The Data Enhancement and Analysis of the REF2021 Impact Case Studies: Data Delivery Specification

1 Introduction

To support the analysis of the REF2021 Impact Case Study (ICS) data, several data enhancements were made using the text content of the impact Case Studies, contextual data collected during the submission process, and bibliographic data relating to the underpinning research mentioned. This document accompanies a series of data files that contain the output of these enhancement steps to facilitate further analysis of the UK research system and minimise duplication of effort. For a more comprehensive description of the methodology, please refer to the main report.

2 Data

Data files are provided as Comma Separated Values (CSV) using UTF-8 encoding. Each row associates a particular Impact Case Study (identified by *ics_id*) to a set of fields and values, as described below. This data can be linked with the public data made available by UKRI.[1]

2.1 Fields of Research

Based on the items referenced in the underpinning research section (or manual tagging if no research was linked), each ICS is associated with up to three group codes from the Australia and New Zealand Fields of Research Classification (2008 edition; FOR) and maps ICS to a hierarchical structure of research field classifications. This classification scheme was used to provide comparability with the REF2014 impact case studies database.

File: ref2021_ics_for_group_codes.csv

Columns:

- ics_id - the unique case study identifier
- for_group_code - the four-digit FOR group code assigned
- group_label - the label of the corresponding FOR group code

N.B. Be sure the load the for_group_code column as a string to prevent loss of the leading zero in some codes!

2.2 Impact Topics

To provide a granular taxonomy of impact types, topic modelling was performed on Section Four text (details of the impact). Coherence testing was used to pick 79 target topics, which are grouped under 12 main clusters. Each ICS is associated with up to three topics that are designated primary (the most relevant), secondary, and tertiary.

[1] https://results2021.ref.ac.uk/impact

File: ref2021_ics_impact_topics.csv

Columns:

- ics_id - the unique case study identifier
- Cluster - cluster identifier for the primary topic
- Cluster label - corresponding label for cluster
- Primary topic - topic identifier for the primary topic assigned
- Secondary topic - topic identifier for the secondary topic assigned
- Tertiary topic - topic identifier for the tertiary topic assigned
- Primary topic label - corresponding label for primary topic
- Secondary topic label - corresponding label for secondary topic
- Tertiary topic label - corresponding label for tertiary topic

2.3 Geotagging

In order to provide comprehensive coverage of the locations mentioned in case studies, Section four text (details of the impact) was processed using the Open Source Edinburgh Geoparser[2] - a system to automatically recognise place names in text and disambiguate them with respect to a gazetteer. For the purposes of this analysis, the open-source Geonames[3] gazetteer was used as it provides global coverage and contains an extensive list of place names. We also made use of the *limiting geographical area* feature that allows users to provide a rectangular locality box. The geoparser will prefer places in the area specified but will still choose locations outside it if other factors give them higher weighting. For this analysis, a bounding box surrounding the UK was used, helping to properly disambiguate common place names that appear in multiple geographies.

Following the automatic tagging process, a series of manual curation steps were used to ensure high-quality, accurate data. Custom spreadsheets showing the matched tokens (i.e. the words referring to a place name), their context (a text fragment included 10 tokens before and after the match), and basic gazetteer information (e.g. country name, region, population) were created for manual review. In particular, tokens containing location names that are part of a longer proper noun were flagged for review. For example, the text "research conducted at the University of X" was often incorrectly matched to the location X by the geoparser. Other examples of this filtering include project names, strategies, report titles, television station names, charity names, governmental departments, prisons, military facilities, hospitals, and NHS trusts.

The data provided are extracted from Geonames and include regional classifications for European locations according to the Nomenclature of territorial units for statistics[4].

The file uses UTF-8 encoding – please import the data correctly into Excel otherwise some characters (for example apostrophes, letters with accents, symbols) will not be displayed properly.

File: ref2021_ics_geotagging.csv

---

[2] https://www.ltg.ed.ac.uk/software/geoparser/

[3] http://www.geonames.org/

[4] https://ec.europa.eu/eurostat/web/nuts/background

Columns:

- ics_id - the unique case study identifier
- match - the token that was matched to a geographic location
- fragment - the fragment of text surrounding the location mention
- geonamesid - the unique gazetteer id for the location matched
- asciiname - name of geographical point in plain ascii characters (from Geonames)
- latitude - latitude in decimal degrees (wgs84) (from Geonames)
- longitude - longitude in decimal degrees (wgs84) (from Geonames)
- country code - ISO-3166 2-letter country code (from Geonames)
- country label - (from Geonames)
- continent_code - (from Geonames)
- continent label - (from Geonames)
- admin1 - code for the first administrative division (from Geonames)
- admin2 - code for the second administrative division (from Geonames)
- fcode - feature code (from Geonames, see **Error! Bookmark not defined.** for definition)
- fclass - feature class (from Geonames, see **Error! Bookmark not defined.** for definition)
- nuts1_id - level 1 NUTS classification code
- nuts1_name - label for NUTS1 classification
- nuts2_id - level 2 NUTS classification code
- nuts2_name - label for NUTS2 classification
- nuts3_id - level 3 NUTS classification code
- nuts3_name - label for NUTS3 classification

2.4 Impact Reach

Based on Geotagging data, the impact reach of each ICS is classified across four dimensions:

- local - locations mentioned within <15km of the submitting institution.
- regional - locations mentioned in the same NUTS1 region as the submitting institution.
- national - any impact in the UK.
- international - locations mentioned outside the UK.

File: ref2021_ics_impact_reach.csv

Columns:

- ics_id - the unique case study identifier
- Impact Reach Local - True / False
- Impact Reach Regional - True / False
- Impact Reach National - True / False
- Impact Reach International - True / False

2.5 Time periods

Information provided in the ICS template on time periods (start, end) for underpinning reach and impact are provided for each ICS. These values have been extracted automatically by parsing text from the original submission documents.

File: ref2021_ics_time_periods.csv

Columns:

- ics_id - the unique case study identifier
- research_start - year when research started
- research_end - year when research ended
- impact_start - year when impact started
- impact_end - year when impact ended

2.6 Funding Acknowledgements

Using information collected as contextual data[5], funder names mentioned (after disambiguation) have been classified in one of four categories:

- URKI - one of the UKRI funders
- European Union - From the European Commission (e.g. EU, ERC, FP7, H2020)
- Industry - Public and Private companies
- Other - any other funder type

File: ref2021_ics_funder_acknowledgement_types.csv

Columns:

- ics_id - the unique case study identifier
- funding_ack_count - total number of funders acknowledged
- UKRI_funder_count - count of UKRI funders
- EU_funder_count - count of EU funders
- Industry_funder_count - count of Industry funders
- Other_funder_count - count of other funders

2.7 Underpinning research

Based on the text provided in Section 3 (underpinning research) of the case studies, references to research artefacts were identified, either by matching text fragments resembling academic references, or via explicit mentions of Digital Object Identifiers[6] (DOIs). These can be easily matched to your bibliographic database of choice. For convenience, we also provide data from OpenAlex - a free and open catalogue of the world's scholarly papers, researchers, journals, and institutions. The accuracy of the data sourced from OpenAlex has not been verified and may contain errors.

File: ref2021_ics_dois.csv

- ics_id - the unique case study identifier
- doi - the Digital Object Identifier
- cited_by_count - number of citation (OpenAlex)
- publication_year - year of publication (OpenAlex)
- title - title (OpenAlex)
- host_venue_name - where the research was published (OpenAlex)
- host_venue_publisher - name of the host publisher (OpenAlex)

---

[5] See question: What is contextual information ? https://ref.ac.uk/guidance-on-results/impact-case-study-database-faqs/

[6] https://www.doi.org/

- volume - volume of publication, when available (OpenAlex)
- issue - issue of publication, when available (OpenAlex)
- id - The OpenAlex identifier for the corresponding record in OpenAlex

2.8 Interdisciplinarity metric

To provide insights into the relative disciplinarity of research underpinning the case studies, the widely used Rao-Stirling metric was utilised. It defines disciplinarity using three aspects: variety (how many different subjects were referenced), balance (how skewed towards certain subjects the references are), and disparity (how unusual the combination of subjects referenced is). The value produced ranges from 0 (least interdisciplinary) to 1.0 (the most interdisciplinary). Calculations were made using reference data collected from OpenAlex, based on the mapping of referenced journals to FOR groups (4-digit level).

File: ref2021_ics_interdisciplinarity.csv

Columns:

- ics_id - the unique case study identifier
- percentile_rs_idr - panel-normalised percentile for Rao-Stirling metric

The IDR metric is only provided when sufficient bibliographic data (at least one doi) could be linked to the ICS. Unfortunately, there are 305 ICS with no assigned value because they are not linked to bibliographic items.

2.9 Return on Investment

To support various analyses on the potential return on investment (ROI) of UK research, regular expression searches were used to identify two types of indicators:

1. Mentions of the phrase “return on investment” or “financial return”
2. Mentions of currency amounts (e.g. £100 Million)

Searches were performed across all Sections of the ICS.

File: ref2021_ics_roi.csv

Columns:

- ics_id - the unique case study identifier
- roi total – number of mentions of “return on investment” or “financial return”
- currency total – number of mentions of a currency unit







Data enhancement and analysis of the REF 2021 Impact Case Studies
Cagla Stevenson, Jonathan Grant, Martin Szomszor, Cecilia Ang,
Devika Kapoor, Salil Gunashekar, and Susan Guthrie

For more information on this publication, visit www.rand.org/t/RRA2162-1

About RAND Europe
RAND Europe is a not-for-profit research organisation that helps improve policy and decision making through research and analysis. To learn more about RAND Europe, visit www.randeurope.org.

About Electric Data Solutions
Electric Data Solutions provides bespoke analysis to universities, funders and publishers to help them understand their unique contribution to the global research system. To learn more about Electric Data Solutions, visit electricdata.solutions.

Different Angles
Different Angles Ltd is a consultancy that focuses on the social impact of universities and research. To learn more about Different Angles, visit www.differentangles.co.uk.

Research Integrity
Our mission to help improve policy and decision making through research and analysis is enabled through our core values of quality and objectivity and our unwavering commitment to the highest level of integrity and ethical behaviour. To help ensure our research and analysis are rigorous, objective, and nonpartisan, we subject our research publications to a robust and exacting quality-assurance process; avoid both the appearance and reality of financial and other conflicts of interest through staff training, project screening, and a policy of mandatory disclosure; and pursue transparency in our research engagements through our commitment to the open publication of our research findings and recommendations, disclosure of the source of funding of published research, and policies to ensure intellectual independence. For more information, visit www.rand.org/about/principles.

© 2023 Research England, Higher Education Funding Council for Wales, Scottish Funding Council, Department for the Economy Northern Ireland and UK Research and Innovation

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the copyright holders. Web of Science and Clarivate are trademarks of their respective owners and used herein with permission

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif., and Cambridge, UK

RAND® is a registered trademark.