RELEVANT, INCLUSIVE, TRUSTED AND OPEN INDICATORS
FOR RESEARCH AND INNOVATION POLICY

D1.2 Audit of new R&I data

Deliverable No.		D1.2	
Workpackage No.	1	**Workpackage Title**	**Scoping**
Lead beneficiary		Nesta	
Dissemination level		Public	
Type		Report	
Due Date		M5 (31 May 2018)	
Version No.		0.3	
Submission Date		1 June 2018	
File Name		D1.2 Audit of new R&I data Final	
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	5 May 2018	Nesta	Template Creation
0.2	30 May 2018	Nesta (Chantale Tippet) / Fraunhofer (Knut Blind) / DTU (Pedro Parraguez Ruiz)	First Draft Report as a collaborative effort of authors. Input also provided by Cotec.
0.3	1 June 2018	Nesta	Final Report

Reviewers List

Version	Date	Reviewers	Notes
0.2.1	31 May 2018	Nesta (Juan Mateos-Garcia)	Final Comments

Disclaimer

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

Executive Summary

This brief report describes the data auditing framework developed for the EURITO project. The audit framework consists of four interrelated components:

1. Conceptual anchoring phase;
2. Basic audit of data sources allowing for an assessment of minimum viability to explore the concept of interest;
3. Identification of pilot ideas, and
4. Advanced audit framework to be carried out following the selection of pilots.

This four-phase approach aims to provide enough structure to guide the search for new data, while also allowing for adequate flexibility to explore emergent concepts and ideas. The division of the main auditing steps into 'basic' and 'advanced' phases permits a more efficient use of time and resources by ensuring that only those data sources which are linked to a pilot idea are explored in more depth.

The auditing approach outlined here will serve as the basis for the next phases of the EURITO project, underpinning both the conceptual and analytical progression toward the development of indicators that are relevant, inclusive, trusted, timely, and open.

Table of Contents

List of Figures

List of Tables

1. Introduction

The present deliverable aims to carry out an audit of new data sources for R&I Policy, and in doing so, proposes a process that allows for the characterisation of new data sources that may be drawn upon in later stages of the project. According to the Grant Agreement 770420 (GA), the report is to be presented for feedback and validation at the New Data for R&I Workshop, which was originally planned for June but moved to September 2018 to coincide with a workshop organised by the Horizon 2020 project Data4Impact to maximise reach and impact. A revised timeline related to this deliverable, though not affecting the technical description of the project, is included in Section 4 of the present document.

It is emphasised that the data audit structure proposed here is a continuous framework and should be considered as a living document.

2. Data audit framework

As per the approach outlined in the GA, the data audit structure is inspired by the R&I data source and indicator audit conducted by Science-Metrix for the project entitled Data Mining for Research & Innovation Policy (Campbell et al, 2017).

The former data audit focused primarily on traditional data sources, however. To build on this structure for the purpose of identifying new data sources, the team drew on the EURITO literature review (Deliverable 1.1, submitted 30 April 2018) and participant suggestions during the Policy maker Workshop (17 April 2018, Brussels). These were supplemented by internet searches. The structure of the framework that arose from this process is presented in Figure 1, with each of the four components described in turn in the sections that follow (a link to the current excel workbook is included here[1]).

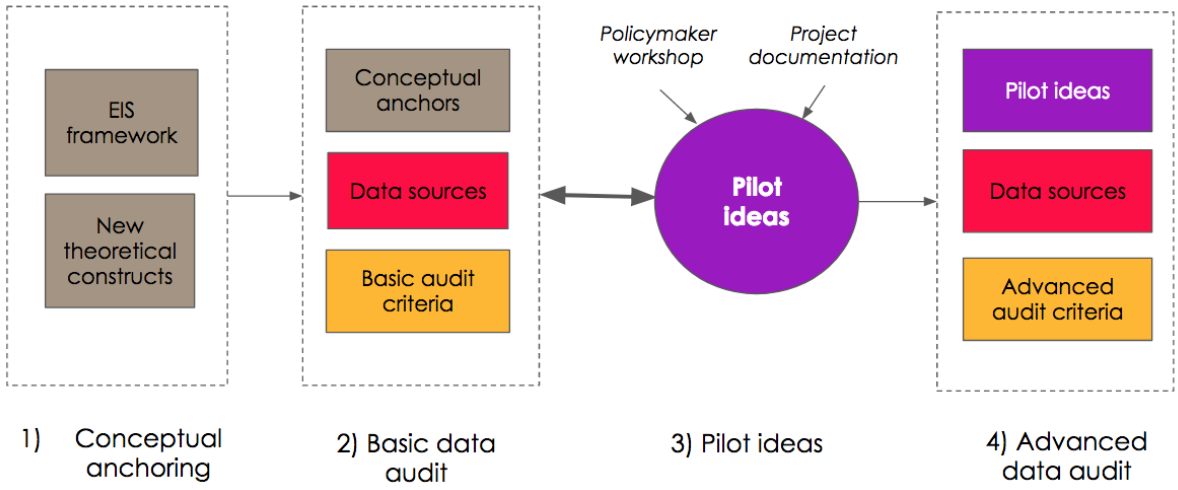

Figure 1: EURITO data audit framework

[1] The address to the excel workbook is: https://docs.google.com/spreadsheets/d/1yOqscIJE8_JDfluU-GEm2gi-_vjt9w53OBteEuNAZbk/edit#gid=695881273

2.1. Conceptual anchoring of data audits

The first component of the data audit framework is what has been termed 'conceptual anchoring'. This component aims to link broad conceptual dimensions of the R&I ecosystem to more precise constructs represented by 'traditional' indicators and new theoretical constructs which may be of interest the R&I policymakers but have not yet been embodied in an indicator (e.g. open innovation). This approach was adopted after the project team determined that bounding the search for new data within higher-level conceptual boundaries would greatly facilitate and guide the search process. For instance, a search for data to capture activity in 'open innovation' permits a focus on a specific subset of platforms and datasets.

Following discussions with consortium members regarding the available options for conceptual anchoring, the decision was taken to anchor the EURITO data audit to the methodological framework of the 2017 European Innovation Scoreboard (Hollanders and Es-Sadki, 2017). Other methodological frameworks, such as the one used in the Oslo Manual, could have been used as a basis for anchoring. The EIS was selected as the framework given that it was reviewed and updated recently, while the Oslo Manual (OECD 2005) has not been updated since 2005.

This conceptual framing allowed the team to anchor some of the new theoretical constructs around the dimensions of the EIS framework (e.g. human resources, attractive research systems, etc.). Concepts that did not fit into a dimension were either discarded for lack of relevance (e.g. 'approaches' such as the sbvIMPROVER quality control measurement for industrial R&D, or company-specific crowdsourcing platforms such as DELL's IdeaStorm) or merged into a dimension of an existing group (e.g. under impacts, the addition of dimensions of environmental and societal), or a new group/dimension combination (e.g. a group entitled cross-cutting, with a dimension entitled media-based and culture-based public).

Table 1: EURITO data audit conceptual anchoring structure

Column name	Description
Group	High-level category, such as Framework Conditions or Investments.
Dimension	Subcategory of the Group. For example, Human Resources is a subcategory of Framework Conditions.
Traditional indicators	Indicators from the European Innovation Scoreboard that fall under each subcategory. For example, new doctorate graduates fall under the Human Resources subcategory of Framework Conditions.
New theoretical constructs and data sources	'Theoretical construct' is used here to describe concepts that may be of interest for R&I policy but may not be easily observable or quantifiable at present. Examples of data sources are also included here for illustrative purposes.

2.1.2. Contested categorisations

Not all of the new theoretical constructs and data sources fell neatly into one of the categories. For example, standards can be considered part of the Framework Conditions (e.g. regulations in Europe are often specified by European standards), but they could also fall under the Firm Investments category. In some cases, therefore, new theoretical constructs and their data sources are listed under more than one sub-category in the framework.

2.2. Basic audit

The Basic Audit provides a link between the boundaries of the conceptual framing and the flexibility to explore a wide variety of data sources. In a sense, it serves as a litmus test to determine whether a given theoretical construct of interest can be linked to data. The Basic Audit can be triggered either from the conceptual anchoring phase that precedes it (see section 2.1), or from the Pilot Ideas phase (see section 2.3).

The categories of the Basic Audit are presented in Table 2 below.

Table 2: Basic Audit criteria

Column name	Description	Example
Group	See Table 1 above	See Table 1 above
Dimension	See Table 1 above	See Table 1 above
Theoretical construct	See Table 1 above	See Table 1 above
Data source	The name of the data source	Mendeley
Short description	Short description of the data source	Mendeley is a desktop and web program produced by Elsevier for managing and sharing research papers, discovering research data and collaborating online.
URL	Link to the source	https://www.mendeley.com/

2.3. Pilot ideas

The third component of the Data Audit Framework is the Pilot Ideas, which serve as the linking point between the basic and advanced data audits. The structure of the Pilot Ideas framework component is described in
Table 3 below.

One could theoretically begin the auditing process at the Pilot Ideas stage, and then work backwards to perform a basic audit to determine whether there are any potentially viable data sources before moving into a more advanced audit.

Alternatively, it would be possible to develop pilot ideas based on the Basic Audit phase.

The pilot ideas captured in the accompanying Excel workbook were collected during the EURITO Policymaker workshop conducted in May 2018, as well as from the project's original vision outlined in the GA. Note that in the accompanying Excel workbook and related Appendix C, the level of granularity of pilot ideas varies. Some are highly developed and suggest potential data sources and analytics, whilst others are high-level descriptions of ideas. This flexibility permits open thinking about pilots that might be carried out, with the potential to continue developing basic ideas over time.

Table 3: Pilot idea structure

Column name	Description	Example
Pilot idea*	A short paragraph describing the aim of a potential problem, and/or the type of issue(s) it seeks to address	Nowcasting business R&D investments
Group	See Table 1 above	See Table 1 above
Dimension	See Table 1 above	See Table 1 above
Source	The provenance of the pilot idea	Grant agreement or workshop
Comments	Any other comments regarding the pilot idea	Pilot idea also addressed at project kickoff meeting

2.4 Advanced audit

The Advanced Audit phase is triggered when a satisfactory combination of Pilot Ideas and Basic Audits occurs. The advanced data audit aims to assess aspects of the data that require deeper exploration and reflection, such as legal access considerations and detailed geographic coverage.

The advanced data audit stage has the potential to be time and resource-intensive, so it is prudent to engage in this phase only if a pilot is considered to be viable (or of high interest) from a policy or otherwise relevant perspective.

The categories of the Advanced Audit are outlined in Table 4 below. Note that at the time of submitting this deliverable, the final decision on which pilots to carry out has not yet been made. As a result, this section of the accompanying Excel spreadsheet has not been completed.

Table 4: Advanced data audit criteria

Column name	Description	Example
Data source	The name of the data source	Mendeley
Pilot idea	Link to pilot idea	Nowcasting business R&D investments
Access (technical)	Specification of how the data can be accessed	Download (e.g. CSV), through an API, scraping, or other.
Access (Legal)	Is the data source legally permitted to be used?	For example, do the terms of service allow or ban certain types of use of the data?
Relevant source links	URLs to pages that may be relevant to accessing the data	Data download page, API documentation
Found or intentional data	Have the data been intentionally collected/ produced for research purposes, or is it "found", i.e. produced for a purpose other than research.	Twitter would be an example of "found" data, as it was not built explicitly to be a data source for research.
Links to existing indicators	Does this data source improve on, or link, to existing indicators?	For example, data from Google Scholar would be related to traditional bibliometric indicators (e.g. Web of Science or Scopus)
Coverage (time)	Time frame covered by the dataset.	The platform launched in June of 2010. Further details around time coverage may also be relevant here, such as how often updates occur or whether historical data are available.
Coverage (geo)	What geographic coverage do the data provide? At what level of resolution?	Covers countries A, B, and C with data available at the level of cities.
Coverage (population)	Considers whether there is likely to be a coverage bias arising from the data source. For example, who is active on the platform in question?	A given social media platform tends to be used primarily by affluent women.
Timestamped records	Whether the metadata contain a timestamp for each record	Yes, all records contain timestamps
Privacy or ethical considerations	Consideration as to whether the data source is likely to raise privacy or ethical considerations.	For example, if individuals are identifiable, consideration would need to be taken to determine appropriate preprocessing,

		anonymisation, etc.
Possible limitations	Possible limitations of the data source	Potential biases or other limitations that the data source may contain (these may have been identified in the other categories)
Date of advanced audit	Date the audit was performed, as changes to some of the above may occur.	20 May 2018
Level of validation	Has the data source undergone any validation, either internal or external	For example, the idea of measuring activity on crowdfunding platforms was presented at a workshop.
Other observations	Any other observations	Open-ended

3. Conclusions and lessons learned

An important lesson learned during the development of the framework is that having clearly defined conceptual boundaries is essential for a productive and fruitful search for new data sources. While there can be some value in undertaking a broader exploration of the data landscape, this approach quickly becomes unproductive if one does not revert to an anchoring point determined either by a theoretical construct or a pilot idea. This lesson—as well as the need to allow ideas to enter the framework in various stages of completeness—is reflected in the framework's four-stage structure.

4. Next steps

As the decisions regarding which pilots to pursue are taken, the manual approach adopted in the search for data sources of interest may be augmented with a data-driven approach using twitter mining or traditional web crawlers.

Mentioned in Section 1 Introduction, although the platform to present the data audit has been delayed to September 2018 (rather than the planned June), EURITO will reach out to stakeholders with the basic audit and pilot ideas with the view to populate the advanced data audit criteria with data sources through feedback and validation in a bilateral manner. Progress on pilot development, as well as on data infrastructure and processes will be used as detailed discussion points for refinement at the September workshop.

References

Bakhshi, H., Davies, J. and Mateos-Garcia, J. 2015. The Net Effect: Using social media data to understand the impact of a conference on social networks. Nesta: London. Available from https://www.nesta.org.uk/report/the-net-effect-using-social-media-data-to-understand-the-impact-of-a-conference-on-social-networks/

Blind, K. 2016. Standardization and Standards as Research and Innovation Indicators: Current opportunities and future challenges. OECD BlueSky Ghent 2016. Available from http://www.oecd.org/sti/049%20-%20BlueSky_Standards_Blind.pdf

Blind, K., Petersen, S. and Riillo, C. 2017. The Impact of Standards and Regulation on Innovation in Uncertain Markets. Research Policy, 46(1), pp. 249–264.

Campbell, D., Tippett, C., Struck, DB., Lefebvre, C., Côté, G. and Archambault, É. 2017. 'Data Mining on Key Innovation Policy Issues for the Private Sector: Technical Report'. Prepared by Science-Metrix for the European Commission.

Hollanders, H. and Es-Sadki, N. 2017. European Innovation Scoreboard 2017 - Methodology Report. Available at: http://ec.europa.eu/DocsRoom/documents/25101

Lima, A., Rossi, L. and Musolesi, M. 2014. Coding together at scale: Github as a collaborative social network. arXiv preprint arXiv:1407.2535.

Mateos-Garcia, J., and Gardiner, J. 2016. From detecting to engaging: An analysis of emerging tech topics using Meetup data. Available from https://www.nesta.org.uk/blog/from-detecting-to-engaging-an-analysis-of-emerging-tech-topics-using-meetup-data/

OECD/Eurostat. 2005. *Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data, 3rd Edition*, The Measurement of Scientific and Technological Activities, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264013100-en.

Rammer, C., Kinne, J. and Blind, K., 2016. Microgeography of innovation in the city: Location patterns of innovative firms in Berlin. Available from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882503