

Classifying scientific publications for the measurement and analysis of R&D undertaken in Australia and New Zealand.

- ARC assesses Australian research excellence every 3-4 years via ERA
- Research metrics are part of the assessment
- Research is categorized into Fields of Research (FoR)
- Elsevier has extensive expertise in deep learning networks and using multiple data sources.
- The Elsevier data science team have categorized the Scopus journals into the FoR codes using a combination of machine learning and natural language processing

Introduction

The Australian Research Council's Excellence in Research for Australia (ERA) is Australia's national research evaluation framework. ERA promotes and identifies excellence across research activity in Australia's higher education institutions. Excellence in research in ERA is identified by comparing Australia's university research against international benchmarks.

Metrics around the scholarly impact of an institution are part of the initiative, which are looked at in distinct disciplines called Fields of Research (FoR). Fields or disciplines are defined as two, four and six digit FoR codes as identified in the Australian and New Zealand Standard Research Classification. Further information about ANZSRC is available on the [Australian Bureau of Statistics website](#) [Australian and New Zealand Standard Research Classification](#).

To assist Universities in their planning and modelling, Elsevier has used internal capabilities, together with multiple data sources to classify the journals in Scopus into the FoR disciplines. Details of the methodology are below.

Problem Formulation and Objectives

In 2020, a new set of FoR codes were released to the research community. In some cases there were one to one matches but in many cases there were one to many matches from the old to the new codes most specifically at the 4 and 6 digit levels.

In the past, a common practice for the realization of such systems has been to perform the classification task at the journal level. The primary method has been the assignment of one or more FoR codes for each journal, and the inheritance of these FoR codes by the individual articles published in each of the respective journals. With respect to the limitations of such an approach Elsevier has implemented Machine Learning (ML) and Natural Language Processing (NLP) automated classification system against FoR codes, at individual article level.

The key objective was to use the below approach to create mapping across all journals currently indexed in Scopus, from the old to the new FoR codes, to then be displayed in SciVal and potentially across other Elsevier Research Intelligence products.

Overall Approach

Elsevier is spearheading one of the most important challenges of classifying, indexing and extracting information from scientific text, for the purposes of biomedical text mining, namely BioAsQ [1]. The art of classifying and indexing scientific text has evolved rapidly over the last decade, especially since the natural language understanding algorithms have become more efficient and accurate, e.g., by means of applying the most effective self-supervised learned embeddings (BERT, SciBERT, BioBERT, word2vec, Glove,) via a variety of deep learning architectures, and by then fine tuning these architectures to the relevant task.

More specifically, for the organization of scientific articles against research fields, Elsevier has fine-tuned the latest deep learning networks for classification against concepts from a large, polyhierarchical taxonomy that covers all of the scientific domains of articles indexed in Scopus, in a very high granularity. This polyhierarchical taxonomy is called Omniscience, and comprises ~50,000 concepts organized in multiple ways (i.e., the notion of polyhierarchy) across up to 17 levels of depth, in order to be able to cover the different views of the world (taxonomies) of the national research agencies across the globe. The advantage of this structure is that, practically, every existing national research taxonomy can be mapped to Omniscience, and the same, trained models can be re-utilized for the classification against that scheme.

As it is often the case, the national research taxonomies, as are the FoRs, have a very specific structure, which can also be utilized efficiently (hierarchical multi-label classification). In cases where there is a need to expand the models with concepts that are new, or not present in the current Omniscience view, the Data Science and Knowledge Representation teams can address this task in-house, with little effort and by using Subject Matter Experts (SMEs), active learning and semi-supervised learning techniques to bootstrap fast training (i.e., labelled) data, as well as validation/testing data. The mapping of national research taxonomies to Omniscience is also being conducted in-house, in a semi-automated manner, using a suite of taxonomy development, merging and management NLP techniques, and subsequent manual validation from in-house SMEs.

Data

Training data size: Elsevier has used roughly, 4 million data points mapping to all the 213 FoR Codes. This training data set continues to grow as the opportunity presents itself. This data was used to train and test a Machine Learning (ML) model that can assign codes to any piece of description (commonly abstract, funding information, grants proposals, etc).

Sources	○ Scopus - Journals mapped to FoR codes. ○ 1Findr - Journals mapped to FoR codes. ○ University data - Data shared by universities for training purposes. ○ University open repository data - Data gathered from multiple open access repositories with relevant FoR codes. ○ ERA Journal list - Mapping based on ISSNs using Scopus ○ SSRN - Mapping of networks/journals to FoR codes
---------	---

Open access repositories	All the open-access content present in these sources were extracted. We also have a semi-automated pipeline to refresh most of these resources • AITASIS • Digital Commons • Trove • New Zealand Research
University data	Collaborated with some universities who shared data with us which had this information.
ERA journal list	ERA journal mapping to ISSNs
Keyword-based search	Search for relevant keywords at the 6-digit code to aggregate the results to the 4-digit code level. Identify topics that can be part of this category, by process of elimination (from other 6 digit categories), search for relevant keywords, and assign the FoR code at the 6 digit level - roll this up to 4 digit code to get the relevant training data.

- The team recognise the complexity of Indigenous codes and “Not elsewhere classified” codes and the work to ensure best possible mapped results will be ongoing.

Success Criteria

By default, Elsevier sets high standards before releasing capabilities into its products, tools and services. We attach high value to quality. In this case, we will be aiming at the highest possible precision (P) and recall (R) and there the F-score. Our standards start at a P/R of 90%/90 and this threshold should apply to the 2- and 4-digits level (or as close as possible).

References

[1] Tsatsaronis et al., An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinform. 16: 138:1-138:28 (2015).

[2] ScienceDirect Topic Pages: <https://www.sciencedirect.com/topics/index>

[3] Kosmopoulos et al. Evaluation measures for hierarchical classification: a unified view and novel approached, Data Min. and Knowl. Discovery 29(3):820-865, 2015.