ARDC
Australian Research Data Commons

Automatically classifying data records using ANZSRC-FoR subject headings

Mingfang Wu

Australian Research Data Commons, Australia

2021-10-13

DCMI 2021 Virtual

The Australian Research Data Commons is enabled by NCRIS.

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative

#ARDC_AU

Outlines

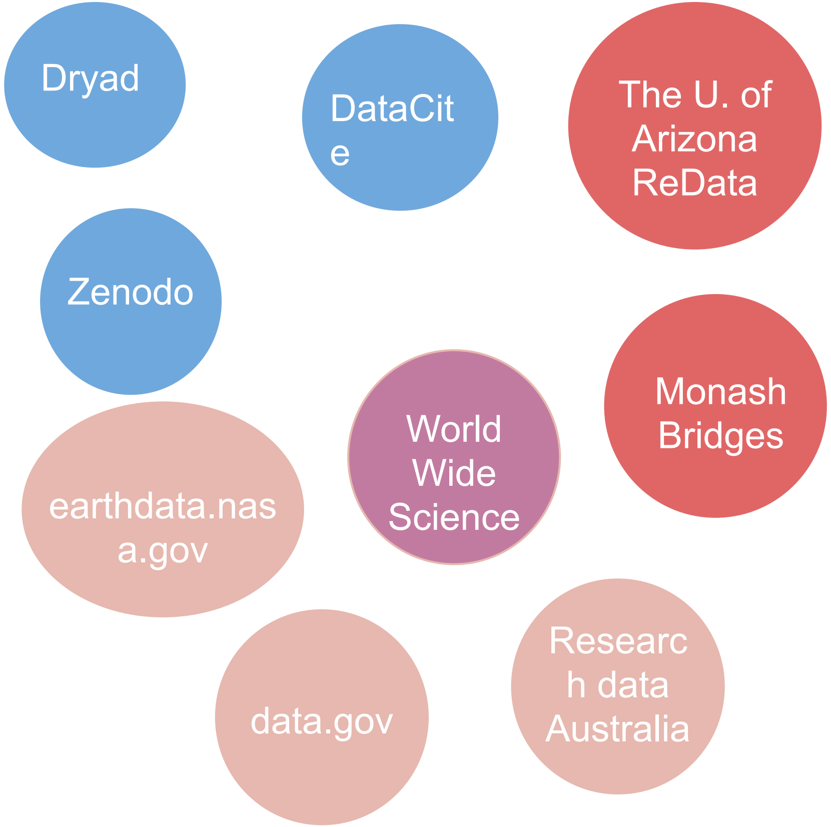
- ANZSRC-FoR subject headings: A background about the use of subject metadata in the data catalogue: Research Data Australia

- Experiment: Automatic labelling/classification of data records with the schema - Approach & Result & Implication

Research data repositories/catalogues

There has been a growing number of data repositories & catalogues for publishing and sharing data

re3data.org, the Registry of Research Data Repositories, had 23 repositories when it went online in 2012; the number quickly increased to over 1,200 data repositories from across the globe in three years and, by February 2020, the registry had more than 2450 repositories.



Dryad

DataCite

The U. of Arizona ReData

Zenodo

Monash Bridges

World Wide Science

earthdata.nasa.gov

data.gov

Research data Australia

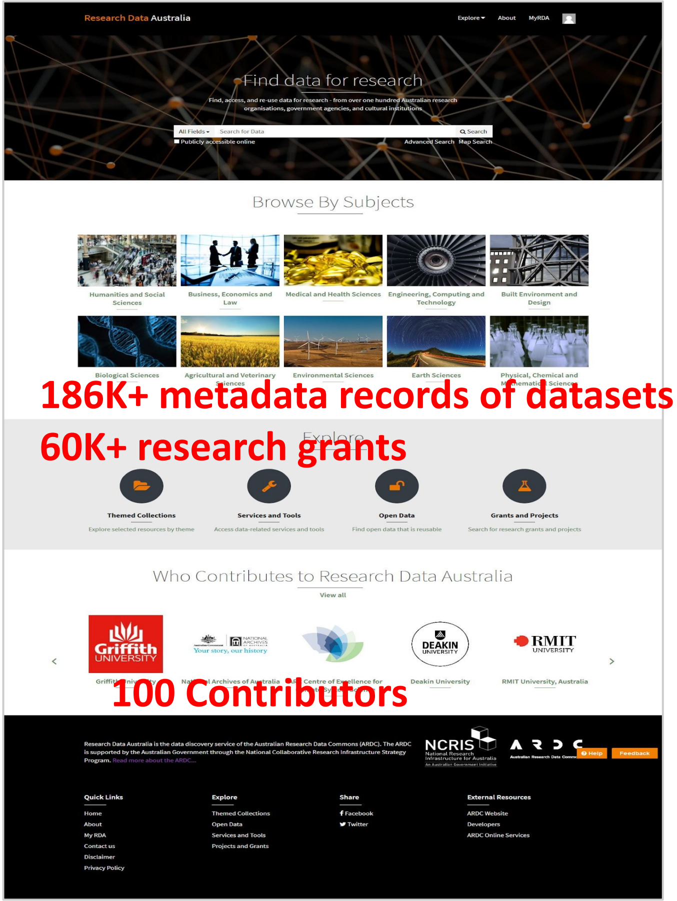
Subject metadata

- Dublin Core Definition:
  - A topic of the content of the resource.
  - Typically, subject will be expressed as keywords, key phrases, or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme (e.g. Faceted Application of Subject Terminology).

- Benefit: Subject metadata is a powerful way of knowledge organisation and linkage of (distributed) resources for interoperability and discovery
  - Subject metadata is included in almost all metadata schemas. For schemas describing dataset, most of them include subject metadata as an optional not mandatory field.

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

Research Data Australia - A National Data Catalogue



Research Data Australia

Explore ▾ About MyRDA

Find data for research

Find, access, and re-use data for research - from over one hundred Australian research organisations, government agencies, and cultural institutions

All Fields ▾ | Search for Data | Search

□ Publicly accessible online

Advanced Search Map Search

Browse By Subjects

Humanities and Social Sciences

Business, Economics and Law

Medical and Health Sciences

Engineering, Computing and Technology

Built Environment and Design

Biological Sciences

Agricultural and Veterinary Sciences

Environmental Sciences

Earth Sciences

Physical, Chemical and Mathematical Sciences

186K+ metadata records of datasets

60K+ research grants

Explore

Themed Collections

Explore selected resources by theme

Services and Tools

Access data-related services and tools

Open Data

Find open data that is reusable

Grants and Projects

Search for research grants and projects

Who Contributes to Research Data Australia

View all

Griffith University

National Archives of Australia

ARC Centre of Excellence for Climate System Science

Deakin University

RMIT University, Australia

100 Contributors

Research Data Australia is the data discovery service of the Australian Research Data Commons (ARDC). The ARDC is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy Program. Read more about the ARDC...

Help Feedback

Quick Links

- Home
- About
- My RDA
- Contact us
- Disclaimer
- Privacy Policy

Explore

- Themed Collections
- Open Data
- Services and Tools
- Projects and Grants

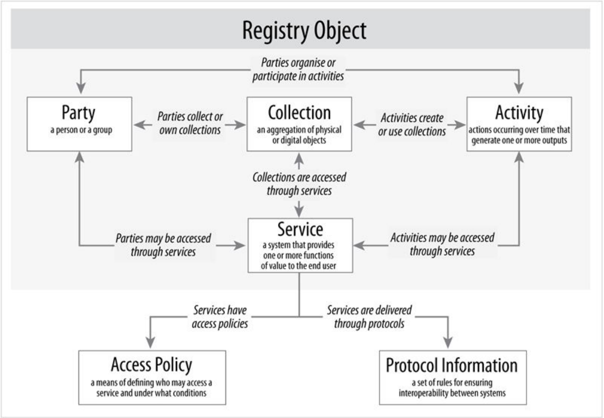
Share

- Facebook
- Twitter

External Resources

- ARDC Website
- Developers
- ARDC Online Services

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



Registry Object

Parties organise or participate in activities

Party
a person or a group

Parties collect or own collections

Collection
an aggregation of physical or digital objects

Activities create or use collections

Activity
actions occurring over time that generate one or more outputs

Collections are accessed through services

Parties may be accessed through services

Service
a system that provides one or more functions of value to the end user

Activities may be accessed through services

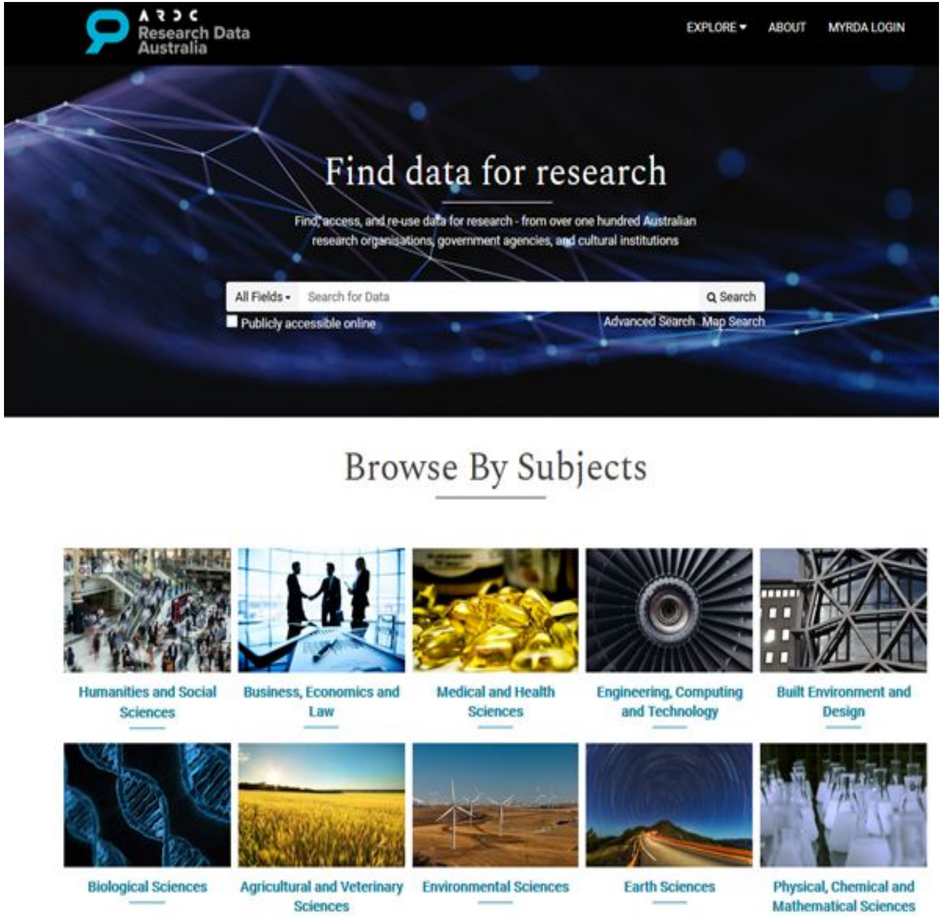
Services have access policies

Services are delivered through protocols

Access Policy
a means of defining who may access a service and under what conditions

Protocol Information
a set of rules for ensuring interoperability between systems

Browse by subjects



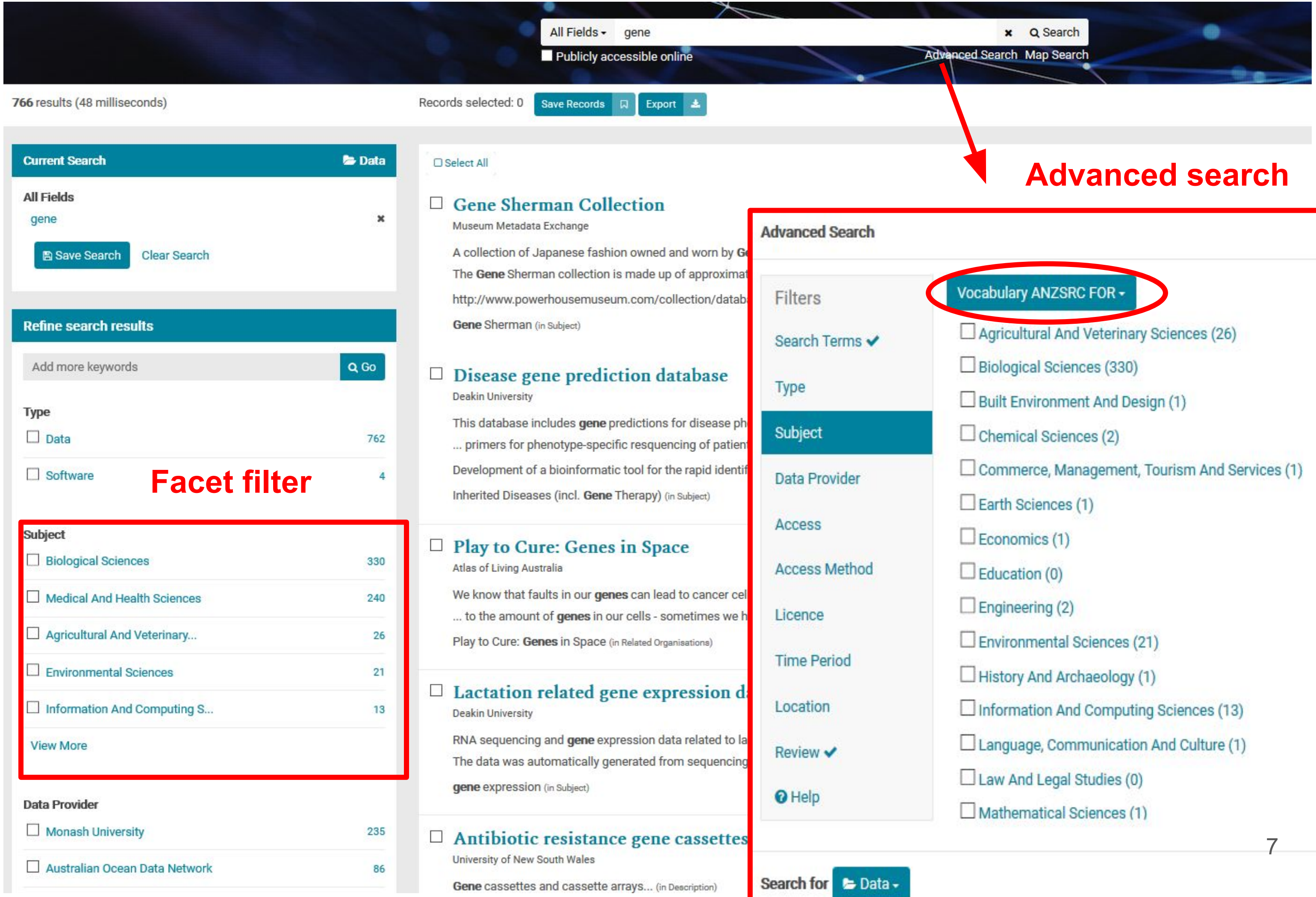
Find data for research

Find, access, and re-use data for research - from over one hundred Australian research organisations, government agencies, and cultural institutions

Browse By Subjects

- Humanities and Social Sciences
- Business, Economics and Law
- Medical and Health Sciences
- Engineering, Computing and Technology
- Built Environment and Design
- Biological Sciences
- Agricultural and Veterinary Sciences
- Environmental Sciences
- Earth Sciences
- Physical, Chemical and Mathematical Sciences

Subject headings



All Fields ▾ gene ✖ Search

□ Publicly accessible online

Advanced Search Map Search

766 results (48 milliseconds)

Records selected: 0 Save Records Export

Current Search Data

All Fields

gene ✖

Save Search Clear Search

Refine search results

Add more keywords Go

Type

□ Data	762
□ Software	4

Subject

□ Biological Sciences	330
□ Medical And Health Sciences	240
□ Agricultural And Veterinary...	26
□ Environmental Sciences	21
□ Information And Computing S...	13

View More

Data Provider

□ Monash University	235
□ Australian Ocean Data Network	86

□ Select All

□ **Gene Sherman Collection**

Museum Metadata Exchange

A collection of Japanese fashion owned and worn by Ge

The Gene Sherman collection is made up of approximat

http://www.powerhousemuseum.com/collection/datab

Gene Sherman (in Subject)

□ **Disease gene prediction database**

Deakin University

This database includes gene predictions for disease ph

... primers for phenotype-specific resquencing of patien

Development of a bioinformatic tool for the rapid identi

Inherited Diseases (incl. Gene Therapy) (in Subject)

□ **Play to Cure: Genes in Space**

Atlas of Living Australia

We know that faults in our genes can lead to cancer cel

... to the amount of genes in our cells - sometimes we h

Play to Cure: Genes in Space (in Related Organisations)

□ **Lactation related gene expression d**

Deakin University

RNA sequencing and gene expression data related to la

The data was automatically generated from sequencing

gene expression (in Subject)

□ **Antibiotic resistance gene cassettes**

University of New South Wales

Gene cassettes and cassette arrays... (in Description)

Advanced Search

Filters

- Search Terms ✓
- Type
- Subject
- Data Provider
- Access
- Access Method
- Licence
- Time Period
- Location
- Review ✓
- Help

Vocabulary ANZSRC FOR ▾

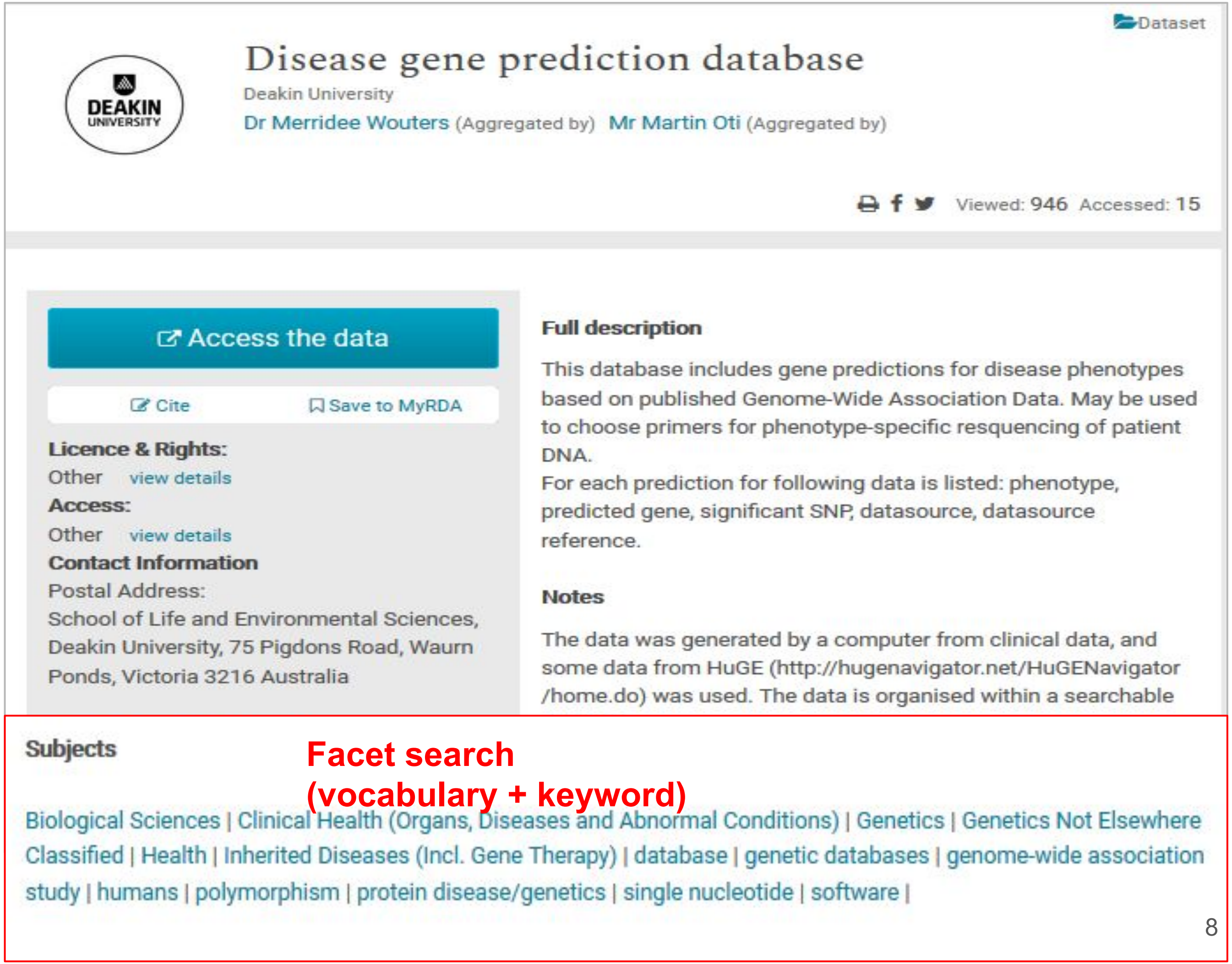
- □ Agricultural And Veterinary Sciences (26)
- □ Biological Sciences (330)
- □ Built Environment And Design (1)
- □ Chemical Sciences (2)
- □ Commerce, Management, Tourism And Services (1)
- □ Earth Sciences (1)
- □ Economics (1)
- □ Education (0)
- □ Engineering (2)
- □ Environmental Sciences (21)
- □ History And Archaeology (1)
- □ Information And Computing Sciences (13)
- □ Language, Communication And Culture (1)
- □ Law And Legal Studies (0)
- □ Mathematical Sciences (1)

Search for Data ▾

Advanced search

Facet filter

Record view



Disease gene prediction database

Deakin University

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

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 (vocabulary + keyword)

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 |

Types of subject classification schemas

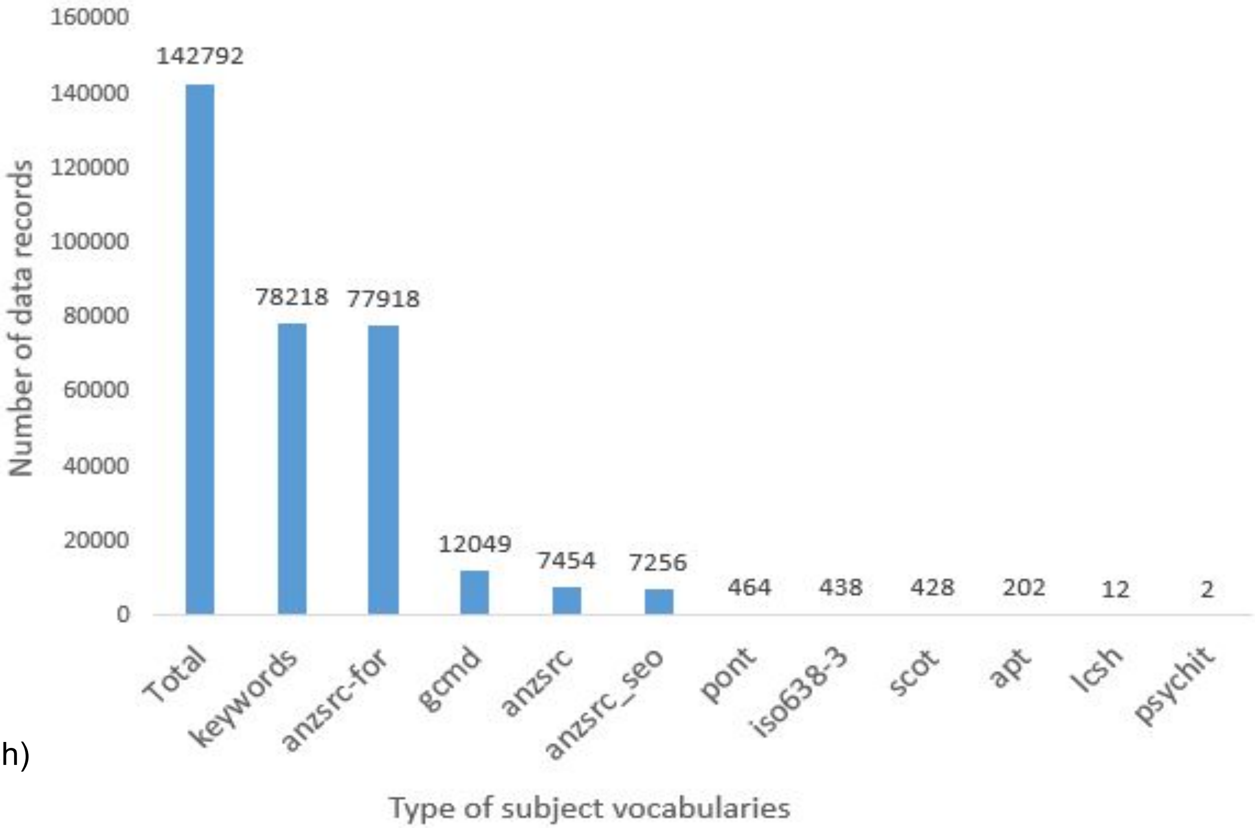
ANZSRC-FoR: The Australian and New Zealand Standard Research Classification (ANZSRC, fields of research)

Global change master directory (GCMD) keywords

Australian Pictorial Thesaurus (apt)

Thesaurus of Psychological Index Terms (psychit)

Library of Congress Subject Headings (lcsh)

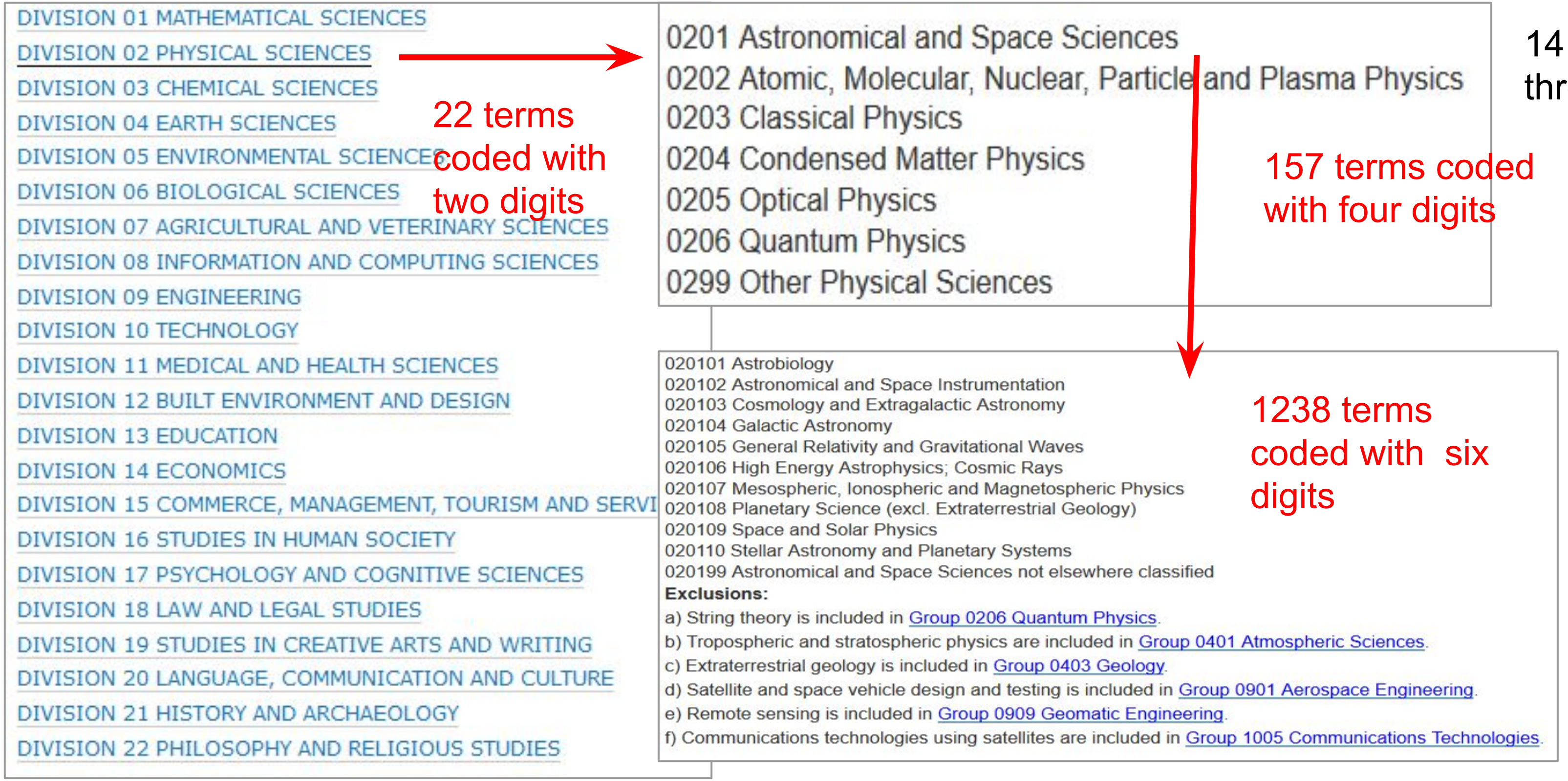


ANZSRC-FoR: The Australian and New Zealand Standard Research Classification - Fields of Research

- 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)



DIVISION 01 MATHEMATICAL SCIENCES
DIVISION 02 PHYSICAL SCIENCES
DIVISION 03 CHEMICAL SCIENCES
DIVISION 04 EARTH SCIENCES
DIVISION 05 ENVIRONMENTAL SCIENCES
DIVISION 06 BIOLOGICAL SCIENCES
DIVISION 07 AGRICULTURAL AND VETERINARY SCIENCES
DIVISION 08 INFORMATION AND COMPUTING SCIENCES
DIVISION 09 ENGINEERING
DIVISION 10 TECHNOLOGY
DIVISION 11 MEDICAL AND HEALTH SCIENCES
DIVISION 12 BUILT ENVIRONMENT AND DESIGN
DIVISION 13 EDUCATION
DIVISION 14 ECONOMICS
DIVISION 15 COMMERCE, MANAGEMENT, TOURISM AND SERVI
DIVISION 16 STUDIES IN HUMAN SOCIETY
DIVISION 17 PSYCHOLOGY AND COGNITIVE SCIENCES
DIVISION 18 LAW AND LEGAL STUDIES
DIVISION 19 STUDIES IN CREATIVE ARTS AND WRITING
DIVISION 20 LANGUAGE, COMMUNICATION AND CULTURE
DIVISION 21 HISTORY AND ARCHAEOLOGY
DIVISION 22 PHILOSOPHY AND RELIGIOUS STUDIES

22 terms coded with two digits

0201 Astronomical and Space Sciences
0202 Atomic, Molecular, Nuclear, Particle and Plasma Physics
0203 Classical Physics
0204 Condensed Matter Physics
0205 Optical Physics
0206 Quantum Physics
0299 Other Physical Sciences

157 terms coded with four digits

1417 terms in three layers

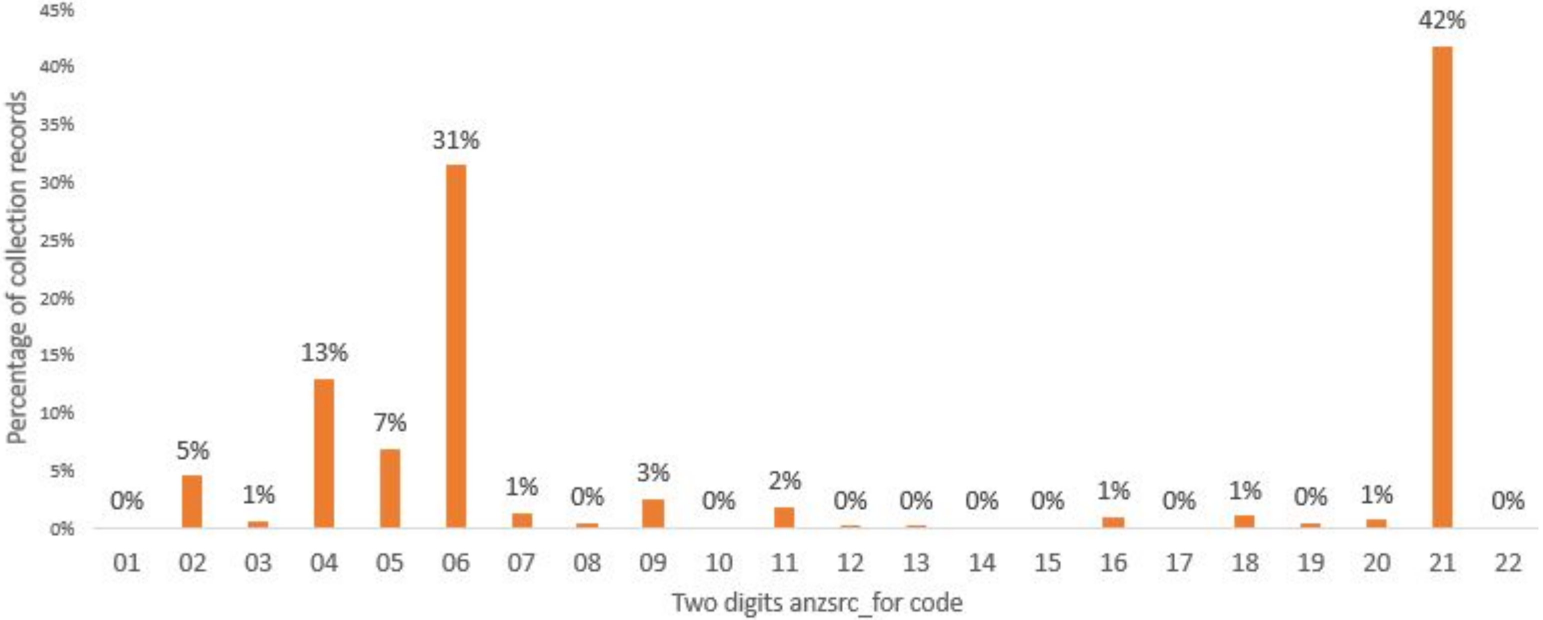
020101 Astrobiology
020102 Astronomical and Space Instrumentation
020103 Cosmology and Extragalactic Astronomy
020104 Galactic Astronomy
020105 General Relativity and Gravitational Waves
020106 High Energy Astrophysics; Cosmic Rays
020107 Mesospheric, Ionospheric and Magnetospheric Physics
020108 Planetary Science (excl. Extraterrestrial Geology)
020109 Space and Solar Physics
020110 Stellar Astronomy and Planetary Systems
020199 Astronomical and Space Sciences not elsewhere classified

Exclusions:

a) String theory is included in Group 0206 Quantum Physics.
b) Tropospheric and stratospheric physics are included in Group 0401 Atmospheric Sciences.
c) Extraterrestrial geology is included in Group 0403 Geology.
d) Satellite and space vehicle design and testing is included in Group 0901 Aerospace Engineering.
e) Remote sensing is included in Group 0909 Geomatic Engineering.
f) Communications technologies using satellites are included in Group 1005 Communications Technologies.

1238 terms coded with six digits

Number of records per anzsrc-for two digits



04: Earth Sciences

06: Biological Sciences

21: History and Archaeology

Machine learning for classifying/labelling subject metadata

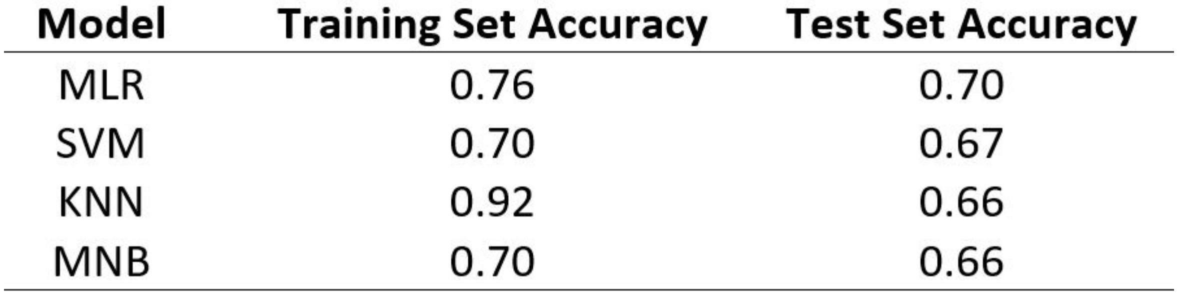
- Assign ANZSRC-FoR code to unlabelled records automatically
  - Aim to improve search experience for both human and machine
  - Understand domain coverage of the collection

- Train models, three components are essential for the training:
  - Classifier - four supervised machine learning methods:
    - multinomial logistic regression (MLR), multinomial naive bayes (MNB), K Nearest Neighbors (KNN), Support Vector Machine (SVM)
  - Labels - top layer 22 labels 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

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

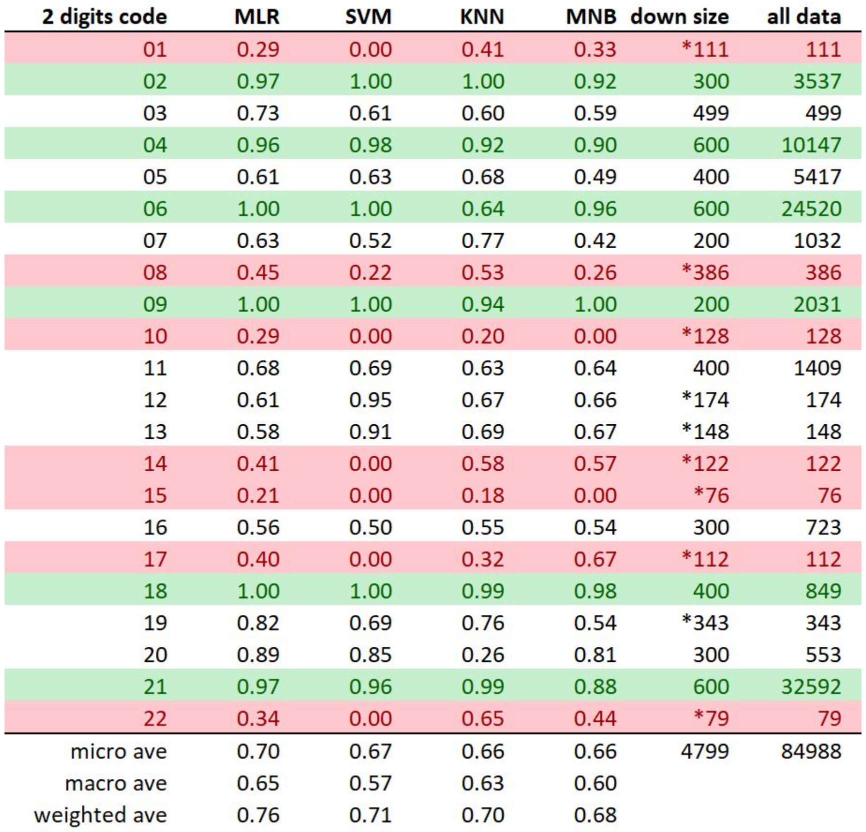
Result

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



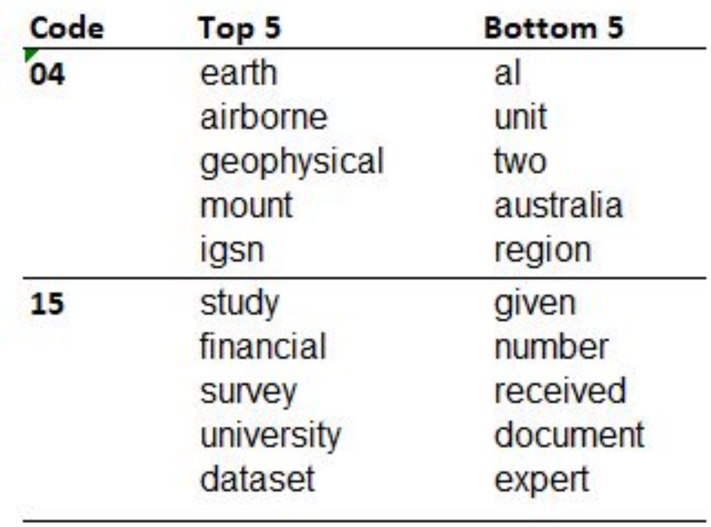
Model	Training Set Accuracy	Test Set Accuracy
MLR	0.76	0.70
SVM	0.70	0.67
KNN	0.92	0.66
MNB	0.70	0.66

Performance per label/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 airborne geophysical mount igsn	al unit two australia region
15	study financial survey university dataset	given number received document expert

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

What we have learnt

- Large proportion of records from the catalogue don’t the subject metadata a known issue
- Those with subject metadata are biased toward a few categories

  - Encourage underrepresented subject areas to publish and share data
- Automatic classification works for some categories

  - Explore correlation and improvement, and other machine learning methods

Discussion

- There does not exist a classifier that fits for all text classification tasks or all collections. This leads to a generalisation issue, it is hard for organisations who want to improve subject metadata but have no resources to train classification models for their collection.
- Traditional ML methods require a good amount of training dataset with quality annotation. Some new technologies (e.g. deep learning and transfer learning) may provide a way, for example, wording embedding technique by learning word association from other similar big collections.
- Practically there is a question about how to use ML outputs: MLs suggest labels based on probability, human expertise may be necessary to look at a range of suggestions and make a decision, this may lead to improve ML intelligence and classification accuracy - we help the ML community to help ourselves.

Acknowledgement

Ying-Hsang Liu, Oslo Metropolitan University, Norway
Rowan Brownlee, Australian Research Data Commons
Jenny Xiuzhen Zhang, RMIT University, Australia

Australian Research Data Commons

Thank you

Contact:
mingfang.wu@ardc.edu.au

www.ardc.edu.au

in australian-research-data-commons

ARDC_AU

The Australian Research
Data Commons is enabled
by NCRIS.

NCRIS
National Research
Infrastructure for Australia
An Australian Government Initiative