

Evaluating utility of subject headings in a data repository: A preliminary finding from a data search log and record classification

Presented by:

Mingfang Wu, Australian Research Data Commons
mingfang.wu@ardc.edu.au

Contributors:

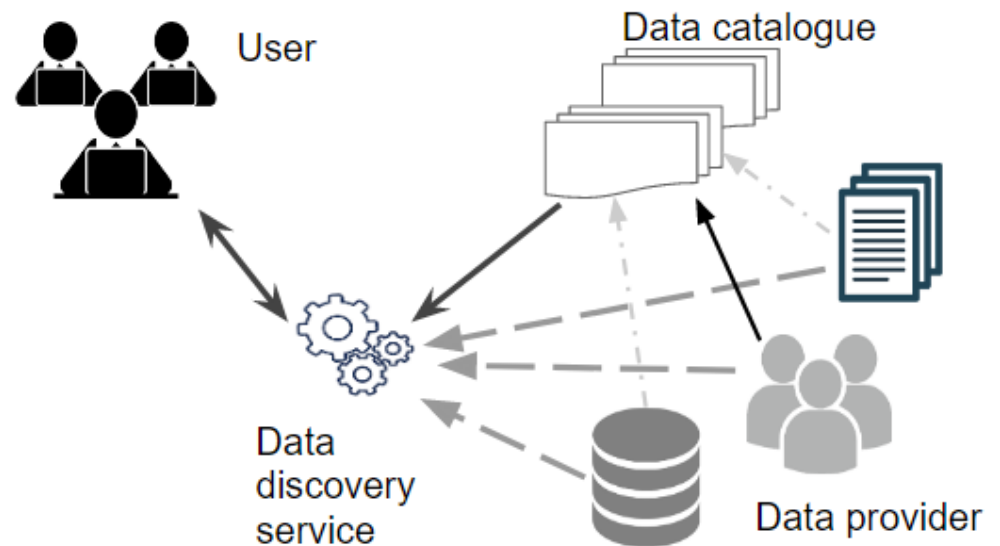
Rowan Brownlee, Australian Research Data Commons
Ying-Hsang Liu, University of Southern Denmark
Jenny Xiuzhen Zhang, RMIT University, Australia

NKOS, 10 Sept. 2020



Outlines

- A background about the studied data catalogue: Research Data Australia
- Log analysis: the usage of subject headings
- Experiments on data record classification
- Future work



Research Data Australia - A National Data Catalogue

Research Data Australia

Find data for research

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

Advanced Search

Reg 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 Environmental Sciences

144K+ metadata records of dataset

60K+ research grants

Themed Collections

Services and Tools

Open Data

Grants and Projects

Who Contributes to Research Data Australia

Griffith University

University of Technology Sydney

ARC Centre of Excellence for

Deakin University

RMIT University, Australia

99 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...](#)

Quick Links

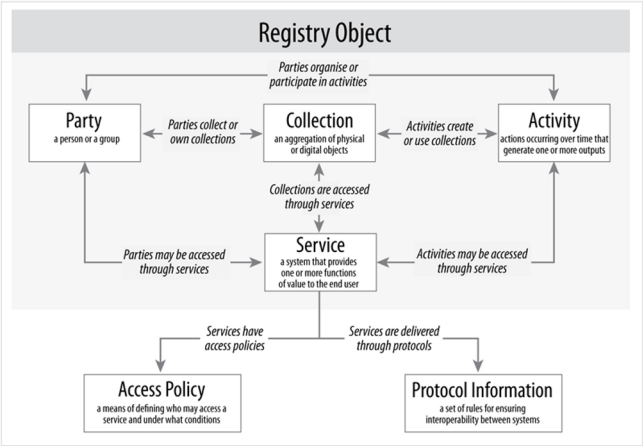
Explore

Share

External Resources



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



Types of subject vocabularies

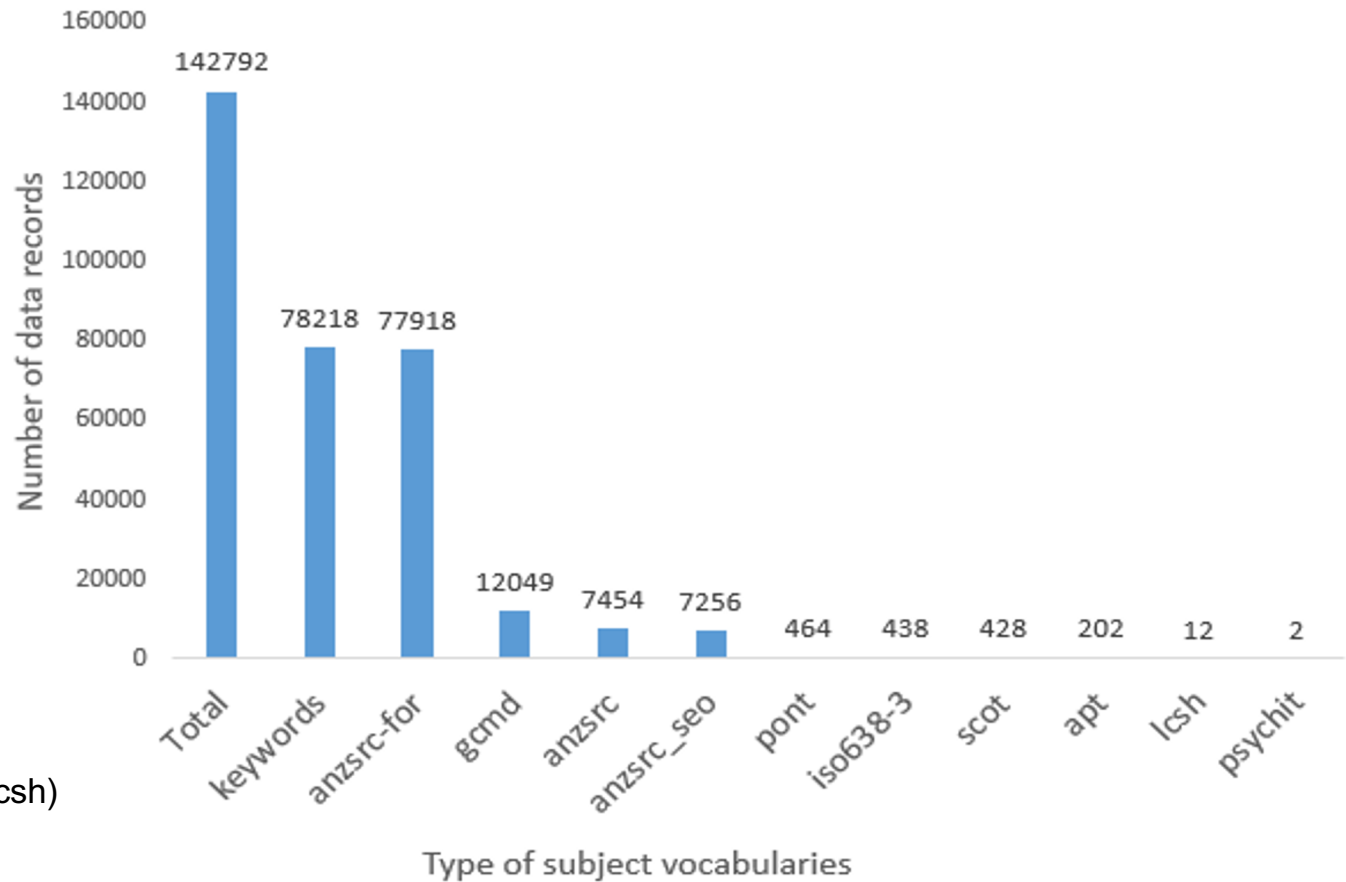
Anzsrc-for: The Australian and New Zealand Standard Research Classification (ANZSRC, fields of research)

Global change master directory (GCMD) keywords

Australian Pictorial Thesaurus (apt)

Thesaurus of Psychological Index Terms (psychit)

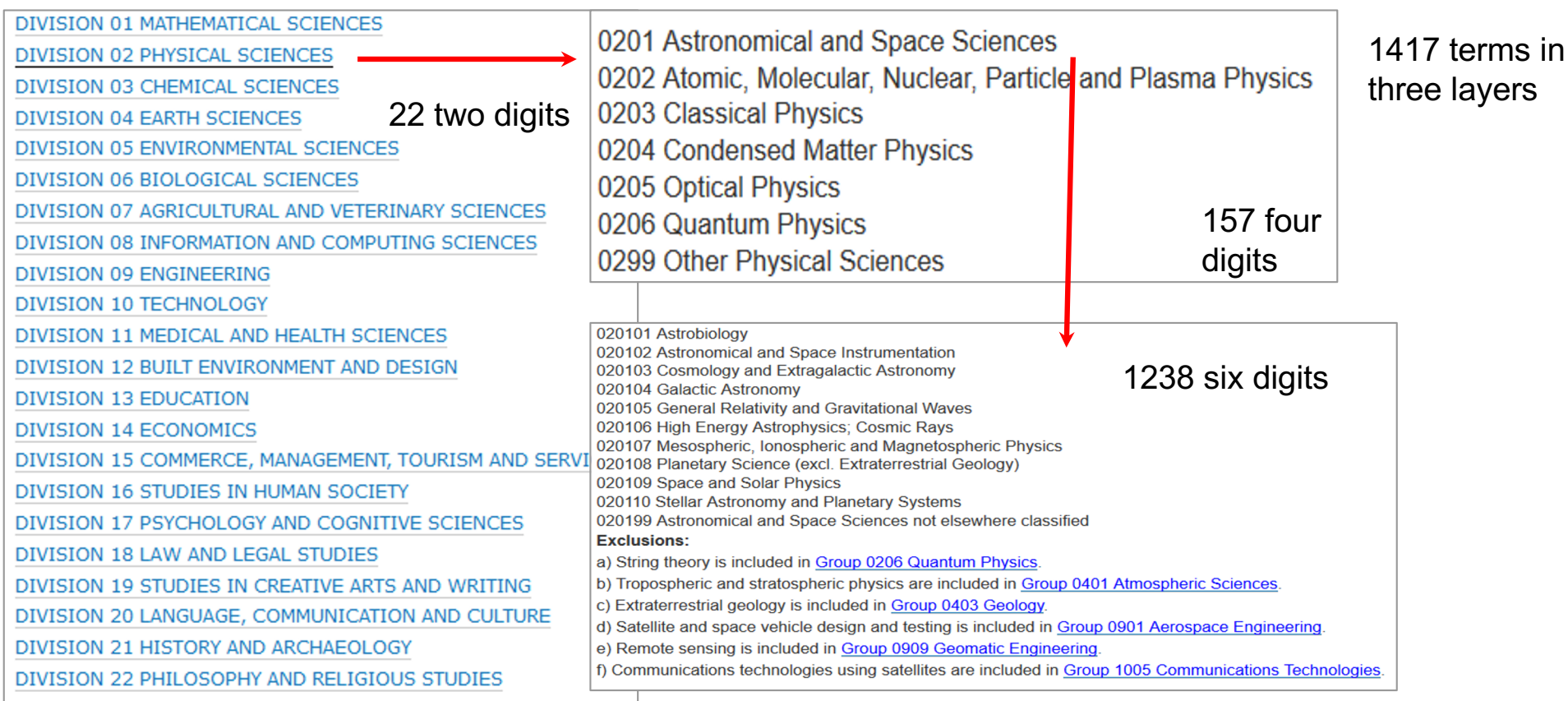
Library of Congress Subject Headings (lcsh)



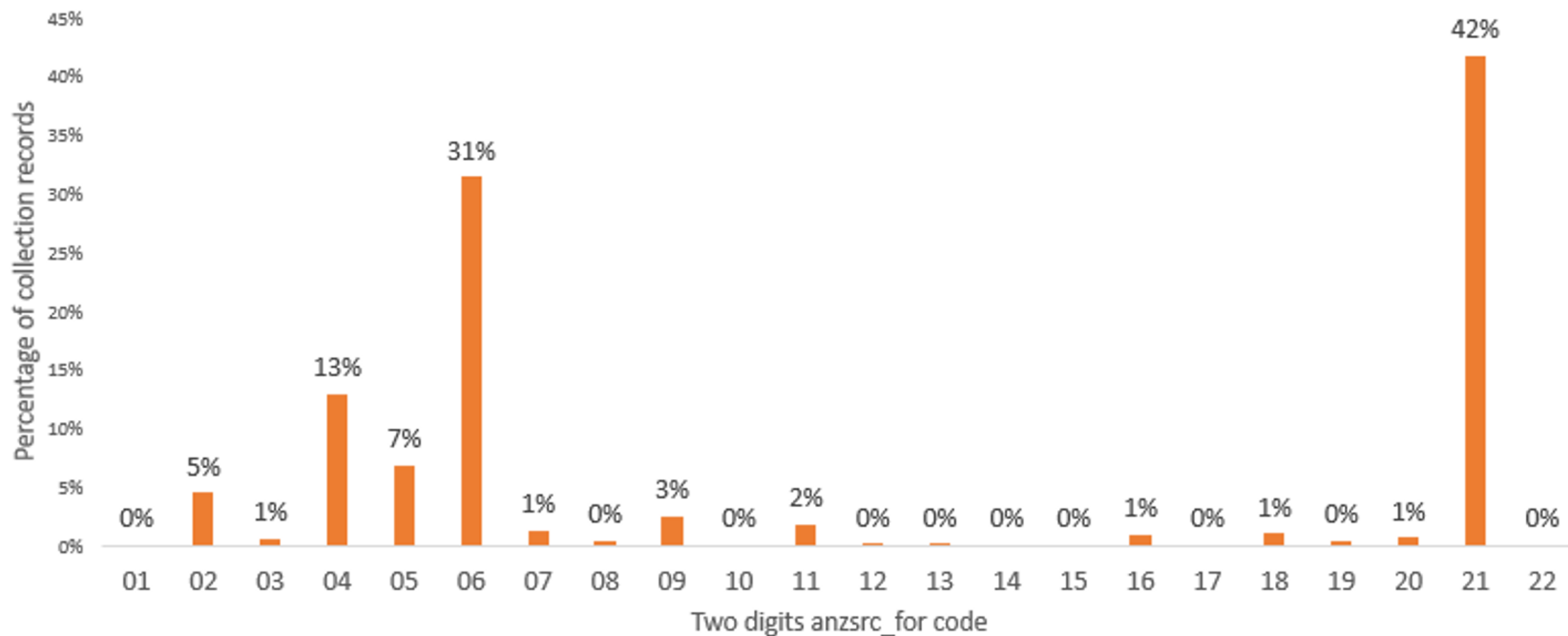
Anzsrc-for: The Australian and New Zealand Standard Research Classification - Fields of Research

- 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)



Number of records per anzsrc-for two digits



04: Earth Sciences

06: Biological Sciences

21: History and Archaeology

Search interface

All text strings
(including subject
headings) are
indexed.

All Fields ▾ gene × Q 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 Q 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 **Gene** Sherman from the late 1980s to 2007. (in Description)
The **Gene** Sherman collection is made up of approximately 60 fashion items, including clothing and accessories. (in Description)
http://www.powerhousemuseum.com/collection/database/collection=**Gene**_Sherman (in Identifier)
Gene Sherman (in Subject)

☐ **Disease gene prediction database**
Deakin University
This database includes **gene** predictions for disease phenotypes based on published Genome-Wide Association Data. (in Description)
... primers for phenotype-specific ressequencing of patient DNA.For each prediction for following data is listed: phenotype, predicted **gene**, significant... (in Description)
Development of a bioinformatic tool for the rapid identification of candidate disease **genes** (in Related Project or Grant)
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 cells forming. (in Description)
... to the amount of **genes** in our cells - sometimes we have more and sometimes we have less. It can take years for scientists to analyze all... (in Description)
Play to Cure: **Genes** in Space (in Related Organisations)

☐ **Lactation related gene expression data**
Deakin University
RNA sequencing and **gene** expression data related to lactation (mammary gland, milk and their sub compartments) obtained in a number... (in Description)
The data was automatically generated from sequencing of **gene** expression, Chips platforms and refined by computational analysis (parts... (in Description)
gene expression (in Subject)

☐ **Antibiotic resistance gene cassettes**
University of New South Wales
Gene cassettes and cassette arrays... (in Description)

8

Subject headings

766 results (48 milliseconds) Records selected: 0 Save Records Export

All Fields ▾ gene Search
Publicly accessible online Advanced Search Map Search

Current Search Data

All Fields
gene

Save Search Clear Search

Refine search results

Add more keywords Go

Type

- ☐ Data 762
- ☐ Software 4

2. Facet filter

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

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)

1. Advanced search

Gene Sherman Collection
Museum Metadata Exchange
A collection of Japanese fashion owned and worn by Gene Sherman. The Gene Sherman collection is made up of approximately 1000 items. <http://www.powerhousemuseum.com/collection/database.php>
Gene Sherman (in Subject)

Disease gene prediction database
Deakin University
This database includes gene predictions for disease phenotypes. It includes primers for phenotype-specific ressequencing of patient genomes. Development of a bioinformatic tool for the rapid identification of disease-associated genes.
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 cells. ... to the amount of genes in our cells - sometimes we have more than we need.
Play to Cure: Genes in Space (in Related Organisations)


Lactation related gene expression data
Deakin University
RNA sequencing and gene expression data related to lactation. The data was automatically generated from sequencing of gene expression (in Subject)

Antibiotic resistance gene cassettes
University of New South Wales
Gene cassettes and cassette arrays... (in Description)

Search for Data ▾


9


Record view



Disease gene prediction database

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

 Dataset

 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

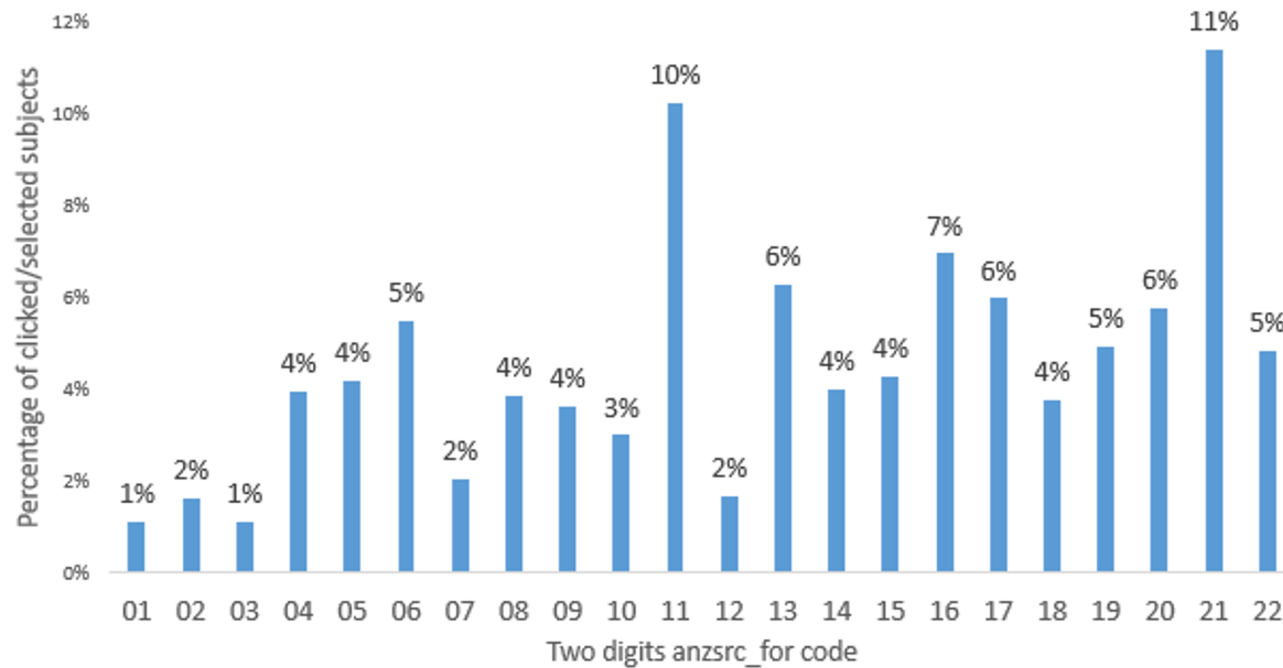
Biological Sciences | Clinical Health (Organs, Diseases and Abnormal Conditions) | Genetics | Genetics Not Elsewhere Classified | Health | Inherited Diseases (Incl. Gene Therapy) | database | genetic databases | genome-wide association study | humans | polymorphism | protein disease/genetics | single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

3. Facet search (vocabulary + keyword)

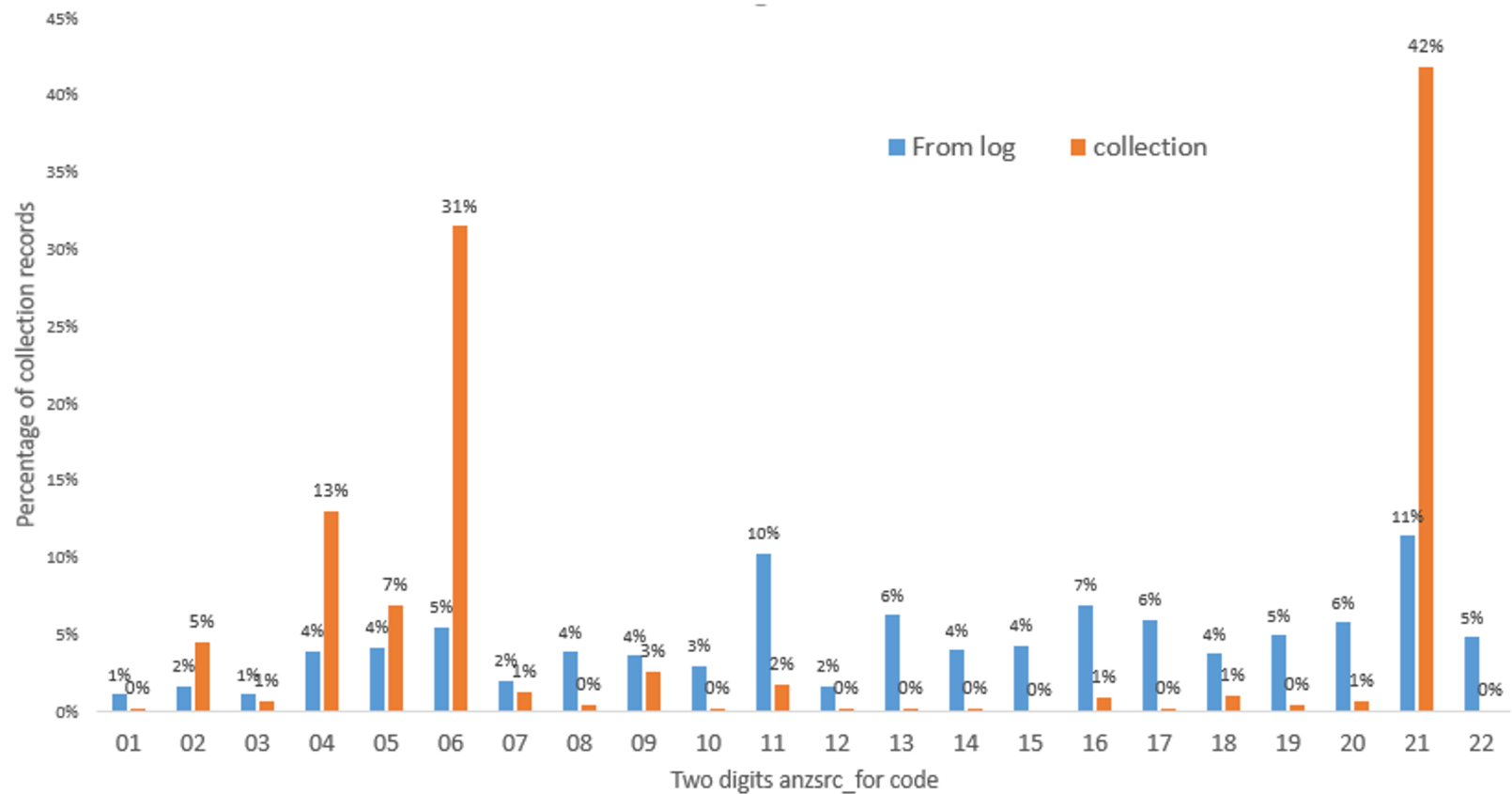
Log analysis: the usage of subject headings

- Transaction log: one year (2019) of activities recorded from the RDA catalogue
- About 2 million entries/activities, 63% from Australia
- About 496,739 sessions (with 30 minutes duration from the same IP address)
- 37,056 sessions have at least a search event (keyword search, advanced search, subject (facet) filter, subject search)
- 4668 (12.6%) of search sessions involved filters/search with the anzsrc-for subjects, only 45 (0.1%) with gcmd subject

Subject usages per anzsrc-for two digits code



Subject distribution among clicks and the collection



Log analysis: the usage of subject headings

- There is less bias in user's behaviour of applying subject headings, compared to the content bias toward a few subject headings.
- However, this log shows low usage of subject headings
- Exploring causes
 - Further log analysis, e.g. correlation between subject usage and
 - query types
 - domain knowledge
 - search quality
 - Interface design
 - At the record level: only half of the indexed records have anzsrc-for codes

Machine learning for record classification

- 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:
 - Labels - anzsrc-for code
 - Classifier - four supervised machine learning methods:
 - multinomial logistic regression (MLR), multinomial naive bayes (MNB), K Nearest Neighbors (KNN), Support Vector Machine (SVM)
 - Data - (~78k) records with anzsrc-for code
 - Split into two sets: training set, test set
- Apply model(s)/best prediction to unlabelled records

Record classification with anzsrc-for code

- Use 77918 records that have an anzsrc-for code for training models
- Step by step: first the top two digits, then move down to four, six digits
- 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
Logistic Regression	0.769149	0.701299
SVM	0.696435	0.676324
Multinomial Naïve Bayes	0.702965	0.659341
KNN	0.906460	0.642358

Acknowledgement:

Adapted [the code](#) from Miguel Frenandez Zafra

Performance per category

Most correlated unigrams:

Code	Top 5	Bottom 5
04	earth airborne geophysical mount ign	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

2 digitis Code	MLR	SVM	MNB	KNN	No. of records
01	0.29	0.00	0.43	0.33	111
02	0.97	1.00	0.95	1.00	*300
03	0.67	0.56	0.55	0.58	499
04	0.98	0.96	0.94	0.96	*600
05	0.68	0.71	0.53	0.54	*400
06	0.98	1.00	0.89	0.78	*600
07	0.63	0.55	0.52	0.79	*200
08	0.42	0.22	0.23	0.41	386
09	0.95	1.00	1.00	0.84	*200
10	0.33	0.00	0.00	0.19	128
11	0.82	0.81	0.83	0.66	*400
12	0.71	1.00	0.77	0.81	174
13	0.58	0.82	0.54	0.56	148
14	0.35	0.00	0.83	0.57	122
15	0.23	0.00	0.00	0.25	76
16	0.49	0.47	0.44	0.50	*300
17	0.47	0.00	0.57	0.50	112
18	1.00	1.00	1.00	1.00	*400
19	0.77	0.62	0.48	0.58	343
20	0.64	0.72	0.48	0.20	*300
21	0.96	0.94	0.88	0.98	*600
22	0.22	0.00	0.33	0.20	79
micro ave	0.70	0.68	0.66	0.64	
macro ave	0.65	0.56	0.60	0.61	

Examples of classification within two-digits code

Method: MLR

06: Biological Sciences (41505 records)

02: Physical Sciences (3533 records)

06: 17268 records (out of 41505) have both 0601 and 0604 labels

	precision	test data		precision	test data
0601	0.58	2859	0201	1.00	752
0602	0.99	652	0202	0.00	1
0603	0.15	22	0203	0.04	2
0604	0.49	2560	0204	0.32	13
0605	0.01	11	0205	0.00	0
0606	1	3	0206	0.00	0
0607	0.1	48	0299	1.00	116
0608	0.52	51			
0699		20			
micro avg	0.5	6226		0.58	884
macro ave	0.43			0.34	
weighted ave	0.58			0.99	

Confusion matrix - MLR (06 - Biological Sciences)										
Predicted	Biochemistry and Cell Biology (601) -	560	0	6	2248	0	1	0	44	0
	Ecology (602) -	4	325	21	0	0	57	0	230	15
	Evolutionary Biology (603) -	2	0	6	0	0	3	0	11	0
	Genetics (604) -	397	1	1	2130	0	15	0	16	0
	Microbiology (605) -	1	1	2	0	0	4	0	3	0
	Physiology (606) -	0	0	0	0	0	1	0	2	0
	Plant Biology (607) -	3	0	2	0	0	15	10	18	0
	Zoology (608) -	2	1	3	0	0	7	0	38	0
	Other Biological Sciences (699) -	1	0	0	0	0	1	0	2	16
		601	602	603	604	605	606	607	608	699
Actual										

Discussion and future work

- User behaviour:
 - Evidence that subject headings are used
 - Why and why not
 - Low usage of subject headings from this log collection
 - Is this unique to this data catalogue and interface?

Log analysis + survey and interview
- Collection characteristics:
 - Large proportion of records from the catalogue without a “standard” vocabulary for the subject headings a known issue
 - Those with subject headings are biased toward a few categories
 - Encourage underrepresented subject areas to publish and share data
 - Record classification works for some categories
 - Explore correlation, improvement

Thanks!