RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE
AND INNOVATION POLICY STUDIES

HOW TO MONITOR AND ANALYSE THE LABOUR-MARKET PLACEMENT OF TRAINED PhDs USING LINKED DATASETS

RISIS Policy Brief Series

Issue #13| June 2023

www.risis2.eu
risis.communication@ircres.cnr.it

CONTENTS

The numbers of doctorates[1] awarded in **OECD countries** have increased dramatically since the mid-1990s. PhD production is currently outpacing the rate needed simply to replace research and teaching competencies in the academic sector[2]. In consequence, a substantial and apparently growing share of PhDs is percolating into other sectors of the economy. This trend is not new. “The perennial shortage-surplus debate” has long roots[3]. Looking back, it stands out as a characterizing feature of the **formative knowledge economy**. Looking ahead, its importance is likely to continue to grow, with manifold implications both for education and innovation polic

A product of the RISIS2 project, the **Doctoral Degree and Career dataset (DDC)** is designed to inform ongoing policymaking on these issues in Europe. Three aims are central:

- to take stock of PhD production on a year-to-year basis for a wide range of European countries

- to identify the proportion that flows into other, non-academic sectors,

- and to build the lens on appropriate scientific framework that appreciates the long-term returns to society of its **investment into PhD** production in terms of new economic activities and innovation in the economy.

Thus, the DDC aims at complementing ongoing activities at the national and international levels (e.g., CDH) and at providing important added functionality.

1. INTRODUCTION

The fact that the ‘academic labor force’ has gone from a state of relative shortage in the postwar years (Blank & Stigler, 1957) to the remarkable surplus we see today is a hallmark of the modern knowledge-economy. Across many OECD countries, universities are producing[4] many **more PhDs than can be absorbed in the ranks of university staff.**

This means that today’s doctorate faces lengthening odds when trying to gain a toehold on an academic ladder and even longer odds when trying to successfully climb that ladder (Cañibano et al, 2023). This important trend also has manifold implications for **public policy**.

Moreover, it has implications for different domains of public policy. The more PhD production exceeds the 'replenishment rate' for academic scientists, the more the issues will extend beyond the traditional boundaries of **science policy** and into policy realms associated with the “Knowledge-Based Economy”[5]. If an increasing share of PhDs no longer enter academia, where do they end up, what effect does this have on the individual's career, and on the role they play in the generation and **spread of knowledge**.

To approach the policy issues, it is useful to distinguish two formally separate domains of public policy involved. A schematic distinction between 'supply-side' and a 'demand-side' helps to bring together two dimensions of an issue that often are treated in isolation from each other. These dimensions should be to a greater degree treated together.

- Science policy, which overlaps with education policy here: it is primary in as far as it influences the increased supply of trained PhDs (aka "PhD Production") and in as far it influences how newly trained PhDs move into 'academic careers'. It is associated with concerns about potential risks of a “PhD glut” (e.g., Bok 2015; Cyranoski et al. 2011) and about the implications this might have for the sustainability and relevance of **doctoral training** (e.g., Auriol 2007, 2010).

- The 'labor-market placement' forms a notional 'demand-side' of the equation and falls under the more formative areas of knowledge or innovation policy. This is a much less developed area of focus among the scientific and policy communities. It however builds on a long tradition that can be traced back to Nelson and Phelps (1966).

These traditions are currently being updated while education policy and innovation policy frames are being brought together to consider a host of questions, such as: How many trained PhDs wind up outside academia?

Where do they go? In what ways is this share changing in different (geographical or FoS) contexts? What **implications does this 'export'** have for the (knowledge) economy?

The question of measurement –how to identify and track PhDs— has been central to the work that links the 'production of PhDs' to the 'deployment' of trained PhDs in the labor market from the start. This is underlined by the fact that studies that combine the scientific with the statistical community are historically well-represented[6].

A separate working paper[7] reviews the various approaches that have been employed (survey-based, register-based, document-based, combinations) and discusses challenges and limitations that have been faced. It particularly focuses on concerted efforts by **UN, OECD, and Eurostat (UOE)** and by the **EU** to improve methods to compile reliable information about the (non-acade ıic) **labor-market placement of trained PhDs**.

Today, those efforts are at something of a crossroads as conditions for data-access and coordination evolve. In this changing landscape, the **RISIS2 project addresses the need for better empirical strategies** by developing a dissertation-based approach to monitor and analyze labor market placement of trained PhDs. This document introduces the Doctoral Degree & Career Dataset (DDC), describes the basic methodology, presents some

2. METHODOLOGY AND DATA

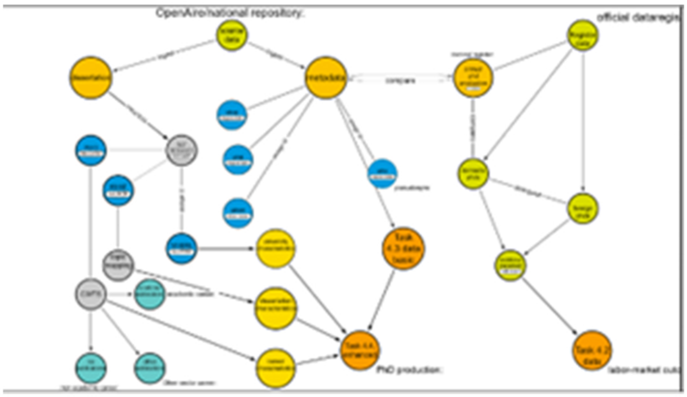
The **Doctoral Degree and Career dataset** (DDC) is an experimental dissertation-centric database that makes two major contributions. Primarily, the DDC offers the user community an enriched "PhD production" dataset. It combines multiple RISIS resources around a core of dissertation metadata. The DDC recruits and integrates richer information about the **dissertation** (e.g., topic mapping), about the **degree-granting university** (e.g., geolocation), as well as basic information about the individual (e.g., gender).

On the 'demand-side', DDC develops an experimental workflow that estimates the post-dissertation labour-market (LM) outcome based on onward publications. This yields a basic indicator of "Career", primarily distinguishing between academic and non-academic variants. This component is designed to provide the community with a tool to **study the rate at which PhD students pursue careers inside (outside) academia** over time while evaluating factors that may shape labour market outcomes. There are currently **93,134 observations** in the pilot of dissertations that were issued in the six countries (AT, DE, IL, NL, ES, NO) for two years (2010, 2014). The observations in the DDC are arranged in an array of variables around the dissertation, which is the unit of analysis.

Deliverable 10.5[8] documents details of how the DDC dataset was compiled. Five main steps were involved: to establish the population frame for the annual PhD production; to ingest the data from identified sources; to clean and diambiguate the underlying data; to link to the **RISIS Core Facility (RCF)**—primarily **ETER**/**OrgReg** and **CWTS Publication data**; and to synthesize a career indicator.

Estimating the career indicator is worth emphasizing. It involves automated linking of cohorts to the publication streams of the dissertation author. A matching algorithm developed by CWTS against its publication dataset is used to proxy the pursuit of an 'academic' vs a non-academic job. It consists of multiple steps. An automated step to link dissertation data against publication data, a supervised step to calculate precision and recall, and an inference & evaluation step. Figure 1 illustrates the steps taken.

Figure 1. Model for the workflow of DDC data



Please note: GDPR compatibility is a precondition and an overriding consideration through all the underlying steps. DDC takes special steps (see RISIS Deliverable 10.4 - Towards Opening the Doctoral Degree and Career Dataset (DDC)) to ensure that the dataset is compliant.

3. FINDINGS

The DDC is a pilot dataset that is being finalized and extended in 2023. Results are preliminary but very promising. In general, findings so far indicate that:

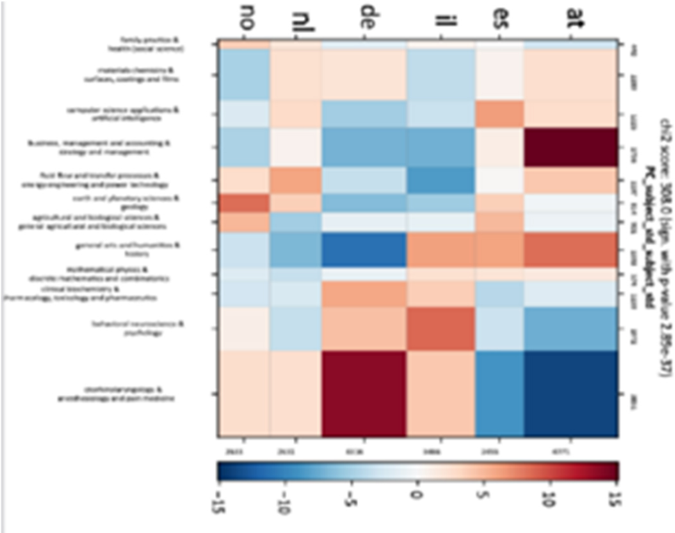
- The PhD production component of DDC provides significantly enriched information relative to a register-based dataset. Relative to self-reporting alone (e.g., survey-based), it provides **near complete coverage of the population**.

- The procedure for Career indicator is more involved, and full results are not yet available. However, the identification procedure (using subsequent publication to predict real labor-market placement) has been validated, and **single country studies have been carried out**.

PhD production component: Enrichment includes at the level of the individual (e.g., gender and age brackets) and of the degree granting university (via ETER/OrgReg on geolocation, annual staff of university, size of student body, funding model). In addition, it includes information from the dissertation. We showcase the application of topic mapping. But the application of GATE in RCF is still being applied on extended sections of the full text of the dissertations.

This preliminary exercise, run by the RISIS Core Facility team, presents a **national specialization contingency matrix for the six countries**.

Figure 2. Cluster of Subjects Against Countries: reporting Chi2 scores

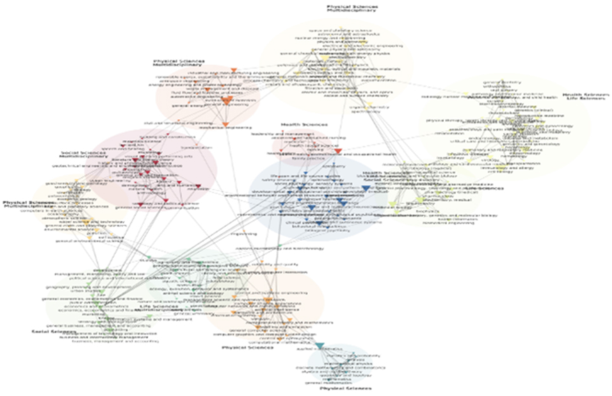


Source: Breuker et al, 2022.

The work illustrates a **clear business management specialization** in **Austria** and **medical specialization** in **Germany**. Norway's petroleum industry is indicated by the relative specialization in geological sciences. The potential of this cross-country exercise would furthermore benefit from including more countries (currently underway). Running the exercise over time would provide the basis to study the evolution of the specialization patterns.

An extension of the possibilities in topic mapping is further illustrated in figure 3. Dissertations tend to be pigeonholed in single 'fields' (e.g., medicine), while they tend to be more heterogeneous and polyvalent.

Figure 3. Co-occurrence of topics: Louvain Clusters of topics (similarity, top 250) and domains (chi2)



Source: Breuker et al, 2022.

The figure demonstrates the real extent of 'multidisciplinarity'. It illustrates how different topics tend to cluster together into galaxies. Some fields are more separate (e.g., Physical sciences at the top). Others, such as the Life Science galaxy (bottom left) link multiple fields.

The Career Component: While the career component of DDC is not yet fully-fledged, work by different national teams attests to its potential. The precision of the **unsupervised identification strategy is high** (**~90%**) and recall of the raw match is encouraging (between 70-85 percent). Work underway in Israel delves into the importance of the Post-doc for subsequent academic positions, work in **Spain** explores career-'hybridization' (~6% of sample) and finds a growth in the propensity to move into a non-academic sector from 2010 to 2014. Preliminary findings in **Norway** indicate that around 35 percent of cohorts continue in academia, while a further 30 percent work public research institutions and/or in public health (hospitals). About 8 percent of graduates in technology fields are traced into the petroleum industry.

4. IMPLICATIONS

The (new) economics of science literature indicates that PhD production traces the scientific frontier (Dasgupta & David, 1994) and improved indicators based on PhD production and job placement can provide a **valuable addition to innovation indicator** (Stephan, 2002).

How, then, should we monitor and analyze the labor-market placement of trained PhDs? To address this question, the Doctoral Degree and Career Dataset (DDC), brings together a **PhD production** and a **Career component** in order to be able **to inform both the science and education policy associated** with the supply-side as well as innovation policy associated with the lemand-side.

The DDC demonstrates the advantage of the design. Using the dissertation as the unit of analysis can provide a hugely informative and quasi-complete population. Using datasets from within a secure environment (RISIS Core Facility) can enrich this basic data with ancillary information (eg about the degree granting university) while also ensuring GDPR compliance. Within the perimeter of this environment, the use of built-in tools (geocoding, topic mapping) **to improve analytical possibilities of the data.** At the same time, the (non) existence of a link to a further publication record is used to predict whether or not the candidate continues in an academic career.

Notes

[1] Following convention, a “PhD” will be used as short-hand for doctorate-degree holder in this paper.

[2] Definitions: we distinguish between research sectors, which includes research institutions, and HES. The role of hospitals is a special case. See Frascati.

[3] See for example the 2001 “Policy and Data Issues of the Scientific Workforce” Conference by the NBER’s Science & Engineering Workforce Project.

[4] Terminology involving the ‘production’ of PhDs is used guardedly to emphasize the investment that lays behind it, both by the individual, the university, and the public purse.

[5] Note that the OECD published “Knowledge-Based Economy” (OECD, 1996) at about the same time as PhD production started to take off.

[6] Examples trace back at least to Blank and Stigler (1957) and include recent work such as that of Zolas et al. (2015).

[7] W10-5.2 CDH-Plus: Building empirical lenses with official statistics.

[8] https://zenodo.org/record/7733595

REFERENCES

Auriol, L. (2007). Labour market characteristics and international mobility of doctorate holders: results for seven countries. OECD Science, Technology and Industry Working Papers, No. 2007/02, Paris: OECD Publishing.

Auriol, L. (2010). Careers of Doctorate Holders: Employment and Mobility Patterns. OECD Science, Technology and Industry Working Papers, No. 2010/04, Paris: OECD Publishing.

Blank, D. M., & Stigler, G. (1957). The Demand and Supply of Scientific Personnel. National Bureau of Economic Research, Inc. https://EconPapers.repec.org/RePEc:nbr:nberbk:blan57-1

Bok, D. (2015). Higher Education in America. Princeton: Princeton University Press.

Cyranoski, D., Gilbert, N., Ledford, H., Nayar, A. & Yahia, M. (2011) Education: The PhD factory. Nature, 472, 276–279.

Dasgupta, P., & David, P. A. (1994). Toward a new economics of science. Research Policy, 23(5), 487-521.

Nelson, R. R., & Phelps, E. (1966). Investment in Humans, Technological Diffusion, and Economic Growth. The American Economic Review, 56, 69-75.

OECD (1996). The Knowledge-based Economy, OECD/GD (96)102. Paris: OECD.

Stephan, P. (2002). Using Human Resource Data to Illuminate Innovation and Research

Utilization. In S. A. Merrill & M. McGeary (Eds.), Using Human Resource Data to Track Innovation: Summary of a Workshop (pp. 80). The National Academies Press. https://doi.org/doi:10.17226/10475

Zolas, N., Goldschlag, N., Jarmin, R., Stephan, P., Smith, J. O., Rosen, R. F., Allen, B. M. Weinberg, B. A. & Lane, J. I. (2015). Wrapping it up in a person: Examining employment and earnings outcomes for Ph. D. recipients. Science, 350(6266), 1367-1371.

RISIS2 - European Research Infrastructure for Science, technology and Innovation policy Studies aims at building a data and services infrastructure supporting the development of a new generation of analyses and indicators on STI fields.

To develop a deeper understanding of knowledge dynamics and policy relevant evidence, RISIS goes beyond established quantitative indicators, developing positioning indicators, in order to reduce asymmetries in actors producing new knowledge, in places where knowledge is generated, and in themes addressed.

RISIS community is dealing with sensitive issues as social innovation, non-technological innovation, the role of PhDs in society, and portfolios of public funding instruments, studying both universities and firms.

__RISIS Policy Brief Series__ aim at disseminating key results coming from RISIS2 to improve the use of data for evidence-based policy making. The outcomes are presented through short documents pointing out the main policy issues at stake, demonstrating the contribution provided by RISIS, and what new avenues for research are now open.

AUTHORS OF THE CURRENT ISSUE:

Eric J. Iversen | NIFU

and the DDC team

EDITORIAL BOARD:

GRAPHIC DESIGN:

Serena Fabrizio | CNR

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

www.risis2.eu