A Case Study on Applying Machine Learning Methods in Annotating Subject Headings to Dataset Records

November 9, 2022
IFLA Subject Analysis and Access Section WG

PRESENTED BY

Mingfang Wu, PhD
Australian Research Data Commons

mingfang.wu@ardc.edu.au

Outlines

- Background information: Australian Research Data Commons (ARDC) and the data catalogue - Research Data Australia (RDA)

- ANZSRC-FoR subject headings: The Australian and New Zealand Standard Research Classification - Fields of Research

- Project: Automatic classification/annotation of data records with ANZSRC-FoR subject headings - Approach & Result & Implication

Background information
Australian Research Data Commons
Data Catalogue: Research Data Australia

Australian Research Data Commons

Purpose

To provide Australian researchers with competitive advantage through data.

Mission

To accelerate research and innovation by driving excellence in the creation, analysis and retention of high-quality data assets.

ARDC Services

Research Data Australia

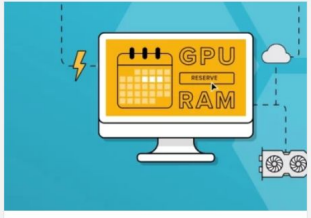
Find, access, reuse and attribute data from Australian research organisations.

Explore >

ARDC Nectar Research Cloud

Your national research cloud – start your free 6-month trial today.

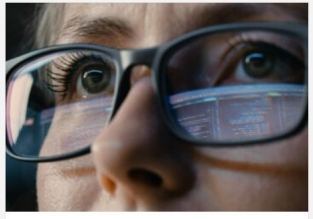
Explore >



ARDC Services Powered by the ARDC Nectar Research Cloud

Enhance your research with quick and easy access to extra computational capabilities.

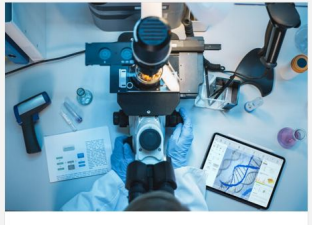
Explore >



Research Vocabularies Australia

Share and combine your data with confidence using ARDC Research Vocabularies Australia.

Explore >



ARDC Identifier Services

We provide a range of services for research organisations to create and manage persistent identifiers.

Explore >

Advisory Services

Supporting your data and digital research challenges.

Explore >

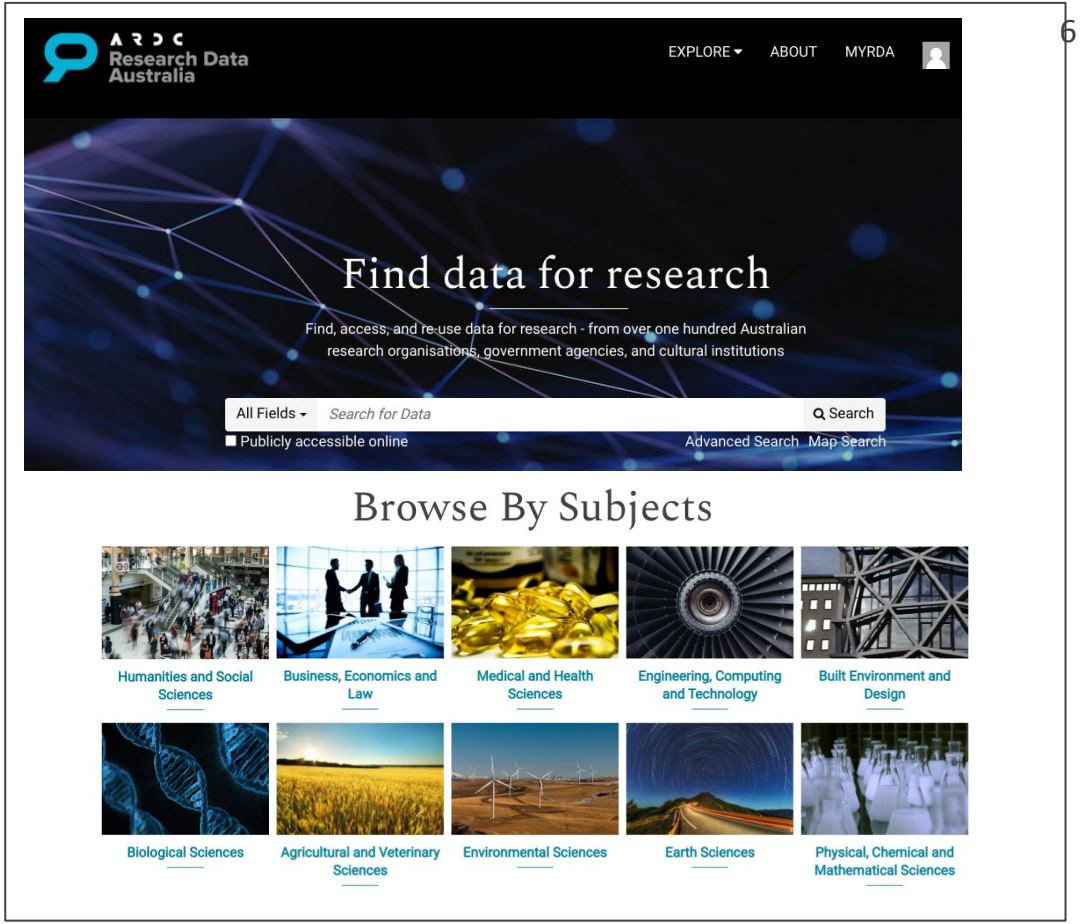
Communities and Groups

Explore and join the many digital research and data communities and groups we facilitate for...

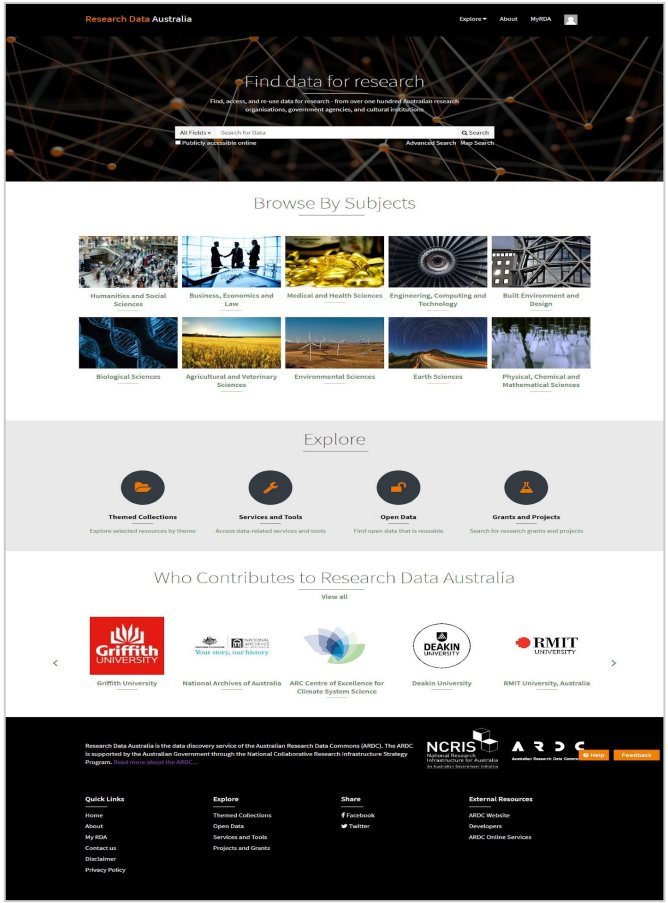
Explore >

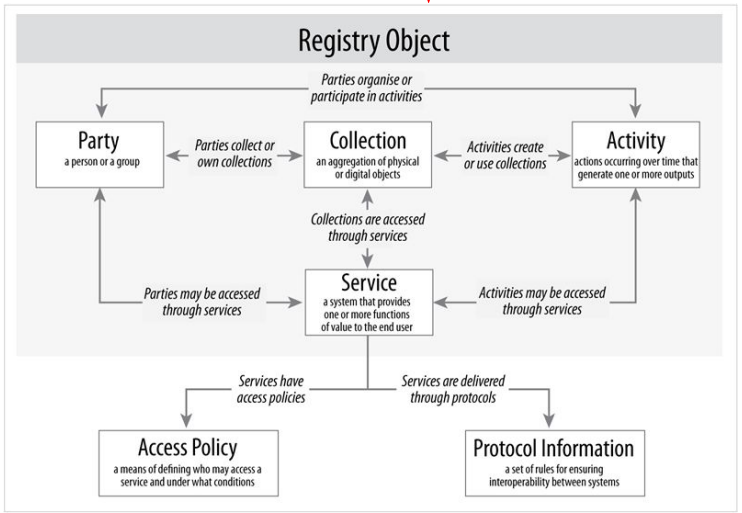
Research Data Australia

- Discovery portal for data (services and grants)
- Multidisciplinary
- Metadata only
- Over 190K datasets
- 104 Australian contributors (organisations)
- https://researchdata.edu.au



Research Data Australia - Schema and supporting services





Schema: The Registry Interchange Format - Collections and Services (RIF-CS, ISO 2146:2010)

Australian Research Data Commons

NCRIS National Research Infrastructure for Australia

ARDC is enabled by NCRIS

ANZSRC-FoR subject headings: A background about the usage of subject headings in the data catalogue

Subject metadata

- Benefit: Subject metadata is a powerful way of knowledge organisation and linkage of (distributed) resources for interoperability and discovery
  - RDA scheme offers the description of subject metadata, an optional but not mandatory field.
  - RDA data provider guides suggest to use subject metadata from controlled vocabulary, and offer vocabulary services for either publishing or accessing vocabularies.

- Cost: Manually labelling resources with subject metadata is not efficient and may introduce inconsistency and omission.

Types of subject headings

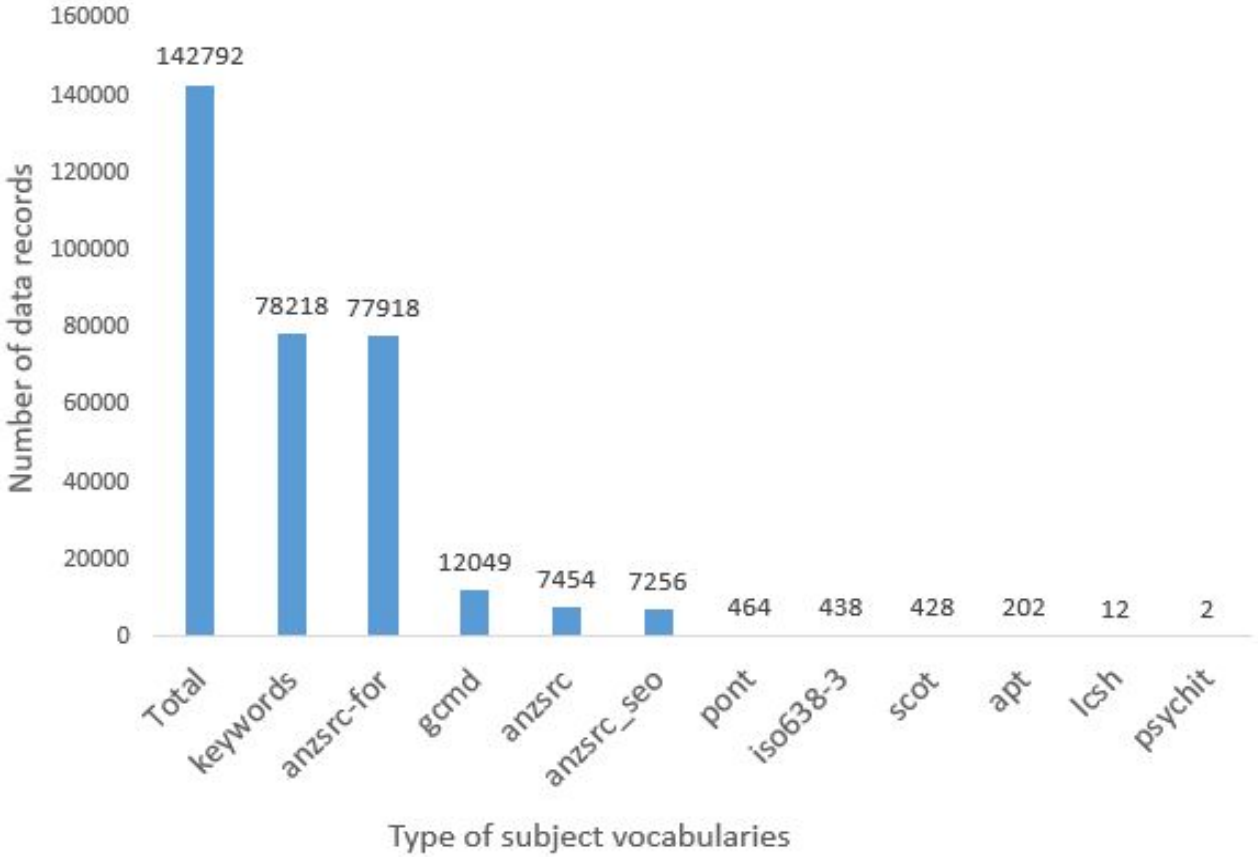
ANZSRC-FoR: The Australian and New Zealand Standard Research Classification (ANZSRC, fields of research)

Library of Congress Subject Headings (lcsh)

Australian Pictorial Thesaurus (apt)

Global change master directory (GCMD) keywords

Thesaurus of Psychological Index Terms (psychit)

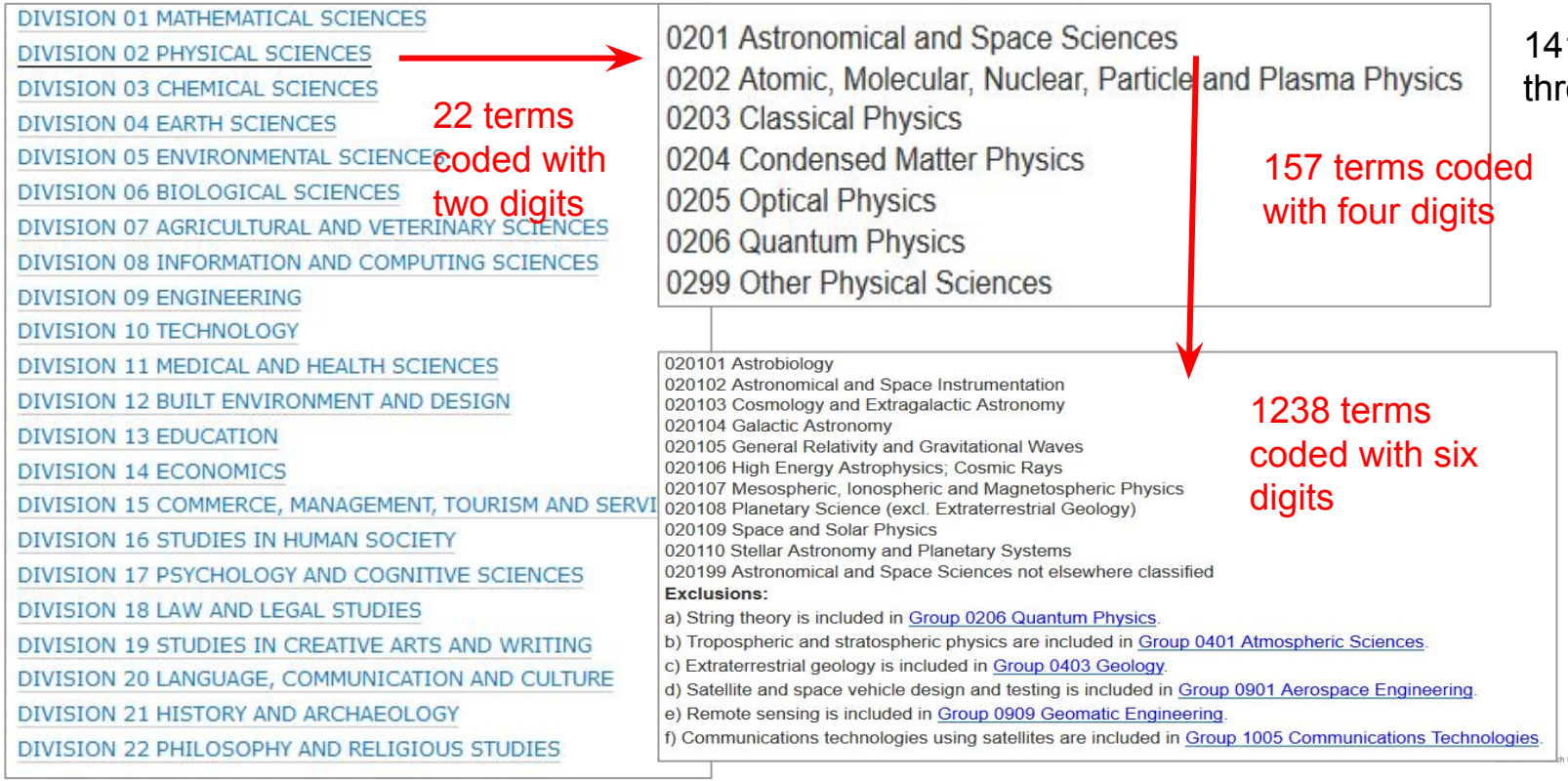


ANZSRC-FoR: The Australian and New Zealand Standard Research Classification - Fields of Research

- ANZSRC ensures that R&D statistics collected are useful to governments, educational institutions, international organisations, scientific, professional or business organisations, business enterprises, community groups and private individuals in Australia and New Zealand.

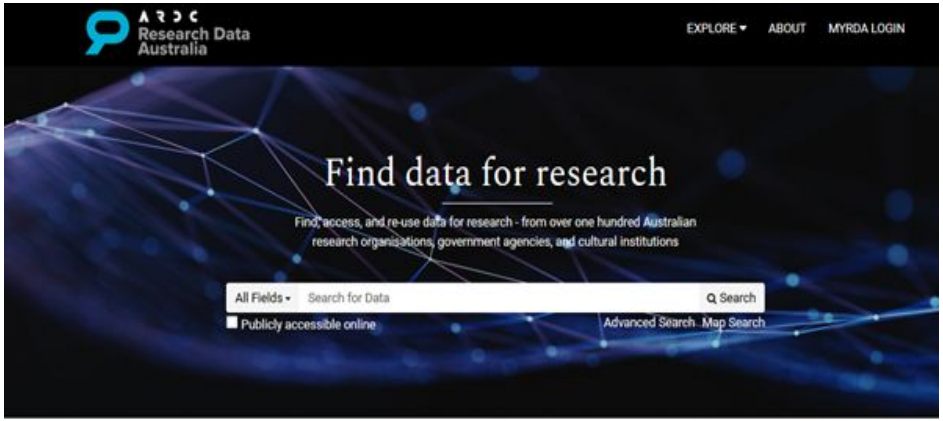
- ANZSRC-FoR include major fields and related sub-fields of research and emerging areas of study investigated by businesses, universities, tertiary institutions, national research institutions and other organisations.

ANZSRC-FoR: The Australian and New Zealand Standard Research Classification - Fields of Research) (2008 version)

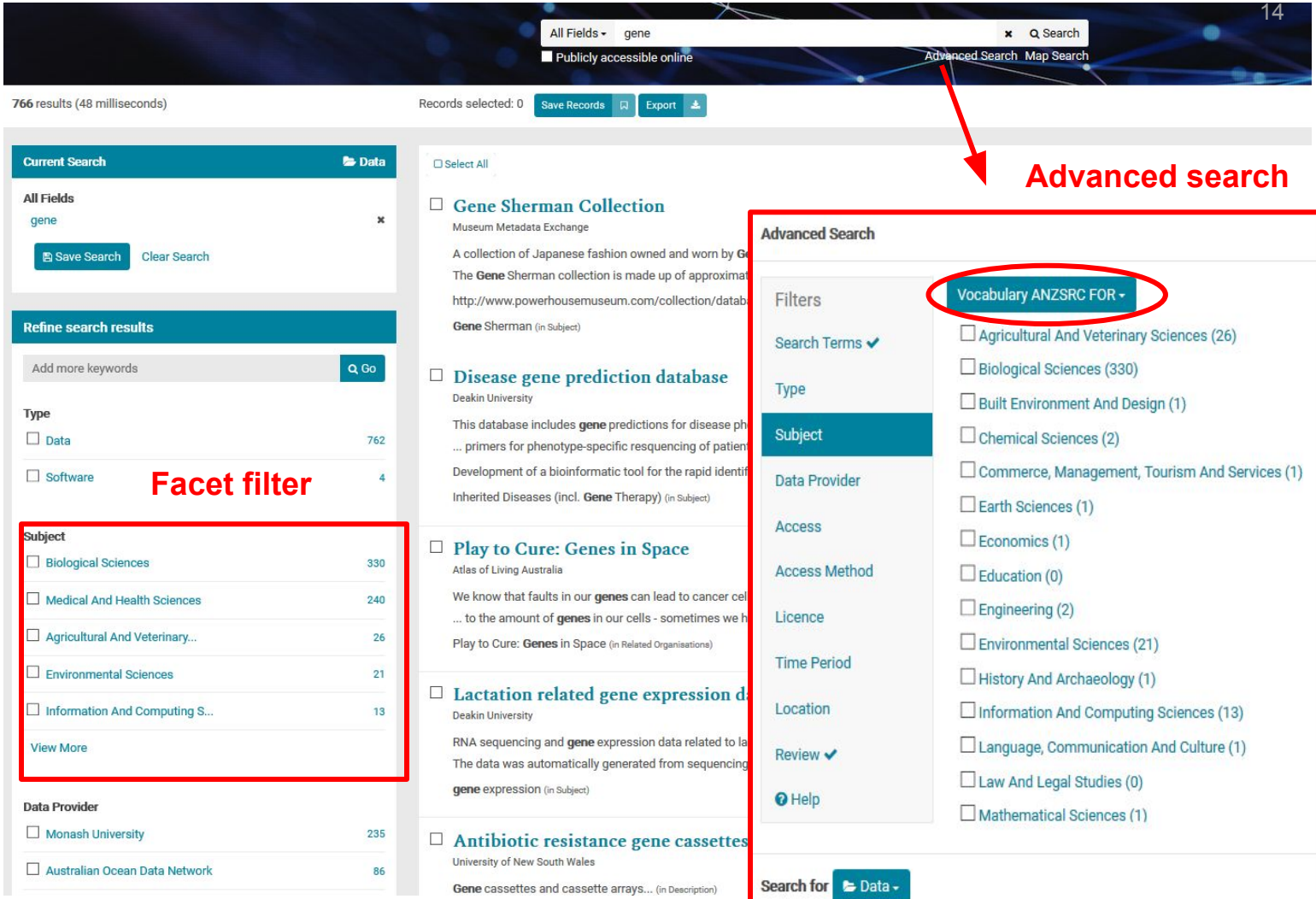


1417 terms in three layers

Subjects headings: browse



Subject headings: advanced search and facet filter



Subject headings: dataset Record

Dataset

Disease gene prediction database

Deakin University

Dr Merridee Wouters (Aggregated by) Mr Martin Oti (Aggregated by)

Viewed: 946 Accessed: 15

Access the data

Cite Save to MyRDA

Licence & Rights:
Other view details

Access:
Other view details

Contact Information
Postal Address:
School of Life and Environmental Sciences, Deakin University, 75 Pigdons Road, Waurn Ponds, Victoria 3216 Australia

Full description

This database includes gene predictions for disease phenotypes based on published Genome-Wide Association Data. May be used to choose primers for phenotype-specific resquencing of patient DNA.
For each prediction for following data is listed: phenotype, predicted gene, significant SNP, datasource, datasource reference.

Notes

The data was generated by a computer from clinical data, and some data from HuGE (http://hugenavigator.net/HuGENavigator/home.do) was used. The data is organised within a searchable

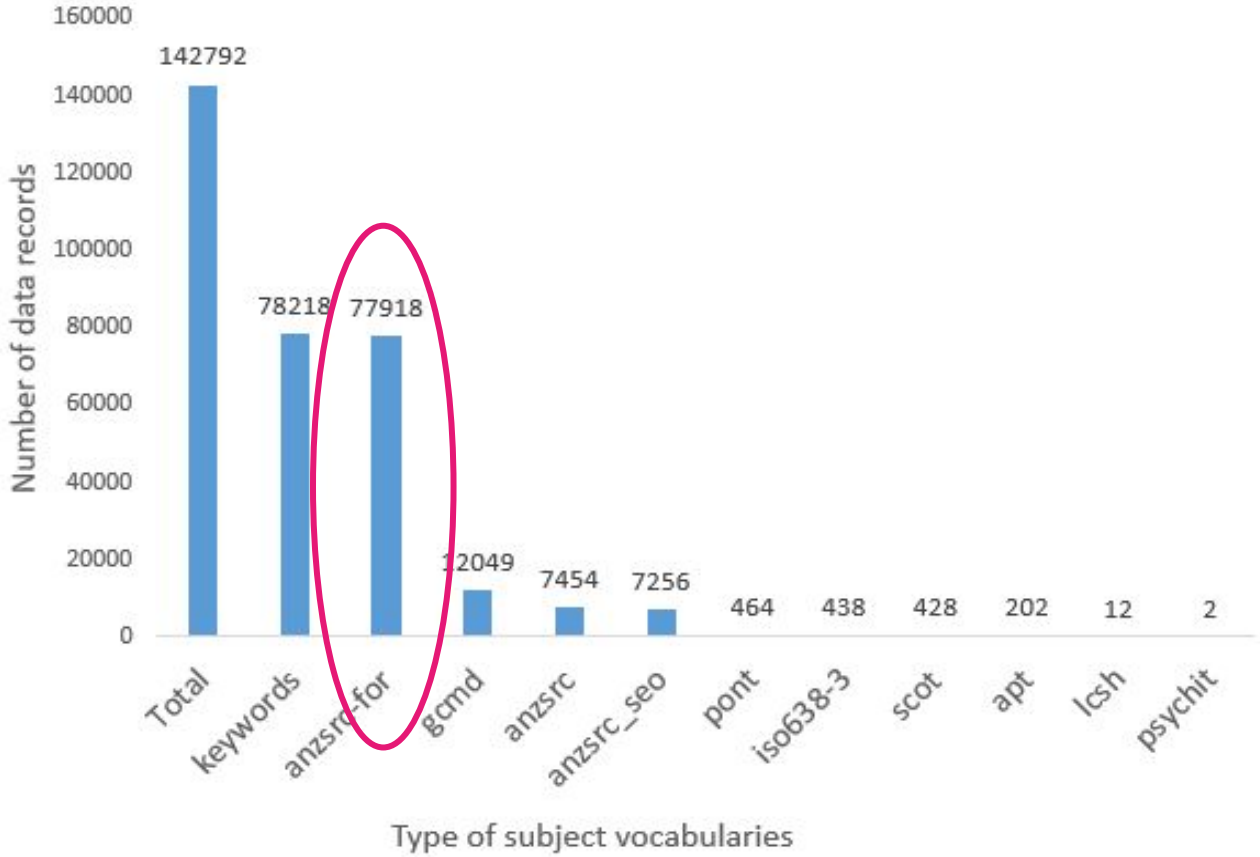
Subjects

Facet search

Biological Sciences | Clinical Health (Organs, Diseases and Abnormal Conditions) | Genetics | Genetics Not Elsewhere Classified | Health | Inherited Diseases (Incl. Gene Therapy) | database | genetic databases | genome-wide association study | humans | polymorphism | protein disease/genetics | single nucleotide | software |

Motivation for annotating subject headings

About half of the catalogue records have an ANZSRC-FoR heading/code



Project: Automatic classification/annotation of data records with the ANZSRC-FoR top layer headings

Machine learning for classifying/annotating subject metadata

- Assign ANZSRC-FoR headings to unlabelled records automatically
  - Aim to improve search experience for both human and machine
  - Understand domain coverage of the catalogue

- Train models, three components are essential for the training:
  - Labels - top layer 22 headings from the ANZSRC-FoR code
  - Data - (~78k) records with anzsrc-for code
    - Extract title & description from each metadata record
    - Split into two sets: training set, test set
  - Classifier - four supervised machine learning methods:
    multinomial logistic regression (MLR), multinomial naive bayes (MNB),
    K Nearest Neighbors (KNN), Support Vector Machine (SVM)

- Apply model(s)/best prediction to test set

Result

Four models: multinomial logistic regression (MLR), multinomial naive bayes (MNB), K Nearest Neighbors (KNN), Support Vector Machine (SVM)

Model	Training Set Accuracy	Test Set Accuracy
MLR	0.76	0.70
SVM	0.70	0.67
KNN	0.92	0.66
MNB	0.70	0.66

Performance per heading/category

2 digits code	MLR	SVM	KNN	MNB	down size	all data
01	0.29	0.00	0.41	0.33	*111	111
02	0.97	1.00	1.00	0.92	300	3537
03	0.73	0.61	0.60	0.59	499	499
04	0.96	0.98	0.92	0.90	600	10147
05	0.61	0.63	0.68	0.49	400	5417
06	1.00	1.00	0.64	0.96	600	24520
07	0.63	0.52	0.77	0.42	200	1032
08	0.45	0.22	0.53	0.26	*386	386
09	1.00	1.00	0.94	1.00	200	2031
10	0.29	0.00	0.20	0.00	*128	128
11	0.68	0.69	0.63	0.64	400	1409
12	0.61	0.95	0.67	0.66	*174	174
13	0.58	0.91	0.69	0.67	*148	148
14	0.41	0.00	0.58	0.57	*122	122
15	0.21	0.00	0.18	0.00	*76	76
16	0.56	0.50	0.55	0.54	300	723
17	0.40	0.00	0.32	0.67	*112	112
18	1.00	1.00	0.99	0.98	400	849
19	0.82	0.69	0.76	0.54	*343	343
20	0.89	0.85	0.26	0.81	300	553
21	0.97	0.96	0.99	0.88	600	32592
22	0.34	0.00	0.65	0.44	*79	79
micro ave	0.70	0.67	0.66	0.66	4799	84988
macro ave	0.65	0.57	0.63	0.60		
weighted ave	0.76	0.71	0.70	0.68		

Most correlated unigrams:

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

04: Earth Science
15: Commerce, Management, Tourism and Services

What we have learnt

- Large proportion of records from the catalogue don't have a subject heading
- Automatic classification/subject annotation works for some subject headings
- There are many options could be explored further - ML models, expand training data or train better feature representations with external resources, …
    - More human resource is required

Discussion

- How can we effectively apply trained models due to performance variations among subject headings, lack of training data, and catalogue content changes over time
  - Seek large and rich resources outside of the catalogue for reliable pre-trained models
  - Set up a workflow that periodically trains and updates models

- How to interpret and evaluate machine annotated subject headings? How to use ML outputs?
  - Should machine annotated subject headings be treated differently?
  - Collect user feedback for increasing training data and improve annotation accuracy

- Could some resources required for ML be shared among (data) catalogues and ML communities?
  - Training datasets, word associations, source codes or platforms, skills training …

Questions?

THANK YOU

ardc.edu.au

contact@ardc.edu.au

+61 3 9902 0585

@ARDC_AU

Australian-Research- Data-Commons

Subscribe to the **ARDC CONNECT** newsletter