RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

10.3
Manual & Documentation of Doctoral Degree and Career Dataset (DDC) – first version

Jakob Tesch (DZHW); Eric Iversen, Asgeir Skålholt (NIFU); Thomas Franssen, Jeroen van Honk (CWTS); Carolina Cañibano-Sanchez, François Perruchas (ingenio); Eran Leck, Daphne Getz, Ella Barzani (SNI); Daniel Wagner-Schuster (Joanneum)

31.12.2020

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 824091

Outline

1 Preface and Basic Characteristics

As a worldwide development, the doctorate today has shifted from qualifying doctoral students primarily for academic jobs towards teaching broader qualifications that prepare students for a variety of tasks also beyond academic research particularly in research and innovation. Already at the beginning of their training doctoral students in many countries aim for non-academic- rather than academic careers (Ambrasat & Heger, 2020; Gemme & Gingras, 2012, Sauermann & Roach 2012) and leave academia shortly after conferral of the degree. Consequently, the overwhelming majority of those remaining post-docs are striving to stay in academia, often with fixed term contracts and usually outnumbering the available positions for professors (van der Weijden et al., 2016, Woolston, 2020). The discussion about reasons for these developments highlighted a lack of permanent positions in academia at the post-doctoral level as well as statistically worsening chances of becoming professors due to a lack of positions at the highest academic career stage. In addition, studies mention a growing demand for highly skilled staff for research and development in the private sector and a high attractiveness of private sector jobs while highlighting the high attractiveness of jobs in the academic sector apart from the satisfaction with career prospects and job insecurity (Roach & Sauermann, 2010).

In academic as well as non-academic sectors, doctorate holders carry out research related work and contribute to research and innovation outcomes directly (Lissoni et al., 2013) and indirectly (Haapakorpi, 2017). Recognizing the contribution to knowledge production in both the academic and the non-academic research sector, research and innovation policy at the EU-level now emphasizes openness of labor markets and free flows of human capital between sectors encouraging sectoral and regional mobility. Yet somewhat surprising country comparative statistics on labor flows between sectors for doctorate holders are scarce within the EU/ERA as a whole. A recent report by the European University Association (2020) surveyed different approaches for tracking careers of doctorate holders and lists a number of methodological possibilities, while revealing at the same time that none of the data produced is available for a broader range of countries and that regional or university specific approaches dominate the patchy landscape of tracking careers of doctorate holders. An exception to this is the specific use of Labor Force Surveys and their aggregation within the Careers of Doctorate Holders project (CDH). However, the CDH project has seen a resource shortage lately and carried out the last round of data collection in 2016 in form of CDH-light with limited country participation and analysis. Thus, the availability of data for doctorate holders is not in line with demand, as it has been articulated in recent resolutions by the European Council (2017) calling for tracking the outcomes of graduates.

The RISIS Doctoral Degree and Career Dataset (DDC) takes a novel approach towards tracking careers of doctorate holders by providing researchers and policy makers data and tools necessary for analyzing research careers. Based on a variety of data sources, DDC allows linking forms and outputs of doctoral education with their outcomes on a country comparative level thus producing solid indicators grounded in a conceptual framework. The DDC focusses on analyzing research careers, understood as “work lives lived through the performance of scientific research” (Cañibano et al., 2019, p. 1971) here specifically by doctorate holders. Important dimensions of research careers are the degree to which scientific research characterizes the employing organizations as well as the degree of research work performed by doctorate holders within these organizations. The Research Career Conceptual Framework (RCCF) behind the DDC assumes that “research careers condition the type and volume of knowledge outcomes that are produced by researchers in different social and institutional contexts” (Cañibano et al. 2019, p. 1964) thus substantially shaping a countries’ or regions’ research and innovation output. The DDC applies this framework to a curated set of newly combined data sets thus providing a database for linking individual, organizational as well as institutional characteristics to the outcomes from research careers.

This document describes the status of the DDC at halfway through the RISIS project.

2 Database content

2.1 Database Topics

The DDC addresses an important gap in research & innovation statistics by allowing the study of the specific situation of doctorate holders in different data across countries. DDC mobilizes or adds to a number of different RISIS facilities and services to supply standardized, curated data and indicators for understanding research careers, particularly labor flows between academic-/non-academic sectors across countries thus informing on the openness of the European labor market for researchers across several dimensions. The following Figure 1 summarizes the main topics of the database.

Openness of the European labor market for researchers - Knowledge-transfer from academia and doctoral education - Mobility trajectories	**Transnational co-operation and competition** - PhD production and academic opportunity structures - Contribution of PhDs to scientific output
Diversity of non-academic careers - Institutional characteristics and PhD careers - Business sector uptake of PhDs	**Dissertation topics and cognitive careers** - Dissertations in the broader knowledge context - Utilization of PhD knowledge in non-academic careers

Figure 1 Main topics of DDC database

2.2 Database structure

Conceptually, the core of the DDC dataset consists of micro-level person data for cohorts of trained PhDs identified through dissertation metadata. This core is used to "fix" the study of empirical phenomena in other RISIS resources to identified individuals thus disambiguating information from a variety of other data sources and allowing the study of individual trajectories such as research careers. The core can also be used to compare to existing aggregated-level statistics (education statistics, HRST statistics etc.).

In terms of database structure and shape, the DDC is a data hub making use of several RISIS and external data sources, which allow an aggregated population view (aggregated-level) as well as a cohort based micro-level view. The micro-level view commences from the core of individual doctorate holders identified through a collection of dissertation metadata. This dissertation meta-data links to other data on the micro-level as well as other RISIS resources to study research careers. The population DDC allows viewing the total doctoral population within countries for selected years/cohorts based on education statistics and results from CDH. This aggregated population view also establishes a common official reference frame for the micro level DDC-data to be compared upon. Currently, the DDC contains data for the two cohorts 2010 and 2014 from six pilot countries (AT, DE, IL, NL, ES, NO). The approach will be extended and automatized further during the remaining project.

Technically speaking the DDC is a multi-table dataset with linking variables on different levels of aggregation such as scientific fields, HEIs, countries, and years (Figure 2).

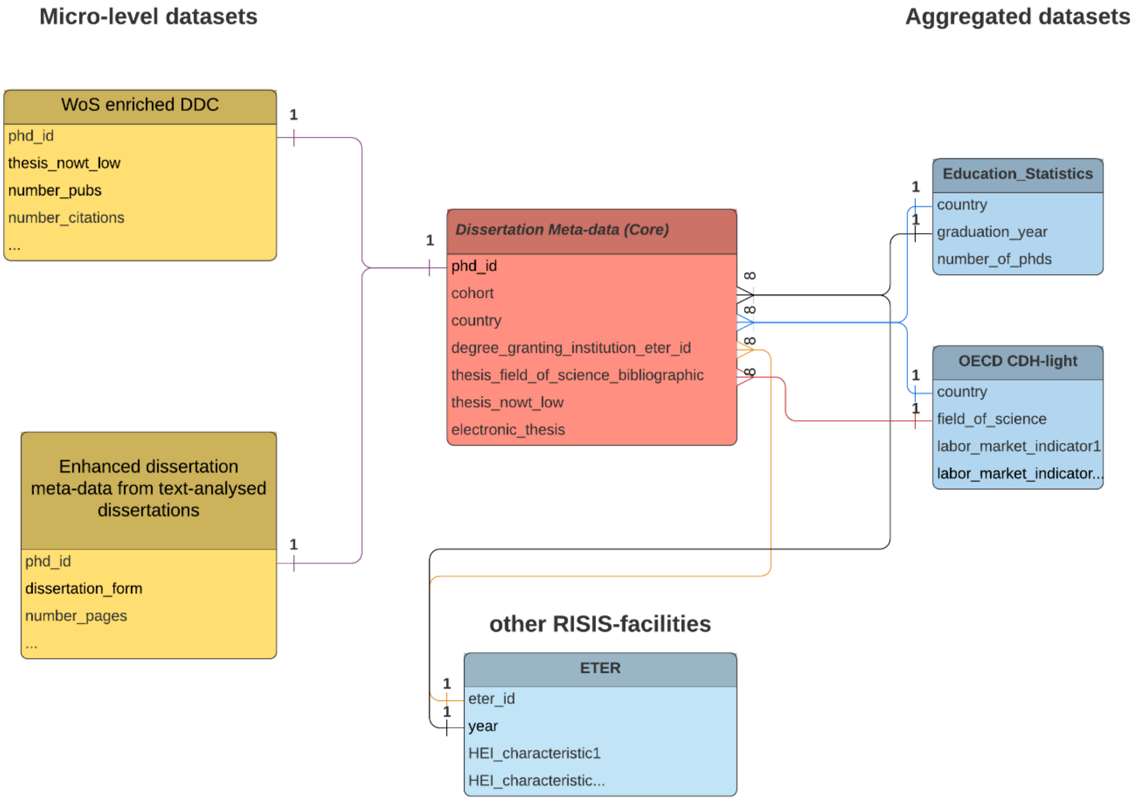


Figure 2 Structure of the dataset at M24

2.3 Sectorial, temporal and geographical coverage

Currently the core of the DDC comprises dissertation metadata for the two cohorts 2010 and 2014 from six pilot countries (AT, DE, IL, NL, ES, NO). The core metadata form a full population of trained PhDs from these countries (see section 3.3.5).

3 Data acquisition and processing of data sets included in DDC data-hub

Aggregated Datasets

3.1 OECD CDH-light

The inclusion of OECD-CDH-light type studies allows links to further regular statistics apart from assessing the characteristics of the 'true population' of doctorate holders who are 'deployed' in the national economy, whether in academia or elsewhere in the labor-stock. The most straightforward approach is to link the educational attainment variables directly to the data frame for active members of the labor-market that countries regularly compile. Currently, this direct approach that bridges register (administrative) microdata for education attainment with that for labor-market participation is not yet widespread. It is however being piloted in Norway, in order to check the accuracy of these data and to demonstrate how they may be used to analyze the 'deployment' of doctorate holders, both into the labor-market (the case of 'newly minted' PhDs) and through the labor-market (career). Thus, it provides standardized information on type of jobs, satisfaction, and occupation as researchers etc.

The need to better understand the labor-market placement of doctorate holders has long been recognized by the statistical community (and public administration more generally), by the policy community, and by the scientific community. A remarkable collaboration between OECD, EUROSTAT and UNESCO is currently trying to move this important work along at an international level. Their work involves coordinating developing a light-weight method (since 2016, dubbed 'CDH-Light') that can be deployed in different countries using statistics that are regularly collected. A core issue has been to develop and implement a common nomenclature for the educational attainment of a country's population in terms of the fields-of-science (FoS) of the granted degree and of the level of the granted degree (e.g. bachelors, masters, doctorate).

One byproduct of this quest for a compatible and comparable headline data on the educational component of human-capital is that it can potentially be applied to study the labor-market placement of doctorate holders across time and across country. Since this quest is complementary to the focus of the DDC activity, we address whether/in which cases the underlying CDH approach could provide a frame of the 'true population' of doctorate holders who are 'deployed' in the national economy, whether in academia or elsewhere in the labor-stock. A separate report (*)[1] investigates the scope to enrich the data underlying the CDH-Light exercise in order to provide a common official population frame for the DDC-data, it demonstrates the potential of a common approach for a "CDH-Plus" dataset, and it discusses some of the challenges that are faced in adapting it as a population frame for DDC. In this section, we present basic dimensions of the CDH light.

The combined efforts of the OECD, EUROSTAT and UNESCO (UOE) have developed and implemented a common nomenclature for the educational attainment of a country's active workforce. The most recent published edition is the dataset of "Career of Doctorate Holders" (CDH-Light, 2017). It departs from earlier attempts at generating a custom dataset (2009); instead, it concentrates on using data that national statistical agencies already compile about educational attainment.

By following classification procedure, data about educational attainment can then be linked to other data collected for national populations that involve labor-market placement. The most straightforward approach is to link the educational attainment variables directly to the data frame for active members of the labor-market that countries regularly compile. Currently, this direct approach that bridges register (administrative) microdata for education attainment with that for labor-market participation is not yet widespread. It is however being piloted in some countries: the separate report showcases one country that does so, Norway, in order to check the accuracy of these data and to demonstrate how they may be used to analyze the 'deployment' of doctorate holders, both into the labor-market (the case of 'newly minted' PhDs) and through the labor-market (career).

By the end of 2021, the current collection of data for over 25 OECD countries should be completed according to the UOE approach. Seventeen OECD countries have already carried out an official survey of doctorate holders in 2019 or 2020, seven of which for full populations. A further nine countries are planning to do so in 2021. These exclude countries that (also) are carrying out analyses of mainstream labour force statistics. A separate report (Nifu, 2020) showcases what type of information about non-academic careers can be derived for countries that combine official labour force statistics with full-count education and demographic statistics.

[1] See report W10-5.2 CDH-Plus: building empirical lenses with official statistics.

3.1.1 Definition and description of observations

Since 2016, the CDH approach has become decentralized, depending on national statistical offices to apply common concepts, definitions, and classifications on domestic data in a harmonized way. A core step involves the implementation of a common nomenclature ("ISCED") for the educational attainment of a country's population in terms of the fields-of-science (FoS) of the granted degree and in terms of the level of the granted degree (e.g. bachelors, masters, doctorate). The ISCED categorization, described here[2], follows a joint data collection methodology, dubbed the UOE. The UOE approach involves two prongs: one at the level of educational attainment ('degree': ISCED 2011) and the second at the level of field-of-science ('fields': ISCED-f, 2013) in the case of higher degrees.[3] The approach has started to yield some headline data on the educational component of human-capital across a range of countries and across time. The most up-to-date results can be found at the OECD, which publishes comprehensive information about the CDH-Light exercise, including technical guidelines as well as other relevant indicators and analysis[4].

3.1.2 Data Sources and Availability

The CDH-Light includes four of the five pilot countries in DDC: Norway, Spain, Germany and the Netherlands. However, comparisons in outcomes across fields of science are only available for three of them: Norway, Germany and the Netherlands. There are also some differences in data collection periods. For all but the Netherlands, the data is from 2016.

There are several features of the CDH-light data that should be appreciated in the quest for better data to better understand the labor-market placement of trained PhDs. A first problem is that few (European) countries have publicly published data from the CDH-light exercise so far.
A second problem is that of those countries that have participated, do not (yet) compile the underlying data for the CDH exercise from the same sources in the national statistics. In addition to the register (administrative) based approach, two other approaches have so far been used.
custom surveys (census) to this project, (e.g. Netherlands),
and other countries relied on existing labor force surveys, such as Germany
It is clear that the different sources of data can affect their compatibility (in cross-country comparisons), their utility (the Dutch and German approaches provide snapshots, and therefore cannot provide insights into how doctorate holders move through the labor-market over time), and potentially, their reliability (e.g. non-response bias in the case of custom surveys). Furthermore, although the CDH data are open and available, the fact that they are aggregated limits their analytic possibilities.
Going forward, the hope is that CDH platform that is being rolled out will move to a common framework. In order to be most useful in providing timely, reliable and rich empirical data about the labor-market placement of trained PhDs across countries and across time, this framework should strive to use the most direct approach that bridges register (administrative) microdata for education attainment with that for labor-market participation.

The current collection of data among the current 25 + OECD countries participating is not expected before the end of 2021 at the earliest.

3.1.3 Data Cleaning

The statistical offices of each country provide cleaned data which are compiled and presented by the OECD.

[2] A manual (UNESCO-UIS / OECD / EUROSTAT, 2016) describes the common strategy, see https://ec.europa.eu/eurostat/statistics-explained/index.php/International_Standard_Classification_of_Education_(ISCED)#ISCED_1997_.28fields.29_and_ISCED-F_2013

[3] The ISCED-f approach is broadly in line with the Fields of Science Classification in OECD's Frascati manual, although not completely, see https://www.oecd-ilibrary.org/docserver/9789264239012-en.pdf?expires=1592571353&id=id&accname=ocid177226&checksum=94E7802BD667A409E9AE6D47927B4532

[4] https://www.oecd.org/innovation/inno/careers-of-doctorate-holders.htm

3.1.4 Information/Guidance on variables/indicators

The CDH provide the following variables across fields of science: age group, country of birth, citizenship, residency status, country of highest educational attainment award, activity, employment, working time, researcher status, industry of employment (main job), sector of employment (main job), changed employer last year, earnings.

3.2 Education Statistics

Education statistics, in particular graduation statistics are among standard information national statistical agency collect. The purpose of inclusion is to situate the number of doctorate holders identified through dissertation metadata in the reference frame of graduation statistics that each country collects regularly.

3.2.1 Definition and description of observations

The observations in this data set are aggregated units of doctorate holders. Available aggregation levels are countries, fields of science, and individual universities.

3.2.2 Data Acquisition

Data was retrieved directly from the national statistical offices.

3.2.3 Information/Guidance on variables/indicators

- Gender
- Fields of science
- Age groups
- Non-domestic PhDs

Micro-Level Datasets

3.3 Dissertation Meta-data (Core)

The DDC-dissertation Meta-data dataset includes doctorate holders (people) and their dissertations from the two dissertation cohort years 2010 and 2014. DDC includes the full population of dissertations, which are contained in the respective (central/decentralized) national repositories for these years from six pilot countries.

The DDC-dissertation meta-data is the core dataset for linking additional micro- and macro-level datasets. Thus, it can be enriched through additional data that match based on the names of doctorate holders from the pilot countries, their degree granting universities and fields of science of the thesis as main linking variables.

The following subsections describe how the original data were acquired and the steps that were undertaken to clean, process and align the original data.

3.3.1 Definition and description of observations

As stated before, the DDC-dissertation Meta-data dataset includes full-count populations of individual doctorate holders as defined by national repositories responsible for publishing dissertation metadata. This includes Medical doctors as well as MDs in the case of Israel, where MDs earn a separate degree.

The current pilot encompasses data from Austria, Germany, Israel, Norway, Spain, and the Netherlands. As a principal rule, all cases which after cleaning could be identified to belong to these pilot countries, were kept and are part of the final dataset. As of now, the dataset is limited to the two cohorts 2010/2014 with their final exam/ the publication year of the thesis in case of Israel, in these years. Norway and Spain have made additional efforts and collected dissertations for the years 2000-2018 and the Netherlands for the period 2010-2016. The dataset consists of PhDs nested in cohorts, countries and degree granting universities as well as fields of science.

3.3.2 Data Acquisition

In many countries around the globe, the criterion for publishing PhD theses led to the emergence of entities now responsible for collecting and publishing dissertations/meta-data. Sometimes these, particularly centralized national registers are backed up by law and failure to submit data on the dissertation to the entity can result in fines. Increasingly universities discover the value of publishing thesis and making them accessible as part of their Third Mission. Thus multiple sources for accessing dissertation metadata exist and inconsistencies between centralized repositories and individual libraries are likely. Recent developments have seen a growing number of harvesting websites retrieving dissertation metadata algorithm-based from individual libraries, which open their catalogues.

Thus, in principle, the three existing access options for acquiring dissertation metadata are

1) access through centralized harvesters (e.g. OpenAire, DART-Europe),
2) central national repositories (e.g. national libraries) and
3) access through individual university libraries.

As a rule regarding the pros and cons of the individual approaches, it can be said that harvesters usually face problems of undercoverage due to their focus on electronic thesis, thus ignoring dissertations published as books. In addition, they provide a limited amount of data fields. While centralized national repositories usually can be regarded as quite complete, they usually contain noise due to deliveries from multiple sources (individual PhD, publishing house, faculty etc.) making deduplication necessary. The high-level of completeness of data makes university libraries an attractive even the preferred source but also challenges the collection of metadata due to the variety of existing meta-data formats. Thus, in countries with a limited number of universities the collection from single universities might be feasible whereas in larger countries with a high number of degree granting universities this will soon become a resource intensive task, outweighing the effort needed for cleaning data from centralized national repositories.

In two cases within the DDC-dissertation metadata however, even with the presence of a centralized national repository, partners deliberately decided to collect data from single universities instead. The team in the Netherlands, for example, retrieved data directly from the 18 universities, mostly via email and in one case via direct download from the university library. One of the reasons here is that the centralized publisher contains electronic thesis only. In Israel, the most comprehensive data is found in the Israel Union List of libraries in Israel (ULI), which contains more than nine million bibliographic records from the catalogues of university libraries, colleges and several other major libraries. Although the use of ULI records would have been most ideal for data retrieval due to its inclusive nature, two main problems have prevented the Israeli partner from using this platform. The first and most significant obstacle was the inability to separate master theses from doctoral dissertations in a simple manner (both were under the same field) and the second difficulty was that the search results were limited to 1,000 entries. Nevertheless, as the comparison with education statistics in Section 3.3.5 illustrates accessing data through individual universities and a centralized repository both produce results, which are comparable with regard to the numbers reported in education statistics.

Table 1 lists the different approaches undertaken to acquire dissertation metadata. All in all, data acquisition is a manageable task yet sometimes complicated by download limits.

Table 1 General overview of data acquisition of dissertation meta-data, by country

COUNTRY	SOURCE	METHOD	RETRIEVAL DATE	LINK TO CENTRALIZED PUBLISHER WHERE APPLICABLE
AUSTRIA	Austrian Library Network (OBV)	Manual download	March 2020	https://search.obvsg.at/primo-explore/search?vid=OBV
GERMANY	German National Library (DNB)	Manual download	October and August (Medicine) 2019	https://portal.dnb.de/opac.htm?method=showSearchForm#top
ISRAEL	Individual university libraries	Crawler for ex-libris university library websites (N=6/7)	May 2019	NA
NORWAY	The Norwegian Doctoral Register (NIFU) and individual university websites	Manual Download	-	https://www.nifu.no/en/statistics-indicators/doktorgrader/
SPAIN	TESEO	Crawler for TESEO	February 2019	https://www.educacion.gob.es/teseo/irGestionarConsulta.do
THE NETHERLANDS	Individual universities libraries via email (N=17) or direct repository access (N=1)	Email delivery (N=17)/ manual download (N=1)	2017	NA

3.3.3 Data Cleaning – missing information, duplicates, language issues

The individual partners were responsible for data cleaning, which in most cases involved development of routines for the discovery and removal of duplicate entries. To a lesser extent, this also involved removal of incomplete entries or obviously wrong entries. As can be seen from Table 2, the amount of cleaning differs between the countries. Norway is on one extreme: the Norwegian Doctoral Register ("Doktorgradsregister") publishes doctorate production by HEI, field-of-science, gender etc each year. However, the Register does not include all types of information (eg it lacks information about the supervisor for example).

Much more cleaning is necessary to derive the metadata for Germany. A few incomplete entries were removed from the original data, e.g. those where author name and thesis title are missing (Israel) or when university, author, title or date were missing (ES) or when the name of the university was missing or unprecise and could not be identified from other sources (DE). A few wrong entries (DE:N=3) were removed, e.g. where the name of the university was a "Fachhochschule" and thus did not have the right to award a doctorate at that point in time or where web-search of cases with incomplete information revealed that they were in fact not dissertations. Additionally, in Germany 1,363 individual theses were identified to not belong to German universities, which corresponds to around 2 % of all individuals per cohort for Germany. Another four cases in Germany turned out to not belong to the target cohorts. These were also deleted.

Another issue for cleaning are joint dissertations. In some German universities, regulations allow dissertations to be joint work between two or more doctoral students. As the goal of the cleaning procedure is to extract all individual doctorate holders, based on the dissertation title it was checked, whether a second version of the thesis exists in the data and both authors are thus included. In one case the dissertation was duplicated because it was identified as a joint dissertation between two doctoral students and the export lacked individual versions for both authors. The cleaning procedure for joint dissertations resulted in 78 thesis belonging to joint dissertations. Some of the so identified dissertations are across cohorts e.g. because one part of the joint dissertation was published before the other. In this case only thesis belonging to the target cohorts were kept.

Apart from wrong and missing entries, manual cleaning was necessary for some of the fields in the data, e.g. standardization of author names (NL), correcting university names (DE) and manual retrieval of author names due to the failure of the web crawler to account for them (due to very complex and irregular structure of the bibliographic records, IL).

A main challenge for cleaning, particularly identifying duplicates is the existence of different languages. According to the law in Spain, dissertations can be written in English, Spanish, Catalan/Valencian, Galician, Basque and other languages if their use is important in the field. The German data included dissertations with Spanish and Italian titles apart from German and English. Thus, identifying duplicates is contingent on pre-determining the language of the title. This is something that could be improved in future versions of the cleaning procedures.

Table 2 Amount of cleaning required for dissertation metadata- Number of observations before and after cleaning, by country.

COUNTRY	NUMBER OF ENTRIES IN RAW DATA	NUMBER OF ENTRIES IN CLEANED DATA
AUSTRIA	4,772	4,772
GERMANY	72,406	55,255
ISRAEL	3,738	3,475
NORWAY	2644	2633
SPAIN	20,099	19,886
THE NETHERLANDS	7,995	7,953

3.3.4 Information on variables

The version of the dataset described here includes the following variables. The country datasets sometimes contain an extended set of variables. In order provide a standardized picture, we report here on the agreed variables only.

Table 3 Description of dissertation Meta-data content variables

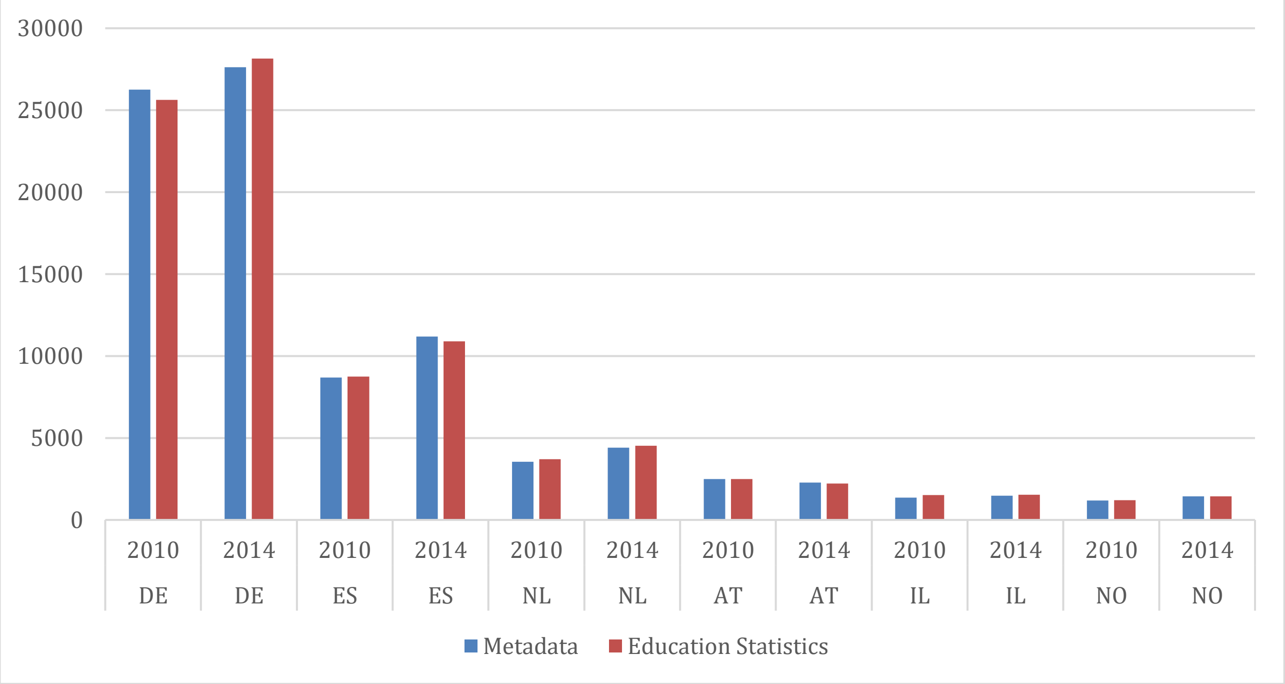
Content	Name of corresponding data fields	Additional Remark
unique identifier	ID	-
Cohort based on date of exam	Cohort	2010 or 2014
Publication year of thesis	thesis_publication_year	Maybe multiple years if different versions exist
Degree granting university	degree_granting_institution_EN, degree_granting_institution_national_name degree_granting_institution_eter	OrgReg/ETER IDs, one university only
Additional affiliation during/after doctorate	subsequent_institution	OrgReg/ETER IDs, incomplete
Name of doctorate holder, author name	author_first_name, author_prefix_last_name, author_last_name	Including birth names where available
Name of supervisor /reviewer	reviewer_supervisor_names	Taken from metadata where available

Field of Science (bibiographic)	thesis_field_of_science_bibliographic	Classification used by metadata source
Field of science (NOWT)	thesis_NOWT_Low, thesis_NOWT_Medium, thesis_NOWT_High	Classification used by CWTS for WoS
Title of the thesis	title_thesis_en, title_thesis_[national_language]	English/national language
Abstract of thesis	thesis_abstract_en	English only
Keywords	thesis_keywords_en	According to national classification
Hyperlink to full text	thesis_link_full_text	Link to university repository

3.3.5 Number of observations, quality and accuracy of data

In all the countries covered here, legislation requires the publication of dissertations. In addition, local funding incentives for universities are sometimes built on the number of granted doctorates. Thus, it can be assumed that the data are generally reported. The only reference to compare the number of dissertations identified in the national datasets using the procedures described above is their comparison to the number of doctorate holders from education, in particular graduation statistics. However, there is a major limitation for this comparison, namely, the different time periods that both sources cover. While graduation statistics usually refer to the academic year the dissertations are reported in calendar years. Figure 3 compares the numbers of doctorate holders as in national graduation statistics usually covering the winter semester and the following summer semester of a year with the numbers retrieved through the collection of metadata. The differences between the two sources are rather small, averaging to 2.7 percent over all cases. Currently the DDC includes 43,540 doctorate holders from 2010 and 48,417 from 2014.

Figure 3 Comparison between number of doctorate holders from dissertation metadata and education Statistics by country and Cohorts



3.4 WoS-enriched DDC

This dataset links the doctorate holders from the DDC core to their publication outputs and derives information on publication output, citations and affiliations and thus mobility. The link can also be used to identify doctorate holders entering into certain career tracks, e.g. a position at a higher education institution, a position at a research institute or a publishing position at a firm or government organization (incl. entrepreneurs). If the doctorate holder is not found to publish in the subsequent time-frame this can result from different employment outcomes such as employment in private or public sector in a non-research position or unemployment. Thus, this approach provides an indication of general tendencies and labour flows based on details of the underlying degree. In short, the approach will yield relative measures for the probability of the cohort members to move to a given track in a given year at least for those disciplines well covered in WoS. It yields information on the value of doctoral training and subsequent usage of competencies in these disciplines.

3.4.1 Definition and description of observations

The result from the matching are WoS author profiles that are linked to the doctorate holders from the DDC core.

3.4.2 Data Acquisition

The data was acquired through matching the names of the doctorate holders from the DDC core with the CWTS in-house version of the Web of Science (WoS) database produced by Clarivate. This in-house version consists of a number of citation indices such as Science Citation Index Expanded (SCIE), the Social Sciences Citation Index (SSCI), the Arts & Humanities Citation Index (AHCI), and the Conference Proceedings Citation Index (CPCI), thus the main albeit not the full WoS. CWTS developed a matching algorithm that determined which publication author profile identified in the CWTS-inhouse version by the CWTS belongs to the respective doctorate holder based on PhD-names, degree granting university, field of science, and name of supervisor. The CWTS provided an online interface where each partner manually validated 450 cases for their respective country. During validation each individual doctorate holders' publications were checked and publications could be added or removed. These 450 cases were used as a gold dataset to train a matching algorithm to link doctorate holders from the DDC core to the author profile identified in the CWTS-inhouse version by the CWTS.

3.5 Enhanced dissertation meta-data from text-analysed dissertations

This dataset results from the joint approach between U Sheffield, U Gustave Eiffel, NIFU and DZHW and offers information retrieved from text analysed dissertations. Building this dataset follows two independent goals. First, it provides tools to prepare dissertations - which are a valuable source of information for Science Studies - through the extraction of desired information from dissertations such as individual characteristics of the doctorate holder (funding, discipline etc.) as well as their supervisors, insights into discipline-specific processes of knowledge production (through the cited literature, methods, facilities used etc.) as well as self-citations which identify further works published by the doctorate holders during their training. This will result in data which researchers can use to study the characteristics of dissertations, variations between fields/countries, and over time. We forecast interest in the following fields: e. g. in history of ideas, economics of science, labor-economics, public policy. The second goal of this approach lies in automatization of the manual routines developed through the collection of dissertation metadata (see section 3.3.). It is planned to extend the methodology beyond the pilot countries/cohorts and include dissertations from further national/international repositories and years. This will ultimately lead to a large extension of the DDC core. Currently, this pilot is exploratory to achieve a Demonstrator and also experiential for the development of the RCF and the Scenario approach.

3.5.1 Definition and description of observations

The object of analysis is the dissertation described through a table of attributes. This Core Table is the result of various layers of information to be captured, retrieved or created that can be subsumed under the three dimensions of what, where and how. The purpose is to benefit from existing sources in the RISIS environment (e.g. GATE) or elsewhere (OpenAire, ORCID etc.) in order to establish the best and complete Core Table of a dissertation. The "what" dimension of the Core Table defines the dissertation in a given context (country, field of science, year), its form (monograph, by article, as art-work, model, etc.) and its extent (number of pages, sections, articles, references, figures/tables, etc.) as well as its research topic (Jel codes, faculty, references, research questions). The "where" dimension identifies entities involved in the dissertation production, e. g. university affiliation (dissertation locus), collation with other research entities (labs, companies), funding agencies. The "who" dimension concerns primarily the candidate, but also the supervisor, and any co-authors. The "how" dimension addresses the models/methods used in the dissertation, including equipment (e. g. tele/microscopes, software, etc.).

3.5.2 Data Acquisition

In a preliminary version, data are ingested using the DDC core.

3.5.3 Data Cleaning

Cleaning and disambiguation routines for the dissertations are currently under development.

4 References

Ambrasat, J., & Heger, C. (2020). *Barometer für die Wissenschaft. Ergebnisse der Wissenschaftsbefragung 2019/20*. Berlin: DZHW.

European University Association. (2020). *Tracking the careers of doctorate holders EUA-CDE Thematic Peer Group Report*. European University Association. https://eua.eu/downloads/publications/eua-cde%20tpg_web.pdf?utm_source=flexmail&utm_medium=e-mail&utm_campaign=euacdepublishesthematicpeergroupreport591trackingthecareersofd20201102t1010&utm_content=tracking+the+careers+of+doctorate+holders

European Council. (2017). *Council Recommendation of 20 November 2017 on tracking graduates*. The Council of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017H1209(01)&from=EN.

Gemme, B., & Gingras, Y. (2012). Academic careers for graduate students: A strong attractor in a changed environment. *Higher Education*, *63*(6), 667–683. https://doi.org/10.1007/s10734-011-9466-3

Lissoni, F./Montobbio, F./Zirulia, L. (2013): Inventorship and authorship as attribution rights. In: Journal of Economic Behavior & Organization, 95, S. 49–69

Roach, M., & Sauermann, H. (2010). A taste for science? PhD scientists' academic orientation and self-selection into research careers in industry. Research Policy, 39(3), 422–434. https://doi.org/10.1016/j.respol.2010.01.004

Sauermann, H., & Roach, M. (2012). Science PhD Career Preferences: Levels, Changes, and Advisor Encouragement. *PLoS ONE*, *7*(5), e36307. https://doi.org/10.1371/journal.pone.0036307

van der Weijden, I., Teelken, C., de Boer, M., & Drost, M. (2016). Career satisfaction of postdoctoral researchers in relation to their expectations for the future. *Higher Education*, *72*(1), 25–40. https://doi.org/10.1007/s10734-015-9936-0

Woolston, C. (2020). Postdoc survey reveals disenchantment with working life. *Nature*, *587*(7834), 505–508. https://doi.org/10.1038/d41586-020-03191-7