D3.1 RRI and Open Science Datasets

Observing and Negating Matthew Effects in Responsible Research and Innovation Transition

Version 1.0
Public

This deliverable presents ON-MERRIT's RRI and Open Science Datasets. More specifically, it describes two datasets, the Promotion, Review and Tenure Dataset (PRT) and the Research Papers Dataset. It also provides detailed information about the contents of the dataset and the need for its creation.

ON-MERRIT - Grant Agreement 824612

The research leading to these results has received funding from the European Union's Horizon 2020 Research and Innovation Programme, under Grant Agreement no 824612.

Document Description

D3.1 - RRI and Open Science Datasets

D3.1 - RRI and Open Science Datasets			
WP3 - Research Cultures, support and incentives			
Due date	31.05.2020	Actual delivery date:	31.05.2020
Nature of document	Report	Version	1.0
Dissemination level	Public		
Lead Partner for deliverable	The Open University		
Authors	Nancy Pontika, Bikash Gyawali, Petr Knoth, Antonia Correia, Pedro Principe, Thomas Klebel, Helene Brinken, Hannah Metzler, Tony Ross-Hellauer		
Reviewers	Ilaria Fava, Birgit Schmidt (UGOE), Ilire Hasani-Mavriqi (KNOW), Thed van Leeuwen (CWTS - external), Juan Pablo Alperin (SFU - external)		

Revision History

Issue	Item	Comments	Author/Reviewer
V 0.1	Draft version		Nancy Pontika, Bikash Gyawali, Petr Knoth, Antonia Correia, Pedro Principe, Thomas Klebel, Helene Brinken, Hannah Metzler, Tony Ross-Hellauer
V0.2	Revised	Feedback and comment	Revised Nancy Pontika, Bikash Gyawali, Tony Ross-Hellauer, based on reviewer comments from Ilaria Fava, Ilire Hasani-Mavriqi, Thed van Leeuwen, Juan Alperin, Birgit Schmidt
V1.0	Submitted version		Nancy Pontika, Thomas Klebel, Tony Ross-Hellauer

Table of Contents

Tables

Abbreviations

EC – European Commission
MAG - Microsoft Academic Graph
ON-MERRIT - Observing and Negating Matthew Effects in Responsible Research and Innovation
PRT - Promotion, Review and Tenure
RRI - Responsible Research and Innovation
THEWUR - Times of Higher Education World University Rankings
WP – Work Package

Executive summary

The Observing and Negating Matthew Effects in Responsible Research and Innovation (ON-MERRIT) project aims to bring equity and inclusivity to research. ON-MERRIT studies the "Matthew effects" of cumulative advantage on Open Science and Responsible Research and Innovation across research, industry and policy-making, through a mix of sociologic, bibliometric and computational approaches. Where such effects are discovered, ON-MERRIT will make policy recommendations to mitigate or negate these effects.

Work Package (WP) 3 focuses on identifying the research cultures, support and incentives of academics with regards to Responsible Research and Innovation (RRI) and Open Science. It does that by creating datasets on career promotion policies and research papers, by investigating the Matthew effects in Science and exploring the uptake of RRI and Open Science principles concerning policy training. The WP leader is the Open University, and the WP partners are the Know-Center, TU Graz, University of Minho and University of Göttingen.

The work conducted in WP3 is directly fed to WP6, "Synthesis, validation and policy recommendations" and more specifically to its first task, "Task 6.1 RRI and Open Science Incentives and Indicators". Work on T6.1 will take place between months 7 - 21. It will provide the analysis of policies collected in T3.1 as well as making the connection between the policies analysis results and the incentives, traditional or related to RRI and Open Science, to career progression.

The aim of Task 3.1 (Months 1 to 8) is to perform research and collect data to provide a mapping between indicators, including the MoRRI indicators, to answer specific questions relating to RRI and Open Science. More specifically, this task gathers data on promotion, review and tenure (PRT) and produces datasets from research papers, using scholarly resources corpora from CORE and Microsoft Academic Graph (MAG). It then analyses all the collected information to draw conclusions and more specifically show the relationship and impact of RRI and Open Science effects in academia. All the aforementioned activities aim to answer a series of research questions relating to how the assessment criteria affect academic promotion and the adoption of certain RRI and Open Science practices. In particular, the questions investigated in this research activity relate to the motivations underlying academics' research practices and publishing behaviours.

This deliverable presents the process used to create a dataset composed of 1) a collection of promotion review and tenure (PRT) university policies (42) from seven countries (Austria, Brazil, Germany, India, Portugal, United Kingdom and the United States) and 2) a corpus of scholarly research outputs processed from MAG (Sinha et al., 2015), a freely available database of scholarly information about research, and CORE (Knoth and Zdrahal, 2012), the world's largest aggregator of Open Access content which was created by the WP3 leaders. The purpose of the dataset is to enable the quantitative analysis of the links between criteria for academic progression and the productivity of academics, both in terms of traditional metrics as well as in terms of their Open Science practices. The significance of this dataset lies in its potential to answer a range of questions (see Section 1.4.3.2) that are key to our understanding of what motivates academics with regards to their research practices and publishing behaviours, using various indicators, including the MoRRI indicators (Ravn, Nielse and Mejlagaard, 2015). From a policy perspective, the dataset could also be used to analyse changes in institutional PRT policies towards Open Science, which are likely to result in much-needed adoption of Open Science practices across academia. This dataset constitutes, to our knowledge, one of the

first efforts in delivering a large machine-readable dataset enabling quantitative analysis on these aspects, as much work in this area has been carried out only through surveys and qualitative analysis.

The deliverable's structure is as follows:

1. Introduction: sets forth the urgency for conducting this research and makes clear the potential significance of the dataset.
2. Literature Review: offers a collection of existing research output infrastructures, universities' ranking tools and shows the uptake of Open Access about policies.
3. Research Methodology: provides a description of how the two datasets, the PRT and Research Papers, were created. In addition, it explains the challenges of integrating them.
4. Results: presents in detail the contents of the two datasets, such as the number of PRT policies and the countries investigated, and the percentage of content from MAG and CORE.
5. Discussion: considers alternative research questions that these datasets could be used for and provides a limitation for this research.
6. Conclusion: concludes the datasets creation work.

The uniqueness of these datasets lies in the fact that after their public release in a machine-readable form, they can be reused by the scientific community to answer questions relating to academic productivity, RRI and Open Science.

The project partners consider the possibility of further developing the datasets work presented in this deliverable into a paper, which will be submitted for publication in a peer-reviewed scientific journal or conference.

In the final section of this deliverable, there is a working plan for Task 3.3. This task, which starts in Month 7 (this deliverable covers up to Month 8), has a focus on policy creation and training. T3.3 aims to examine the adoption and provides an insight into the level of researchers' familiarity and application of RRI and Open Science in a variety of geographical areas.

1. Task 3.1: Research Data Collection and preparation

1.1 Introduction

Truly effective policies require an accurate assessment of their efficacy. Yet compiling the data to enable such analysis is often highly problematic. Take the case of Open Science, for instance, which is increasingly mainstream policy for institutions, research funders and even nations (Burgelman et al., 2019). However, even in the scholarly publishing and information domain where analytics companies proliferate, there is low availability of large corpora of data and challenges with regards to their data collection (Squazzoni et al., 2020). There is a particular dearth of centralised, machine-readable datasets for Open Science policy information.

This deliverable presents the process used to create a dataset composed of 1) a collection of promotion review and tenure (PRT) university policies (42) from seven countries (Austria, Brazil, Germany, India, Portugal, United Kingdom and the United States) and 2) a corpus of scholarly research outputs processed from Microsoft Academic Graph (MAG), a freely available database of scholarly information about research and CORE (Knoth and Zdrahal, 2012), the world's largest aggregator of open access content. This data is collected and combined to enable the quantitative analysis of the extent to which institutional Open Science policies, as reflected in promotion and tenure criteria, influence researcher behaviour, e.g. changing career paths or progressing in their roles.

By describing the process used to create this dataset, we hence illuminate the difficulties in compiling the data required to judge the efficacy of institutional policies. This work leads to the potential benefits of gathering policy-information in a standardised, machine-readable way.

1.2 Dataset Significance

This deliverable presents a unique digitised dataset available in a machine-readable form to enable quantitative analysis of links between criteria for academic progression (as defined in promotion policies of academic institutions) and the productivity of academics (as measured by their research outputs and Open Science practices). Such data was previously available in an unstructured way and distributed fashion. Here we are curating and processing it into a structured machine-readable dataset to enable quantitative analysis. Although there can be a variety of factors influencing academic progression which were not investigated and are not discussed in this deliverable, e.g. other university internal processes and policies, university strategies and culture, this deliverable focuses only on a possible relationship between academic progression and RRI and Open Science factors. The deliverable succeeds in documenting universities' adoption of Open Science by assessing various indicators relating to their research performance and assessment. This includes number of citations, journal metrics, peer review, publication quality, etc. In addition, some MoRRI indicators (Ravn, Nielse and Mejlagaard, 2015) are also examined, for example, gender balance, citizen science, public engagement, etc. This data is obtained by processing the promotion policies of universities in addition to their research outputs, as obtained from MAG and CORE.

More specifically, this dataset consists of two subsets linked via the institution entity:

1. Promotion, Review and Tenure (PRT) dataset: a CSV file and several pdf documents of PRT policies, which were manually collected either by downloading them from university webpages or by requesting them via email when the policies were not accessible online.
2. Research Papers dataset: a large corpus of scholarly publications of universities extracted from the MAG database and processed for linking it to the corresponding universities' promotion policies. It collectively encodes the data on research papers as well as the career profile of their authors as determined from the MAG. Other enrichments from external data sources such as the information on Open Access status of publications as given by the CORE Discovery service are also included.

Both the PRT and the Research Papers dataset explore the current situation in seven countries: Austria, Brazil, Germany, India, Portugal, UK and the United States. The PRT dataset contains data collected in three languages: English, German and Portuguese, while the Research Papers dataset is in English. We conduct the analysis of the promotion policies in their original language and produce the analysis results in English.

The significance of this dataset lies in the fact that this information combined allows for the analysis of whether certain types of policies are associated with practical effects. For example, what is the level of Open Access output adoption in a university with requirements or incentives for publishing Open Access in PRT policies and how does this compare to another university that does not provide such requirements or incentives. We envision this dataset to enable answering a wide range of questions correlating promotion policies with research outputs in a machine-readable form.

Apart from the contributions mentioned above, this deliverable calls for machine accessibility of PRT policies demonstrating the utility of linking them to other scholarly datasets enabling quantitative analysis of policy instruments and their likely effects. Such data are needed to improve our understanding of what incentivises academics to practice Open Science.

1.3 Existing data-sources

The study compiles information on universities in terms of the promotion policies they adopt and their academic performance as measured by their research outputs. In this context, we survey the existing infrastructures which allow for extracting and processing the information we need to build our dataset.

1.3.1 Existing research output infrastructures

Large corpora have been collected and become available from organisations, projects, commercial and non-commercial services. These include services that:

- discover and deliver Open Access content,
- disciplinary repositories where Open Access content is submitted,
- harvesters that aggregate Open Access content available elsewhere,
- databases that use both Open Access and closed access content, and
- registries that provide essential information about the identification of the research outputs.

More specifically, Table 1 below shows a non-exhaustive overview of the various resources relevant to our research.

Resource name	Resource Type	Description	Free to use? Yes /No[1]	Collection Covered[2]
Core Discovery[3]	Service	CORE Discovery is a tool that finds links to freely accessible research papers.	Yes	> 24,936,921
OA Button[4]	Service	A service that delivers links to an open access version of research articles (complemented by a request mechanism) by looking at thousands of resources.	Yes	This service uses a variety of data sources, which perform a "live" discovery and there is not a number available online
Unpaywall[5]	Service	An open database that harvests thousands of open access contents and delivers links to open access research articles.	Yes	26,009,865 free scholarly articles
Microsoft Academic Graph	Database	A multidisciplinary database consisting of scientific papers, demonstrates connections between papers and citations and offers rich metadata information, such as authors, institutions, journals, conferences and fields of study - free of cost service.	Yes	233 Million Paper Records
OpenAIRE Research Graph[6]	Database	A multidisciplinary database of openly available scientific papers with additional information such as related datasets and software with funders, projects and communities.	Yes	450 million metadata records
Scopus[7]	Database	A multidisciplinary database of scientific papers in various disciplines such as social, life and health sciences.	No	> 34,000 journal titles
Web of Science[8]	Database	Multidisciplinary databases with exhaustive citation data.	No	> 21,000 journal titles, 1.7 billion cited references.

[1] Some of these services may support products where payment is required. For this research, we used the free of cost products only.
[2] The numbers in the collection column are as of April 2020
[3] CORE Discovery https://core.ac.uk/services/discovery/
[4] Open Access Button https://openaccessbutton.org/
[5] Unpaywall: An open database of free scholarly outputs https://unpaywall.org/
[6] OpenAIRE Research Graph https://www.openaire.eu/openaire-research-graph-open-for-comments
[7] Scopus https://www.scopus.com/home.uri
[8] Web of Science https://clarivate.com/webofsciencegroup/solutions/web-of-science/

CrossRef	Registry	A registry of Digital Object Identifiers (DOIs).	Yes	Around 100 million metadata records from more than 4,500 publishers

Table 1 - Research Output Infrastructures

In this work, we mainly use the CORE Discovery service because it hosts the largest collection of outputs, is the most recently released service as compared to alternative discovery tools and performs better both in content coverage and precision (Knoth and Cancellieri, 2019). In addition to that, some of the authors are affiliated with the service, are very familiar with it and are comfortable working with it. We use the 2018 release of the MAG database to extract the publication dataset for the Research Papers Dataset. This limits the scope of our study to publications made until 2018 which is a fair choice considering the duration needed for discovery of open access content from repositories.

Both MAG and CORE Discovery are freely available for use and extracting data from both is straightforward, which makes them an ideal choice for our work.

1.3.2 Existing Universities' Ranking Tools

The performance and quality of universities are often measured by using a mixture of factors. These focus on a variety of teaching, research performance and excellence components, collaborations with third parties, e.g., enterprise and industry, academic reputation, income, international student numbers, subject of field, and many more. Combining these elements yields comprehensive lists which, in turn, provide the universities' ranking (Eccles, 2010). Currently, there is a wide variety of national and international rankings, often supported by governments, newspapers and websites. Table 2 provides some examples:

Resource name	Description	Topics
Times Higher Education World University Rankings[9] (THEWUR)	A global university ranking list, examining institutions in five areas	1. teaching, 2. international outlook, 3. industry income, 4. research and 5. citations
Academic Ranking of World Universities[10] (ARWU)	A global university ranking list, examining institutions in six areas	1. number of alumni, 2. total number of staff winning Nobel Prizes, 3. number of highly cited researchers, 4. number of papers published in Nature and Science, 5. number of papers indexed in Science Citation Index- Expanded and Social Science Citation Index and 6. weighted scores of the above five indicators divided by the number of full-time equivalent academic staff.

[9] THE World University Rankings https://www.timeshighereducation.com/world-university-rankings/2020/world-ranking#!/page/0/length/25/sort_by/rank/sort_order/asc/cols/stats

[10] ARWU World University Rankings http://www.shanghairanking.com/

QS World University Rankings[11]	A global university ranking list, examining institutions in six areas	1. academic peer review, 2. faculty/student ratio, 3. citations per faculty, 4. employer reputation, 5. international student ratio and 6. international staff ratio.
UMultirank[12]	A global university ranking list, examining institutions in five areas	1. teaching and learning, 2. research, 3. knowledge transfer, 4. international orientation and 5. regional engagement.
UniversityRankings.ch[13]	A global university ranking list, examining institutions in two areas	1. academic and 2. research performance.
Round University Ranking[14]	A global university ranking list, examining institutions in four areas	1. teaching, 2. research, 3. international diversity and 4. financial sustainability.
The Carnegie Classification of Institutions in Higher Education[15]	A university ranking tool focusing on U.S. Higher Education Institutions.	N/A
Macleans University Rankings[16]	A university ranking tool focusing on Canada.	N/A

Table 2 - University Ranking Tools

All rankings weigh their indicators according to their own private algorithms to create the final ranking for each university.

Several studies have been conducted about university rankings. Aguillo et al. (2010) compared various university ranking tools and discovered that despite the variety in their algorithms, they make use of similar attributes to compute the rankings. Pusser and Marginson (2016) viewed the power of university rankings from a critical and theoretical point of view and found that these lists have an essential role in university power shaping. Saisana, d'Hombres and Salteli (2011) discovered that, although at a country level the rating conclusions may not be as accurate, the results for larger scale areas are stronger.

The study by Alperin et al. (2018), that relates to the dataset presented in this deliverable, investigated the value of academia's work by looking into USA and Canadian promotion review and tenure (PRT) policies, using the Carnegie Classification of Institutions in Higher Education and the Macleans University ranking to

[11] QS World University Rankings https://www.topuniversities.com/university-rankings/world-university-rankings/2020
[12] UMultirank https://www.umultirank.org/
[13] Universityrankings.ch https://www.universityrankings.ch/en
[14] Round University Ranking https://roundranking.com/ranking/world-university-rankings.html#world-2019
[15] Carnegie Classifications https://carnegieclassifications.iu.edu/
[16] Rankings Archives https://www.macleans.ca/education/unirankings/

measure university scoring. The authors found that the current metrics relate more to "classic and traditional" evaluation components, such as publishing in subscription channels and citation metrics and call for a shift change in the current assessment procedures. The work conducted in this study, uses a different university sample with regards to the countries investigated and combines results from qualitative and quantitative analysis of data collected from a variety of indicators and from a large corpus of open access publications.

1.4 Research Methodology

1.4.1 Representative Countries and Universities Selection

To conduct this research, there is the need to choose a single university ranking tool that would provide seamless access to the results of various countries globally. At the same time, its ranking should be normalised across countries, since the same indicators with the same weight would be applied. For these reasons, we use THEWUR as it includes "global performance tables that judge research-intensive universities across all their core missions: teaching, research, knowledge transfer and international outlook" (World University Rankings 2019: Methodology, 2018). In contrast to other similar tools, for example ARWU or QS world ranking, THEWUR provides a direct indicator of how universities rank with regards to research and citations. Since the scope of this research is to look at PRT policies and connect them to research excellence and assessment, we choose THEWUR and two out of its five categories:

1. **Research:** Collecting data from the annual Academic Reputation Survey, this indicator is 30% of the total THEWUR. That excludes universities with less than 1,000 relevant publications between 2013 and 2017 and universities with 80% or more of their research outputs in a single subject area.
2. **Citations:** This indicator is 30% of the total THEWUR; the purpose of this category is to investigate the research impact of a publication, based on the number of times it is cited.

The following section describes how THEWUR is utilised for compiling our dataset, the reasoning behind the countries' selection and how we choose the universities in each country.

As the amount of national and institutional Open Access[17], Open Data[18] and in general Open Science[19] policies vary per country[20], this work aims to investigate universities from a representative mix of countries from around the world. At the same time, this research is conducted by an international group of researchers, who speak and understand a variety of languages and could consequently collect policies in languages other than English. As a result, this dataset includes university policies written in English, German and Portuguese from seven countries: Austria, Brazil, Germany, India, Portugal, United Kingdom and the United States.

The number of institutions per country varies significantly, i.e. there are countries with a large number of universities, e.g. Brazil, but also countries with much smaller numbers, e.g. Austria. We intend to select the

[17] Open-access (OA) literature is digital, online, free of charge, and free of most copyright and licensing restrictions. What makes it possible is the internet and the consent of the author or copyright-holder. https://legacy.earlham.edu/~peters/fos/brief.htm

[18] Open Data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike. https://opendatahandbook.org/guide/en/what-is-open-data/

[19] Open Science is the movement to make scientific research, data and dissemination accessible to all levels of an inquiring society. https://www.fosteropenscience.eu/taxonomy/term/7

[20] ROARMap http://roarmap.eprints.org/

same number of universities per country listed in THEWUR and investigate how they perform concerning the "Research" and "Citation" categories. We select a total of six universities per country. The total number of universities is limited to six since the manual collection and analysis of policies is time-consuming. As a trade off, we decide that this number would be sufficient to gain an understanding of our research question while making it feasible to conduct the research at the same time.

Our methodology, as outlined below, ensures that the selection of universities is reproducible and that the outliers, i.e. universities performing extremely high or low in research and citations, are not included in the sample. Furthermore, the total number of universities is divided into three categories ("High", "Medium" and "Low"), and the median for each group is calculated. Within each category, we divide the total number of universities by 3. In case the remainder is different from 0, we do the following:

- if residual is 1, the first subcategory (high) will have 1 university more than the other two (medium, low)
- if the residual is 2, then two subcategories (high and medium) will have 1 university more.

When within a subcategory there is an even number of universities, and their median is a decimal, to select the median we round down to the smaller integer value. When a university would rank at the same position for both "Research" and "Citations" and to have a larger sample of policies, we choose the next available university, i.e. the one with the next highest rank in the category. If that university is already included in the dataset or no policy is available for it, we pick the next lower ranking one.

1.4.2. PRT Database

The PRT policies are manually collected using a search engine. Table 2 shows the set of keywords identified and used for the policies identification in the three languages: English, German and Portuguese.

English	German	Portuguese
Policy	Satzung, Richtlinie, Verfahren	política de seleção, procedimento de seleção, procedimento, recrutamento
Review	Qualifikationsprüfung, Review, Beurteilung, Leistungsevaluation, Regelung, Richtlinie, Strategie	revisão
Academic, Researcher, Professor	wissenschaftliche Mitarbeiter, (Junior-) Professor	académico, universitário, investigador, professor
Promotion	Beförderung, Promotion, Berufung	promoção

Table 3 - Search key terms in English, German and Portuguese to retrieve related PRT policies.

In the PRT dataset we include only university-level policies due to difficulties in identifying specific departmental policies in some countries, for example USA and UK. It is also challenging to assign policies on

the universities' websites to their specific departments. To ensure a consistent set of policies, we define the following exclusion rules: first, we do not collect advertisements of job descriptions even though these could include some insightful requirements applicable to the PRT policies. Second, PRT policies are not examined in conjunction with other policies that could relate to PRT, such as university policies about Ethics or Diversity and institutional Open Access policies.

The universities' policies are then matched to a set of indicators. These include a selection of MoRRI, some indicators mentioned in the Alperin at al. (2018) paper and some that the researchers believed to be prominent in the policies' description during the research pilot. As a result, 18 different indicators are collected and examined (more information on the specific indicators can be found in the Results Section). After the identification of both policies and indicators, we proceed with the cross-matching of these two components. To succeed in that, we go through each policy's document, read it and highlight the areas where an indicator appears. We also retain a local copy of that sentence (and the sentence before and after if the content is related and useful) in the related spreadsheet.

1.4.2.1 Challenges in Collecting PRT Policies

The dataset includes academic faculty policies by recognising and taking into consideration the diversity and the academic structures of each country. In the UK, some universities have separate policies for associate research fellows, readers, professors and full professors. In contrast, in the United States, oftentimes separate policies are created for tenured and non-tenured staff. In Austria, policies either refer to "habilitation", i.e. a qualification for teaching that is essential for promotion to professor, or to qualification agreements (tenure track) for associate professors, whereas calls do not include promotions to full professors. The German policies refer to the English term of "tenure track" and specify the evaluation process, including details on the committee and the evaluation criteria. Both in Portugal and Brazil there are sometimes separate policies for tenured and non-tenured academic staff, and all these variations were taken into consideration.

Policies not openly accessible, i.e. which require log-in using institutional specific credentials, such as four policies from universities in Austria, two in Brazil, and one in the UK, are obtained after contacting universities via email requesting for a copy of their policy.

1.4.3 Research Data Papers Dataset

1.4.3.1 Data Source

For this dataset, we primarily make use of two existing data sources; MAG and CORE Discovery. MAG is organized into database tables that provide a variety of information on scholarly publications such as citations, author names, institution names (universities as well as other publishing bodies) and publication years[21]. The second data source we use is the CORE Discovery, a service that finds links to freely accessible copies of research articles from across the web. We use the CORE Discovery to determine the Open Access status of the scholarly publications retrieved from MAG. Given a DOI for a paper as input, CORE Discovery

[21] The complete list of all MAG database tables as well as their schema is available at https://tinyurl.com/v4r5tfv

returns a flag which indicates whether the paper is known to be Open Access or of unknown status[22]. Since we use the 2018 release of the MAG database, our data contains records of publications as recent as 2018 and this comprises any publication type as available on MAG; i.e. conference or journal articles, review papers, book chapters etc. Further, this dataset records data for all universities identified in MAG for the countries in our sample and not just the universities used for the PRT dataset. This happens to facilitate the dataset reuse for analysis of universities with and without promotion policies in comparison to their publication output. With regards to the career profile of academics, we obtain the information on authors profile by consuming the information directly available on MAG (e.g., authors rank, total papers count, total citation counts) as well as by further processing of related information (e.g., determining the seniority of authors based on the number of years since first publication until their last publication). Other potential sources for obtaining career profiles (e.g., LinkedIn, Google Scholar, ORCID) have not been explored in the course of this deliverable and will be studied in the upcoming months.

1.4.3.2 Dataset Creation Methodology

We define seven research questions which could be used to analyse the publications coming from the universities quantitatively. They are:

1. What percentage of papers coming from a university is Open Access?
2. How are papers published by the universities distributed across the three scientific disciplines (i.e. Agriculture, Climatology and Medicine) of our choice as outlined in the DoW?
3. What is the gender distribution in the authorship of papers published by the universities?
4. What is the distribution of seniority, i.e. number of years since first publication until last publication, of staff in the universities?
5. What is the distribution of incoming citations for Open Access vs other papers published by the university? In other words, if University ***A*** publishes ***X*** number of OA papers and ***Y*** number of papers for which their OA status is unknown; what is the count of citations received for ***X*** vs ***Y***.
6. What is the distribution of references made for Open Access articles vs other papers in articles published by the universities, i.e. if University ***A*** publishes ***X*** papers which reference ***M*** papers; how many of those ***M*** papers are OA vs unknown.
7. How does the distribution of references made for Open Access articles vs other papers (question 6) evolve from 2007 to 2017? In other words, if university ***A*** publishes ***X*** papers per year which reference ***M*** papers; what is the median per university per year for the proportion of M_{OA} vs. $M_{unknown}$?

To address these questions, we create a dataset combining information from CORE Discovery, MAG and other external sources as outlined in Table 3 below. The dataset contains records for each of our selected universities and makes use of the various data sources.

Question no.	MAG Tables Used	CORE Discovery Used (Y/N)?	External Data Used and its Purpose	Output Data Schema
1	Papers, Affiliations, PaperAuthorAffiliations	Yes	Natural Earth Dataset[23] -- to map institutions' geographic	PaperID, Univ_name, Country_name, OA_flag

[22] CORE Discovery may not have a 100% open access discovery success, for example in case of a missing DOI; hence we can only say that such papers have an "unknown" status.

[23] Based on the implementation provided in https://github.com/datasets/geo-countries

			coordinates to the countries they belong to.	
2	Papers, Affiliations, PaperAuthorAffiliations, PaperFieldsofStudy, FieldsofStudy	No	Same as for Question No. 1	PaperID, Univ_name, Country_name, FieldofStudy
3	Papers, Affiliations, PaperAuthorAffiliations, Authors	No	Same as for Question No. 1 + Gender API[24] for automatic gender detection of authors.	PaperID, Univ_name, Country_name, Author_Name, Gender
4	Papers, Affiliations, PaperAuthorAffiliations	No	Same as for Question No. 1	PaperID, Univ_name, Country_name, AuthorID, Author_Rank
5	Papers, Affiliations, PaperAuthorAffiliations	Yes	Same as for Question No. 1	PaperID, Univ_name, Country_name, Citation_Count, OA_flag
6	Papers, Affiliations, PaperAuthorAffiliations, PaperReferences	Yes	Same as for Question No. 1	PaperID, Univ_name, Country_name, Count_OA_References, Count_Unknown_References
7	Same as for Question Nb. 6	Yes	Same as for Question No. 1	Table 1: year, quantile, value Table 2: univ_name, year, median_oa_perc

Table 4 - Data Sources Used and Output Data

1.4.3.3 Challenges in Creating the Dataset

There are several challenges involved in creating this dataset. To begin with, processing the MAG database calls for techniques in big data processing and needs to be supported with appropriate hardware in a cluster computing environment. Analysing such data in conjunction with CORE Discovery and other external resources is a multi-step task which requires efficient resource planning and software optimization. There are issues with the universities in THEWUR not matching to the names in MAG, and we perform text normalisation (lowercase, punctuation removal and ASCIIfication) on the universities' names to look for a match. We also take proper care to discard duplicates seen within the collection for the same university. That applies, for example, to the same paper (PaperID) being recorded twice in our dataset with the OA_flag set to true because there were two entries in MAG for that paper, each of them associated to one of the two distinct co-authors from the same university for that paper. On the contrary, a paper could have multiple authors affiliated with different universities. In such cases, we included a single instance of the paper for each of the universities concerned.

[24] https://pypi.org/project/gender-guesser/

1.5 Results

1.5.1 PRT Dataset Details

The PRT dataset consists of data from seven countries, with six universities per country. This information is presented in a CSV file, which contains the following fields:

1. University ID: a unique number that identifies a university, which is created by combining the country code and a serial number.
2. Status: whether the institution is listed in the "Research" or "Citations" THEWUR list.
3. Level: whether the institution belongs to the "high", "medium" or "low" tier.
4. Policy saved file name: the file name the policy is saved.
5. Type of access: whether the policy is available on the internet or behind a username and password.
6. Notes: any additional notes the researchers would like to add.

The specific PRT policies in pdf are included in a single folder named after the "university id" as defined in the CSV file.

The policies indicators' file consists of the selected indicators per each country. Where an indicator is marked with zero, it would not apply and one when it would be applicable. Table 5 below reports the number of policies analysed per country and the quantities of each indicator. (Some universities had different policies for certain evaluated positions, resulting in analysing more than one policy per university.)

Country	Austria (n = 13)	Brazil (n = 4)	Germany (n = 6)	India (n = 6)	Portugal (n = 7)	UK (n = 8)	USA (n = 8)	Total % (n = 50)
Gender Equality	5	0	2	0	0	0	0	14%
Gender Reviewers	4	0	2	0	0	0	0	12%
Gender Balance Reviewers	2	0	3	0	0	0	0	10%
Citizen Science	0	0	1	0	0	0	0	2%
Impact	3	3	3	1	6	7	3	52%
Public Engagement	2	5	3	1	6	6	0	46%
Policy Makers	1	3	1	0	0	5	0	20%
Industry	3	2	4	1	4	7	0	42%
Open Access	0	0	0	0	0	0	0	0%
Data	0	0	0	0	0	0	0	0%
Software	0	6	1	0	0	0	0	14%
Journal Metrics	3	2	1	5	1	2	1	30%
Citations	1	0	0	0	0	2	3	12%
Number of Publications	13	6	5	0	4	2	4	68%
Publication Quality	4	0	3	1	1	8	2	38%

Peer Review	1	6	3	0	3	3	3	38%
Pastoral Work	3	6	4	2	6	7	5	66%
Patents	2	4	6	4	4	0	0	44%

Table 5 - Indicator quantities per country

1.5.2 Research Paper Dataset Details

Table 6 shows the count of total universities in THEWUR and the corresponding matches we found in MAG. This accounts for a total of 379 universities included in our study. The lowest coverage is observed for Germany (70.83%) while others have fairly good coverage, with the UK having the highest (91.0%). In total, the dataset contains information on 126,795 distinct papers published from the universities in Austria, 682,819 from Brazil, 664,165 from Germany, 272,784 from India, 139,983 from Portugal, 1,490,843 from the UK, and 4,844,193 from the USA. The difference in coverage can be linked to the difference in the data sources THEWUR uses (possibly Scopus and others) compared to our choice of MAG.

	Austria	Brazil	Germany	India	Portugal	UK	USA
THE WUR Count	11	46	48	56	13	100	172
MAG Count	9	34	34	46	11	91	154
% Coverage	81.81	73.91	70.83	82.14	84.61	91.0	89.53

Table 6 - Count of Universities found in THEWUR and MAG for each country analysed

1.6 Discussion and Future Work

Promotion policies are an important lever when trying to change researcher behaviour. They make explicit the current norms of the scientific system and determine who will be able to continue their career within academia by rewarding certain practices. Current initiatives like the Hong Kong Principles for Assessing Researchers (Moher et al., 2019) aim at increasing the trustworthiness of research through recognizing practices such as responsible research, transparent reporting, Open Science via promotion policies. Our dataset enables researchers to investigate how the mention of some of these aspects in promotion policies is associated with researcher productivity. Our dataset opens new avenues for future research; we present some potential research questions below:

1. If a university mentions Open Access publications as a requirement in its promotion policy,
    a. How high does it rank in terms of the overall share of Open Access publication it produces?
    b. How often do papers published by that university reference papers which are Open Access?
    c. Do authors affiliated with that university get a higher number of citations, on average?
2. For a university that has a gender equality requirement in its promotion policy, what is the gender distribution in the authorship of papers published by that university?
3. Does the number of years that authors have to remain at a specific grade level before being eligible to apply for a higher-grade level correlate with the number of years they usually stay at a university?
4. How do the representative universities rank in terms of research output performance compared to other universities for which the promotion policy couldn't be found?

Question 1.a, for example, could be answered by analysing the proportion of OA papers (Dataset B) published by universities having Open Access based requirements in their promotion policy (PRT Dataset). Identifying

further questions and interpreting their results lie within the scope of future work which can be carried out with our datasets.

Research achievement should be assessed in two ways: 1. in terms of its impact on the research community and society, and 2. with regards to the originality of ideas and research integrity. Unfortunately, other external factors, e.g. personal characteristics or commonly held opinions of a person's character, could erroneously also be currently applicable. The new concepts, Responsible Research and Innovation[25] (RRI) and Open Science aim to promote the conduction of research, through public participation and make it more inclusive, participatory, understandable, accessible and re-usable. These concepts can bring a change and have an effect only when countries, universities and policy-makers are in the position to absorb and integrate them in the current research assessment policies.

Our dataset could have been much richer in terms of the PRT data if the policies were originally available in a structured machine-readable format, e.g. XML. In that case, the file's fields could carry specific information about core aspects of the policies. For example, current policies often do not specify when the policy was created, applied or amended. Should such information be available, then the higher level of detail would give us more complete data and enable the drawing of more informed conclusions, producing more thorough and compelling results.

Similarly, a structured format could be extended into making connections and identifying the emergence of new fields in the policies relating to RRI and Open Science. When these components are added in the policies, these could then be easily processed by machines, enabling an automatic comparison among them. In that case, conclusions could be easily drawn as to whether a PRT policy from university A promotes RRI and Open Science components as compared to a policy from university B.

A limitation has to be taken into account when analysing our datasets. That relates to the fact that the Open Access availability of certain outputs could be delivered with an embargo period. In this research, we took into consideration only the current status of the research outputs (as of April 2020), without an effort to determine their Open Access availability at the time of their first publication. Because of this, estimates for the propensity of Open Access are likely higher than if they were based on data on Open Access status at the publication date. An additional limitation lies in the diversity of processes relating to career progression per institution, but also per country and how these are connected to a university's strategy and even tradition. A third and final limitation can be found in the use of PRT policies as the sole factor of promotion, which may not be the case in some universities and country cultures.

1.7 Conclusion

In this deliverable, we presented the creation of a dataset which consists of two parts: 1. a collection of promotion, review and tenure policies from university in seven countries; 2. a research papers dataset, that combines the world's largest aggregator of Open Access content, CORE, and the largest database of scholarly publications, MAG. The significance of this dataset is that it allows others to quantitatively study questions related to academic productivity, Open Science practices and incentives. Through the collection of the

[25] Responsible Research and Innovation is: "Involving society in science and innovation 'very upstream' in the processes of R&I to align its outcomes with the values of society. A wide umbrella connecting different aspects of the relationship between R&I and society: public engagement, open access, gender equality, science education, ethics, and governance." (Definition from https://www.rri-tools.eu/about-rri)

sources for this dataset, it became evident that there is a need for machine accessibility of university policies, so that the policies' descriptions can be classified more easily and taken into account when investigating RRI and Open Science uptake based on large scholarly corpora.

2. Task 3.3 Uptake of RRI and Open Science principles in relation to policy and training

2.1 Rationale

In this task we will assess the participation of researchers in Responsible Research and Innovation and Open Science training across geographical areas and RRI and Open Science sub-topics, providing an understanding as to what extent researchers are familiar with these concepts and practices.

This task will build upon the training initiatives and activities carried out by other projects like FOSTERplus, OpenAIRE and FIT4RRI. In parallel, we intend to develop and launch a survey and interviews with the aim to identify which RRI and Open Science principles are supported or driven by institutional policies.

Using available datasets from surveys and desk research based on the Open Science overview in Europe compiled by the OpenAIRE's National Open Access Desks (NOADs), and in collaboration with Task 3.2, this task will analyse, in coordination with other factors, which RRI and Open Science principles are more frequently adopted by academics and which are not.

2.2 Methodology

The task team will develop the work in three fronts:

2.2.1 Assess the participation of researchers in RRI and Open Science training

Task 3.3 team will identify training activities carried out by other projects and institutions, in order to compare contexts, evaluate activities and results. The survey recipients will be identified from partner projects and institutions, e.g. ERC researchers, Eurodoc affiliates, former participants in training events carried out by FOSTER and FOSTER Plus[26], FIT4RRI[27] and OpenAIRE[28], widening the scope to a large number of participants, from all European countries.

With this survey Task 3.3 aims to:

1) understand the general knowledge of the OS and RRI concepts,
2) access the OS and RRI training offer in the subject's respective country,
3) incorporate OS and RRI practices within the daily workflows relates with the university's policies, and
4) understand in what degree the training offer matches the needs of the subjects.

[26] FOSTER Plus (Fostering the practical implementation of Open Science in Horizon 2020 and beyond) was a 2-year EU-funded project, aiming to contribute to a real and lasting shift in the behaviour of European researchers to ensure that Open Science would become the norm. https://www.fosteropenscience.eu/.

[27] FIT4RRI (Fostering Improved Training Tools For Responsible Research & Innovation) is a 3-year H2020 funded project aiming to analyse trends, barriers and drivers in the implementation of RRI and Open Science and to enhance competencies and skills through an improvement of the training offer. https://fit4rri.eu/project/

[28] OpenAIRE is a socio-technical infrastructure for scholarly communication and Open Science. For over ten years it has been supporting Open Science at national levels via its network of experts (National Open Access Desks – NOADs) who support policy development for Open Science within the research realm. https://www.openaire.eu/support

The survey will be conducted via the survey tool Lime Survey, hosted through participant organisations and distributed across the participants via email. Extra care will be devoted to ensuring compliance with European data protection provisions (GDPR).

2.2.2. Identify which RRI and OS principles are supported by institutional policies

In parallel, the task team will identify which RRI and OS principles are supported/driven by institutional policies. In order to do so, we will elaborate a list of the practices and try to match them with the institutional policies, selecting a sample of countries and institutions. The task team will interview selected participants from Germany, UK, Austria, Portugal, and other countries to be determined, based on the researchers' countries covered within the OS training assessment survey and in a way that ensures representation of European Union regions (North, South West, East countries).

For the interviews we will select survey participants taking into account their countries, type of institution, age, gender, to understand the context factors that may influence the impact of OS and RRI training. Respondents will be asked to describe their role in their respective organization and to reflect on concrete examples of incentives their research institutions may have to support OS and RRI practices and policies. We will then analyse which RRI and Open Science principles are more frequently adopted by academics, and which are not, and try to relate this with the training offer and the institutional policies.

2.2.3 Desk research

All across this task we will rely on desk research and studies previously made by project stakeholders, namely the European University Association, the Scholarly Publishing and Academic Resources Coalition, and OpenAIRE, as well as available research on institutional policies and open access national initiatives and bodies.

3. References

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-june Paul Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. *Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion*, 243–246. https://doi.org/10.1145/2740908.2742839

Aguillo Isidro, Bar-Ilan Judit, Levene Mark and Ortega Hose. 2010. Comparing University Rankings. *Scientometrics*, 85:243-256. https://doi.org/10.1007/s11192-010-0190-z

Alperin Juan Pablo, Muñoz Carol, Schimanski Lesley, Frischman Gustavo E., Niles Meredith T. and McKierran Erin. 2018. *Humanities Commons*. https://hcommons.org/deposits/item/hc:21015

Burgelman Jean-Claude, Pascu Corina, Szkuta Katarzyna, Schomberg Rene Von, Karalopoulos Athanasios, Repanas Konstantinos, Schouppe Michel. 2019. Open Science, Open Data, and Open Scholarship: European Policies to make science fit for the twenty-first century. *Front.Big Data,* 43(2). https://doi.org/10.3389/fdata.2019.00043

Eccles Charles. 2010. The use of University Ranking in the United Kingdom. *Higher Education in Europe*, 4:27. https://doi.org/10.1080/0379772022000071904

Knoth Petr and Cancellieri Matteo. 2019. Analysing the performance of open access papers discovery tools. PowerPoint presentation. In *Open Repositories 2019*. https://www.slideshare.net/petrknoth/analysing-the-performance-of-open-access-papers-discovery-tools

Knoth Petr and Zdrahal Zdenek. 2012. CORE: Three access levels to underpin open access. *D-Lib Magazine*, 18.11/12: 1-13. https://doi.org/10.1045/november2012-knoth

Moher, D., Naudet, F., Cristea, I. A., Miedema, F., Ioannidis, J. P. A., & Goodman, S. N. (2018). Assessing scientists for hiring, promotion, and tenure. *PLOS Biology*, 16(3), e2004089. https://doi.org/10.1371/journal.pbio.2004089

Moher David, Bouter Lex, Kleinert Sabine, Glasziou Paul, Sham Mai Har, et al. 2019. The Hong Kong Principles for Assessing Researchers: Fostering research integrity. *Open Science Framework Preprints*. 10.31219/osf.io/m9abx

Pusser Brian and Marginson Simon. 2016. University Rankings in Critical Perspective. *The Journal of Higher Education*, 84:4. https://doi.org/10.1080/00221546.2013.11777301

Ravn Tine, Nielse Mathias W. and Mejlagaard Niels. 2015. *Metrics and Indicators of Responsible Research and Innovation. Progress Report D3.2. Monitoring the Evolution and Benefits of Responsible Research and Innovation (MoRRI)* https://www.rri-tools.eu/documents/10184/47609/MORRI-D3.2/

Saisana Michaela, d' Hombres Béatrice and Salteli Andrea. 2011. Rickety numbers: Volatility of university rankings and policy implications. *Research Policy*, 40:1. https://doi.org/10.1016/j.respol.2010.09.003

Squazzoni Flaminio, Ahrweiler Petra, Barros Tiago, Bianchi Federico, Birukou Aliaksandr et al. 2020. Unlock ways to share data on peer review: Journals, funders and scholars must work together to create an infrastructure to study peer review. *Nature*, 578(7796), 512–514. https://doi.org/10.1038/d41586-020-00500-y

World University Rankings 2019: Methodology. 2018. *Times of Higher Education*, September. https://www.timeshighereducation.com/world-university-rankings/world-university-rankings-2019-methodology