AUTOMATING ERA BENCHMARKS SYSTEM TECHNICAL REPORT

AN ON-DEMAND PILOT SYSTEM FOR CALCULATING ERA-LIKE BENCHMARKS USING OPEN DATA AND TRANSPARENT ANALYSIS.

November 2022

MAKE TOMORROW BETTER
OPENKNOWLEDGE.COMMUNITY

CURTIN OPEN KNOWLEDGE INITIATIVE

EXECUTIVE SUMMARY

To raise confidence in decision making, research administrators and funding agencies require insight into the performance of research-active institutions. Focusing on 42 Australian higher education providers and 236 fields of research, the Excellence in Research for Australia process (ERA) reports on research activity relative to local and global benchmarks. An ERA report is compiled for release every three to five years and employs citation-focused methodology in the analysis of research-output data, self-reported by the participating institutions. It is of interest to complement the ERA process by exploring additional data sources and analysis methods. To facilitate this, the Curtin Open Knowledge Initiative (COKI) has constructed a pilot system named RIES (Research Impact Evaluation System), demonstrating the feasibility of conducting an on-demand, ERA-like analysis for research-active institutions (locally and globally), using journal-level metadata from the ARC and article-level metadata from publicly available datasets.

Given a sufficiently comprehensive database of research-output metadata (that includes citation metrics, institutional affiliations, and linkages to journals planned for inclusion in ERA 2023), we show that RIES is able to generate ERA-like benchmarks and select indicators, based on ERA 2018 and proposed ERA 2023 methodology. The default analysis aims to align with ERA 2023; being applied between the years 2016 - 2021 (inclusive); using ANZSRC 2020 fields of research codes; and including calculation of benchmark boundaries for the dynamic RCI, the proposed high-performance indicator, and for citation centile groupings.

Determining institutional performance metrics (used to inform the citation-based ERA panels and ratings), is also feasible given a dataset that resembles those submitted by institutions for previous ERA rounds; containing lists of outputs (per institution) with assigned and apportioned fields of research to each output. We demonstrate that this is possible in principle using open data for institutional affiliation of outputs (a 'byline approach'), together with the ERA 2018 and ERA 2023 journal lists to assign outputs to fields of research. In a fully automated system these demonstration data would be replaced by either institutional submissions, based on a census date, or by an algorithmic field-of-research assignment process.

AUTHORS

Dr. Julian Tonti-Filippini, Curtin Institute for Computation, Curtin University
Dr. Kathryn Napier, Curtin Institute for Computation, Curtin University
Prof. Cameron Neylon, Centre for Culture and Technology, Curtin University

CONTRIBUTOR STATEMENTS

Conceptualization: Julian Tonti-Filippini and Cameron Neylon.
Data curation: Julian Tonti-Filippini.
Formal analysis: Julian Tonti-Filippini.
Funding acquisition: Cameron Neylon.
Investigation: Julian Tonti-Filippini.
Methodology: Julian Tonti-Filippini and Cameron Neylon.
Project administration: Kathryn Napier and Cameron Neylon.
Resources: Cameron Neylon.
Software: Julian Tonti-Filippini and Cameron Neylon.
Supervision: Kathryn Napier and Cameron Neylon.
Validation: Julian Tonti-Filippini.
Visualisation: Julian Tonti-Filippini.
Writing - original draft: Julian Tonti-Filippini and Cameron Neylon.
Writing - review & editing: Julian Tonti-Filippini, Kathryn Napier, and Cameron Neylon.

CONTENTS

LIST OF ABBREVIATIONS

ANZSRC Australian and New Zealand Standard Research Classification.

ARC Australian Research Council.

COKI Curtin Open Knowledge Initiative.

CPP Citations Per Paper.

DOI Digital Object Identifier.

ERA Excellence in Research for Australia.

ETL Extract, Transform and Load.

FoR ANZSRC Field of Research classification.

HEP (Australian) Higher Education Provider.

HPI High Performance Indicator.

ISSN International Standard Serial Number.

JSON JavaScript Object Notation. A text format for serialising data objects.

JSON-L A text-file format in which each line has one JSON-encoded record.

LVT Low Volume Threshold.

RCI Relative Citation Impact.

RIES Research Impact Evaluation System.

ROR Research Organization Registry.

INTRODUCTION

BACKGROUND

Excellence in Research for Australia (ERA) is a periodic assessment that is conducted by the Australian Research Council (ARC). In its upcoming iteration (ERA 2023), the assessment focuses on the activity of 42 Australian higher education providers (HEPs) across 236 fields of research (FoRs), defined by the Australian and New Zealand Standard Research Classification (ANZSRC). Under the ERA process, institutions are ranked by activity per field, then compared against local and world benchmarks to generate measures of relative performance and assign ratings. The analytical methods have a citation-focus and rely upon the participating HEPs self-reporting their affiliated research-outputs, with apportioned FoRs.

The Curtin Open Knowledge Initiative (COKI, based at Curtin University) aggregates publication metadata from publicly available sources such as Crossref, Unpaywall, OpenCitations, Microsoft Academic Graph, and OpenAlex. These datasets form a foundation for further analysis by the COKI team (with a focus on Open Access publication). Using published ERA methods as a guide, and journal-level metadata from the ERA 2023 Journal List, COKI has developed a Google BigQuery-based pilot Research Impact Evaluation System (RIES) that can deliver on-demand, ERA-like reporting for any set of research institutions against ANZSRC fields of research. Although the system currently focuses on citation metrics, it is intended to be modified in the future, to explore how different analysis methods may impact performance rankings and ratings.

This document describes the system's components and methods, without providing an in-depth analysis of results or a detailed discussion. Further documentation (and the project's source code), may be found in the project's GitHub source code repository, which has now been made freely available under the Apache 2.0 licence.

MOTIVATION

As noted in the Minister's letter to the ARC of 26 August 2022, the ERA process imposes a significant reporting burden on the sector. There has long been an interest in automating parts of this process to reduce this burden. Additionally, a 2021 consultation on the ERA process noted an interest in enhancing transparency regarding the construction of benchmarks and performance measures.

Pertaining to research outputs and performance metrics, the increasing availability of open data, over the past decade, has been transformational. Alongside this, computational tools have improved in performance to the extent that large scale analyses of massive datasets are cost-effective and accessible to small analysis teams. Open data sources are competitive against proprietary counterparts and offer the potential for greater transparency, access, accuracy and completeness, provided that they are used correctly by analysts with contextual knowledge.

At this time, national assessment exercises, as well as many higher education providers, continue to rely on traditional, proprietary data sources for performance evaluation. In the lead-up to the planned ERA 2023 round, we wondered whether such an exercise could be partially automated (to reduce administrative workload), could be conducted transparently (using exclusively public and openly licensed data), and might motivate the voluntary enhancement of underlying data sources over time, by institutions that could benefit from a system that used these data sources.

To test this, we sought to construct a pilot system that would implement analysis methods guided by published ERA documentation, and would be able to automatically analyse data, on-demand, drawn from the COKI database: an open-knowledge dataset that aggregates bibliographic and bibliometric data from over 120 million published research outputs.

OVERVIEW OF RESULTS

Using publicly available datasets, we have shown that it is feasible to implement an automated workflow for the production of ERA 2018 and ERA 2023-like benchmarks (and some indicators). Starting with global research output metadata (aggregated in the COKI database), an analytical set of research outputs is filtered by linking to journals in the ERA 2023 Journal List. Using ANZSRC 2020 FoR assignments (inherited from the Journal List), sets of outputs are then linked to fields of research, enabling calculation of citation benchmarks (such as RCI category boundaries, global CPP threshold, HPI threshold, ranks and percentile boundaries). By linking outputs to research institutions, benchmarks and indicators can then be calculated for particular subsets of global institutions, enabling focused analyses to be conducted, with the primary analysis being to rank Australian HEPs against global benchmarks.

Benchmark calculations are implemented in a workflow that may be adapted to utilise different data sources including, but not limited to: alternate journal lists; externally defined FoR assignments and apportionments to research outputs; alternative sources of citation data; and alternative sources of affiliation data for research outputs (for example, lists of ORCID profiles as the basis for linking outputs to institutions). The system has been designed with intent to support the calculation of new performance metrics and indicators using datasets that link research-outputs to statistically useful metadata (such as citation counts).

The system makes use of standards-based, persistent identifiers (PIDs) to enhance extensibility and compatibility with external data sources. These include: DOIs to identify research outputs, ROR codes to identify institutions, ISSNs to identify journals, and ANZSRC FoR codes to identify fields of research. The system is intended to support future testing and sensitivity-analysis of alternative methods such as: machine assignment and apportionment of FoR codes; calculation of non citation-based performance metrics; inclusion of datasets from new data providers; comparison of different institutional groupings, (for example by geography or economic classification); comparison of by-year versus census-period approaches; and testing of new benchmark proposals at scale, encompassing all ROR-listed institutions.

Examples of interactive visualisations, that have been generated from RIES output tables, are presented as figures 1-5.

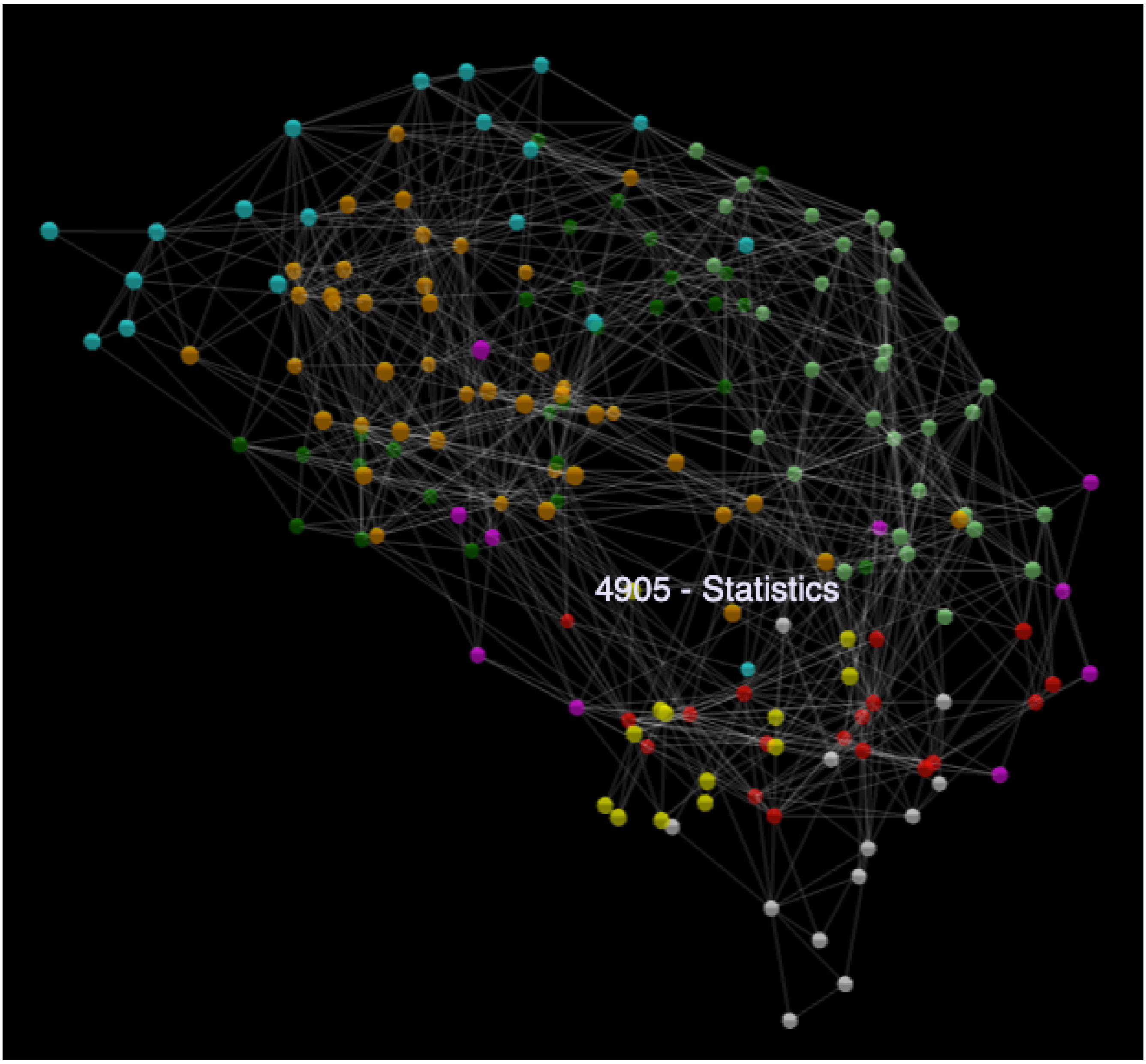

Figure 1: An interactive, 3d network plot assembled from computed FoR co-assignment data. Fields are colour coded by theme: physics & chemistry (cyan), mathematics & engineering (orange), earth & biology (dark green), health & human biology (light green), art & design (magenta), finance & economics (yellow), law & philosophy (grey), and culture & society (red). Individual nodes may be selected (eg, 4905 - Statistics).

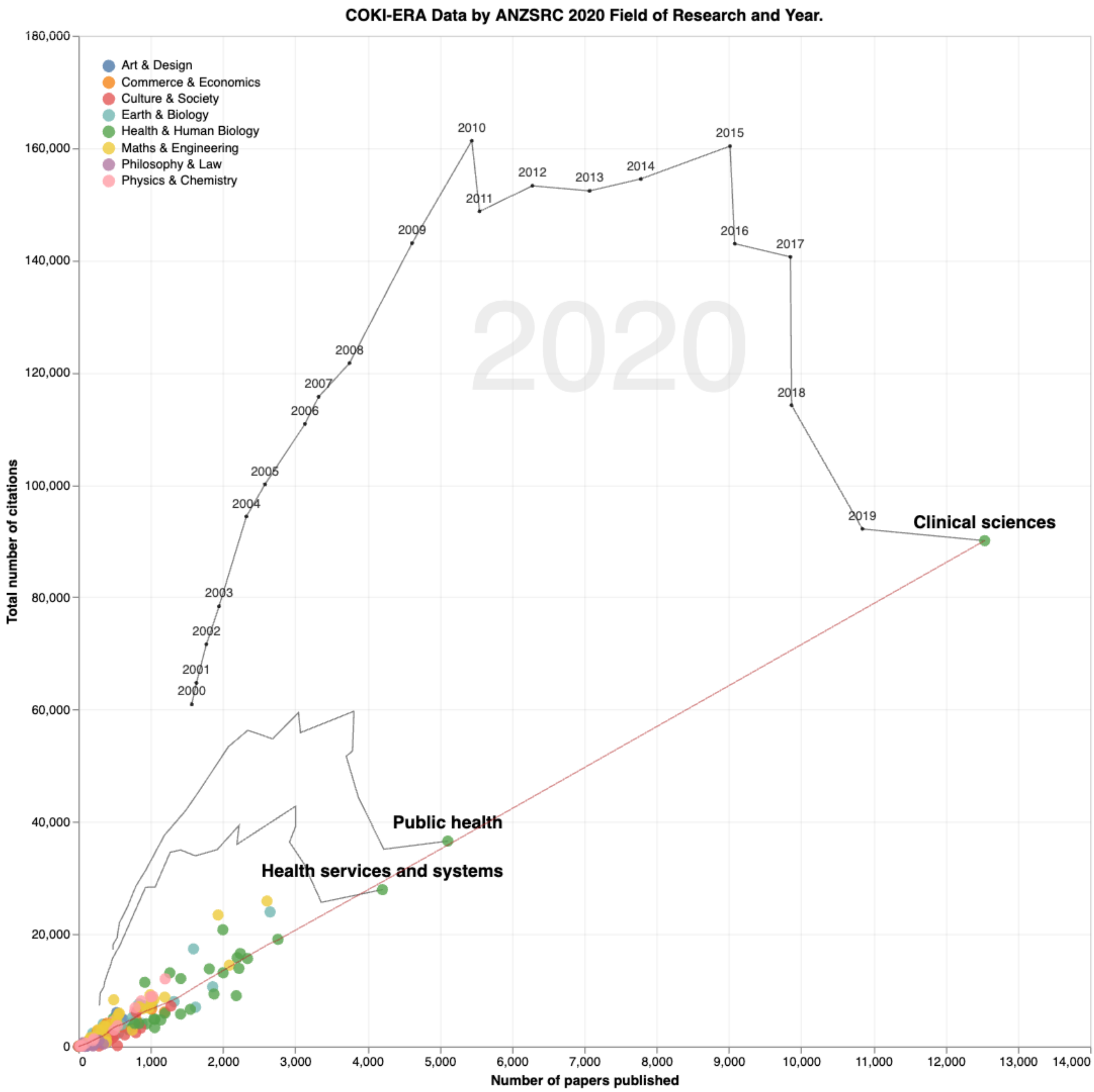

Figure 2: For each field of research (colour-coded by theme), an [interactive time-trace plot]() shows the number of papers published in 2020 versus the number of citations accrued to date. Time-trace lines show the migration of three selected data points between 2000 and 2020.

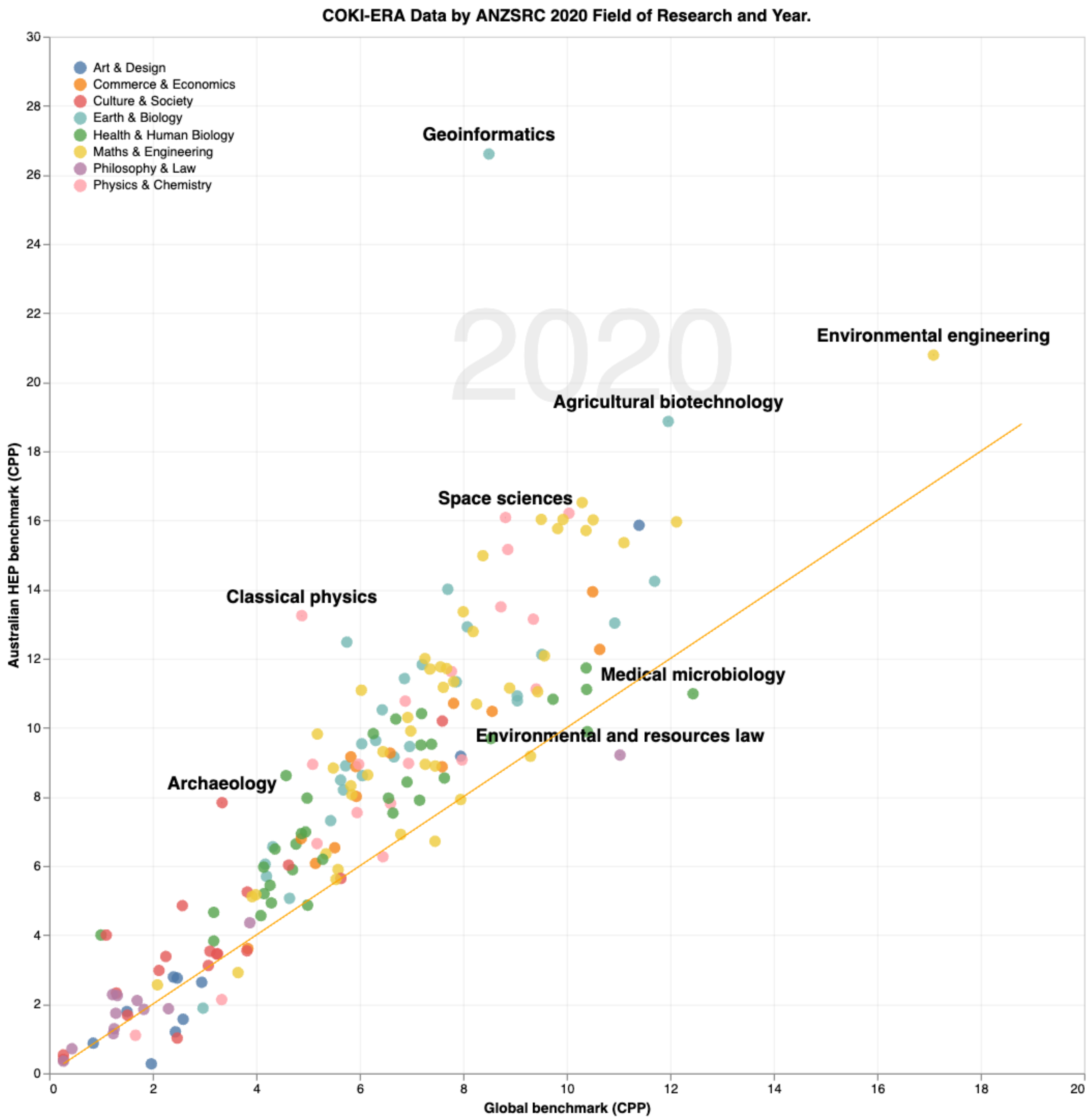

Figure 3: [RIES output example](). For the year 2020, global average citations per FoR (x-axis) are shown against Australian HEP average citations per FoR (y-axis). Points that are above the diagonal line represent FoRs where Australian activity is above the global average (ie, RCI > 1).

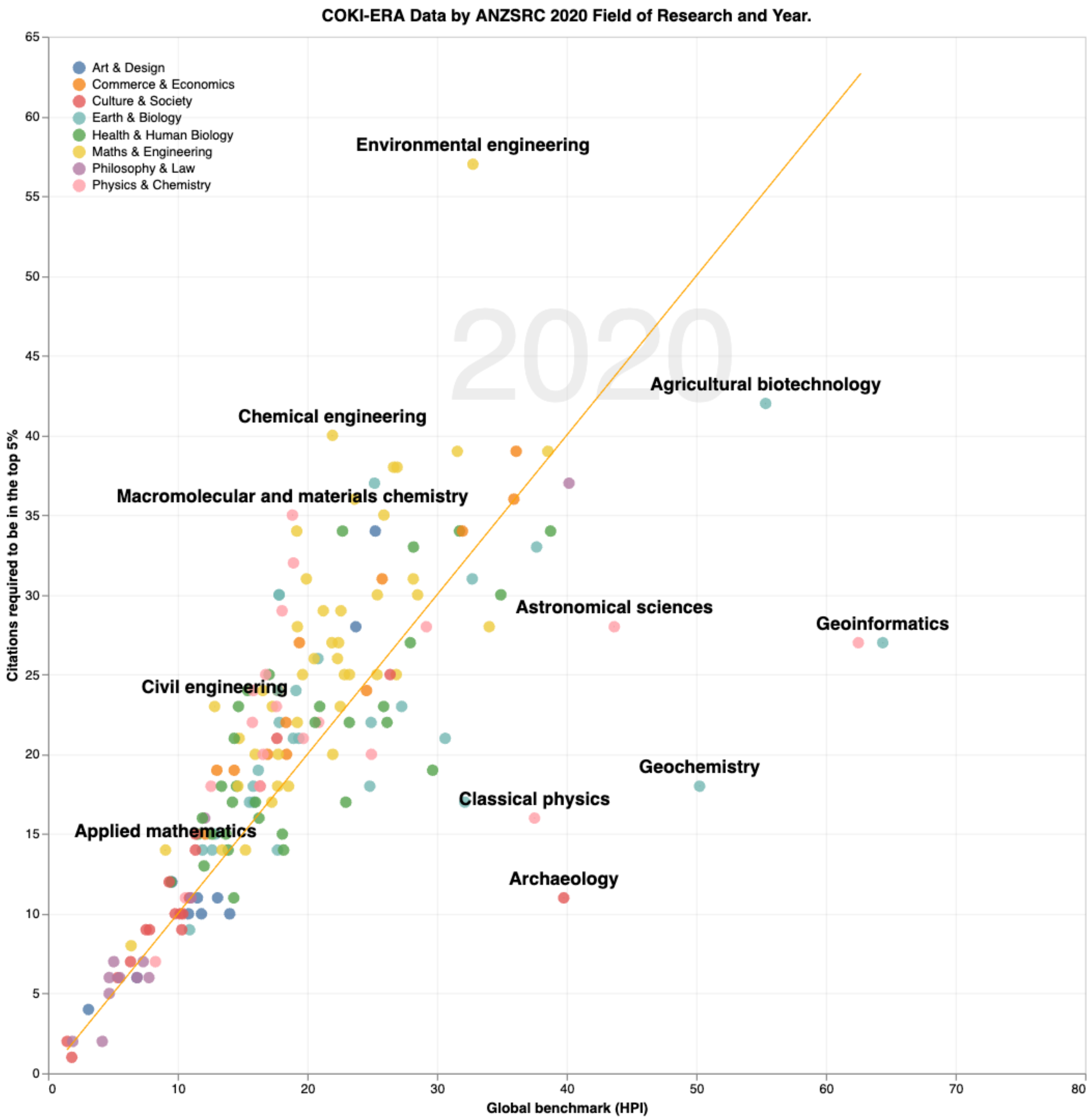

Figure 4: [RIES output example](). On the x-axis, the high-performance indicator (HPI) shows the number of citations required for a paper to qualify as high-performing. The y-axis shows the number of citations needed to be in the top 5% of all papers in that field (for the year 2020). Being above the diagonal line indicates that being in the top 5% of papers qualifies as high-performing. For fields of research that are below the diagonal line, it is more difficult to qualify using this metric and suggests that the HPI is susceptible to outliers..

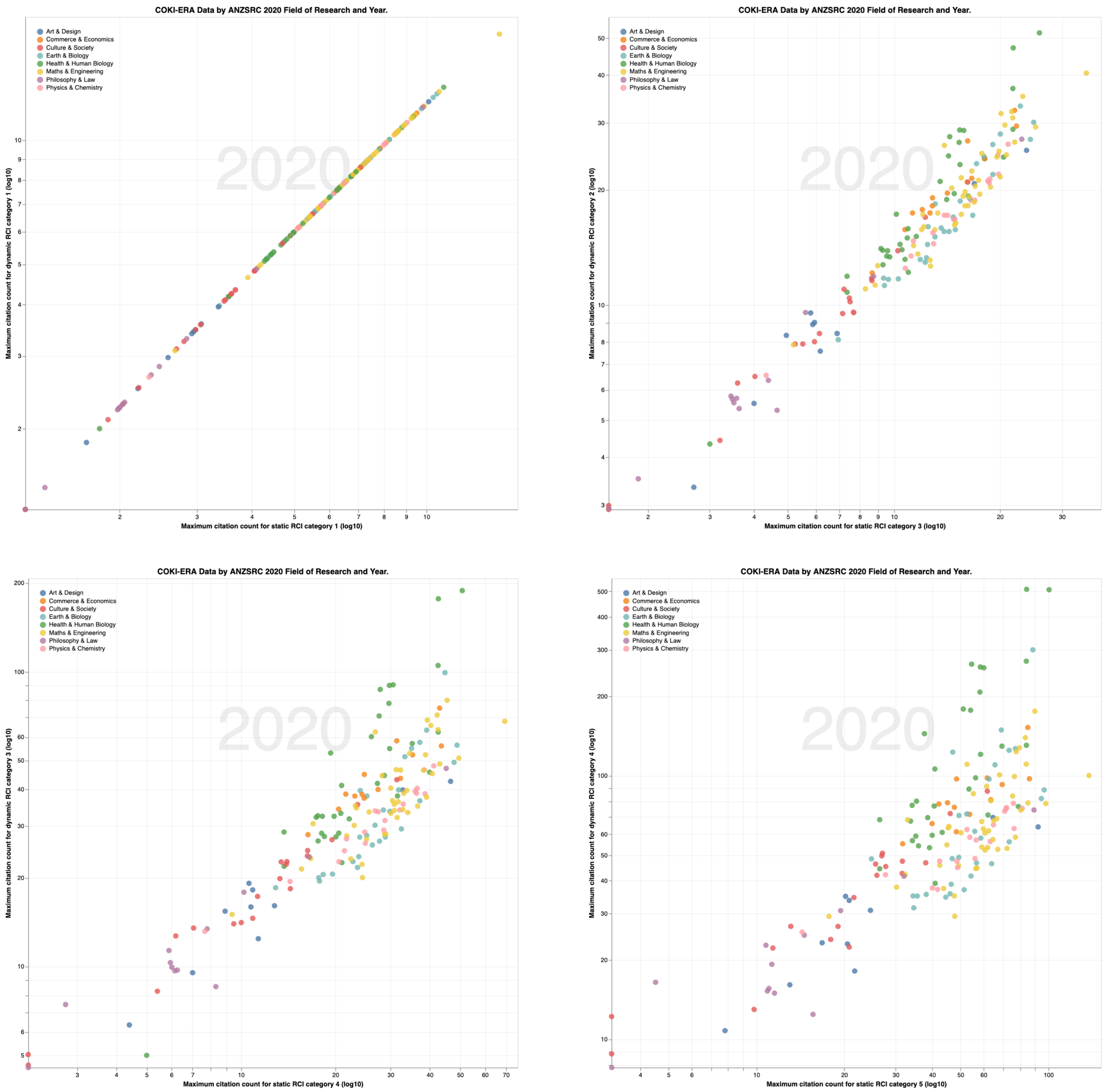

Figure 5: RIES output examples. A panel of interactive plots that compare the static RCI to the dynamic RCI for each FoR in the year 2020. X-axis: the maximum number of citations ($\log_{10}$) for the selected static RCI category (or lower bound for the next higher category). Y-axis: the maximum number of citations ($\log_{10}$) for the selected dynamic RCI category (or lower bound for the next higher category). Top-left: static-1:dynamic-1. Top-right: static-3:dynamic-2. Bottom-left: static-4:dynamic-3. Bottom-right: static-5:dynamic-4.

TECHNICAL DESCRIPTION

REQUIREMENTS

The project stores data in a Google BigQuery database with most of the analytical code written in Google Standard SQL. Several ETL processes are written in NodeJS and are intended to run in Unix-like environments such as Linux, OS X, WSL or Docker. Although direct access to the COKI database is not currently available, a data subset has been prepared and made available via Google Cloud Storage. The project's source code is freely accessible and may be used with the data subset to run an instance of RIES for demonstration purposes..

Detailed technical instructions for installing the code, configuring an instance, and using the interface are available in the project's GitHub repository. To run the code, a workstation is required that either runs Docker, or has the following utilities available at the command line: NodeJS, NPM, bash, git curl and unzip. It is possible to compile SQL queries without requiring a database, however if query execution is desired then access to a BigQuery instance is required.

COMPONENTS AND WORKFLOW

The basic workflow of the project involves ETL of external datasets, construction of benchmarks, then execution of multiple analysis streams to compile data for reporting purposes.

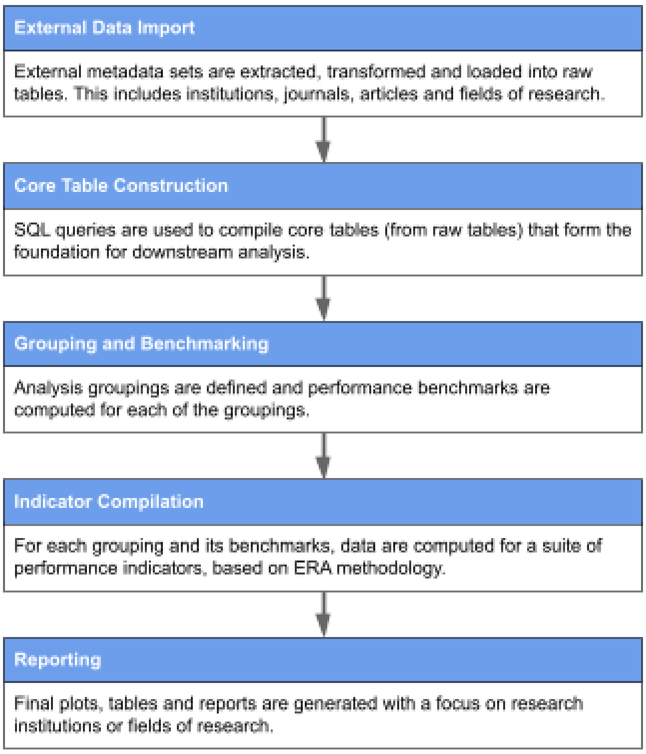

Figure 6: Basic RIES workflow.

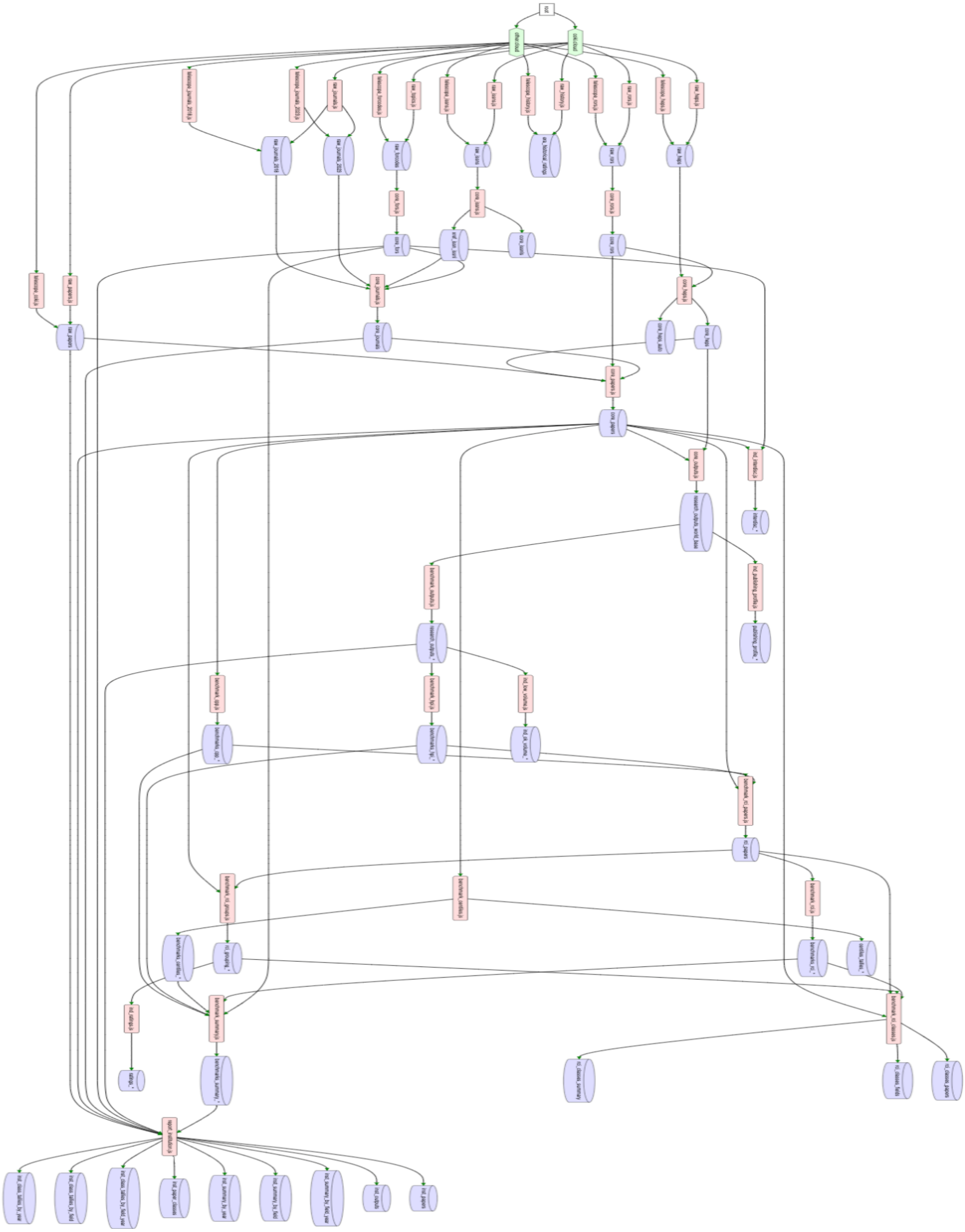

Figure 7: Complete RIES workflow. Green: external data. Red: executable script. Blue: SQL table. After core table construction, sections of the workflow are re-run for dozens of analysis streams.

ANALYTICAL GROUPINGS

The analysis process has three three major dimensions (institution, field of research, year). These dimensions are arranged into various groupings and the analysis flow is executed for each grouping, yielding separate sets of result tables. Some analysis is also conducted using (journal) in place of (institution).

In this document, groupings will be referred to using braced and italicised syntax; for instance {institution,field,year} refers to an analysis stream in which data have been grouped by institutional ID, field of research code, and year of publication. The unit of analysis in this grouping would be an institution's research output, for a specific field of research, during a specific year. Using three primary dimensions, there are seven groupings, each of which are analysed separately. Replacing (institution) with (journal) yields an additional set of seven tables.

Table 1: Primary analysis groupings.

{institution,field,year}	For analysis of research activity per institution, per field of research and per year. This is the highest resolution grouping available in the workflow.
{institution,field}	For analysis of research activity for each field of research in which an institution is active. Analysis is aggregated across all years of a defined analysis time frame (such as an ERA window). This is the primary unit of analysis in the ERA process.
{institution,year}	For analysis of each institution's research output, per year, across all fields of research.
{institution)	For analysis of each institution's research output, across all fields of research and all years of a defined time period.
{field,year}	For analysis of each field's total research activity, per year, combining output from all institutions.
{field}	For analysis of each field's total research activity, combining output from all institutions across all years of a defined time period.
{year}	For analysis of all research output, by year, combining activity from all institutions and all fields of research.

Further expanding the analysis, each dimension may be defined by independent datasets. For example, institutions may refer to a set of only the Australian HEPs, some other subset of institutions, or all global research institutions. Fields may refer to ANZSRC 2008 codes, 2020 codes, or some other set of research classifications. It is intended that the workflow enables a user to switch sets as desired, yielding a new series of seven output streams for each distinct combination of sets. The current system has been tested with two sets of institutions, four sets of research fields, and three sets of years. To clarify which is being referred to, subscripts may be used as follows.

Table 2: Primary analysis dimensions with subscripts to indicate the dataset being used.

$\text{institution}_{local}$	A set of institutions, limited to the 42 Australian higher education providers analysed by the ERA process (see Appendix II).
$\text{institution}_{world}$	All research institutions listed in the Research Organization Registry.
$\text{field}_{2020,2}$	2020 ANZSRC 2-digit field of research codes.
$\text{field}_{2020,4}$	2020 ANZSRC 4-digit field of research codes.
$\text{field}_{2008,2}$	2008 ANZSRC 2-digit field of research codes.
$\text{field}_{2008,4}$	2008 ANZSRC 4-digit field of research codes.
year_{era18}	Years that encompass the ERA 2018 analysis period (2011-2016).
year_{era23}	Years that encompass the ERA 2023 analysis period (2016-2021).
year_{all}	All available years in the COKI dataset.

These set options create the potential for up to 168 different analysis streams (2 institution sets * 4 field sets * 3 year sets * 7 dimensional combinations), further expanded by substituting institution for journal in the case of a journal-centric analysis. The choices of input sets may be defined as parameters when generating analysis queries.

By default, the remainder of this document will refer to a workflow that primarily focuses on the ERA 2023 round and methodology. This yields 28 analysis streams from two institutional sets ($institution_{local}$ and $institution_{world}$), two FoR sets ($field_{2020,2}$ and $field_{2020,4}$) and one date set ($year_{era23}$). Output table names are suffixed according to the choice of sets, shown below.

Table 3: The 28 primary analysis streams that each produce data for a series of indicators.

Grouping	Description
*world_4_institution_field_year	Global research output by institution, 4-digit field of research and year.
*world_4_institution_field	Global research output by institution and 4-digit field of research (summing across all years).
*world_4_institution_year	Global research output by institution and year (summing across all 4-digit fields of research).
*world_4_field_year	Global research output by 4-digit field of research and year (summing across institutions).
*world_4_institution	Global research output by institution (summing across all 4-digit fields of research and all years).
*world_4_field	Global research output by 4-digit field of research (summing across all institutions and years).
*world_4_year	Global research output by year (summing across all 4-digit fields of research and institutions).
*world_2_institution_field_year	Global research output by institution, 2-digit field of research and year.
*world_2_institution_field	Global research output by institution and 2-digit field of research (summing across all years).
*world_2_institution_year	Global research output by institution and year (summing across all 2-digit fields of research).
*world_2_field_year	Global research output by 2-digit field of research and year (summing across institutions).
*world_2_institution	Global research output by institution (summing across all 2-digit fields of research and all years).
*world_2_field	Global research output by 2-digit field of research (summing across all institutions and years).
*world_2_year	Global research output by year (summing across all 2-digit fields of research and institutions).
*local_4_institution_field_year	Australian research output by institution, 4-digit field of research and year.
*local_4_institution_field	Australian research output by institution and 4-digit field of research (summing across all years).
*local_4_institution_year	Australian research output by institution and year (summing across all 4-digit fields of research).
*local_4_field_year	Australian research output by 4-digit field of research and year (summing across institutions).
*local_4_institution	Australian research output by institution (summing across all 4-digit fields of research and all years).
*local_4_field	Australian research output by 4-digit field of research (summing across all institutions and years).
*local_4_year	Australian research output by year (summing across all 4-digit fields of research and institutions).
*local_2_institution_field_year	Australian research output by institution, 2-digit field of research and year.
*local_2_institution_field	Australian research output by institution and 2-digit field of research (summing across all years).

Grouping	Description
*local_2_institution_year	Australian research output by institution and year (summing across all 2-digit fields of research).
*local_2_field_year	Australian research output by 2-digit field of research and year (summing across institutions).
*local_2_institution	Australian research output by institution (summing across all 2-digit fields of research and all years).
*local_2_field	Australian research output by 2-digit field of research (summing across all institutions and years).
*local_2_year	Australian research output by year (summing across all 2-digit fields of research and institutions).

PRIMARY DATASETS

The first stage of the analysis workflow is to collect raw data from external sources, conduct some minor transformations, then load data into BigQuery tables prefixed by "***raw_***". This process is managed by a series of ETL scripts that may be found in the code/loaders directory of projects GitHub repository. It is not necessary to run these scripts to create a RIES instance, but they have been included to make the process transparent.

Rather than run the ETL scripts for each RIES instance, the COKI team periodically pre-runs the scripts and deposits output files into a publicly accessible cloud storage bucket. By default, the raw-table construction queries will import data directly from the COKI-generated data files.

Following construction of raw tables, another set of SQL scripts transforms data into a set of tables prefixed by "**core_**". These tables form the foundation of downstream analysis. A RIES instance requires access to six core datasets:

- **ISSNs** A mapping between ISSN and ISSN-L values. Source: ISSN International Centre.
- **FoRs** A list of ANZSRC field-of-research codes used in ERA. Source: ARC.
- **RORs** A list of institutional identifiers. Source: Research Organization Registry.
- **HEPs** A list of Higher Education Providers used in ERA. Source: ARC.
- **Journals** A list of journals used in ERA. Source: ARC.
- **Papers** A set of publication metadata, indexed by DOI. Source: COKI.

INTERNATIONAL STANDARD SERIAL NUMBERS

To connect journals in the ERA Journal List with journal-articles in the COKI DOI dataset, ISSN values are used as foreign keys. Although not strictly required, an authoritative list of mappings between ISSNs and linking ISSNs is sourced from issn.org. This mapping is then used to upgrade ISSN data in the COKI and ERA datasets where possible. The ETL method may be re-run at any time and is implemented in telescope_issns.js, raw_issns.js and core_issns.js.

Table 4: raw_issns. A list of ISSN to ISSN-L mappings sourced from issn.org (raw data). Created by: telescope_issns.js or raw_issns.js. Requires: issn.org.

Field	Type	Description
issn	STR	ISSN
issnl	STR	Linking ISSN

FIELDS OF RESEARCH

In ERA, and in the default RIES workflow, research disciplines are categorised using the ANZSRC fields of research scheme. These codes are arranged in a three-level hierarchy with each level of the hierarchy adding an additional two digits. For example:

- **31** - Biological Sciences
- **3101** - Biochemistry and cell biology
- **310101** - Analytical biochemistry

In the ERA process, the primary focus of analysis is on four-digit codes, although additional, aggregated analysis is also reported at the two-digit level (six-digit codes are not analysed). During each ERA round (defined by a year range), participating HEPs provide the ARC with metadata for all research outputs produced during the time frame. In the case of journal-articles, each article may be assigned up to three FoR codes, the values of which are constrained by the FoRs that have been assigned (by the ARC) to the parent journal. Analysis is restricted to a set of journals defined by the ERA Journal List. The following rules for FoR assignment apply:

- An output must have 1, 2 or 3 FoR code assignments.
- Each FoR assignment is apportioned an integer percentage (1-100), summing to 100%..
- If the journal specifies one or more 2-digit FoR codes, then the assignment may use any 4-digit codes that are encompassed by the 2-digit codes.
- If the journal specifies only 4-digit FoR codes, then the assignment may use any of these codes, but no other 4-digit or 2-digit code.

During the ERA 2018 round, the process used ANZSRC 2008 version codes. For the ERA 2023 round, the process will use ANZSRC 2020 version codes. Within the ANZSRC schema, these codes have been defined so that code numbers are mutually exclusive, enabling concurrent use if desired. Note that there is not a simple one to one conceptual mapping between fields-of-research in the 2008 and 2020 versions of the code sets. The lists of 2-digit and 4-digit codes are provided for both versions of the ANZSRC codes in Appendices III & IV.

Like ERA, the RIES workflow runs separate analysis streams for 2-digit and 4-digit codes. Unlike ERA, the RIES workflow does not have access to high-resolution FoR apportionment data (submitted to the ARC by individual HEPs). Although the workflow is configured to be able to ingest such data, in its absence the FoR assignments are instead inferred (inherited) from the containing journal with apportionment being uniformly distributed. The COKI analysis is therefore not able to provide as high a resolution analysis as ERA, for Australian HEPs, but does provide an expanded analysis for all global institutions.

The ETL method may be re-run at any time and is implemented in telescope_forcodes_2008.js, telescope_forcodes_2020.js, raw_fors.js and core_fors.js.

Table 5: raw_forcodes_*. These tables contain ANZSRC field-of-research data sourced from the Australian Bureau of Statistics. Two tables are generated, one for the 2008 version of the codes (used in ERA 2018), the other for the 2020 version of the codes (used in ERA 2023). Created by: telescope_forcodes_2008.js and telescope_forcodes_2020.js or by raw_fors.js. Requires: abs.gov.au, stats.govt.nz

Field	Type	Description
code	STR	Field of research code (either 2-digit, 4-digit or 6-digit)
name	STR	Field of research

Table 6: core_fors. This table is created from raw_fors and becomes the ground truth reference for all fields of research used in any workflow. Although there is a current focus on ANZSRC codes, the table is intended to be generic and work with any code set. Created by: core_fors.js. Requires: raw_forcodes.

Field	Type	Description
vers	STR	The version of ANZSRC codes being used (either 2008 or 2020).
len	INT	ANZSRC FoR hierarchy (either 2 or 4)
code	STR	Field of Research (FoR) subject short-code. Either a zero-padded 2,4 or 6 digit number, or 'MD' for multidisciplinary
name	STR	Field of Research (FoR) subject title

RESEARCH INSTITUTIONS

The Research Organization Registry is a website that maintains a database of registered research organisations from across the globe. Each organisation has a publicly accessible database entry that is defined and accessible by an unique URL. For instance, the ROR ID for Curtin University is **02n415q13** and may be accessed at https://ror.org/02n415q13.

Within COKI's databases, ROR IDs are used to connect between individual papers and organisations. This is dependent upon accurate authorship and affiliation data being captured by external data providers. Linkage data are stored in the COKI DOI dataset under an *affiliations* subset.

The ETL method extracts a full list of ROR IDs and names from ror.org and loads it into RIES. Special consideration is given to the Australian HEPs with an additional field being used as a boolean flag to indicate whether or not a specific ROR ID belongs to an Australian HEP.

The ETL method may be re-run at any time and is implemented in telescope_rors.js, raw_rors.js and core_rors.js.

Table 7: raw_rors. A list of Research Organization Registry identifiers, sourced from ror.org. During the ETL process, not all available fields are imported. Created by: telescope_rors.js or raw_rors.js. Requires: ror.org.

Field	Type	Description
ror	STR	Research Organization Registry ID for the institution
since	STR	The year that the institution was established
status	STR	Current status of the institution (filter for active)
type_0	STR	Primary activity type (ie, types[0])
country	STR	Country in which the research institution is located
name	STR	Name of the research institution
link_0	STR	Primary URL (ie, links[0])
types	[STR]	All activity types that the institution is engaged in
links	[STR]	All links associated with the institution

Table 8: core_rors. This table is created from raw_rors and becomes the ground truth reference for any workflow that refers to institutions. There is little difference between this table and raw_rors due to the high quality of the source data from ROR. The construction is included for consistency with other ETL processes. Created by: core_rors.js. Requires: raw_rors.

Field	Type	Description
ror	STR	Research Organization Registry ID for the institution
since	STR	The year that the institution was established
status	STR	Current status of the institution (filter for active)

country	STR	Country in which the research institution is located
name	STR	Name of the research institution
type_0	STR	Primary activity type (ie, types[0])
link_0	STR	Primary URL (ie, links[0])
types	[STR]	All activity types that the institution is engaged in
links	[STR]	All links associated with the institution

HIGHER EDUCATION PROVIDERS

The ERA analysis focuses on 42 Australian Higher Education Providers (HEPs), listed in Appendix II along with ROR identifiers. This list is simply loaded directly into the system. The ETL method may be re-run at any time and is implemented in telescope_heps.js, raw_heps.js and core_heps.js.

Table 9: raw_heps. A list of the 42 Australian Higher Education Providers (raw data). Created by: telescope_heps.js or raw_heps.js. Requires: arc.gov.au.

Field	Type	Description
ror	STR	Research Organization Registry ID for the institution
name	STR	Name of the research institution

Table 10: core_heps. Constructed from raw_heps, this table becomes the ground truth reference in any workflow that refers to an Australian Higher Education Provider. During construction, the source data are intersected with the set of ROR IDs from core_rors. Created by: core_heps.js. Requires: raw_heps.

Field	Type	Description
ror	STR	Research Organization Registry ID for the institution
name	STR	Name of the Australian higher education provider (from ROR)
era_name	STR	Name of the Australian HEP (according to ERA)

JOURNALS

For each ERA round, the ARC publishes a list of approved scientific journals, from which published articles can be included for analysis. For the purpose of the RIES analysis, two of these lists are downloaded as Excel files, then transformed and imported into the BigQuery database. Two files are sourced from the ARC: ERA 2018 Journal List and ERA 2023 Journal List.

Data are automatically extracted from the Excel files, converted to JSONL, and then uploaded to BigQuery. ISSN values, in the ERA data, are cross-referenced against the official ISSN set and checked for possible duplicates.

The ETL method may be re-run at any time and is implemented in telescope_journals_2018.js, telescope_journals_2023.js, raw_journals.js and core_journals.js.

Table 11: raw_journals_*. These tables contain data extracted from an ARC's ERA Journal List. The source is an Excel file. Two tables are created, one for the ERA 2018 list, the other for the ERA 2023 list. During the ETL, FoR codes are extended to include a uniformly distributed apportionment. Created by: telescope_journals_2018.js and telescope_journals_2023.js or raw_journals.js. Requires: arc.gov.au.

Field	Type	Description
era_id	STR	A unique identifier assigned by the ARC
title	STR	Journal title (English)

foreign_title	STR	Journal title (non-English) for foreign journals
issns	[STR]	List of ISSNs for the journal
forcodes	[OBJ]	Field of Research (FoR) codes assigned by ARC to the journal.
code	STR	FoR code (a STR because the values are zero padded and can be "MD")
name	STR	FoR name
weighted	[OBJ]	Field of Research (FoR) codes assigned by ARC to the journal.
code	STR	FoR code (a STR because the values are zero padded and can be "MD")
weight	INT	Portional assignment of this FoR code (should sum to 100 for the 2-digit codes and 100 for the 4-digit codes)

Table 12: core_journals. This table is created from raw_journals and becomes the ground truth reference for all journals referred to in any workflow. The list is currently coupled to the ERA Journal List, but this dependency may be removed in the future.. Created by: core_journals.js. Requires: raw_journals_*, core_fors, xref_issn_issnl.

Field	**Type**	**Description**
era_round	STR	Year of the ERA round
era_id	STR	A unique identifier assigned by the ARC
title	STR	English title of the journal
ftitle	STR	Non-English title of the journal
issns	[STR]	A list of ISSN codes associated with the journal
fors	[OBJ]	A list of Field of Research codes that have been assigned to the journal by the ARC
fors.vers	STR	ANZSRC FoR code version (either 2008 or 2020)
fors.len	INT	ANZSRC FoR hierarchy (either 2 or 4)
fors.code	STR	ANZSRC FoR code
fors.name	STR	ANZSRC FoR name
fors.weight	INT	FoR apportionment (1-100)

JOURNAL ARTICLES

The Curtin Open Knowledge Initiative (COKI) aggregates publication metadata into a DOI-based table from publicly available sources such as Crossref, Unpaywall, OpenCitations, Microsoft Academic Graph, and OpenAlex. This table provides source data for RIES journal-article records. In the context of this work, a research output refers to a published research work that must be:

- present in the internal COKI DOI dataset,
- classed as a journal-article, and
- assigned a valid DOI

Linking between a COKI record and an ERA Journal record is achieved by using an ISSN value as a foreign key. Optionally, further filters may be applied to ensure that all articles must be:

- listed as published within the ERA analysis period (2016-2021),
- linked (via ISSN) to a journal in the ERA 2023 Submission Journal List, and
- linked (via ROR) to at least one recognised research institution.

ISSN values for these papers are also intersected with the core_issn table, to assign ISSN-L values where possible and to check for duplication.

The research outputs that meet the above requirements are collected into a core_papers table and unnested. Following unnesting, the primary key is a composite of {paper,institution,field} defined by {DOI, ROR ID, FoR code}. For example, if a published work (with a unique DOI) has a set of authors that affiliate with five different institutions, and the work has been assigned two FoR codes, then there will be ten unique rows for this work in the basis table (1 DOI x 5 RORs x 2 FoRs).

As COKI does not have access to individual paper FoR apportionment data (provided by HEPs), it is during this component of the workflow that a research paper inherits FoR assignments from the linked journal. Weighting for these FoR codes is assigned uniformly between the number of codes (apportionment). Should high resolution apportionment data become available to COKI in the future, then this inheritance of values will be discarded.

The ETL methods are implemented in telescope_coki.js, raw_papers.js, core_papers.js and core_outputs.js. They may be re-run at any time.

Table 13: raw_papers. This table is created as a subset of the COKI DOI table, and uses joins to bring in additional data such as FoR code assignment and weighting. As COKI does not have access to HEP-assigned FoR codes and weights, these values are inherited from the linked journal in core_journals. Created by: core_papers.js. Requires: coki.doi, xref_issn_issnl, core_heps, core_rors, core_journals.

Field	Type	Description
doi	STR	Digital Object Identifier for the paper
year	INT	Year of publication for the paper
cits	INT	The number of citations the paper has accumulated since publication
is_oa	BOOL	True if the output has been identified as Open Access
type	STR	The type of publication, currently only journal-articles
issns	[STR]	List of ISSNs for the journal associated with this publication
rors	[STR]	List of ROR codes for institutions affiliated with this publication

Table 14: core_papers. This table is created as a subset of the COKI DOI table, and uses joins to bring in additional data such as FoR code assignment and weighting. As COKI does not have access to HEP-assigned FoR codes and weights, these values are inherited from the linked journal in core_journals. Created by: core_papers.js. Requires: coki.doi, xref_issn_issnl, core_heps, core_rors, core_journals.

Field	Type	Description
doi	STR	Digital Object Identifier for the paper
year_published	INT	year of publication for the paper (source: crossref)
num_citations	INT	the number of citations the paper has accumulated since publication (source: crossref & opencitations)
is_oa	BOOL	true if the output has been identified by Unpaywall as Open Access
era_id	STR	the ERA Journal ID of the journal that published this paper
rors	[STR]	Research Organization Registry IDs for all institutions associated with the work.
heps	[STR]	unique ROR identifiers for Australian higher education providers associated with the work
fors	[OBJ]	field of research codes with weightings (currently) inherited from the publishing journal
fors.vers	STR	ANZSRC FoR code version (either 2008 or 2020)
fors.len	INT	ANZSRC FoR hierarchy (either 2 or 4)
fors.code	STR	ANZSRC FoR code
fors.name	STR	ANZSRC FoR name
fors.weight	INT	FoR apportionment (1-100)

Table 15: research_outputs_world_base. This table contains the final set of research outputs that will be used for analysis, after all filters have been applied. Created by: core_outputs.js. Requires: core_papers, core_heps.

Field	Type	Description
year	INT	year of publication for the paper
journal	STR	the ERA Journal ID of the journal that published this paper
paper	STR	Digital Object Identifier for the paper
cits	INT	the number of citations the paper has accumulated since publication
inst	STR	ROR code of the affiliated institution.
is_hep	BOOL	true if the institution is an Australian higher education provider.
field	STR	Field of research code
field2	STR	First two digits of the field of research code
frac	INT	Apportioned weighting for this field of research (1-100)

BENCHMARKS

In order to assess relative performance, assign rankings and performance categories, the workflow compares grouped citation metrics against benchmarks. Guided by published ERA methods, the system constructs three sets of benchmarks, each of which computes average citations per paper as the metric,, depending on the grouping. The difference is determined by the set of institutions from which outputs are drawn to compute the benchmarks:

1. **Local:** calculated using only local institutions (42 Australian HEPs).
2. **World:** calculated using all active institutions (globally).
3. **HPI:** calculated using only the highest performing global institutions.

CENTILE BENCHMARKS

This method is based on ERA centile analysis (ERA 2018 Evaluation Handbook, section 5.5.2), and is implemented in benchmark_centiles.js (available in the project GitHub source code repository). It is parameterised by:

- **institution set:** either "local" (HEPs) or "world" (all institutions).
- **concept set:** either "2-digit" or "4-digit" ANZSRC FoR codes.

Centile analysis is based on raw citation counts and does not involve comparison to benchmarks. When grouping by field of research, the fractional apportionment of FoR codes is not involved in centile analysis.

For each {field, year}, research outputs are sorted by citation counts and divided into centiles. Citation counts are selected that define the boundaries of the following centiles: 1%, 5%, 10%, 25%, 50% (median) and 100% (total).

Table 16: benchmarks_centiles_*. These tables sort core_papers by citation count, then determine the number of citations required to bound a set of centile groups that have been specifically defined by the ERA process. Created by: benchmark_centiles.js. Requires: core_papers.

Field	Type	Description
field	STR	ANZSRC field of research code
year	INT	Year of analysis / publication
c1	INT	Number of citations required to be in the top 1% of papers
c5	INT	Number of citations required to be in the top 5% of papers
c10	INT	Number of citations required to be in the top 10% of papers

Field	Type	Description
c25	INT	Number of citations required to be in the top 25% of papers
c50	INT	Number of citations required to be in the top 50% of papers (median)
num_outputs	INT	Total number of published papers

Table 17: centile_tallies_*. These tables report summary statistics for the various centile groups defined in ERA. Outputs, citations and portions are tallied. For the typical ERA-like process, 4 tables are generated (two sets of institutions and two sets of research fields). Created by: benchmark_centiles.js. Requires: core_papers.

Field	Type	Description
field	STR	ANZSRC field of research code
year	INT	Year of analysis / publication
papers_1	INT	Number of papers in the top 1%
papers_5	INT	Number of papers in the top 5%
papers_10	INT	Number of papers in the top 10%
papers_25	INT	Number of papers in the top 25%
papers_50	INT	Number of papers in the top 50% (median)
papers_100	INT	Number of papers in the top 100% (total)
papers_uncited	INT	Number of uncited papers
papers_all	INT	Total number of papers (including uncited)
citations_1	INT	Sum of citations for all papers in the top 1%
citations_5	INT	Sum of citations for all papers in the top 5%
citations_10	INT	Sum of citations for all papers in the top 10%
citations_25	INT	Sum of citations for all papers in the top 25%
citations_50	INT	Sum of citations for all papers in the top 50%
citations_100	INT	Sum of citations for all papers in the top 100%
citations_uncited	INT	Sum of citations for uncited papers (qc, should be zero)
citations_all	INT	Sum of citations for all papers
portions_1	DEC	Sum of apportionments for all papers in the top 1%
portions_5	DEC	Sum of apportionments for all papers in the top 5%
portions_10	DEC	Sum of apportionments for all papers in the top 10%
portions_25	DEC	Sum of apportionments for all papers in the top 25%
portions_50	DEC	Sum of apportionments for all papers in the top 50%
portions_100	DEC	Sum of apportionments for all papers in the top 100%
portions_uncited	DEC	Sum of apportionments for uncited papers
portions_all	DEC	Sum of apportionments for all papers

CITATIONS PER PAPER (CPP) BENCHMARK

ERA performance ratings are influenced by citation counts. A highly cited work is considered to have a higher impact than an uncited work and will positively impact the assignment of a performance rating. In computing an impact score for a particular output, or group of outputs, a relative citation impact (RCI) is calculated by dividing citations by a benchmark value. When computing the CPP benchmark, all research outputs are included, even if they belong to groups that failed to meet the low-volume threshold (LVT).

This method is implemented in benchmark_cpp.js (available in the project GitHub source code repository) and is parameterised by institutional scope (world or local) and field of research hierarchy (2 or 4). Four sets of benchmarks are created that are used in multiple analysis streams. For each of the sets of benchmark, an individual benchmark value is grouped by {field,year} and is calculated as a simple average (citations per paper):

$$benchmark(field, year)_{cpp} = \frac{\Sigma\, citations(field,year)}{\Sigma\, outputs(field,year)}$$

The local benchmark is computed by restricting the set of research outputs to only those that can be affiliated to Australian HEPs. The world benchmark considers all outputs. Separate benchmark tables are constructed for 2-digit (level 1) FoR codes and for 4-digit (level 2) FoR codes.

Fractional apportionment of research fields is not considered during the computation of the CPP benchmark (it is not a weighted average). Care must also be taken to avoid double-counting articles that have multi-institutional authorship.

Note that, in ERA, if different institutions conflict on the assignment of FoR codes for the same paper, this is flagged and resolved manually. In the COKI workflow, HEP-assigned FoR codes are not available and so this method is not required or implemented.

Benchmarks are not computed solely on a field basis, as the number of outputs and the total number of citations are highly sensitive to time (the number of outputs increases exponentially over time and citations accumulate). Benchmarks are also not computed solely on a time basis, as there is a high degree of variability in activity between fields making it difficult to compare them directly on this basis.

Table 18: benchmarks_cpp_*. These tables calculate average citations per paper as a benchmark, grouped by field of research and year. For ERA-like reporting, four tables are generated to allow for different analysis streams (two institution sets and two FoR sets). Created by: benchmark_cpp.js. Requires: core_papers.

Field	Type	Description
code	STR	ANZSRC field of research code
year	INT	Year of analysis / publication
num_papers	INT	Total number of published papers in the grouping
num_cited	INT	Number of papers with at least one citation
num_uncited	INT	Number of papers that have not been cited
sum_citations	INT	Total number of citations
max_citations	INT	Maximum number of citations for a single paper
avg_citations	DEC	Average number of citations per paper
sdev_citations	DEC	Standard deviation of citations
benchmark	DEC	Citation benchmark (same as average)

HIGH PERFORMANCE INDICATOR (HPI) BENCHMARK

The high performance indicator assesses an institution's activity in a given field of research relative to the performance of the highest performing institutions in the field. The method for constructing the benchmark is implemented in benchmark_hpi.js (available in the project GitHub source code repository) and is parameterised by:

- **Code hierarchy:** benchmark tables are built separately for 2-digit FoR codes and 4-digit FoR codes.
- **Volume threshold:** the minimum number of weighted outputs required for an institution to be included in the calculation of the HPI (default: 50).
- **Centile threshold:** the centile boundary that denotes membership in the high-performance group of institutions (default: 10). For example, a value of 10 indicates that an institution must be in the top 10% of institutions to be considered a high-performer.

Calculation of the HPI benchmark includes all global institutions. Unlike the CPP benchmark, there is no HEP specific (local) benchmark. As is the case with the CPP benchmark, no attempt is made to calculate benchmarks by year or by field, due to the same uncontrolled biases.

The method for computing the HPI proceeds as follows:

1. All valid global outputs are collected and grouped by institutional affiliation.
2. Institutions are dropped from the analysis if they do not meet the volume threshold.

$$institution(field)_{LVT} \subseteq institutions \mid \Sigma\, portions(field) \geq threshold$$

3. The CPP is calculated for each of the surviving institutions.

$$CPP(institution, field, year) = \frac{\Sigma\, citations(institution, field, year)}{\Sigma\, outputs(institution, field, year)}$$

4. The highest performing institutions are selected (sorted by CPP). The cutoff for selecting these institutions is defined by the centile threshold.

$$institutions_{HPI} \subseteq institutions_{LVT} \mid centile(institution, field) \geq threshold$$

5. If there are insufficient institutions to populate the high-performance group with at least one institution, then the calculation of an HPI for this FoR is aborted.
6. All outputs from the members of the high-performance group are pooled and a new CPP score is computed from this pool. This value then becomes the HPI benchmark for the given field of research and year.

$$benchmark_{HPI} = \frac{\Sigma\, citations(institutions_{HPI}, field, year)}{\Sigma\, outputs(institutions_{HPI}, field, year)}$$

RCI BENCHMARKS AND CLASSES (STATIC & DYNAMIC)

The RCI category indicator assigns an integer value, representing a performance class, to each grouping based on RCI score. An individual paper, or group of papers, is assigned to an RCI category for each of its RCI contexts (local, world, HPI). Each performance class is delineated by a pair of RCI boundary values. Under the ERA 2018 methodology, the assignment of RCI categories uses the following seven static class bands. The upper limits of each band are:

- class 0: $RCI \ = \ 0$ (ie: no citations)
- class 1: $RCI \ < \ 0.80$
- class 2: $RCI \ < \ 1.20$
- class 3: $RCI \ < \ 2.00$
- class 4: $RCI \ < \ 4.00$
- class 5: $RCI \ < \ 8.00$
- class 6: $RCI \ unlimited$

Under the ERA 2023 methodology, there are six RCI classes (0-5) and the class boundaries are computed dynamically. The upper limits (inclusive) of each band are:

- class 0: $RCI \ = \ 0$ (ie: no citations)
- class 1: $mean(RCI)$
- class 2: $mean(RCI \ > \ class\ 1)$
- class 3: $mean(RCI \ > \ class\ 2)$
- class 4: $mean(RCI \ > \ class\ 3)$
- class 5: $RCI \ unlimited$

Note that class 1 does include class 0 (uncited) works.

These methods are implemented in benchmark_rci_classes.js (available in the project GitHub source code repository) and assign RCI classes to individual outputs and to {institution,field,year} groups.

Table 19: benchmarks_rci_*. These tables calculate relative citation impact scores (RCI) that set the boundaries for ERA-defined RCI classes. There are two sets, static and dynamic with different boundary values, calculated by different methods between ERA 2018 and ERA 2023. For ERA-like reporting, four tables are generated to allow for different analysis streams (two institution sets and two FoR sets). Created by: benchmark_rci_classes.js. Requires: rci_papers.

Field	Type	Description
field	STR	ANZSRC field of research code
year	INT	Year of analysis / publication
max_rci	DEC	Maximum RCI for a single paper in the group
s_c0	DEC	RCI upper limit for static RCI category 0 (zero)
s_c1	DEC	RCI upper limit for static RCI category 1
s_c2	DEC	RCI upper limit for static RCI category 2
s_c3	DEC	RCI upper limit for static RCI category 3
s_c4	DEC	RCI upper limit for static RCI category 4
s_c5	DEC	RCI upper limit for static RCI category 5
s_c6	DEC	RCI upper limit for static RCI category 6 (unlimited)
d_c0	INT	RCI upper limit for dynamic RCI category 0 (zero)

d_c1	DEC	RCI upper limit for dynamic RCI category 1
d_c2	DEC	RCI upper limit for dynamic RCI category 2
d_c3	DEC	RCI upper limit for dynamic RCI category 3
d_c4	DEC	RCI upper limit for dynamic RCI category 4
d_c5	DEC	RCI upper limit for dynamic RCI category 5 (unlimited)

BENCHMARKS SUMMARY

The benchmarks summary table brings together all benchmarks, RCI category and centile boundaries into a single table. The method is implemented in benchmark_summary.js (available in the project GitHub source code repository). This table is used for all downstream analysis involving benchmarks and boundaries.

Table 20: benchmarks_summary_*. These tables bring together all benchmark and boundary values, for a {field,year} grouping into a convenient helper table. This includes the CPI benchmarks, HPI benchmark, centile boundaries, and RCI class boundaries (static and dynamic). For ERA-like reporting, four tables are generated to allow for different analysis streams (two institution sets and two FoR sets). Created by: benchmark_summary.js. Requires: benchmarks_cpp_*, benchmarks_hpi_*, benchmarks_centiles_*, benchmarks_rci_*.

Field	Type	Description
field	STR	ANZSRC field of research code
name	STR	ANZSRC field of research name
year	INT	Year of analysis / publication
num_papers	INT	total number of papers published in the year
num_uncited	INT	number of papers that have no citations
cpp_local	DEC	benchmark citations per paper for Australian HEPs only
cpp_world	DEC	benchmark citations per paper for all institutions
cpp_hpi	DEC	benchmark citations per paper for high performing global institutions (avg of the top 10%)
ctile_01	DEC	citations needed to be in the top 1% globally
ctile_05	DEC	citations needed to be in the top 5% globally
ctile_10	DEC	citations needed to be in the top 10% globally
ctile_25	DEC	citations needed to be in the top 25% globally
ctile_50	DEC	citations needed to be in the top 50% globally
dynamic_c0	DEC	RCI score upper limit for category 0 (dynamic method)
dynamic_c1	DEC	RCI score upper limit for category 1 (dynamic method)
dynamic_c2	DEC	RCI score upper limit for category 2 (dynamic method)
dynamic_c3	DEC	RCI score upper limit for category 3 (dynamic method)
dynamic_c4	DEC	RCI score upper limit for category 4 (dynamic method)
dynamic_c5	STR	RCI score upper limit for category 5 (dynamic method)
maximum_rci	DEC	The maximum observed RCI score (technically the precise upper limit for dynamic_c5)
static_c0	DEC	RCI score upper limit for category 0 (static)
static_c1	DEC	RCI score upper limit for category 1 (static)
static_c2	DEC	RCI score upper limit for category 2 (static)
static_c3	DEC	RCI score upper limit for category 3 (static)
static_c4	DEC	RCI score upper limit for category 4 (static)
static_c5	DEC	RCI score upper limit for category 5 (static)
static_c6	STR	RCI score upper limit for category 6 (static)

INDICATORS

Following construction of benchmarks, the workflow then proceeds to analysis. The analysis phase builds a subset of ERA indicators (from ERA 2018 and ERA 2023). The ERA process aims to apply a qualitative activity rating to select institutions (HEPs) for each field of research (FoR) in which the institution is active. It additionally reports aggregate statistics for institutions and for fields of research.

Using methods that are primarily guided by the ERA 2018 Evaluation Handbook and the ERA 2023 Benchmarking and Rating Scale – Consultation Paper, the system generates a subset of matching or similar indicators, limited to an analysis of journal-articles only. These indicators are briefly described below with a more detailed description in the Methods section.

Research Outputs	Summary citation statistics with a focus on institutions and fields
Publishing Profile	Summary citation statistics with a focus on journals and fields
Low Volume Threshold	Analysis of which institutions and fields can be considered active
Interdisciplinary Profiles	Analysis of which fields are linked together by co-apportionment
Relative Citation Impact	Analysis of relative citation impact per grouping.
RCI Class	Assignment of RCI classes using a static or dynamic method
Centile Analysis	Assignment of centiles and ranks per grouping.
Performance Rating	Assignment of performance ratings using three different methods: ERA 2018, 2023-A and 2023-B.

RESEARCH OUTPUTS

In the context of this work, a research output refers to a published research work that must be:

- present in the COKI dataset,
- classed as a journal-article,
- assigned a valid DOI,
- published in a journal that is listed in the ERA 2023 Submission Journal List, and
- published within the ERA analysis period (2016-2021).

The articles that meet these requirements are collected as a subset of COKI's DOI dataset, then unnested into a basis table in which the unique key is a composite of paper ID, institution ID and field-of-research ID. For example, if a published work has a set of authors that affiliate with 5 different institutions, and the work has been assigned 2 FoR codes, then there will be 10 unique rows for this work in the basis table.

For each of the 28 groupings, defined previously, the research outputs indicator computes summary statistics for each cohort of journal-articles within the group, then assigns a rank (1-N) and centile (1-100) for each cohort relative to peers, with 1 representing the highest level of output.

The basis table forms the foundation for downstream analysis, according to the groupings described in the previous section. Each grouping flows through to a grouped set of benchmarks and final reports.

Note that:

- In ERA, outputs are sub-divided by output type, such as book, book chapter, journal article, etc. In this analysis, only journal articles are considered.

- In ERA, analysis focuses on 42 Australian HEPs over a 5-year time frame. The COKI workflow is intended to be applied to any grouping of research institutions over an extended time frame.

Table 21: research_outputs_*. These tables are used to determine which units are the most active within each grouping, based on output counts and citation counts. In the typical analysis flow, there will be 28 tables generated for all combinations of {institution,field,year}. In ERA, these numbers are used to rank institutions and assign centile membership. Although ranks can be compared between fields of research and years, caution should be exercised when contrasting raw tallies. In this case, it is better to use RCI as it is normalised. Created by: benchmark_outputs.js. Requires: research_outputs_*_base.

Field	Type	Description
institution	STR	Research Organization Registry ID for the institution
field	STR	ANZSRC field of research code
year	INT	Year of analysis / publication
sum_papers	INT	Total number of papers published in the journal for this grouping
sum_citations	INT	Sum of all citations for all papers in the grouping
sum_portions	INT	Sum of all fractional assignment for all papers in the grouping
avg_citations	DEC	Average number of citations for papers in the grouping
cent_papers	INT	Centile membership (1-100) for this grouping based on number of papers (1 = top 1%)
cent_citations	INT	Centile membership based on total citations
cent_portions	INT	Centile membership based on total apportionments
cent_cpp	INT	Centile membership based on average citations
rank_papers	INT	Rank (1-N) for this grouping based on number of papers (1 = top rank)
rank_citations	INT	Rank based on total citations
rank_portions	INT	Rank based on total apportionments
rank_cpp	INT	Rank based on average citations

Table 22: research_outputs_*_base. These tables form the basis for grouped research_outputs tables (above). There are four tables generated (two sets of institutions and two sets of research fields).. Created by: core_outputs.js. Requires: core_papers, core_heps.

Field	Type	Description
year	INT	Year of publication
journal	STR	ERA ID for the journal
paper	STR	Digital Object Identifier of the paper
cits	INT	Number of citations
inst	STR	Research Organization Registry ID for the institution
is_hep	BOOL	True if the institution is an Australian HEP
field	STR	Assigned ANZSRC code (either 2 or 4 digit)
field2	STR	Encompassing ANZSRC 2-digit code
frac	INT	Fractional apportionment of the assigned code (1-100)

PUBLISHING PROFILE

The publishing profile indicator shows which scientific journals are the most active within each field of research. The method is implemented in ind_publishing_profile.js (available in the project GitHub source code repository) and is essentially identical to the method used to compile the research outputs indicator, but focusing on journals instead of institutions. Where the research outputs indicator produces 28 output tables based on all combinations of {

institution$_{(local, world)}$, field$_{(2, 4)}$, year$_{era23}$ }, the publishing profile indicator produces 14 output tables based on all combinations of { journal, field$_{(2, 4)}$, year$_{era23}$ }.

Table 23: publishing_profile_*. These tables are used to determine which Journals are the most active within each field of research (by year). This can be useful for researchers who are looking for the best sources of information or best journals to submit to when publishing in a given field. Created by: ind_publishing_profile.js . Requires: research_outputs_base_*.

Field	**Type**	**Description**
journal	STR	Unique ID for the journal (from the ERA Journal List)
field	STR	ANZSRC field of research code
year	INT	Year of analysis / publication
sum_papers	INT	Total number of papers published in the journal for this grouping
sum_citations	INT	Sum of all citations for all papers in the grouping
sum_portions	INT	Sum of all fractional assignment for all papers in the grouping
avg_citations	DEC	Average number of citations for papers in the grouping
cent_papers	INT	Centile membership (1-100) for this grouping based on number of papers (1 = top 1%)
cent_citations	INT	Centile membership based on total citations
cent_portions	INT	Centile membership based on total apportionments
cent_cpp	INT	Centile membership based on average citations
rank_papers	INT	Rank (1-N) for this grouping based on number of papers (1 = top rank)
rank_citations	INT	Rank based on total citations
rank_portions	INT	Rank based on total apportionments
rank_cpp	INT	Rank based on average citations

LOW VOLUME THRESHOLD

The low volume threshold (LVT) is an ERA indicator that flags research institutions for low activity in a particular field of research. In the case of journal-articles, the LVT is set to a minimum of 50 weighted research outputs within the timeframe of the ERA analysis and does not consider yearly variability within the timeframe. If an institution does not meet the LVT for a particular field of research, then an ERA rating will not be assigned for that institution and field. The LVT does not take into account year by year variability of output during an ERA time period

The method is implemented in ind_low_volume.js (available in the project GitHub source code repository) and the threshold value is configurable.

Table 24: ind_ok_volume_*. These tables are used to provide a quick indication of whether or not a particular institution meets the ERA-defined Low Volume Threshold in a given field of research. Created by: ind_low_volume.js. Requires: research_outputs_*.

Field	Type	Description
institution	STR	Research Organization Registry ID for the institution
field	STR	ANZSRC field of research code
sum_papers	INT	Total number of papers that pass the low volume threshold

INTERDISCIPLINARY PROFILES

The interdisciplinary profile is an ERA indicator that shows the frequency of pairings between two different fields of research. A pairing is counted when two FoR codes are assigned to a single research output. This indicator is useful for highlighting interdisciplinary activity (or a lack thereof).

The method is implemented in ind_inderdisc.js (available in the project GitHub source code repository) and is parameterized by scope (world or local) and FoR hierarchy (2 or 4).

Table 25: interdisc_*. These tables are used to show the relationship between pairs of research fields, specifically how many times the fields have been co-listed during assignment of FoRs to a research output. Created by: ind_interdisc.js Requires: core_fors, core_papers.

Field	Type	Description
code1	STR	ANZSRC field of research code (basis)
name1	STR	ANZSRC field of research name (basis)
code2	STR	ANZSRC field of research code (other)
name2	STR	ANZSRC field of research name (other)
num	INT	Total number of papers with this FoR pairing
weight	DEC	Sum of apportionments for this pairing
pct_num	DEC	Of all papers in the basis group, the percentage with this pairing
pct_weight	DEC	Of all papers in the basis group, the percentage of total portions

RELATIVE CITATION IMPACT

The relative citation impact is a score that is computed for an individual research output, or a grouped set of outputs. It is intended to provide a quick indication of where an output (or group of outputs) stands relative to an appropriate benchmark. An RCI of 1.0 indicates that the output is exactly at the average. An RCI of 2.0 indicates that the output is double the average. Under the ERA methodology, if an RCI is >= 8 then a warning is triggered and a percentile analysis may be more appropriate.

RCI scores are used downstream to assign RCI categories (static and dynamic) and these categories are used to instruct the assignment of ERA performance ratings.

The ERA method for computing the RCI for an individual output is not the same as the method for computing the RCI for a set of outputs, with the latter using a weighted average that is sensitive to the apportionment of FoR codes.

The method for individual outputs is implemented in benchmark_rci_papers.js (available in the project GitHub source code repository) and is parameterised only by the code hierarchy (either 2 or 4 digit FoR coding), specifying which benchmark table to use. For each pairing of paper and field of research, three RCI values are computed against the appropriate benchmarks (local, world and HPI)

$$RCI_{local}(paper, FoR) = \frac{citations(paper)}{benchmark_{local}(FoR, year(paper))}$$

$$RCI_{world}(paper, FoR) = \frac{citations(paper)}{benchmark_{world}(FoR, year(paper))}$$

$$RCI_{hpi}(paper, FoR) = \frac{citations(paper)}{benchmark_{hpi}(FoR, year(paper))}$$

The method for calculating the RCI for a set of outputs is implemented in benchmark_rci_groups.js (available in the project GitHub source code repository) and is parameterised by:

- **code hierarchy:** benchmark tables are built separately for 2-digit FoR codes and 4-digit FoR codes.

- **grouping:** RCI scores can be computed for seven different groupings: {institution,field,year}, {institution,field}, {institution,year}, {field,year}, {institution}, {field}, or {year}

As is the case for individual outputs, three RCI values are computed (local, world and HPI). Unlike individual outputs, the group RCI score is computed as the weighted average of the individual RCI scores of outputs in the grouping. Weighting is defined by the apportioned fraction of the field of research, therefore where the field is not involved in the grouping, the field weighting sum to 1.0 and the average is a simple average.

$$RCI_{local}(group) = \frac{\Sigma_{papers}(RCI_{local} \times weight)}{\Sigma_{papers}(weight)}$$

$$RCI_{world}(group) = \frac{\Sigma_{papers}(RCI_{world} \times weight)}{\Sigma_{papers}(weight)}$$

$$RCI_{hpi}(group) = \frac{\Sigma_{papers}(RCI_{hpi} \times weight)}{\Sigma_{papers}(weight)}$$

Additional notes:

- In ERA, if a FoR is linked to less than 75 indexed papers (across the analysis period), then a warning is generated, suggesting that the user considers centile and RCI class analysis instead. This is not currently implemented in the RIES.

- In ERA, where a 4-digit FoR is linked to less than 250 articles, between all HEPs combined (across the analysis period), then a low-volume warning is generated. This is not currently implemented in the RIES.

- In ERA, if a benchmark value is zero, then the paper's RCI will not be included in the calculation of the weighted average RCI calculation for a field of research. This is implemented in the RIES.

These methods are implemented in benchmark_rci_groups.js (available in the project GitHub source code repository) and are parameterised only by field set. Class boundaries are computed for FoR fields at the two-digit code level and the 4-digit code level.

Table 26: rci_grouping_*. These tables show the (weighted average) performance for each institution in each field of research (by year). Relative scores are provided against the CPP benchmarks and the HPI benchmark. These scores are used to rank institutions by performance and to assign performance classes. Created by: benchmark_rci_groups.js
Requires: rci_papers, core_papers.

Field	**Type**	**Description**
institution	STR	Research Organization Registry ID for the institution
field	STR	ANZSRC field of research code
year	INT	Year of analysis / publication
rci_local	DEC	Weighted average RCI for Australian HEPs
rci_world	DEC	Weighted average RCI for all institutions
hpi_world	DEC	Weighted average HPI for all institutions

Table 27: rci_papers. This table assigns relative citation impact scores to each paper in the dataset. Scores are assigned based on citation count relative to a benchmark for a given field or research and year (of publication). This table can be used to rank and identify highly cited works and compare across fields and time (due to the normalisation

effects of calculating RCIs). Created by: benchmark_rci_papers.js. Requires: core_papers, benchmarks_cpp_*, benchmarks_hpi_*.

Field	Type	Description
doi	STR	Digital Object Identifier for the paper
year	INT	Year of analysis / publication
field	STR	ANZSRC FoR code
weight	INT	FoR apportionment (1-100)
rci_local	DEC	Relative Citation Impact (RCI) against the local benchmark
rci_world	DEC	RCI against the world benchmark
hpi_world	DEC	RCI against the high-performance benchmark

PERFORMANCE RATINGS

A key outcome of ERA analysis is to assign a relative performance rating to each {institution,field} pairing. These ratings are not the same as RCI classes. By our understanding, the assignment of performance ratings in ERA is a qualitative assessment, made by ERA committee members and is not strictly formulaic, although guided by RCI and centile metrics. Nevertheless, this workflow does attempt to formulaically assign a rating to all {institution,field} pairings based on RCI boundaries and is not limited to Australian HEPs.

The ERA methodology for assigning ratings is under revision for ERA 2023 with two new ratings schemes proposed, referred to as Option A and Option B. The COKI workflow assigns ratings that approximate all three methods, described below.

Note that:

- The COKI workflow uses RCI and HPI bands to assign performance ratings. These bands represent our interpretation of ERA boundaries and likely do not match the official ERA method.
- The proposed new ratings are not directly comparable, 1:1, to the ERA 2018 ratings.
- The proposed ratings schemes, for ERA 2023, do not use integer values (possibly to reduce the chance of confusion with RCI classes). However, in the following tables, integer values have been inserted to simplify the approximate comparison of ratings between the three methods.

The method is implemented in ind_ratings.js (available in the project GitHub source code repository) and is parameterised by institution set and field set.

Table 28: ERA 2018 Ratings. Performance ratings are influenced by field-level RCI, against the world benchmark.

Rating	Assessment	~2023 Ratings	RCI Band*
5	Well above world standard	A:5,4 B:6,5,4	>= 1.6
4	Above world standard	A:3 B:3	>= 1.2 to < 1.6
3	At world standard	A:2 B:2	>= 0.8 to < 1.2
2	Below world standard	A:1 B:1	>= 0.4 to < 0.8
1	Well below world standard	A:1 B:1	< 0.4
n/a	Not assessed due to not meeting the low-volume threshold		
n/r	Not rated due to other factors such as data quality concerns		

*RCI bands are our estimations and may not reflect actual ERA methodology.

Table 29: ERA 2023 Option A Ratings. The top three ratings are influenced by the HPI and world benchmarks. For lower ratings, only the world benchmark is used.

	Rating	~2018 Ratings	RCI Band*	HPI Band*
5	World leading	5	>= 1.6	>= 1.2
4	Well above world standard	5	>= 1.6	>= 0.8 to < 1.2
3	Above world standard	4	>= 1.2 to < 1.6	< 0.8
2	World standard	3	>= 0.8 to < 1.2	
1	Not at world standard	2,1	< 0.8	

*RCI & HPI bands are our estimations and may not reflect actual ERA methodology.

Table 30: ERA 2023 Option B Ratings. The top three ratings are influenced by the HPI benchmark. The lower three ratings are influenced by the world benchmark.

	Rating	~2018 Ratings	RCI Band*	HPI Band*
6	AAA	5		>= 1.6
5	AA	5		>= 1.2 to < 1.6
4	A	5		>= 0.8 to < 1.2
3	B	4	>= 1.2	
2	C	3	>= 0.8 to < 1.2	
1	D	2,1	< 0.8	

*RCI & HPI bands are our estimations and may not reflect actual ERA methodology.

Table 31: era_historical_ratings. This table contains a summary of ERA ratings, assigned in prior ERA rounds. The table is sourced directly from the ARC and may be used to assess relative workflow outcomes. Created by: telescope_era_history.js Requires: arc.gov.au

Field	Type	Description
hep_code	STR	short-code for the Australian institution (higher education provider)
hep_name	STR	institution name
for_vers	STR	ANZSRC FoR code version (either 2008 or 2020)
for_code	STR	field of research code
for_name	STR	field of research name
era_2010	STR	ERA rating assigned in 2010 (NA = not assessed)
era_2012	STR	ERA rating assigned in 2012 (NA = not assessed)
era_2015	STR	ERA rating assigned in 2015 (NA = not assessed)
era_2018	STR	ERA rating assigned in 2018 (NA = not assessed)

RATINGS SUMMARY

The ratings summary indicator produces summary statistics for each analysis grouping, to show counts and ranks for each bounded category: centile bands, RCI classes and performance ratings. Tallies encompass all years of the analysis window. This is considered acceptable because RCI values are normalised by comparing to year-specific benchmarks. The method is implemented in report_institution.js (parameterised by ROR) and is available in the project GitHub source code repository. An additional set of tallies and rankings are computed using FoR portions.

Note that the following summary statistics were defined in ERA 2018 but have been removed from ERA 2023. The workflow does not currently compute them but will add them in the future.

- the percentage of papers (from all HEPs) in each 2018 RCI class
- the percentage of papers (from each HEP) of all HEPs' papers in each RCI class
- tally of papers in ERA 2018 RCI classes 0,1 (low) using FoR fractions
- tally of papers in ERA 2018 RCI classes 4,5,6 (high) using FoR fractions
- ratio of 2018 RCI low to high

KNOWN ISSUES

Missing institutional affiliations

Within the aggregated COKI dataset, there is a known lack of linking between authors and institutional affiliations. This can occur, for example, if an automated metadata collector fails to expand an *et al.* authorship listing (thereby failing to identify a link between an author and institution), or if there is insufficient institutional affiliation data to establish an institution's ROR identifier. This results in some data loss as papers with missing ROR data cannot contribute to analysis streams with an institutional focus.

Inheritance of FoR assignment

In the ERA process, the ARC ingests publication metadata from HEPs in which FoR codes have been carefully assigned and apportioned by the institutions' staff. As COKI does not have access to these data, the assignment and apportionment of FoR codes to works is instead inherited from the encompassing journal. FoR assignment, at the journal level, is provided by the ARC as part of the ERA Journal List, however this does not include apportionment information. Consequently, the COKI workflow apportions each of the inherited field codes uniformly. This results in FoR analyses that may be overly generic. Should these data become available in the future, the RIES workflows have been pre-designed to ingest higher-resolution FoR assignment and apportionment data.

Ambiguous ISSNs

Following cleaning of ISSN values (and disambiguation with ISSN-L values), erroneous input data may result in some papers linking to more than one journal. These papers are logged and excluded from analysis pending manual correction, resulting in minor data loss. At this time, there is no attempt to correct these records.

No valid ISSN

The workflow validates ISSN values (for journals and papers) by mapping against an official ISSN to ISSN-L dataset from issn.org. Following sanitation, some journals or papers may have no remaining ISSN-L value assigned. These records are logged and excluded from analysis pending manual correction, resulting in some data loss. At this time, there is no attempt to correct these records.

Author weighting

Like the ERA process, no attempt is made to assign analytical weighting based on authorship (and institutional affiliation). For example, if a paper has nine authors from university A and one author from university B, then the citation metrics will currently be assigned equally between A and B. The ROR ID for institution A will not be counted 9 times or receive a 90% weighting, it will be counted only once and both institutions will receive a 100% weighting.

APPENDIX I - LIST OF TABLES

Raw Tables: created by ETL scripts that import data from external sources.

raw_forcodes%	Original list of field-of-research categories (source: ABS).
raw_heps	Original list of Higher Education Providers (source: ARC).
raw_issns	Original list of ISSN-ISSNL mappings (source: issn.org).
raw_journals%	Original ERA Journal List(s) (source: ARC).
raw_papers	Original list of published research outputs (source: aggregated by COKI).
raw_rors	Original list of Research Organization Registry identifiers (source: ror.org).

Core Tables: created by transforming raw_% tables into an analysis-ready form.

core_fors	Analysis-set of fields of research.
core_heps	Analysis-set of higher education providers.
core_journals	Analysis-set of scientific journals.
core_outputs	Analysis set of unnested units of evaluation (aka research_outputs_world_base)
core_papers	Analysis-set of journal articles, filtered from the COKI DOI dataset.
core_rors	Analysis-set of organisation identifiers.

Benchmark Tables: created through analysis of core_% tables.

benchmarks_centiles%	Calculated centile boundaries to support the ERA centiles indicator.
benchmarks_cpp%	Calculated citation benchmarks to support ERA RCI indicators.
benchmarks_rci%	Calculated RCI boundaries to support ERA RCI category indicators.
benchmarks_summary%	Aggregation of all other benchmarks, grouped by year and field of research.

Indicator Tables: result tables for specific reports / indicators.

centile_tallies%	Data for reporting performance by centile(s).
era_historical_ratings	Historical performance ratings (assigned by ERA), by institution, field of research and ERA round.
ind_ok_volume%	Data to support the ERA low-volume threshold indicator.
interdisc%	Data to support the ERA interdisciplinary / co-publishing indicator.
publishing_profile%	Data to support the ERA journal profile indicator.
ratings_*	Assigned ERA-style ratings emulating ERA 2018 and proposed ERA 2023 methods.
rci_grouping%	Calculated RCI values for groups of outputs (weighted average).
rci_papers	Calculated RCI values for individual outputs.
research_outputs%	Summary statistics for outputs, grouped by all variants of institution, year and field of research.
research_outputs%base	Base table for calculating research output summary statistics.

APPENDIX II - COMPONENTS & DEPENDENCIES

Stage	Source Code	Tables Required	Tables Created
External	telescope_forcodes_2008.js		raw_forcodes_2008
External	telescope_forcodes_2020.js		raw_forcodes_2020
External	telescope_heps.js		raw_heps
External	telescope_issns.js		raw_issns
External	telescope_journals_2018.js		raw_journals_2018
External	telescope_journals_2023.js		raw_journals_2023
External	telescope_rors.js		raw_rors
Core	core_fors.js	raw_forcodes	core_fors
Core	core_heps.js	raw_heps core_rors	core_heps core_heps_auto
Core	core_issnls.js	raw_issns	core_issnls xref_issn_issnl
Core	core_journals.js	raw_journals_2018 raw_journals_2023 xref_issn_issnl core_fors	core_journals xref_for_journal
Core	core_outputs.js	core_papers core_heps	research_outputs_world_base
Core	core_papers.js	coki.doi xref_issn_issnl core_heps core_rors core_journals	core_papers
Core	core_rors.js	raw_rors	core_rors
Benchmarks	benchmark_centiles.js	core_papers	benchmark_centiles_* centiles_tallies_*
Benchmarks	benchmark_cpp.js	core_papers	benchmarks_cpp_*
Benchmarks	benchmark_hpi.js	research_outputs_*	benchmarks_hpi_*
Benchmarks	benchmark_rci.js	rci_papers	benchmarks_rci_*
Benchmarks	benchmark_summary.js	benchmarks_cpp_* benchmarks_hpi_* benchmarks_centiles_* benchmarks_rci_*	benchmarks_summary_*
Indicators	ind_interdisc.js	core_papers core_fors	interdisc_*
Indicators	ind_low_volume.js	research_outputs_*	ind_ok_volume_*
Indicators	ind_publishing_profile.js	research_outputs_world_base	publishing_profile_*
Indicators	ind_ratings.js	core_papers core_heps core_fors benchmarks_hpi_* benchmarks_rci_*	rci_scores_* ratings_*
Indicators	rci_classes.js	core_papers rci_papers rci_grouping_*	rci_classes_papers rci_classes_fields rci_classes_summary
Indicators	rci_groups.js	core_papers rci_papers	rci_grouping_*
Indicators	rci_papers.js	core_papers benchmarks_cpp_* benchmarks_hpi_*	rci_papers

Stage	Source Code	Tables Required	Tables Created
Indicators	single_institution.js	coki.doi core_papers core_journals research_outputs_* benchmarks_summary_*	*_papers *_outputs *_summary_by_field_year *_summary_by_field *_summary_by_year *_paper_classes *_class_tallies_by_field_year *_class_tallies_by_field *_class_tallies_by_year
Indicators	hep_vs_global_centiles.js	benchmarks_centiles_*	centiles_tallies_local_world*

APPENDIX III - HIGHER EDUCATION PROVIDERS

ROR Identifier	Institution
https://ror.org/04cxm4j25	Australian Catholic University
https://ror.org/019wvm592	The Australian National University
https://ror.org/03n0gvg35	Batchelor Institute of Indigenous Tertiary Education
https://ror.org/006jxzx88	Bond University
https://ror.org/023q4bk22	Central Queensland University
https://ror.org/048zcaj52	Charles Darwin University
https://ror.org/00wfvh315	Charles Sturt University
https://ror.org/02n415q13	Curtin University
https://ror.org/02czsnj07	Deakin University
https://ror.org/05jhnwe22	Edith Cowan University
https://ror.org/05qbzwv83	Federation University
https://ror.org/01kpzv902	Flinders University
https://ror.org/02sc3r913	Griffith University
https://ror.org/04gsp2c11	James Cook University
https://ror.org/01rxfrp27	La Trobe University
https://ror.org/01sf06y89	Macquarie University
https://ror.org/02bfwt286	Monash University
https://ror.org/00r4sry34	Murdoch University
https://ror.org/03pnv4752	Queensland University of Technology
https://ror.org/04ttjf776	Royal Melbourne Institute of Technology
https://ror.org/001xkv632	Southern Cross University
https://ror.org/031rekg67	Swinburne University of Technology
https://ror.org/0351xae06	Torrens University Australia
https://ror.org/03r8z3t63	University of New South Wales
https://ror.org/00892tw58	University of Adelaide
https://ror.org/04s1nv328	University of Canberra
https://ror.org/02xn8bh65	University of Divinity
https://ror.org/01ej9dk98	University of Melbourne
https://ror.org/04r659a56	University of New England
https://ror.org/00eae9z71	University of Newcastle
https://ror.org/02stey378	University of Notre Dame Australia
https://ror.org/00rqy9422	University of Queensland
https://ror.org/01p93h210	University of South Australia
https://ror.org/04sjbnx57	University of Southern Queensland
https://ror.org/0384j8v12	University of Sydney
https://ror.org/01nfmeh72	University of Tasmania
https://ror.org/03f0f6041	University of Technology, Sydney
https://ror.org/047272k79	University of Western Australia
https://ror.org/00jtmb277	University of Wollongong
https://ror.org/016gb9e15	University of the Sunshine Coast
https://ror.org/04j757h98	Victoria University
https://ror.org/03t52dk35	Western Sydney University

APPENDIX IV - ANZSRC FOR CODES (2008)

Code	Field of Research (v2008)
01	**Mathematical sciences**
0101	Pure mathematics
0102	Applied mathematics
0103	Numerical and computational mathematics
0104	Statistics
0105	Mathematical physics
0199	Other mathematical sciences
02	**Physical sciences**
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
03	**Chemical sciences**
0301	Analytical chemistry
0302	Inorganic chemistry
0303	Macromolecular and materials chemistry
0304	Medicinal and biomolecular chemistry
0305	Organic chemistry
0306	Physical chemistry (incl. structural)
0307	Theoretical and computational chemistry
0399	Other chemical sciences
04	**Earth sciences**
0401	Atmospheric sciences
0402	Geochemistry
0403	Geology
0404	Geophysics
0405	Oceanography
0406	Physical geography and environmental geoscience
0499	Other earth sciences
05	**Environmental sciences**
0501	Ecological applications
0502	Environmental science and management
0503	Soil sciences
0599	Other environmental sciences
06	**Biological sciences**
0601	Biochemistry and cell biology
0602	Ecology
0603	Evolutionary biology
0604	Genetics
0605	Microbiology
0606	Physiology
0607	Plant biology
0608	Zoology
0699	Other biological sciences
07	**Agricultural and veterinary sciences**
0701	Agriculture, land and farm management
0702	Animal production
0703	Crop and pasture production
0704	Fisheries sciences
0705	Forestry sciences
0706	Horticultural production
0707	Veterinary sciences
0799	Other agricultural and veterinary sciences
08	**Information and computing sciences**
0801	Artificial intelligence and image processing
0802	Computation theory and mathematics
0803	Computer software
0804	Data format
0805	Distributed computing
0806	Information systems
0807	Library and information studies
0899	Other information and computing sciences
09	**Engineering**
0901	Aerospace engineering

Code	Field of Research (v2008)
0902	Automotive engineering
0903	Biomedical engineering
0904	Chemical engineering
0905	Civil engineering
0906	Electrical and electronic engineering
0907	Environmental engineering
0908	Food sciences
0909	Geomatic engineering
0910	Manufacturing engineering
0911	Maritime engineering
0912	Materials engineering
0913	Mechanical engineering
0914	Resources engineering and extractive metallurgy
0915	Interdisciplinary engineering
0999	Other engineering
10	**Technology**
1001	Agricultural biotechnology
1002	Environmental biotechnology
1003	Industrial biotechnology
1004	Medical biotechnology
1005	Communications technologies
1006	Computer hardware
1007	Nanotechnology
1099	Other technology
11	**Medical and health sciences**
1101	Medical biochemistry and metabolomics
1102	Cardiorespiratory medicine and haematology
1103	Clinical sciences
1104	Complementary and alternative medicine
1105	Dentistry
1106	Human movement and sports science
1107	Immunology
1108	Medical microbiology
1109	Neurosciences
1110	Nursing
1111	Nutrition and dietetics
1112	Oncology and carcinogenesis
1113	Ophthalmology and optometry
1114	Paediatrics and reproductive medicine
1115	Pharmacology and pharmaceutical sciences
1116	Medical physiology
1117	Public health and health services
1199	Other medical and health sciences
12	**Built environment and design**
1201	Architecture
1202	Building
1203	Design practice and management
1204	Engineering design
1205	Urban and regional planning
1299	Other built environment and design
13	**Education**
1301	Education systems
1302	Curriculum and pedagogy
1303	Specialist studies in education
1399	Other education
14	**Economics**
1401	Economic theory
1402	Applied economics
1403	Econometrics
1499	Other economics
15	**Commerce, management, tourism and services**
1501	Accounting, auditing and accountability
1502	Banking, finance and investment
1503	Business and management
1504	Commercial services
1505	Marketing
1506	Tourism
1507	Transportation and freight services
1599	Other commerce, management, tourism and services
16	**Studies in human society**

Code	Field of Research (v2008)
1601	Anthropology
1602	Criminology
1603	Demography
1604	Human geography
1605	Policy and administration
1606	Political science
1607	Social work
1608	Sociology
1699	Other studies in human society
17	**Psychology and cognitive sciences**
1701	Psychology
1702	Cognitive sciences
1799	Other psychology and cognitive sciences
18	Law and legal studies
1801	Law
1802	Maori law
1899	Other law and legal studies
19	**Studies in creative arts and writing**
1901	Art theory and criticism
1902	Film, television and digital media
1903	Journalism and professional writing
1904	Performing arts and creative writing
1905	Visual arts and crafts
1999	Other studies in creative arts and writing
20	**Language, communication and culture**
2001	Communication and media studies
2002	Cultural studies
2003	Language studies
2004	Linguistics
2005	Literary studies
2099	Other language, communication and culture
21	**History and archaeology**
2101	Archaeology
2102	Curatorial and related studies
2103	Historical studies
2199	Other history and archaeology
22	**Philosophy and religious studies**
2201	Applied ethics
2202	History and philosophy of specific fields
2203	Philosophy
2204	Religion and religious studies
2299	Other philosophy and religious studies

APPENDIX V - ANZSRC FOR CODES (2020)

Code	Field of Research (v2020)
30	**Agricultural, veterinary and food sciences**
3001	Agricultural biotechnology
3002	Agriculture, land and farm management
3003	Animal production
3004	Crop and pasture production
3005	Fisheries sciences
3006	Food sciences
3007	Forestry sciences
3008	Horticultural production
3009	Veterinary sciences
3099	Other agricultural, veterinary and food sciences
31	**Biological sciences**
3101	Biochemistry and cell biology
3102	Bioinformatics and computational biology
3103	Ecology
3104	Evolutionary biology
3105	Genetics
3106	Industrial biotechnology
3107	Microbiology
3108	Plant biology
3109	Zoology
3199	Other biological sciences
32	**Biomedical and clinical sciences**
3201	Cardiovascular medicine and haematology
3202	Clinical sciences
3203	Dentistry
3204	Immunology
3205	Medical biochemistry and metabolomics
3206	Medical biotechnology
3207	Medical microbiology
3208	Medical physiology
3209	Neurosciences
3210	Nutrition and dietetics
3211	Oncology and carcinogenesis
3212	Ophthalmology and optometry
3213	Paediatrics
3214	Pharmacology and pharmaceutical sciences
3215	Reproductive medicine
3299	Other biomedical and clinical sciences
33	**Built environment and design**
3301	Architecture
3302	Building
3303	Design
3304	Urban and regional planning
3399	Other built environment and design
34	**Chemical sciences**
3401	Analytical chemistry
3402	Inorganic chemistry
3403	Macromolecular and materials chemistry
3404	Medicinal and biomolecular chemistry
3405	Organic chemistry
3406	Physical chemistry
3407	Theoretical and computational chemistry
3499	Other chemical sciences
35	**Commerce, management, tourism and services**
3501	Accounting, auditing and accountability
3502	Banking, finance and investment
3503	Business systems in context
3504	Commercial services
3505	Human resources and industrial relations
3506	Marketing
3507	Strategy, management and organisational behaviour
3508	Tourism
3509	Transportation, logistics and supply chains
3599	Other commerce, management, tourism and services
36	**Creative arts and writing**
3601	Art history, theory and criticism

Code	Field of Research (v2020)
3602	Creative and professional writing
3603	Music
3604	Performing arts
3605	Screen and digital media
3606	Visual arts
3699	Other creative arts and writing
37	**Earth sciences**
3701	Atmospheric sciences
3702	Climate change science
3703	Geochemistry
3704	Geoinformatics
3705	Geology
3706	Geophysics
3707	Hydrology
3708	Oceanography
3709	Physical geography and environmental geoscience
3799	Other earth sciences
38	**Economics**
3801	Applied economics
3802	Econometrics
3803	Economic theory
3899	Other economics
39	**Education**
3901	Curriculum and pedagogy
3902	Education policy, sociology and philosophy
3903	Education systems
3904	Specialist studies in education
3999	Other education
40	**Engineering**
4001	Aerospace engineering
4002	Automotive engineering
4003	Biomedical engineering
4004	Chemical engineering
4005	Civil engineering
4006	Communications engineering
4007	Control engineering, mechatronics and robotics
4008	Electrical engineering
4009	Electronics, sensors and digital hardware
4010	Engineering practice and education
4011	Environmental engineering
4012	Fluid mechanics and thermal engineering
4013	Geomatic engineering
4014	Manufacturing engineering
4015	Maritime engineering
4016	Materials engineering
4017	Mechanical engineering
4018	Nanotechnology
4019	Resources engineering and extractive metallurgy
4099	Other engineering
41	**Environmental sciences**
4101	Climate change impacts and adaptation
4102	Ecological applications
4103	Environmental biotechnology
4104	Environmental management
4105	Pollution and contamination
4106	Soil sciences
4199	Other environmental sciences
42	**Health sciences**
4201	Allied health and rehabilitation science
4202	Epidemiology
4203	Health services and systems
4204	Midwifery
4205	Nursing
4206	Public health
4207	Sports science and exercise
4208	Traditional, complementary and integrative medicine
4299	Other health sciences
43	**History, heritage and archaeology**
4301	Archaeology
4302	Heritage, archive and museum studies

Code	Field of Research (v2020)
4303	Historical studies
4399	Other history, heritage and archaeology
44	**Human society**
4401	Anthropology
4402	Criminology
4403	Demography
4404	Development studies
4405	Gender studies
4406	Human geography
4407	Policy and administration
4408	Political science
4409	Social work
4410	Sociology
4499	Other human society
45	**Indigenous studies**
4501	Aboriginal and torres strait islander culture, language and history
4502	Aboriginal and torres strait islander education
4503	Aboriginal and torres strait islander environmental knowledges and management
4504	Aboriginal and torres strait islander health and wellbeing
4505	Aboriginal and torres strait islander peoples, society and community
4506	Aboriginal and torres strait islander sciences
4507	Te ahurea, reo me te hītori o te māori (māori culture, language and history)
4508	Mātauranga māori (māori education)
4509	Ngā mātauranga taiao o te māori (māori environmental knowledges)
4510	Te hauora me te oranga o te māori (māori health and wellbeing)
4511	Ngā tāngata, te porihanga me ngā hapori o te māori (māori peoples, society and community)
4512	Ngā pūtaiao māori (māori sciences)
4513	Pacific peoples culture, language and history
4514	Pacific peoples education
4515	Pacific peoples environmental knowledges
4516	Pacific peoples health and wellbeing
4517	Pacific peoples sciences
4518	Pacific peoples society and community
4519	Other indigenous data, methodologies and global indigenous studies
4599	Other indigenous studies
46	**Information and computing sciences**
4601	Applied computing
4602	Artificial intelligence
4603	Computer vision and multimedia computation
4604	Cybersecurity and privacy
4605	Data management and data science
4606	Distributed computing and systems software
4607	Graphics, augmented reality and games
4608	Human-centred computing
4609	Information systems
4610	Library and information studies
4611	Machine learning
4612	Software engineering
4613	Theory of computation
4699	Other information and computing sciences
47	**Language, communication and culture**
4701	Communication and media studies
4702	Cultural studies
4703	Language studies
4704	Linguistics
4705	Literary studies
4799	Other language, communication and culture
48	**Law and legal studies**
4801	Commercial law
4802	Environmental and resources law
4803	International and comparative law
4804	Law in context
4805	Legal systems
4806	Private law and civil obligations
4807	Public law
4899	Other law and legal studies
49	**Mathematical sciences**
4901	Applied mathematics
4902	Mathematical physics
4903	Numerical and computational mathematics

Code	Field of Research (v2020)
4904	Pure mathematics
4905	Statistics
4999	Other mathematical sciences
50	**Philosophy and religious studies**
5001	Applied ethics
5002	History and philosophy of specific fields
5003	Philosophy
5004	Religious studies
5005	Theology
5099	Other philosophy and religious studies
51	**Physical sciences**
5101	Astronomical sciences
5102	Atomic, molecular and optical physics
5103	Classical physics
5104	Condensed matter physics
5105	Medical and biological physics
5106	Nuclear and plasma physics
5107	Particle and high energy physics
5108	Quantum physics
5109	Space sciences
5110	Synchrotrons and accelerators
5199	Other physical sciences
52	**Psychology**
5201	Applied and developmental psychology
5202	Biological psychology
5203	Clinical and health psychology
5204	Cognitive and computational psychology
5205	Social and personality psychology
5299	Other psychology

HTTPS://OPENKNOWLEDGE.COMMUNITY
CRICOS Provider Code 00301J