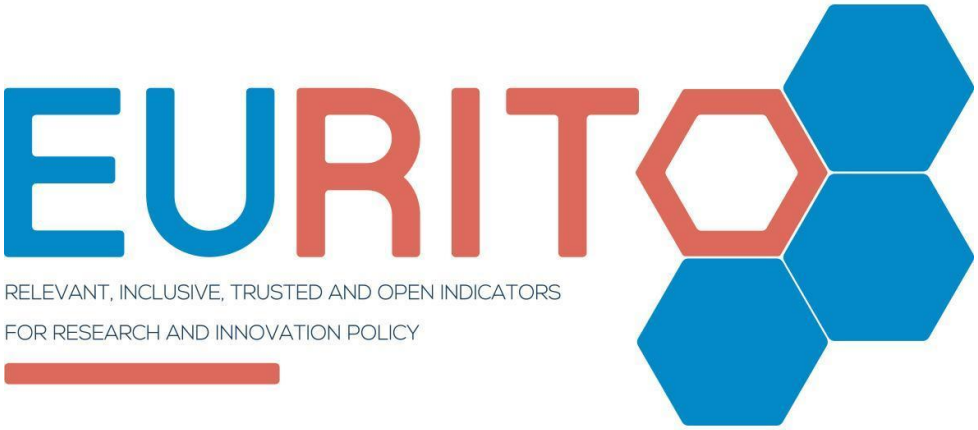

D3.3 R&I Performance Indicators

Deliverable No.		D3.3	
Workpackage No.	3	**Workpackage Title**	Data Collection and Analysis
Lead beneficiary		DTU	
Dissemination level		Public	
Type		Demonstrator	
Due Date		M20 (31 August 2019)	
Version No.		0.3	
Submission Date		30/08/2019	
File Name		D3.3 R&I Performance Indicators 0.1	
Project Duration		36 Months	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 770420.

Version Control

Version	Date	Author	Notes
0.1	2 April 2018	Nesta	Template Creation
0.2	9 August 2019	DTU	Outline creation, detailing scale-up #2
0.3	22 August 2019	Nesta	Detailing scale-ups #1 and #3
0.4	29 August 2019	DTU	Submitted for internal review

Reviewers List

Version	Date	Reviewers	Notes
0.2.1	29 August 2019	Fraunhofer	Overall review
0.2.2	30 August	Nesta	Final review

Disclaimer

This document has been produced with the assistance of the European Union. The contents of this publication are the sole responsibility of the author and can in no way be taken to reflect the views of the European Union.

Executive Summary

Present report outlines the R&I performance indicators developed during the work performed during WP3. Indicators are discussed in connection to the scale-ups they are based on, the use of underlying datasets, quantitative algorithms and policy questions they address.

Table of Contents

1. Introduction

The present report builds upon the previous deliverables in the project and uses a combination of data sources, quantitative methods and data infrastructure to produce R&I indicators that satisfies RITO criteria. Main aim of the report is to systemise and dissect developed indicators to provide their clearer definitions.

In WP2 “Exploration” we have completed eight exploratory data pilots, where we examined a variety of datasets and quantitative methods. After validation with policy makers, further work included combining related elements of these data pilots into four scale-up themes:

1. Emerging Technology and Mapping
2. New Research Funding Analytics
3. Inclusive and Mission-Oriented R&I
4. Predictive Analytics

In D3.1 we have defined requirements to the data infrastructure the purpose of which is to make sure that the developed indicators (and connected data) are easily accessible to users. These requirements will guide the technical implementation of an API, which will serve two purposes: a) later visualisation development in WP5 and b) access to raw and processed data for external authorised users.

In D3.2 we have outlined quantitative methods that are used to produce indicators from datasets. Descriptive and prescriptive analysis methods, such as machine learning, clustering, network analysis were discussed with respect to each scale-up.

The current report, in combination with code, documentation and data produced in D3.4, will serve as a basis for the further work packages: WP4 “Validation” and WP5 “Visualisation”. The API, which will be finalised during D3.4 will act as a main source of data for the outlined indicators and the respective interactive visualisations.

Inspired by Horizon 2020 program report on indicators (available at https://ec.europa.eu/programmes/horizon2020/en/news/horizon-2020-indicators-assessing-results-and-impact-horizon), we provide a summary of indicators in Table 1 and then discuss data, methods and policy relevance in connection to respective scale-ups.

2. Set of R&I Indicators

Table 1 provides an overview of EURITO indicators, their definitions, levels of granularity and the required datasets. Each indicator is then discussed in detail in connection with the respective scale-up.

Table 1. EURITO Indicators

#	Indicator	Definition	Granularity
1	Level of technological activity	number of term occurrences in the dataset. Technology terms are extracted from a search query specified by the user.	per specified technology, per data source (e.g. number of term occurrences in arxiv data), per country or region (e.g. Europe, Asia)
2	Concentration of technological activity	level of activity in a specified technology within a specified location compared to its global share in all technological activities	per specified technology, per data source (e.g. number of term occurrences in arxiv data), per country or region (e.g. Europe, Asia)
3	Structural change	difference in distributions of cluster sizes over time, where *a cluster* is defined as a group of technological terms in the overall body of data that have frequent co-occurrence compared to other clusters	per data source (e.g. number of term occurrences in arxiv data), per country or region (e.g. Europe, Asia)
4	Project centrality	betweenness centrality of a node in the R&I network that represents a research project, where R&I network refers to a network of research projects, organisations and research outputs linked together.	per year, per funding program, per country
5	Organisation centrality	betweenness centrality of a node in the R&I network that represents a research organisation, where R&I network refers to a network of research projects, organisations and research outputs linked together.	per year, per funding program, per country
6	Publication K-factor	percentage increase in cumulative keyword cooccurence counts compared to the keyword	per project, per funding program, per country

		cooccurence count in the overall set of publications in the previous time period,	
7	Level of mission activity	number of mission term occurrences in the dataset. Mission terms are extracted from a search query specified by the user	per specified mission, per data source, per country or region (e.g. Europe, Asia)
8	Mission configuration	distribution of shares for each mission component term in the overall dataset	per specified mission, per data source, per country or region (e.g. Europe, Asia)
9	Gender diversity in the mission field	mission is defined as a percentage distribution of genders within a set of documents related to the mission component terms.	per specified mission, per data source, per country or region (e.g. Europe, Asia)

2.1 Scale-Up 1: Emerging Technology and Mapping

2.1.1 Scale-Up Description

This scale-up combines *Pilot 1: Emerging Technology Ecosystems (Artificial Intelligence) and Pilot 3: Technological Change Indicators* to generate indicators about emerging technology Research and Development (R&D) and its technological innovation system. To achieve this, we will use a collection of unstructured data sources that will be queried using an informational retrieval tool that takes an initial seed of keywords and expands it using measures of semantic similarity estimated with machine learning methods. We envisage that such an approach will allow us to measure levels of technological activity, concentration of technological activities within countries (regions) and the rate of technological structural change.

2.1.2 Indicators

2.1.2.1 Levels of activity, evolution and geography

Having identified R&D outputs (papers, patents, projects) related to an emerging technology, we will measure their levels of activity, their evolution over time and their geographical distribution. In addition to measuring overall levels of activity, we will also estimate revealed comparative advantage indices that take into account a country's specialisation in an emerging technology.

Definition: *Level of activity* of a technology is defined as a number of term occurrences in the dataset. Technology terms are extracted from a search query specified by the user.

Granularity: per specified technology, per data source (e.g. number of term occurrences in arxiv data), per country or region (e.g. Europe, Asia).

2.1.2.2 Sub-national concentration

We will calculate the level of subnational concentration of R&D activity in emerging technologies using indices of concentration such as Herfindahl and Gini indices. We will compare those indicators with concentration for non-emerging R&D activities and the distribution of the population, and also consider if subnational concentration is increasing or decreasing over time.

Definition: *Concentration of an activity* is defined as a level of activity in a specified technology within a specified location compared to its global share in all technological activities.

Granularity: per specified technology, per data source (e.g. number of term occurrences in arxiv data), per country or region (e.g. Europe, Asia).

2.1.2.3 Structural change

We will measure discontinuities in the development of an emerging technology by comparing the topical composition of the field over time. This topical composition will be calculated using topic modelling algorithms trained on textual descriptions of activity in various datasets.

Definition: *Structural change* is defined as a difference in distributions of cluster sizes over time, where a *cluster* is defined as a group of technological terms in the overall body of data that have frequent co-occurrence compared to other clusters.

Granularity: per data source (e.g. number of term occurrences in arxiv data), per country or region (e.g. Europe, Asia).

2.1.3 Datasets

We will use the following datasets:

- arXiv: an open repository of research widely used by scientists, engineers and technology companies to disseminate the findings of their research.
- Microsoft Academic Graph (MAG): A publication database.
- CORDIS: A database with information about EU-funded research projects.
- PATSTAT: A patents database.

2.1.4 Quantitative Methods

We will analyse the data using the following data science methods:

- Natural Language Processing to process textual descriptions of activity in various scientific and technological databases. More specifically, we use the word2vec algorithm to expand lists of seed keywords that can be used to

identify activities related to emerging technologies in the data. Word2vec uses a shallow neural network to learn a vector representation of words based on their ability to predict other words in their contexts. This vector representation captures the meaning of words in a condensed form, and can be used to measure similarity between them using geometrical definitions of distance.

- Topic modelling to identify the topical composition of research corpora. There are various topic modelling algorithms that generally identify clusters of related words or n-grams ('topics') based on their co-occurrence in documents, and then estimate the relative importance of topics in documents based on the presence of those words. We will use 'grid search' over combinations of parameters in order to identify the model configurations that generate more consistent topics.
- Supervised machine learning trained on labelled corpora of text to enrich our data with other policy-relevant information such as the industries that research or technology relate to, or the Sustainable Development Goals it is linked with.

2.1.5 How do these indicators answer the policy questions?

These indicators will provide a detailed and timely view of levels of activity related to emerging technologies in the EU - an area of significant policy interest. They will also help to understand spatial disparities in the development of emerging technologies, potentially informing policies to minimise the risks that the benefits of emerging technologies are captured by a small number of regions. Our analysis of discontinuities in R&D activity will help us to identify 'breaks' in the evolution of a technology area that might have been brought about by policy interventions, or demand changes in policy approaches.

2.2 Scale-Up 2: New Research Funding Analytics

2.2.1 Scale-Up Description

This Scale-up combines Pilot 6: "Advanced Research & Innovation Funding Analytics", Pilot 8: "Linkages and Knowledge Exchange Indicators" and Pilot 4 "Standards As Innovation Diffusion Indicators" to develop novel indicators of research diversity, novelty and impactfulness. Complex network analysis techniques are applied to model the linkages between research grants, organisations, researchers, and publications and to analyse correlations between allocated funding and collaborations and knowledge exchange in Europe. In particular, research projects of Horizon 2020 program is analysed.

2.2.2 Indicators:

2.2.2.1 Project centrality

Project centrality shows how central is a project in the R&I ecosystem.

Definition: *Project centrality* is the betweenness centrality of a node in the R&I network that represents a research project, where R&I network refers to a network of research projects, organisations and research outputs linked together.

2.2.2.2 Organisation centrality

Organisation centrality shows how central is an organisation in the R&I ecosystem.

Definition: *Organisation centrality* is the betweenness centrality of a node in the R&I network that represents a research organisation, where R&I network refers to a network of research projects, organisations and research outputs linked together.

2.2.2.3 K-Factor

The K-Factor is a measure of the novelty of a research project. It represents the degree to which a project contains atypical combinations of topics. High K projects are dominated by topics that have previously appeared together only a few times or not at all.

Definition: *K-Factor* of a publication is a percentage increase in cumulative keyword cooccurence counts compared to the keyword cooccurence count in the overall set of publications in the previous time period,

Keyword cooccurrence count here refers to the number of times when two keywords appear in the same publication.

2.2.3 Datasets

CORDIS dataset includes data about research projects and related organisations funded under the EC Horizon 2020 programme from 2014 to 2020. Each project in the dataset has the following relevant fields: project ID, project start and end dates, amount of funding received, coordinator and participant organisations with respective countries. The dataset is updated monthly.

Dataset URL: https://data.europa.eu/euodp/en/data/dataset/cordisH2020projects

OpenAIRE is a curated dataset of research output produced using public funding in Europe. Adhering to the principles of Open Science, OpenAIRE contains data about research projects, related publications, produced datasets, software and other research products.

Dataset URL: https://develop.openaire.eu/api.html

2.2.4 Quantitative Methods

Complex Network Analysis methods are used to determine network centrality for projects and organisations. First, research projects, participating organisations with researchers linked to these projects and organisations were retrieved from CORDIS dataset. Then, we linked this list of projects to research outputs retrieved from the OpenAIRE dataset.

After, a list of projects, publications, organisations and researchers was converted into nodelist and edgelist structure, suitable for the further construction of a network. The resulted dataset was converted to network representation.

Once the network was constructed, for each node in the network we have calculated network centrality measures, such as betweenness and eigenvector centrality (for definitions of the measures please refer to D3.2). Moreover, for research project nodes, these measures were compared to the amount of funding these projects have received, which was calculated using a CFR score (please see D3.2 for more information).

2.2.5 How do these indicators answer the policy questions?

The main insight illustrated by these indicators is the impact of a particular research funding program on the development and support of scientific collaboration networks in Europe. At a more granular level, indicators allow to quantify the role of each project or organisation in the R&I ecosystem and respective research outputs. Using the developed indicators, policy makers can analyse parts of networks with low network connectivity of projects, especially with respect to the received funding. Similarly, projects and organisations with higher connectivity indicate a better knowledge flow.

As the indicators are defined at several levels of granularity, various comparisons and analysis are enabled. One possibility is to study the collaboration not only on the country level, but on the inter-organisational level. Moreover, the dynamics of these linkages could be studied to understand how certain policy measures have affected the R&I ecosystem and respective outputs.

2.3 Scale-Up 3: Inclusive and Mission-Oriented R&I

2.3.1 Scale-Up Description

This scale-up will further develop analyses conducted in Pilot 7: “Inclusive Innovation” in combination with Pilot 5: “Evidence Base For Mission-Oriented Research & Innovation” in order to generate indicators about the levels of activity and configuration of mission fields and the gender diversity of the researchers participating in them compared to baselines. This will provide policy- and decision-makers intending to implement mission-oriented policies with indicators to support the design of more bold and directed initiatives.

2.3.2 Indicators:

2.3.2.1 Level of activity, evolution and geography

Similarly to the approach that we outlined in Scale-up 1 (‘emerging technology’), we will use an expanded keyword search strategy to identify papers, projects and technologies in an active ‘mission field’ (the specific missions we focus are yet to be determined). By this, we mean activities that combine the key components of the active mission field, such as a technology and a societal challenge (e.g. AI and chronic diseases). The resulting indicators of activity and its evolution will provide a baseline about the state of the mission field that can be used to monitor and evaluate the impact of mission-oriented policies aimed at catalysing those fields. We will also consider the

geographical distribution of the active mission field and its components, hopefully helping to identify opportunities for collaboration between nations/regions with different capabilities and challenges.

Definition: *Level of activity* of a mission is defined as a number of term occurrences in the dataset. Mission terms are extracted from a search query specified by the user.

Granularity: per specified mission, per data source, per country or region (e.g. Europe, Asia).

2.3.2.2 Configuration of the mission field

We will compare the levels of interdisciplinarity in the active mission field with those of its constituent components in order to establish the extent to which it involves crossover between fields, and estimate measures of novelty for the mission field such as those developed in Scale up #2. This will help us to determine whether the active mission field is indeed nurturing novel innovation activities by comparison to the intra-disciplinary status quo. In addition to this, we will also calculate measures of diversity (e.g. entropy) in the mission field based on the evolution of its topical composition. This will help us to measure if the mission field is nurturing exploration or converging on a single solution for its 'challenge'.

Definition: *Mission configuration* is defined as a distribution of percentage shares for each mission component term in the overall dataset.

Granularity: per specified mission, per data source, per country or region (e.g. Europe, Asia).

2.3.2.3 Gender diversity in the mission field

Finally, we will use the names and surnames of individuals participating in the mission field in order to estimate their gender and calculate indicators of gender diversity and its evolution. This will help us explore whether the disciplinary diversity that mission-oriented policies seek to spawn are reflected in socio-demographic and cognitive diversity of the researchers involved.

Definition: *Gender diversity* of a mission is defined as a percentage distribution of genders within a set of documents related to the mission component terms.

Granularity: per specified mission, per data source, per country or region (e.g. Europe, Asia)

2.3.3 Datasets

The main dataset we will use in our analysis is CORDIS (see above for its description).

2.3.4 Quantitative Methods

Most of the methods that we will use have already been outlined above. These include word2vec-enabled keyword expansion to identify relevant documents in our data, supervised machine learning to predict the disciplinary mix of papers, network-based indicators of novelty to measure whether projects in the mission field involve novel or existing combinations of terms, and topic modelling to measure the topical composition of the mission field and its evolution over time.

2.3.5 How do these indicators answer the policy questions?

It is hoped that mission-oriented innovation policies will bring together various fields and research communities in order to deploy science and technology to tackle big societal challenges. This scale-up will contribute to the evidence base helping to inform, target and evaluate those policies by establishing what are the current levels of activity and their evolution, the extent to which mission-related activities currently underway are truly novel, and the gender and topical diversity of the field (two important determinants of its ability to generate inclusive outcomes).

2.4 Scale-Up 4: Predictive Analysis

This scale-up builds upon the indicators developed in Scale-Ups 1-3 (summarised in Table 1) and enables predictive analysis of these indicators. The scale-up focuses on the combination and juxtaposition of the developed novel indicators and traditional ones (e.g. business R&D expenditure). This will serve as a validation for the developed novel indicators.

Continuing further the work performed in Pilot 2: Nowcasting Business Research & Development, this scale-up leverages on the developments in the field of machine learning to address the problems of laggy and inconsistent data behind the indicators.

3. Demonstrator

A brief demonstration of the way how some of the indicators above will be developed are described in the article below: https://www.nesta.org.uk/blog/bringing-arxiv-data-life/. In particular, two web-applications are being developed at Nesta:

1. **arXlive** is a tool for monitoring of innovation activity within arXiv publication data in the realtime
2. **Rhodonite** enables to identify novel industries or technologies for the provided publication dataset

4. Conclusion

In this report, we outline the performance indicators in R&I that will are developed as a result of the EURITO project. This report serves as a basis for the next deliverable D3.4, which will consist of API code, input and output data and the related documentation.

While in this report we have defined the indicators, in D3.4 we divide the API development into robust and exploratory indicators, where exploratory indicators may require further refining.