in Section 4.1, nearly half of the catalogue records do not include a subject term in the subject field; this is a well-known issue facing cataloguers, as manually assigning each data record with terms from a subject vocabulary is a laborious activity. Hence, automatically assigning subject terms to catalogue records has been widely researched (e.g. Golub et al., 2020). The automatic classification experiment presented in this section tests how successful is the automatic assignment of terms to records missing a subject term. Here we present an experiment with ANZSRC-FoR's top level of 22 two digit codes/terms.

There is no classifier that will work for all collections, as each collection may have different characteristics in terms of content coverage, word distribution and association, thus a classifier trained and tested with one collection may have different results when applied to another collection. In this preliminary study, we apply four supervised machine learning classification models to the RDA collection, with the aim of testing the feasibility of automatically assigning ANZSRC-FoR terms as labels to RDA data records. The four models are:

- Multinomial logistic regression (MLR)[7],
- Multinomial naive bayes (MNB),
- K Nearest Neighbors (KNN), and
- Support Vector Machine (SVM).

All these models are widely used for text classification in the machine learning literature (Kowsari et al. 2019). We imple mented the above models using the Python Scikit-learn package (2021) by adapting source codes as introduced by Zafra (2019).

4.3.1 Experiment collection

A supervised learning method is one in which the model learns from records that already have subject metadata, as training data. We have 77918 records that have been assigned with at least one ANZSRC-FoR term. The top level of two digit codes/terms have been assigned to 84988 records in total, as a record may have been assigned more than one term. The distribution of records per code is highly biased toward a few codes (e.g. 21, 06 and 04 in Table 1), and this may introduce bias in the process of developing training models, and may negatively affect the classification accuracy of the trained models for the rare class (Zhang et al., 2017). We thus randomly downsized sample records from the terms that have a large number of records in order to balance class distribution, a widely used strategy for classification in cases of imbalanced class distribution (Zhang et al., 2017). The furthest right column of Table 1 shows the number of records per two digit code as used in the experiment, and the star (*) indicates that these categories are under-represented, thus all records from these categories are included in the experiment, i.e. downsize does not apply to these categories.

After we built a collection of records for classification, we first extracted and combined the title and the description from each record, removed words from the stop-words list included in the python NLTK library (words like "the", "a", "in") to reduce noise, then applied the Lemmatizer method to stem the remaining words into tokens (e.g lemmatising the words "playing", "plays", "played" to "play"). Each token is given a numeric value based on the tf*idf, where tf represents term frequency in a record, and where idf represents the inverse frequency of a token in all records in the collection. The remaining tokens and their values per record are the input to the four classification codes that are the focus of this experiment.

4.3.2 Classification performance

Each classification method used three quarters of records from each category for the training model, and the remaining one quarter was used for testing. Table 2 shows test performance of each classifier in terms of precision, i.e. the number of records that are correctly assigned to a category, where the micro-average aggregates the true predictions of all classes, and where the macro-average computes precision per class and then takes the average. For unbalanced numbers of labelled data per category, the micro-average shows a bias toward bigger categories, the macro-averaging treats all categories equally, and the weighted average takes a proportion of each category when performing the macro-averaging. We can see that the four classifiers have very close performance, with the MLR being slightly better than the other three (which is also the most efficient model). The performance varies from category to category; 6 categories colored in green have their precision closer to 1, while 7 in red are under 0.5, and the rest are in between. All the poor performing categories (in red) are from those under-represented categories, as shown in Table 1.

The more distinct features representing a category, the higher classification precision can be achieved. We find the

2 digits code	all data	down size
01	111	*111
02	3537	300
03	499	499
04	10147	600
05	5417	400
06	24520	600
07	1032	200
08	386	*386
09	2031	200
10	128	*128
11	1409	400
12	174	*174
13	148	*148
14	122	*122
15	76	*76
16	723	300
17	112	*112
18	849	400
19	343	*343
20	553	300
21	32592	600
22	79	*79
Total	84988	4799

Table 1. The number of records per top two digit code (* indicating under-represented categories, thus no down size applied to the category)

2 digits code	MLR	SVM	KNN	MNB
01	0.29	0.00	0.41	0.33
02	0.97	1.00	1.00	0.92
03	0.73	0.61	0.60	0.59
04	0.96	0.98	0.92	0.90
05	0.61	0.63	0.68	0.49
06	1.00	1.00	0.64	0.96
07	0.63	0.52	0.77	0.42
08	0.45	0.22	0.53	0.26
09	1.00	1.00	0.94	1.00
10	0.29	0.00	0.20	0.00
11	0.68	0.69	0.63	0.64
12	0.61	0.95	0.67	0.66
13	0.58	0.91	0.69	0.67
14	0.41	0.00	0.58	0.57
15	0.21	0.00	0.18	0.00
16	0.56	0.50	0.55	0.54
17	0.40	0.00	0.32	0.67
18	1.00	1.00	0.99	0.98
19	0.82	0.69	0.76	0.54
20	0.89	0.85	0.26	0.81
21	0.97	0.96	0.99	0.88
22	0.34	0.00	0.65	0.44
micro ave	0.70	0.67	0.66	0.66
macro ave	0.65	0.57	0.63	0.60
weighted ave	0.76	0.71	0.70	0.68

Table 2. Classification precision per top two-digit code

Code	Top 5	Bottom 5
04	earth	al
	airborne	unit
	geophysical	two
	mount	australia
	igsn	region
15	study	given
	financial	number
	survey	received
	university	document
	dataset	expert

Table 3. Top 5 most and least correlated features for the category 04 (Earth Science) and 15 (Commerce, Management, Tourism and Services).

features from those under-represented categories do not well represent their category. Table 3 shows an example of the top 5 most correlated features and the bottom 5 least correlated features for the categories 04 and 15 respectively. The top 5 and even some bottom 5 terms represent the subject 04 "Earth Science" well, while the top 5 highly correlated terms that represent the subject "(Commerce, Management, Tourism and Services)" are general, and do not distinguish this category from the rest. This may indicate that some manual intervention is required to boost the la-

belling of the under-represented categories in the collection. Data providers from these categories should be encouraged to use more distinct words to describe their data records, before an automatic classification is applied to the collection.

It is well documented in the literature that no classifier exists that works well

for different text classification tasks (Kowsari et al. 2019). Even for the same task, the performance of an ML classification model largely depends on the nuances of training data such as class distribution, features and data volume. In a relevant study by Golub et al. (2020), two models -- MNB and SVM -- are applied for a similar but not the same task of automatically assigning subject terms to digital resources, and the Dewey Decimal classification system is used for subject classification. In this study, several training datasets with different number of classes (ranging from 29 to 806 classes), class distribution (data rich classes of at least 1000 data records and data-poor classes of one single record), and different text (title and keywords), were curated to evaluate the performance of automatic assignment of subject classes. Their models also show varying performance with accuracy (result biased by frequent categories) in the range of 34%--80% and SVM generally shows better performance over MNB.

5.0 Discussion and ongoing work

This paper presents a case study of how well subject metadata are presented and utilised from a national data catalogue. Our initial study reveals that:

1. there is an inconsistent distribution of subject metadata across metadata records;
2. users from 12.6% of search sessions make use of subject filters or searches;
3. automatic classification of data records performance differs from category to category, with better performance by those well-represented categories and well described records.

Those records with zero or poor subject metadata may not be discovered if a subject filter is applied. Indeed, 20% of zero hits could be attributed to the poor quality of metadata records in library catalogue systems (Schultheiß et al., 2020). We will undertake further investigation to determine if there are any differences between search sessions with subject filters involved and those without, or if some types of queries (e.g., with broad meaning or scope) may make greater use of subject filters. We may also need to explore how to map subject metadata from other classification scheme to ANZSRC-FoR to address semantic gaps (MacFarlane, 2016), as mapping between subject metadata is important for interoperability especially for a catalogue aggregator that harmonises content from multiple distributed resources (Tudhope and Binding, 2016). We will continue our observation through log analysis, to explore whether the overall usage of subject search and filter correlates with an increase in the number of records containing subject metadata.

Our initial classification experiment result shows that, on average, the simple machine learning method MLR performs better than the other three methods; and all four experimented methods work well for records with categories that are well represented in the collection and which have a good number of high-correlated terms representing their designated categories. That is, machine learning methods can work well for records with high-quality metadata in well-represented categories within the collection. This finding can guide data repository managers in selection of subject vocabularies and may guide developers in creating appropriate functions to make best use of subject vocabularies for enhancing the data discovery experience of data seekers.

Note that this classification experiment was conducted on the top-level categories of the ANZSRC-FoR hierarchy. For automatic subject classification to be of practical use, it is more desirable to explore more granular categories with four digit or six digit codes. Towards this end, more records with granular codes need to be collected for training machine learning models to achieve better classification performance.

Our future work will also seek to understand more about the characteristics of different types of repositories (e.g., a generic one like RDA, a discipline oriented one, for example, an earth science data portal that supports the GCMD vocabulary), their intended users, users' search behaviour and relationships between search behaviour and metadata types/attributes. Based on this understanding, we intend to investigate how the RDA, as a discipline agnostic data repository, can balance subject metadata from both discipline agnostic and specific vocabularies.

Acknowledgements

Special thanks to Mr. Joel Benn, ARDC Manager of Development Operations for extracting and providing log data and the RDA metadata collection and to Dr. Adrian Burton for his support and many fruitful discussions of the project.

The views and the research findings expressed herein are those of the authors and are not necessary those of the Australian Research Data Commons.

Notes

1. Australian Research Council: Classification Codes - FoR, RFCD, SEO and ANZSIC Codes

2. Research Data Australia (RDA): https:// researchdata.edu.au/
3. Australian Research Data Commons: http://ardc.edu.au
4. Subject metadata as accommodated by RDA: https://documentation.ardc.edu.au/display/DOC/Subject#Subject-Subjectattributes
5. Global Change Master Directory (GCMD): https://earthdata.nasa.gov/earth-observation-data/find-data/gcmd/gcmd-keywords
6. Powerhouse Museum Object Name Thesaurus (pont): https://maas.museum/research/object-name-thesaurus/
7. Despite the name, Multinomial logistic regression is a classification model also known as maximum-entropy classification (MaxEnt) or the log-linear model.

References

Australian and New Zealand Standard Research Classification (ANZSRC). 2008. Archived Issue, Released on 31.03.2008 by Australian Bureau of Statistics. Retrieved 08.09.2021, from: https://www.abs.gov.au/ausstats/abs@.nsf/Previousproducts/1297.0Main%20Features52008?opendocument&tabname=Summary&prodno=1297.0&issue=2008

Bogaard, T., l. Hollink, J. Wielemaker, J. van Ossenbruggen and L. Hardman. 2019. "Metadata Categorization for Identifying Search Patterns in a Digital Library." *Journal of Documentation*, 75(2): 270-86. https://doi.org/10.1108/JD-06-2018-0087

Bowker, G. C., and S. L. Star. 2000. *Sorting Things Out: Classification and its Consequences*. Cambridge, MA.: MIT Press.

Carevic, Z., D. Roy amd P. Mayr. 2020. "Characteristics of Dataset Retrieval Sessions: Experiences from a Real-Life Digital Library." *Lecture Notes in Computer Science* 12246 LNCS: 185–93. https://doi.org/10.1007/978-3-030-54956-5_14

Chapman, A., E. Simperl, L. Koesten, G. Konstantinidis, L.-D. Ibáñez, E. Kacprzak, and P. Groth. 2020. "Dataset Search: a Survey." *VLDB Journal* 29(1): 251–72. https://doi.org/10.1007/s00778-019-00564-x

Cousijn, H. and M. Fenner. 2020. "German Research Foundation to Fund New Services". *DataCite Blog*, Feb. 6, 2020. Available: https://doi.org/10.5438/fwkt-3t12

Dai S., R. You, Z. Lu, X. Huang, H. Mamitsuka and S. Zhu. 2020. "FullMeSH: Improving Large Scale MeSH Indexing with Full Text." *Bioinformatics* 36(5):1533-1541. doi:10.1093/bioinformatics/btz756

Ding, W., X. Lin and M. Zarro. 2017. "Information Architecture: The Design and Integration of Information Spaces." *Synthesis Lectures on Information Concepts, Retrieval, and Services* 9 (2): i–152. https://doi.org/10.2200/S00755ED2V01Y201701ICR056.

de Waard, A., S. J. Khalsa, F. Psomopoulos and M. Wu. 2017. *RDA IG Data Discovery Paradigms IG: Use Cases Data* [Data set]. Zenodo. DOI: https://doi.org/10.5281/zenodo.1050976

Fellbaum, C. ed. 1998. *WordNet: an Electronic Lexical Database*. Cambridge MA: MIT Press

Golub, K., J. Hagelbäck and A. Ardö. 2020. "Automatic Classification of Swedish Metadata Using Dewey Decimal Classification: A Comparison of Approaches." *Journal of Data and Information Science* 5(1). Published online 22 April 2020. DOI: https://doi.org/10.2478/jdis-2020-0003

Gross, T., A. G. Taylor and D. N. Joudrey. 2015. "Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching." *Cataloging & Classification Quarterly* 53 (1): 1–39. https://doi.org/10.1080/01639374.2014.917447.

Hearst, M. 2006. "Design Recommendations for Hierarchical Faceted Search Interfaces." Paper presented at the ACM SIGIR Workshop on Faceted Search, August, 2006

Hjørland, B. 2016. "Does the Traditional Thesaurus Have a Place in Modern Information Retrieval?" *Knowledge Organization* 43(3): 145–59. https://doi.org/10.5771/0943-7444-2016-3-145

Hjørland, B. 2018. "Indexing Concepts and Theory." In *ISKO Encyclopedia of Knowledge Organization* edited by Birger Hjorland and Claudio Gnoli. Available: https://www.isko.org/cyclo/indexing#2.1

Hodge, G. 2000. *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Washington, DC: Council on Library and Information Resources. Available: https://www.clir.org/wp-content/uploads/sites/6/pub91.pdf

Jansen, B. J. 2009. *Understanding User-Web interactions via Web analytics*. San Rafael, CA: Morgan-Claypool.

Kacprzak, E., L. Koesten, L.-D. Ibáñez, T. Blount, J. Tennison and E. Simperl. 2019. "Characterising Dataset Search: An Analysis of Search Logs and Data Requests." *Journal of Web Semantics* 55: 37–55. https://doi.org/https://doi.org/10.1016/j.websem.2018.11.003.

Kato, M. P., H. Ohshima, Y.-H. Liu and H. Chen. 2020. "Overview of the NTCIR-15 Data Search Task." In *Proceedings of the 15th NTCIR Conference: Evaluation of Information Access Technologies*. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings15/pdf/ntcir/01-NTCIR15-OV-DATA-KatoM.pdf

Khalsa, S. J., P. Cotroneo and M. Wu. 2018, "A Survey of Current Practices in Data Search Services." *Mendeley Data*. http://dx.doi.org/10.17632/7j43z6n22z.1

Kemman, M., M. Kleppe and J. Maarseveen. 2013. "Eye Tracking the Use of a Collapsible Facets Panel in a Search Interface.". In *Research and Advanced Technology for Digital Libraries: Proceedings of the International Conference on Theory and Practice of Digital Libraries, TPDL 2013, Valletta, Malta, September 22-26, 2013*, edited by T. Aalberg, C. Papatheodorou, M. Dobreva, G. Tsakonas and C. Farrugia. *Lecture Notes in Computer Science, Vol 8092*. Berlin, Heidelberg: Springer, 405-8. https://doi.org/10.1007/978-3-642-40501-3_47

Koesten, L., K. Gregory, P. Groth and E. Simperl, E. 2021. "Talking Datasets: Understanding Data Sensemaking Behaviours." *International Journal of Human-Computer Studies* 146: 102562. https://doi.org/10.1016/j.ijhcs.2020.102562

Kowsari, K., K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes and D. Brown. 2019. "Text Classification Algorithms: a Survey". *Information* 10(4): 50.

Kules, B. and R. Capra. 2011. "Influence of Training and Stage of Search on Gaze Behavior in a Library Catalog Faceted Search Interface." *Journal of the American Society for Information Science and Technology* 63: 114–38. https://doi.org/10.1002/asi.21647.

Liu, Y. H. and N. Wacholder. 2017. "Evaluating the Impact of MeSH (Medical Subject Headings) Terms on Different Types of Searchers." *Information Processing and Management* 53(4): 851–70. https://doi.org/10.1016/j.ipm.2017.03.004

Löffler, F., A. Schuldt, B. König-Ries, H. Bruelheide, H. and F. Klan. 2021. "A Test Collection for Dataset Retrieval in Biodiversity Research." *Research Ideas and Outcomes* 7 (May): 67887. https://doi.org/10.3897/rio.7.e67887.

MacFarlane, A. 2016. "Knowledge Organization and its Role in Multimedia Information Retrieval." *Knowledge Organization* 43(3): 180-3. DOI: https://doi.org/10.5771/0943-7444-2016-3-180

Nelson, D. and L. Turney. 2015. "What's in a Word? Rethinking Facet Headings in a Discovery Service." *Information Technology and Libraries* 34(2): 76-91.

Pratt, W., M. Hearst and L. Fagan. 1999. "A Knowledge-Based Approach to Organizing Retrieved Documents." *AAAI-99: Proceedings of the Sixteenth National Conference on Artificial Intelligence, Orlando, Florida, 1999*, edited by Jim Hendler and Devika Subramanian. Cambridge, MA: MIT Press, 80-5.

Rosenfeld, L., P. Morville and J. Arango. 2015. *Information Architecture: For the Web and Beyond*. 4th ed. Sebastopol, CA.: O'Reilly.

Schultheiß, S., A. Linhart, C. Behnert, I. Rulik and D. Lewandowski. 2020. "Known-Item Searches and Search Tactics in Library Search Systems: Results from Four Transaction Log Analysis Studies." *Journal of Academic Librarianship* 46(5) 102202. https://doi.org/https://doi.org/10.1016/j.acalib.2020.102202

Scikit-Learn: a Set of Python Modules for Machine Learning and Data Mining. Retrieved 2021-09-16 from https://scikit-learn.org/stable/

Svenonius, E. 1986. "Unanswered Questions in the Design of Controlled Vocabularies." *Journal of the American Society for Information Science* 37 (5): 331–40. https://doi.org/10.1002/(SICI)1097-4571(198609)37:5<331::AID-ASI8>3.0.CO;2-E.

Tudhope, D. and C. Binding. 2016. "Still Quite Popular After All Those Years - The Continued Relevance of the Information Retrieval Thesaurus." *Knowledge Organization* 43(3): 174-9. DOI: https://doi.org/10.5771/0943-7444-2016-3-174

Walsh, D., P. Clough, M. M. Hall, F. Hopfgartner, J. Foster and G. Kontonatsios. 2019. "Analysis of Transaction Logs from National Museums Liverpool." In *Digital Libraries for Open Knowledge*. *TPDL 2019*, *Oslo, Norway, September 9-12,* edited by Antoine Doucet, Antoine Isaac, Koraljka Golub, Trond Aalberg, and Adam Jatowt. Cham: Springer, 84–98. https://doi.org/10.1007/978-3-030-30760-8_7.

Wilkinson, M. D. et al. 2016 "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3:160018. doi: 10.1038/sdata.2016.18

Wu, M., F. Psomopoulos, S. J. Khalsa and A. de Waard. 2019. "Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories." *Data Science Journal* 18 (1): 3. https://doi.org/10.5334/dsj-2019-003

Wu, M., M. Fuller and R. Wilkinson. 2001. "Using Clustering and Classification Approaches in Interactive Retrieval." *Information Processing & Management* 37(3): 459-84.

Zafra, M. F. 2019. "Text Classification in Python: Learn to Build a Text Classification Model in Python". Retrieved in August 2020 from: https://towardsdatascience.com/text-classification-in-python-dd95d264c802

Zhang, X., Y. Li, R. Kotagiri, L. Wu, Z. Tari and M. Cheriet. 2017. "KRNN: k Rare-class Nearest Neighbour Classification." *Pattern Recognition* 62: 33-44.