Australian Council for Educational Research (ACER)

ACEReSearch

Information Management | Cunningham Library

2-10-2016

Introducing an automated subject classifier

Pru Mitchell
Australian Council for Educational Research (ACER)

Tine Grimston
Australian Council for Educational Research (ACER)

Robert Parkes
Australian Council for Educational Research (ACER)

Follow this and additional works at: https://research.acer.edu.au/information_management

 Part of the Cataloging and Metadata Commons

Recommended Citation

Mitchell, P., Grimston, T., & Parkes, R. (2016). Introducing an automated subject classifier. VALA. https://research.acer.edu.au/information_management/3

This Conference Paper is brought to you by the Cunningham Library at ACEReSearch. It has been accepted for inclusion in Information Management by an authorized administrator of ACEReSearch. For more information, please contact repository@acer.edu.au.

Introducing an automated subject classifier

Pru Mitchell
Manager, Information Services
Australian Council for Educational Research
pru.mitchell@acer.edu.au
http://orcid.org/0000-0001-9824-5425

Tine Grimston
Senior Librarian, Technical Services
Australian Council for Educational Research
tine.grimston@acer.edu.au
http://orcid.org/0000-0002-4020-8442

Robert Parkes
Librarian
Australian Council for Educational Research
robert.parkes@acer.edu.au
http://orcid.org/0000-0001-9376-3901

Abstract:
The library community understands the value of controlled vocabularies in enhancing resource discovery. There is however ongoing tension between that value and the cost of maintaining and applying specialist vocabularies. This paper presents the outcomes of a 2014-15 trial of automated subject indexing at the Australian Council for Educational Research. The integration of a machine learning classification tool has resulted in streamlined workflows and increased use of machine-readable data. Insights were gained into the decisions human indexers make in using a controlled vocabulary, and into the importance of quality abstracts and metadata.

Background

Established in 1930, the Australian Council for Educational Research (ACER) is a not-for-profit organisation providing educational research services and products. The Cunningham Library at ACER (the Library) has a research level collection (Australian Libraries Gateway, 2015) in Australian education.

> The ACER Library has for many years helped administrators, teachers and students to find information about Australian education. The sources are numerous, the task is growing as years pass, and the indexing of information is an increasingly onerous task.

This quote from the preface to the first edition of the *Australian Education Index* (Radford, 1958, p.1) remains just as true for ACER's Cunningham Library as it continues to maintain this Index in 2015. The *Australian Education Index* (AEI) is a bibliographic database containing over 200,000 entries and abstracts. A rigorous selection process ensures comprehensive coverage of significant Australian education research. The challenge of curating and indexing the literature on Australian education required by administrators, teachers and students is one of even greater complexity and cost, as types and sources of literature increase and the topics related to education expand.

The *Australian Thesaurus of Education Descriptors* (ATED) (ACER, 2013) is a hierarchically-structured thesaurus of concepts across all levels of Australian education. It is used to index and search the subject matter of the AEI and its subsidiary databases, as well as the Library's catalogue. While ATED's primary use is as an in-house subject vocabulary, it is also searchable free of charge online, and can be purchased in hard copy or as an electronic dataset to be embedded into a third party organisation's own information services. First published in 1984, ATED is now in its fourth edition and is updated on a six-monthly basis. As at August 2015, it contained 10,348 terms, around half of which are preferred terms, and half are references.

The process of producing the *Australian Education Index* involves the following information tasks.

1. Identification of potential sources
2. Acquisition of identified sources
3. Selection of relevant material from these sources
4. Cataloguing or indexing of selected material
5. Quality assurance of indexed records
6. Dissemination of records to users

While these six components of production for the AEI have been constant since 1958, there have been changes in the way they are performed over the intervening years. This has been in response to both the changing formats of the resources being indexed, and the format of the Index itself. It was originally a print-only publication, then moved to print and CD-ROM, and is currently a purely online product licensed via Informit, ProQuest, and Transmission Books & Microinfo Taiwan. The Index has been moved between metadata platforms several times in its lifetime. Currently the Index, the Cunningham Library's catalogue and related services are developed in Inmagic's DB/TextWorks, which provides the flexibility to

readily modify metadata schema, redesign forms and edit screens to suit regular changes to processes.

Selecting and indexing Australia's education literature is labour intensive and thus an increasingly expensive activity. Curating the ever-growing range of documents and assigning thesaurus terms to metadata records are intellectually demanding processes, as well as being time consuming. While the value of providing the Index is not disputed, increasing costs, as well as a decrease in indexing output, meant support for the professional indexers was required. One strategy was to investigate ways of automating the indexing process.

Automated indexing

The quest for automated indexing is not new. A 1965 monograph by Stevens, entitled *Automatic indexing: a state-of-the-art report* contains almost 200 pages of experiments in 'automatic assignment indexing, automatic classification and categorisation, computer use of thesauri, statistical association techniques, and linguistic data processing' (p.1). 'Automatic indexing' as a concept was added to ATED in 1984, with a related term 'computational linguistics', which is described as:

> A branch of linguistics concerned with the use of computers for the analysis and synthesis of language data - for example, in machine translation, word frequency counts, and speech recognition and synthesis (ATED, 2013, p. 25).

A search of Australian education and librarianship literature reveals minimal local work on automated indexing in traditional library catalogues or indexes, although there have been projects related to automated metadata in web-based education services (Leibbrandt et al., 2010). Windsor (2015) uses two high profile cases of unfortunate and unacceptable automated tagging to conclude that while automated metadata generation might offer benefits, that use should come with caveats such as 'they cannot be left unattended and need to be checked by human beings.'

A catalyst for ACER's investigation of automated indexing was a presentation at the ALIA Online Conference 2011 from the Parliament of Australia Parliamentary Library. Hutchinson, Missingham & Anderson (2011) outlined an automated classifier project to both select and index news items, using the Parliamentary Library Thesaurus. Their system provider considered a number of tools and techniques including Bayesian probability, decision trees, and support vector machines, ultimately developing and implementing a package for their needs based on a variant of Bayesian categorisation.

After attending this presentation, the idea of finding a tool to help assign subject headings from ATED took hold. Speed was of particular importance to the Parliamentary Library with a requirement for same day selection and classification of news. This was less of an issue for ACER, with indexing records delivered in batches over one or two month periods. However, the functionality that showed relative weighting of the recommended categories was considered very useful by the ACER project team. The Parliamentary Library was working with a single pool of consistently structured news media metadata in digital form, from a single supplier. The Cunningham Library was working with print journals from many different publishers. However, as journals converted to digital, acquisitions practices changed and many publishers no longer provided 'free for indexing' print journals, but rather a

stream of metadata for articles. These files were seen as source data for feeding an automated indexing workflow.

Another example of an indexing service implementing a classifier at that time was the National Agriculture Library (NFAIS). A case study by NFAIS (2014), entitled *Automated indexing: A case study from the National Agricultural Library,* outlined their implementation. Of particular interest was that a significant proportion of their indexing was of journal articles, and they also used a thesaurus. However, their change strategy was to stop indexing altogether for a year whilst choosing and implementing their automated indexing system, which was not a preferred option for ACER. As existing services, both the Parliamentary Library and the National Agriculture Library had a collection of documents that had been indexed manually, and a mature thesaurus from which index terms had been selected. ACER was in the same position. This differs from projects that start with a set of documents and generate bespoke taxonomy from these (Lyte et al., 2009; Randtke, 2003).

Given the similarities between the Parliamentary Library's indexing requirements (Hutchinson, Missingham & Anderson, 2011) and those of ACER's indexing work, the same provider was approached about a solution. Initial discussions were held with Phil Anderson of SAIC (now Leidos Pty Ltd) about a machine learning classification tool (the classifier) that would work with ACER's metadata. This led to ACER commissioning an initial feasibility study in 2013, which extended in 2014 to installation and ongoing use of the classifier software in-house.

Training the classifier

In the investigation phase of the project, Leidos received an export of all records from the AEI master database in Extensible Markup Language (XML) format, and processed this using the TeraText classifier program. They took 90% of the data as a training set and learnt rules from those records. Then, the remaining 10% of records were processed using those rules to attempt to assign categories in a manner similar to that of a human indexer. By setting different thresholds and maximum number of categories assigned, and using different sets of data for training, the precision and recall of the suggested terms could be improved. Full-text resources were also supplied for a proportion of the records. However, tests found better results were obtained by using just the data from the article title, journal title and abstract rather than the full-text document. This was attributed in part to the consistent style used in an abstract, the difficulty of obtaining 'clean' information from the PDF documents, the vastly larger vocabulary the classifier had to deal with across full-text documents and the fact that journal articles vary more in length than abstracts do.

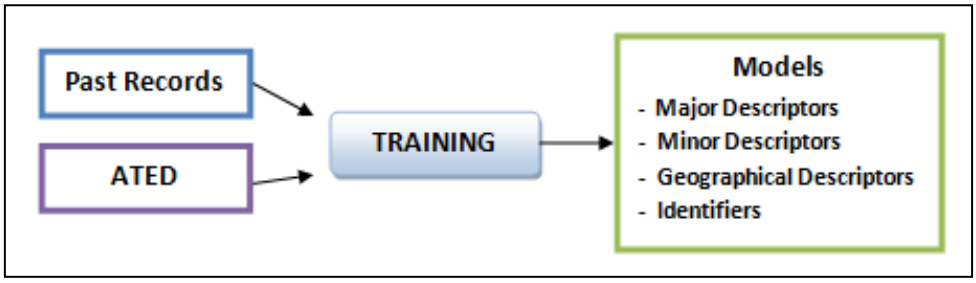

Figure 1 Training: The training program creates models for each of the descriptor fields using information from past records

Initial training provided models for ‘subjects’ - a combined field encompassing four separate fields in the AEI metadata profile, each related to ‘about-ness’.

- Major descriptors: Subject terms from ATED.
- Minor descriptors: Restricted list of terms from ATED used to indicate research methodology and educational level.
- Geographical: Terms for countries, regions, Australian states, cities and towns.
- Identifiers: Other keywords that are principally either proper names, or natural language concepts not yet represented by ATED terms. One way of identifying candidate terms for ATED is to monitor these Identifiers for usage and upgrade them to ATED descriptors when warranted. A threshold of around 20 instances of a term as an Identifier indicates the term should be considered as a new ATED descriptor.

Later, when work was done to implement the classifier into AEI’s production environment, separate statistical models were developed for terms in each of these four fields.

The training is intended to be run periodically in order for the most recent records to be included and therefore improve performance over time. Whenever ATED is updated with new or removed terms, the classifier needs to be trained accordingly. As indexing styles and standards have changed significantly over many years, it has been necessary to experiment to determine which past data to use for the training. For instance, the average number of terms assigned to journal articles has gradually increased from 6.1 in 2000, to 8.7 in 2015. The currency of terms is also important to consider, as the aim is to get the classifier assigning terms that are in common usage today.

Running the classifier

The classifier assigns terms to the four descriptor fields in new records, based on the models created by the training program. It looks at the article title, abstract and journal title.

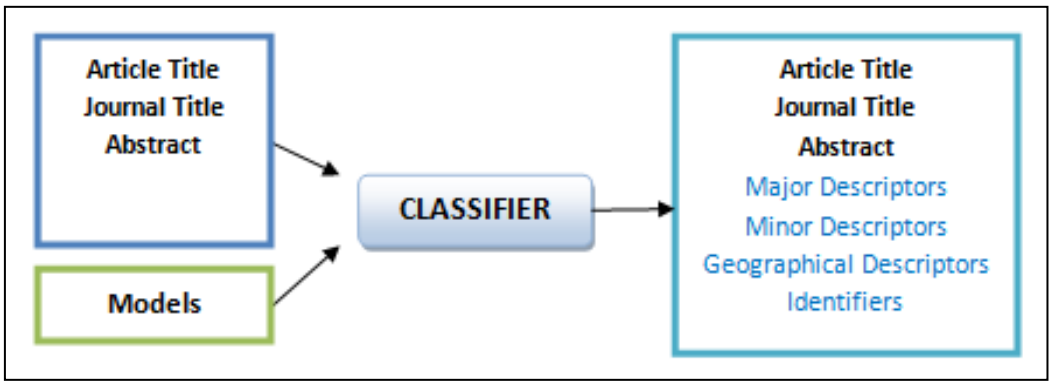

Figure 2 Classifier: Using the models, the classifier suggests descriptors for a new record based on its title, journal title and abstract

For each field, each potential term is given a numerical value (weight) based on the presence of words and phrases within that field that are identified as relevant according to past usage. Any term that has a weight above a threshold value is

assigned to the record, up to a specified maximum number of terms (N). The potential terms are then sorted into order based on their relative weights, and the best 'N' are assigned to the record. The significant words and phrases for each assignment are available for the human indexer to peruse if they wish. The thresholds and maximum number of terms differ for each field and are adjustable.

In the integration phase of the project, Leidos provided a way for the classifier to ingest records exported from DB/TextWorks in XML format. Staff can now run the classifier on single records or batches. When complete, the classifier produces a new XML file containing the records with suggested terms added, which is in turn imported back into DB/TextWorks. A human indexer then completes each record, by removing suggested terms that are incorrect or too general, and adding more relevant thesaurus terms as appropriate.

Implementing the classifier

Unlike the Parliamentary Library project, which switched off their existing system and replaced it with the new system, ACER has continued with business as usual while investigating how and where to best make use of the classifier. ACER's Senior Librarian Technical Services managed the implementation, supported by an IT-qualified in-house indexer, who worked with the software provider during the feasibility stage. Once the classifier was implemented, a collaborative method of working across the indexing team was adopted. Staff were involved in regular discussions concerning bottlenecks in the workflow, quality issues in the records, and refinement of the proposed changes to process.

The publisher metadata means less indexing time is spent on keying in or copying and pasting information into a record. Unfortunately, every publisher uses different data structures, which required creation of Extensible Stylesheet Language Transformations **(**XSLT) stylesheets to import all the XML feed data from different sources into a single DB/TextWorks database. The stylesheets massage the data into a fairly consistent dataset to use for indexing. Tasks like this reinforced the value of a dual-qualified indexer on the team, in terms of technical expertise, immediate availability to troubleshoot and the ability to program solutions.

How the classifier has performed

The terms suggested by the classifier, whether removed by the human indexer or not, are saved in each record. This provides statistics that have been used to evaluate the classifier's performance over the first 707 records completed. The focus of this analysis is the major descriptors field, which is the most important field in terms of evaluating the classifier. The other three subject-related fields do not currently perform as consistently, and the priority thus far has been on optimising the performance of the major descriptors.

Number of major descriptors assigned

The classifier is currently set to assign a maximum of thirteen major descriptors to each record. Of the 707 records, the majority were assigned the maximum of thirteen. The overall average was 11.71.

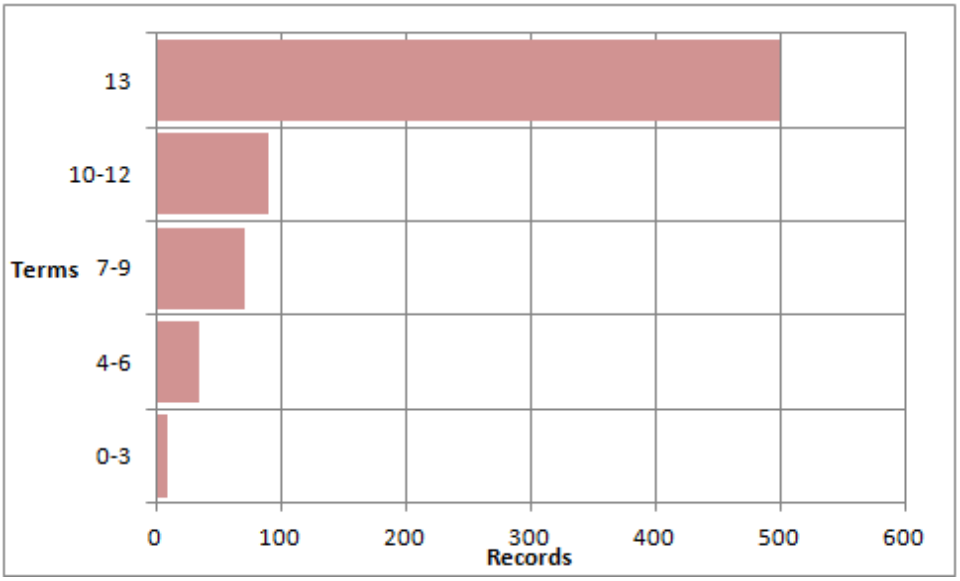

Figure 3 Number of major descriptors assigned by the classifier

The average number of major descriptors used in the completed records (after human indexer input), was 10.32. There was much more variation above thirteen however, as there is no defined maximum number of terms.

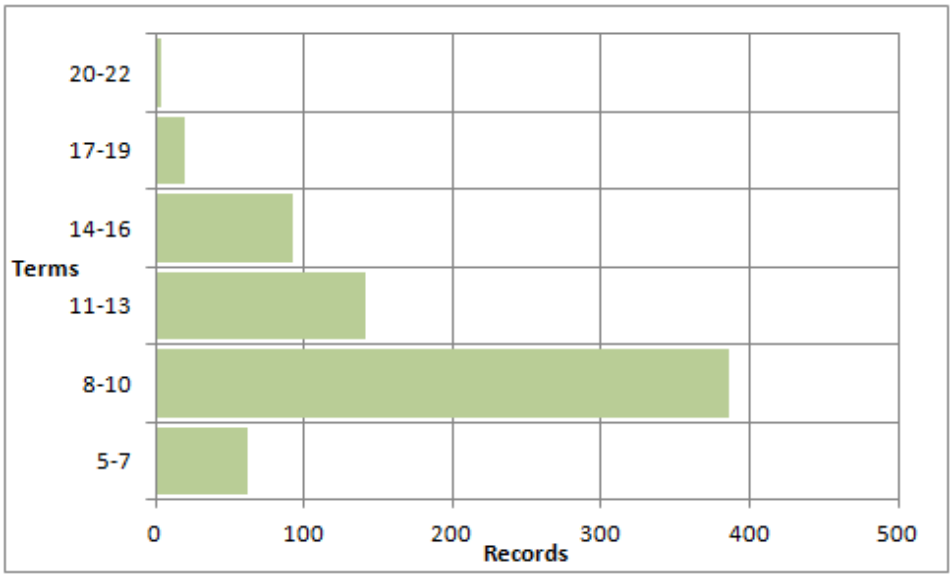

Figure 4 Number of major descriptors in completed record

Classifier suggestions

Each record was also analysed for the following:

- terms suggested by the classifier that were accepted for the final record (categorised as 'correct');
- terms suggested by the classifier that were rejected for the final record ('incorrect'), and
- terms in the final record that the classifier did not suggest and were added by the human indexer ('missed').

The average number of ‘correct’ terms per record was 6.66, and the average number of ‘missed’ terms per record was 3.66.

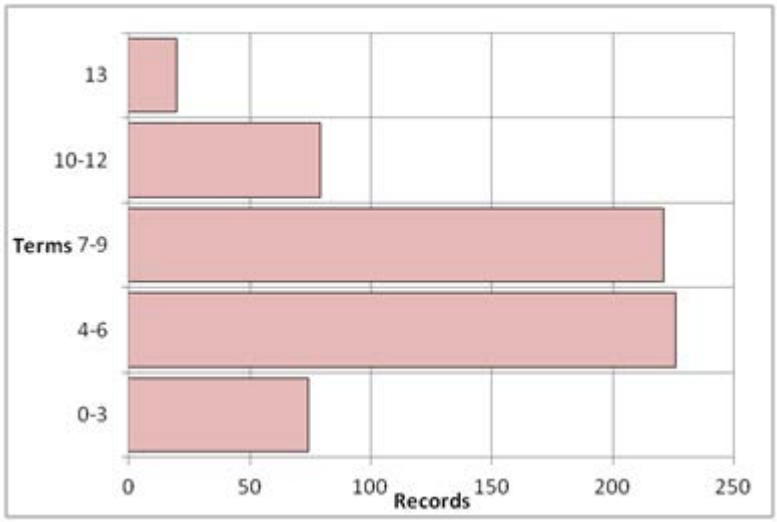

Figure 5 Number of ‘correct’ suggestions by the classifier per record

When evaluating the classifier's performance using these statistics, it is important to consider they are dependent on decisions made by a number of different human indexers, each with differing indexing styles. That is, there was no attempt to control for intra- or inter-indexer variation. These statistics also include the earliest records, since when changes and improvements have been made to the process.

What terms are being assigned?

The classifier assigned 1247 different major descriptors, of which 529 were assigned only once. The completed records contained 1645 different major descriptors, of which 700 of these were assigned only once.

Classifier assigned terms		Completed record assigned terms	
Frequency	**Descriptor**	**Frequency**	**Descriptor**
188	Student attitudes	104	Educational policy
156	University students	90	Teacher attitudes
143	Teacher attitudes	80	Student attitudes
125	Educational policy	70	Young children
110	Teaching methods	69	University students
90	Teaching practice	57	Teaching methods
88	Young children	52	Preschool children
86	Preschool children	50	Child development
80	Professional development	49	Teaching practice Professional development Curriculum development
77	Secondary school teachers	45	Knowledge level

Table 1 Most frequently assigned descriptors

The most common descriptors appear to be similar in both lists but the classifier is assigning them a lot more often. It makes sense that the most common terms keep being assigned, but it can be a problem if less common ‘correct’ terms are being missed.

The most obvious example of this problem is in the Geographical field, where the term ‘Australia’ appears in over 90,000 records in the Australian Education Index (about 9 times as often as the next highest term, ‘Victoria’). The consequence of this in terms of the classifier is that ‘Australia’ tends to be assigned to most records, sometimes incorrectly.

Performance of common terms

The classifier's performance was further analysed using the statistical measures of *precision* and *recall* for each term and overall. In this context, precision refers to the percentage of classifier suggestions that were 'correct', and recall refers to the percentage of terms in completed records that were suggested by the classifier.

ACER had no initial expectations regarding performance, and the classifier's settings (which influence recall and precision) have not been changed from those originally set by Leidos. Over time and with more feedback from indexers it will become more apparent how the classifier is being used and what balance of precision and recall to aim for.

For example, setting the classifier to assign fewer terms, and/or raising the threshold figure may result in fewer 'incorrect' terms being assigned (*higher precision*) but at the cost of a greater number of terms that are 'missed' (*lower recall*). For the indexer completing the record, higher precision and lower recall would mean less time spent removing irrelevant terms, but potentially more time spent adding new terms.

Precision and recall can be combined to produce the F1 score, which is a weighted average that gives an indication of overall performance for each term. If specific terms are performing poorly in these measures, it can indicate where potential adjustments could be made.

Terms that had been assigned 40 or more times by the classifier (35 terms) were ranked by F1 score.

Top 10 Terms		Bottom 10 Terms	
F1 score %	**Descriptor**	**F1 score %**	**Descriptor**
88.61	Young children	29.17	Early childhood education
83.33	Curriculum development	37.11	Secondary school teachers
82.69	Knowledge level	37.33	Primary school teachers
82.10	Educational policy	50.00	Teacher education programs
81.19	Child development	50.79	International students
80.00	Mathematics teaching	51.95	Secondary school students
80.00	Educational leadership	52.46	Outcomes of education
77.42	Student experience	55.97	Student attitudes
76.71	Literacy education	56.29	Teaching methods
75.36	Preschool children	56.67	Teacher improvement

Table 2 Highest and lowest F1 scores for most common terms

The top term 'Young children' appeared in 70 final records and all of these were suggested by the classifier (100% recall). The terms 'Primary school students',

‘Primary school teachers’, ‘Secondary school students’ and ‘Secondary school teachers’ all had very low precision values, indicating that the classifier may be struggling to differentiate between common but similar terms. The most commonly assigned term, ‘Student attitudes’ had one of the lowest precision values.

As an educational level, 'Early childhood education' is more commonly used as a minor descriptor. Initially however, the classifier was regularly assigning it to both fields. It was decided to edit the training models to ensure it would no longer be suggested as a major descriptor. The low F1 score for this term is explained by these early 'incorrect' suggestions.

Generally, the classifier is performing better for recall than precision. This is to be expected given the classifier's settings – thirteen terms is often more than is necessary. Recall is especially high with the most common terms. The overall recall for these 35 terms was over 90%, compared to the precision of 52.1%. This suggests that the classifier is rarely missing these common terms when they are relevant, but is also assigning them too often.

Across all terms and records, 56.9% of the terms suggested by the classifier were ‘correct’ (precision) and 64.5% of terms in the final records were originally suggested by the classifier (recall). ACER did not have any benchmark on which to base targets or expectations prior to the project, but these results are encouraging. As the classifier is used for more and more records, the statistics will become more meaningful and will allow greater insight into how to improve overall performance.

Factors influencing classifier performance

The classifier regularly provides a useful set of terms, but there are instances where its performance is negatively affected by the abstract, the topic of an article, or the hierarchical nature of the thesaurus.

Abstracts

The length, style and level of detail of the abstract are vital to the classifier producing accurate terms. As the abstracts are not checked prior to the classifier being run, it relies on whatever is provided by authors and publishers. This element of inconsistency in source data further highlights the difference between ACER’s project and that of the Parliamentary Library.

Topic of article

While ACER indexes from a wide range of educational and other journals, some articles might have only a peripheral link to education. Some articles that come from areas such as music education, mathematics education, philosophy and psychometrics can be difficult to index using ATED. The thesaurus will have some general concepts that can be used, but not necessarily anything more specific or in-depth.

Examples of titles of articles where the classifier performed poorly for this reason:

> *Playing with performance : The use and abuse of beta-blockers in the performing arts.*
>
> *Detecting distortion : bridging visual and quantitative reasoning on similarity tasks.*

Person proficiency estimates in the dichotomous Rasch model when random guessing is removed from difficulty estimates of multiple choice items.

Also problematic are articles that are from the field of education, but where the main topic is uncommon, new or not adequately covered in ATED. In these cases, the classifier will assign general, common terms, but may miss the specific concept.

Example:

Proposing a comprehensive model for identifying teaching candidates

The classifier actually assigned the maximum thirteen terms to this article, but only two were ‘correct’, indicating that the article’s topic (pre-service teacher candidate selection tests) was too specific. It is generally the case that articles that are a challenge for the human indexer to deal with generally do not yield good results from automatic classification.

Thesaurus structure

Many of the classifier’s suggestions could be considered relevant to the article, but not necessarily at the most appropriate level of specificity. ACER’s indexing guidelines reference two fundamental indexing rules from the Education Resources Information Center (ERIC) (2001, p. xvii):

1. Index only what is in the document
2. Index at the level of specificity of the document

The classifier is not currently designed to understand the hierarchical nature of the ATED thesaurus well enough to follow the second rule as accurately as a human indexer. It does not attempt to read and interpret the explicit hierarchy of the thesaurus terms. This is a known limitation of the current version of the classifier, and is an area of ongoing research. It can only use whatever terminology exists in the record, and often terms are assigned alongside neighbouring terms, such as broader or related terms. An example of this problem is the term 'Leadership', which had both the lowest F1 score and the lowest precision. The classifier assigned this to 25 records but only two of these were considered ‘correct’. In most of those cases, the human indexer would have chosen more specific terms about leadership.

The human indexers used 622 terms that the classifier never assigned. Many of these are terms that have not had much usage historically. The training program excludes a term if there is insufficient evidence to build a good model of how to assign that term in the future. In effect, this means that the human indexer has a significantly wider vocabulary of terms to choose from than the classifier does. This also highlights the importance of the chosen training data set – the classifier's vocabulary can end up looking quite different depending on which past records are used.

Discussion

The introduction of the classifier had an impact on almost every part of the work of indexing. A number of the most significant changes and lessons learned are discussed below.

Indexing from the abstract instead of full text

A major change wrought by the classifier project is the acceptance of abstract-only indexing. In the past, articles were only indexed when access to the full text was available. Some publishers still provide indexers with access to the full text online, but others will not, and the classifier has demonstrated that it assigns terms best when considering just the article title, journal title and abstract. If the abstract is the recommended source for the classifier, this assumes the abstract contains all the relevant information required to either select or reject an article, and to determine the best thesaurus terms. The vital role of abstracts for the classifier's performance strongly indicates that abstracts should be of the informative or structured style rather than the descriptive or unstructured style (Cook et al., 2007, p.1075).

Role of the thesaurus

The fact that ATED is not just a controlled vocabulary, but a hierarchically-structured thesaurus or ontology, helps to explain the issue, with the classifier sometimes assigning a concept that is either too general or too specific, as there is significant overlap between 'subjects'. There is also recognition that 5,000 terms is a large number for an automated system to learn and assign.

The classifier assigns concepts from the thesaurus, so it is vital to ensure that ATED is up-to-date with terminology and references relevant to the literature being indexed. ACER staff and users of ATED recommend thesaurus updates, and the Library welcomes projects that evaluate and enhance the thesaurus such as the recent re-indexing of the Office of Learning and Teaching Resource Library (Hider et al., 2015). Making changes to the thesaurus however means retrospectively updating records already in the AEI and then re-training the classifier. This is because any new terms will never be assigned by the classifier until they have been used in at least 5 records, (but preferably 10 or more records).

As the classifier learns from the thesaurus terms allocated in existing records, it is also important to ensure existing records are as accurate as possible. An agreed subset of the ATED terms has now been implemented for use in the minor descriptors field, and all existing records have been upgraded so they contain only valid thesaurus terms in the major and minor descriptor fields. The geographical field and non-ATED terms in the identifiers field were similarly normalised.

The role of metadata

Arranging for, and customising metadata feeds is not a one-off activity. There are a number of points in the chain where automated online delivery and access can breakdown. For example, a library system software upgrade introduced a change in character encoding that required significant re-coding. There have been occasions where a particular week's data hasn't arrived, or a title has unexpectedly dropped out

of the feed. The metadata from publishers does not always conform to the validation rules present in some fields in our database. This issue was resolved by adding new non-validated fields that allow the data to be imported before being entered in the correct form by the indexers. Not everything to be indexed has metadata available in a form that can be harvested into a feed, so there is still a need for manual data entry. The Library is investigating other ways to extract or obtain metadata, for instance from bibliographic utilities, or from other metadata and discovery services.

Process simplification

The most significant changes in terms of time and effort from this project have centred on simplifying the indexing and cataloguing workflows. Minimising the number of different processes, databases and metadata schemas has involved a number of iterations. For example, separately named databases holding the records for each batch of indexing gave way to a single indexing database. Having consistent database names facilitates search profiles and scripts to automate certain processes. The metadata creation forms for both internal and external indexers were updated, thus enabling external indexers to work in the same database. Several new fields were added to both the catalogue and the indexing database to facilitate the import of suggested terms.

With multiple selection streams, it is possible to index the same item twice. There was a need to design search screens that make it easy to identify:

- what has already been indexed in the master database;
- what needs to be run through the classifier;
- what is waiting for completion; and
- what is completed and ready to be quality checked.

Prioritisation

The classifier trial was conducted as a proof of concept project with the participants involved in both the business as usual and the testing phases. This has had the advantage of permitting immediate research and rapid testing and implementation of new ideas as they are raised. It has the obvious disadvantage of slowing down both the business as usual and project work since they involve the same people. There are ‘classified’ journal article records requiring completion, as well as shelves of print material waiting to be indexed. This poses the dilemma of prioritisation. Do we preference the quicker ‘classified’ indexing or the more labour intensive articles from smaller publishers whose data is less likely to be available elsewhere? How much indexer and technical time should be spent on developing and tuning the classifier?

Indexer experience

Feedback from indexers was an essential component of the project. As in any team, there were varying levels of commitment to the change process and it helped that the champions of the project were well-regarded and trusted team members. As well as collecting detailed notes of the technical issues faced by indexers, and their preferences for interface and workflow, the emotional elements of the project were a topic of discussion. Initially there was a level of discomfort and a loss of productivity due to using new forms that changed the location of fields. Indexers reported that it was more draining to be providing solely intellectual input to an indexing record, rather than a mixture of manual data input combined with thesaurus term selection.

One indexer mentioned that the previous process of routine copying and pasting metadata in fact helped to build their knowledge of the article. There were also feelings of achievement about how many items they could get through. According to feedback from indexers, having a list of suggested terms already in the record makes the process of completing a record faster in most cases than creating it from scratch. It was also valuable that the classifier provided evidence for its choices by displaying a weighting. Indexer experience is an area of research that should continue, perhaps taking the opportunity to track and document any changes of satisfaction over an extended implementation period. It would be interesting to study a new indexer who learns their craft using only the automated system.

Curation

One important realisation from this project was that while there had been great interest and attention paid to streamlining the work of indexing, there was in fact more complexity in the areas of identification, acquisition and selection. Selection for the AEI involves a set of decisions based on knowledge of the priorities in Australian education, knowledge of the scholarly publishing industry, and the needs of multiple audiences. Library staff have come to realise that the curation aspect of the AEI is in fact its prime value, and that perhaps automating the process of identifying and/or selecting candidate documents for indexing is a place to invest future efforts. This reflects the findings of the Parliamentary Library's project, which initially aimed to automate subject classification only, but saw the potential and benefits of using the same Leidos classifier technology for automating selection, thus allowing the library to deal with much larger volumes of incoming data by quickly discarding irrelevant material. The experience of the Parliamentary Library in automated selection was that the categoriser should not use a binary classification of yes or no, but one of three options, yes, no, or maybe, when it has insufficient information (Hutchinson, Missingham & Anderson, 2011, p.9).

Future development

There is a list of items at various stages of design and implementation on the development list, including:

- Create the ability to easily run the classifier from within the catalogue as well as the indexing database;
- Develop a button to run a single record through the classifier to allow indexers to use the classifier on records that have had their bibliographic and abstract details added manually;
- Negotiate to receive feeds from more publishers;
- Investigate replacing the existing minor descriptors field with two new separate fields: Educational level (subject) and Methodology;
- Include a broader document type vocabulary, which requires additional work to set up new models, as well as populating the fields;
- Add an indicator for peer reviewed content and populate it from information currently contained in the notes field (options to be Yes, No or blank);
- Ensure system does not index the same item twice if journals are received in both print and via the publisher feeds;

- Refine XSLT stylesheets to minimise the need for multiple export forms from the publisher feeds database. Currently each publisher requires a different form to overcome the inconsistencies in data;
- Automate selection from the publisher feeds database using canned searches or profiles;
- Improve staff knowledge of how to train the classifier for optimum results;
- Indicate the metadata source as part of the subscriptions database; and
- Enhance the classifier program to take explicit account of the hierarchical nature of the ATED thesaurus.

Conclusion

So has use of the classifier improved the indexing quantity, quality, turnaround time and/or cost effectiveness of production? Any major system change affects productivity, and the Library's findings mirror the experience of other ACER automation projects, such as automated test generation and essay scoring: it takes longer than one year to realise promised productivity benefits. The classifier is a tool for suggesting subject terms from ACER's controlled vocabularies. It does not completely automate the subject classification task, let alone automate the entire indexing process. After ACER's first year of trialling the subject classifier, the Library has found that its conclusions mirror those of the Deutsche Nationalbibliothek.

> It has become very clear that it is not easy to develop and implement a process using a universal controlled vocabulary for automated indexing of a universal collection. But we are convinced it can be done with reasonable results as long as the claim is not that automated indexing produces the same results as rule-guided intellectual indexing (Junger, 2012, p.6).

Introducing the classifier has, however, influenced indexing workflows across all stages, and has greatly increased the use of machine-readable data. While so far the additional time spent in developing and refining new systems outweighs the time saved in indexing individual records, the improvements made to systems have been positive, and show potential for scaling as each solution clears another barrier towards full integration of the classifier.

References

Australian Council for Educational Research (ACER) 2013, *Australian Thesaurus of Education Descriptors,* 4th edn. ACER Press, Camberwell, viewed 16 August 2015, http://www.acer.edu.au/ated

Australian Libraries Gateway 2015, *Definitions of the ALG collecting levels*, National Library of Australia, Canberra, viewed 16 August 2015, http://www.nla.gov.au/libraries/help/subjects_levels.html

Cook, DA, Beckman, TJ & Bordage, G 2007, 'A systematic review of titles and abstracts of experimental studies in medical education: many informative elements missing', *Medical Education*, vol. 41, no. 11, pp. 1074-1081, doi:10.1111/j.1365-2923.2007.02861.x

Education Resources Information Center (ERIC) 2001, *Thesaurus of ERIC descriptors,* 14th edn, Oryx Press, Phoenix AZ.

Hider, P, Spiller, B, Mitchell, P, Parkes, R & Macaulay, R 2015, Towards a New Library of Resources for Higher Education Learning and Teaching*,* paper presented at *The Higher Education Technology Agenda (THETA)* conference*,* 8-10 May, Gold Coast, viewed 16 August 2015, http://theta.edu.au/program/presentations-2015/towards-a-new-library-of-resources-for-higher-education-learning-and-teaching

Hutchinson, J, Missingham, R, & Anderson, P 2011, Revolutionising digital content ingest: building a newspaper clippings collection using practical automation to assist with selection and classification, paper presented at the *ALIA Information Online Conference & Exhibition*, Sydney, viewed 16 August 2015, http://pandora.nla.gov.au/pan/94107/20110823-1043/www.information-online.com.au/sb_clients/iog/data/content_item_files/000001/paper_2011_c9.pdf

Junger, U 2012, *Can indexing be automated? The example of the Deutsche Nationalbibliothek*, Deutsche Nationalbibliothek, Frankfurt, viewed 16 August 2015, http://www.nlib.ee/html/yritus/ifla_jarel/papers/3-4_Junger.docx

Leibbrandt, R, Yang, D, Pfitzner, D, Powers, D, Mitchell, P, Hayman, S & Eddy, H 2010, Smart collections: Can artificial intelligence tools and techniques assist with discovering, evaluating and tagging digital learning resources?, paper presented at the *School Library Association of Queensland and the International Association of School Librarianship Conference incorporating the International Forum on Research in School Librarianship*, 27 September- 1 October, Brisbane, viewed 16 August 2015, http://eric.ed.gov/?id=ED518554

Lyte, V, Jones, S, Ananiadou, S & Kerr, L 2009, *UK institutional repository search: innovation and discovery*, viewed 16 August 2015, http://www.ariadne.ac.uk/issue61/lyte-et-al

National Federation of Advanced Information Services (NFAIS) 2014, *Automated Indexing: A case study from the National Agricultural Library,* webinar viewed 10 April 2014.

Radford, W 1958, 'Preface', *Australian Education Index,* Australian Council for Educational Research, Melbourne.

Randtke, W 2013, 'Automated metadata creation: Possibilities and pitfalls', *The Serials Librarian: From the Printed Page to the Digital Age*, vol. 64, no. 1-4, pp. 267-284, DOI: 10.1080/0361526X.2013.760286

Stevens, M 1965, *Automatic indexing .A state-of-the-art report*, Monograph 91, 30 March, National Bureau of Standards, Washington. D.C.

Windsor, R 2015, 'Google's visual case study of the perils and politics of automated metadata', *Digital Asset Manager News,* 9 July, viewed 16 August 2015, http://digitalassetmanagementnews.org/taxonomy-metadata/googles-visual-case-study-of-the-perils-and-politics-of-automated-metadata

Glossary

AEI	Australian Education Index https://www.acer.edu.au/library/australian-education-index-aei
ATED	Australian Thesaurus of Education Descriptors http://www.acer.edu.au/ated
F1 score	In information retrieval, a measure of document classification performance that provides a weighted average of precision and recall
Precision	In information retrieval, the proportion of results that are relevant
Recall	In information retrieval, the proportion of relevant results that are returned
XML	E**x**tensible **M**arkup **L**anguage – W3C standard for encoding text for storing and transporting data, http://www.w3.org/TR/xml
XSLT	E**x**tensible **S**tylesheet **L**anguage **T**ransformations – used to transform XML documents into other formats, http://www.w3.org/TR/xslt20/