IRPPS
Consiglio Nazionale delle Ricerche
Istituto di Ricerche sulla Popolazione
e le Politiche Sociali

WORKING PAPER

IRPPS WP114

The OpenUP pilot study on research data sharing in Social Science

DICEMBRE 2019

Roberta Ruggieri
Daniela Luzi
Lucio Pisacane

Consiglio Nazionale delle Ricerche

ISSN: 2240-7332

CNR – IRPPS

The OpenUP pilot study on research data sharing in Social Science

Roberta Ruggieri, Daniela Luzi, Lucio Pisacane.

2019, p.46 IRPPS Working papers 114/2019

Abstract: The report presents the results of a pilot study carried out within the European project OpenUP (Opening up new methods, indicators and tools for peer review, dissemination of research results and impact measurement).
The pilot aimed at identifying strong and weak elements in the process of dataset review and validation and intended to outline best practices that facilitate transparency of the process as well as data dissemination, reliability and reuse.
In particular the report reviews data sharing and evaluation practices in Social sciences, on which the selection of the pilot community is based, and reports on the interviews with the management team of the selected community, i.e. Human Mortality Database (HMD) as well as on a questionnaire submitted to HMD users. Lessons learned that can help identifying requisites and best practices for peer review of research data are reported in the conclusions.

Keywords: Data quality, Open data, Open dataset review and validation, Open Peer Review (OPR), Social sciences

CNR – IRPPS

OpenUP studio pilota sulla condivisione dei dati della ricerca nelle Scienze sociali

Roberta Ruggieri, Daniela Luzi, Lucio Pisacane

2019, p.46 IRPPS Working papers 114/2019

Sommario: Il rapporto presenta i risultati di uno studio pilota condotto all'interno del progetto europeo OpenUP (Opening up new methods, indicators and tools for peer review, dissemination of research results and impact measurement). Lo studio pilota aveva come scopo quello di evidenziare i punti di forza e debolezza nel processo di revisione e validazione dei dati della ricerca e, nello stesso tempo, di individuare le buone pratiche che facilitassero la trasparenza del processo, nonché la diffusione, l'affidabilità e il riuso dei dati.
In particolare, il rapporto esamina le pratiche di condivisione e valutazione dei dati della ricerca nelle discipline afferenti le Scienze sociali. Questi primi risultati sono stati utilizzati per selezionare la comunità scientifica sulla quale effettuare lo studio pilota.
Vengono quindi riportati i risultati delle interviste con il team che coordina il repository di dati - Human Mortality Database (HMD) - e quelli del questionario inviato a tutti gli utenti finali di HMD.
Nelle conclusioni vengono identificati i requisiti e le buone pratiche che possono facilitare e migliorare il processo di validazione e revisione paritaria dei dati della ricerca.

Parole chiavi: Qualità dei dati, Dati aperti, Revisione e validazione dei dati, Revisione paritaria aperta, Scienze sociali

Citare questo documento come segue:

Roberta Ruggieri, Daniela Luzi, Lucio Pisacane (2019). The OpenUP pilot study on research data sharing in Social Science. Roma: Consiglio Nazionale delle Ricerche – Istituto di Ricerche sulla Popolazione e le Politiche Sociali. (*IRPPS Working papers n. 114/2019*).

Acknowledgments. This study is part of the Horizon 2020 OpenUP project. Grant agreement no. 710722. The authors acknowledge the support and the collaborative efforts of the Human Mortality Database management team, namely Magali Barbieri (University of California, Berkeley and INED, Paris), Vladimir Shkolnikov (Max Planck Institute for Demographic Research (MPIDR) and Dmitri A. Jdanov, Head of the Laboratory of Demographic Data at MPIDR. A great thanks goes to our CNR colleague Cristiana Crescimbene for the valuable technical support during the OpenUP Pilot.

Redazione: Marco Accorinti, Daniele Archibugi, Sveva Avveduto, Massimiliano Crisci, Fabrizio Pecoraro, Roberta Ruggieri, Tiziana Tesauro e Sandro Turcio.

Editing e composizione: Cristiana Crescimbene, Luca Pianelli e Laura Sperandio

Index

1. Introduction

The results presented in this report are part of the activities carried out within the OpenUP project[1], a European funded project that addresses key aspects and challenges in the scenery of scholarly communication focusing on three pillars: Peer Review, Impact Assessment and Innovative Dissemination. The scope of the project was to produce recommendations & guidelines[2] for policy makers addressing requirements and needs by the different stakeholders (researchers, publishers, innovators, the public and funding bodies) involved in scholarly communication process to support Open Science. The project tested the achieved results in a set of seven pilots (Vignoli 2017; Vignoli 2018; Blümel et al. 2018) that aimed at testing and/or evaluating, selected approaches to innovative peer review, dissemination, and impact measuring applied to specific research communities and areas (Arts & humanities, Social sciences, Energy, and Life sciences).

The Institute for Research on Population and Social Policies (IRPPS) was responsible for the implementation of one of the three pilots on Peer review that was specifically focused on Social sciences. The pilot investigated the applicability of peer review and/or open peer review (OPR) to datasets in Social sciences aiming to identify strong and weak elements in the process of dataset review and validation and to outline best practices that facilitate transparency of the process as well as data dissemination, reliability and reuse.

Data peer review is a quality assessment process of a dataset performed by experts in the field and represents an essential phase in data publishing activities (some authors speak about Publication with capital letter, (Lawrence et al. 2011; Mayernik et al. 2015)). This strengthens the importance of data validation seen as the assessment of technical and scientific quality of data performed in the different phases of the data life cycle, and at the same time outlines a close relationship between this assessment with the one performed in a peer review process, being it traditional or open. To validate this assumption, it is necessary to analyse the different criteria (Kratz and Strasser 2015) that have to be adopted to evaluate data quality as well as the various phases in which the evaluation takes place, so to identify for each phase the suitable criteria, the skills and professionals needed to apply them, procedures, standards and guidelines that may leverage the process and certainly facilitate a high quality and transparent data sharing. Moreover, considering the data life cycle there is a general agreement on the distinction between a pre-publishing phase and a post-publishing one. The first phase concerns all the evaluation activities related to data creation, processing and analysis including the metadata description and the additional documentation as an essential part for sharing and reproducing the research. Activities and criteria for data evaluation generally depend on the type of publication channel chosen as well as on the guidelines and requisites required by the data publisher in the submission phase. Once the data are published, the second phase of validation comprises all the forms of traditional and innovative measures/metrics (Priem et al. 2010) that aim to evaluate the impact and use of the published data. Therefore, besides citations, the possibility to post comments and evaluations by end users may be considered as a trait of open participation in OPR (Ross-Hellauer 2017) performed in a post-publication phase. Other important indicators of impacts can be developed using data citation counts

[1] OpenUP HUB - https://www.openuphub.eu/

[2] OpenUP Policy Recommendations: https://www.openuphub.eu/openup-policy-recommendations. Retrived on 2/10/2019

(Callaghan et al. 2013) and/or statistics of use, as a proxy of a generally not recognized research activity, thus incentivizing researchers to share the data they produce.

Within this framework, the pilot was designed to analyse both validation phases aiming to consider, on the one hand, the internal data quality assessment along with the organizational context of the community that makes data freely available and, on the other, attitudes and preferences in data re-use by end-users as proxy indicator of post-publishing appreciation of the quality of the database.

To achieve this aim different steps were performed during the research activities. In an initial phase we analysed current dataset management and sharing practices in Social sciences. The results allowed us to select the community to be involved in the pilot activities. On this basis a qualitative analysis of data management and validation procedures was carried out through the interviews with the Human Mortality Database (HMD)[3] community management team. Moreover, a survey to HMD users was conducted to capture users' feedback on data availability and reuse.

2. Materials and methods

2.1 Landscape scan analysis on data sharing and selection of the community

As mentioned before, in the initial phase a literature review as well as desktop analysis was conducted to analyse the state of the art on data management and sharing in Social sciences. Data sharing is a wide research topic that includes different features and issues that are influenced by the natures of data produced in the different research areas related to Social sciences in these disciplines, therefore we decided to focus on areas relevant to our objectives and we analysed five wide-ranging topics:

1) researcher's perception and need;
2) types of data produced;
3) types of data providers;
4) modes of diffusing/publishing data and;
5) modes of validating data.

On one side the results of this analysis (Part I) allowed us to provide insight in the general context of dataset lifecycle management in Social sciences and identify specific characteristics as well as problematic issues. On the other, we detected research communities that are sharing Social science data.

On the basis of a pre-defined set of inclusion and exclusion criteria, (Table 1) we listed potential communities to be involved in the pilot study.

[3] Human Mortality Database – HMD: https://www.mortality.org/

Table 1. List of inclusion criteria

A community that provides open and free of charge access to a data repository and/or data journal;
A community that provides a data repository and/or publish a data journal focused on a very specific topic;
A community that collects a specific type of datasets in an identifiable subfield of Social sciences;
A community that involves well-defined profiles of both data providers (data contributor, data manager, etc.) and data users (belonging to different communities could be a plus);
A community that manages a number of datasets large enough to provide useful information about users.

Between these communities, two have voiced their interest. The first one is HMD, which was created to provide detailed mortality and population data to researchers, students, journalists, policy analysts, and others interested in life expectancy and related social implications. The second community is Unidata Bicocca Data Archive[4], which supports the diffusion of data produced by the Italian official statistical office and the dissemination of some important surveys conducted at international level (Eurobarometer, European Social Survey, Word Value Survey etc.). It is also part of Consortium of European Social Science Data Archives (CESSDA)[5], a Pan-European Research Infrastructure in Social Sciences.

Several informal contacts were established with both communities aimed at analysing the fulfilment of the above mentioned criteria as well as their willingness to an active collaboration on the analysis. At the end, we decided to formalize the collaboration with the HMD that had the advantage to deal with a specific set of data and had a clearly identifiable user's community that can be traced through a free of charge subscription procedure.

2.2 Interviews with HMD Managers

According to our pilot design, the perspectives of data managers in data curation, sharing and publication were analysed via interviews.

According to different roles of the interviewers in data curation, we elaborated two different sets of questions to be submitted respectively to HMD managers and data Country Specialists (CSs) who are responsible for the validation of data coming from the contributing countries. Five interviews were conducted at the Max Planck Institute for Demographic Research (MPIDR) in Rostock

HMD managers were asked to better describe some HMD features: origin, motivations and organizational features of the scientific community, as well as opinion on Open access of data.

CSs were interviewed to analyse in detail how they perform the data quality assessment process.

2.3 HMD users'survey

The HMD users' perception of the data availability and reuse was captured through an online questionnaire. The questionnaire was developed in collaboration with the HMD managers

[4] Unidata Bicocca Data Archive: https://www.unidata.unimib.it/?lang=it

[5] Consortium of European Social Science Data Archives – CESSDA : https://www.cessda.eu/

through web calls and a face-to-face meeting in Rostock. A pilot version was created and sent to a small number of respondents (n=10) to pre-test it. Comments and related updates were then incorporated into the final version of the survey. The survey period was March - June 2018 that included a reminder to missing respondents.

The survey made use of a semi-structured questionnaire of 20 questions, most of them were multiple choice, while plain text answers were also included to collect researchers' opinions on specific features of HMD database.

The questionnaire consisted of two main parts. In the first one, respondents were asked for information on sex, age, country of residence, occupational position and related institutional affiliation as well as main field of interest of their work. These elements allowed us to provide a demographic composition of the HMD users. In the second part, questions were specifically focused on the HMD user's practices and attitudes in data access and use. In particular we explored the following aspects:

- General information on access (frequency and length of use - countries of interest)
- Modes of dataset acquisition (manual or automatic downloads - type of datasets)
- Dataset use (purpose in using - ways of processing dataset - software used - other source of information used in the field)
- User's perception of HMD (advantages - comments and suggestions)

The survey adopted the software LimeSurvey, an open source software that also supports invitations, reminders, and makes answers anonymous. All participants were informed that the survey was anonymous and voluntary, that all data would be kept confidential and evaluated anonymously and that the purpose was to improve and enhance the content of HMD database and website. Participants were informed that the study results and underlying data were to be shared with the EU project OpenUP.

To explore whether differences in uses exist among HMD respondents, each question was analysed stratifying respondents according to occupation.

3. Results

3.1 Landscape scan analysis on data sharing and publication

In this section we report the results of landscape analysis on data sharing in Social science. According to our methodology we summarized the mainly characteristics, features and issues related to data sharing, as following:
1) researcher's perception and need;
2) types of data produced;
3) types of data providers;
4) modes of diffusing/publishing data;
5) modes of validating data.

3.1.1 Researchers' motivation and constraints

Several surveys have investigated researchers' attitudes to analyse barriers and facilitators towards data sharing. General surveys allow us to compare social scientists' attitudes with other disciplines. Tenopir et al. (2011) found that social scientists have a lower propensity to share the data they produce compared to the STEM researchers. These results have been confirmed in other surveys (Kim and Adler 2015; Faniel et al. 2015). This may depend on the

nature of data, especially when qualitative data are involved, privacy and confidential issues, lack of technical standards and easy-to use platforms. Social scientists share the same concerns as scientists in other disciplines, such as not being recognised for making the data available, misuse of data, costs and time-consuming activities required.

Concerning peer review, Kratz and Strasser's (2015) survey shows that researchers of different disciplinary fields are still unsure on how data peer review should work and in which context it should occur, even if they expect that data in repositories are subject to validation. The OpenUp survey indicates that social scientists together with ICT researchers are the ones that are less satisfied with the traditional peer review process of publications.

3.1.2 Types of datasets produced

Social Sciences cover a wide range of sub-disciplines, each one with its own practices for managing and diffusing its research results. Each sub-discipline applies a wide spectrum of scientific methods that strongly influence data collection techniques. Generally, research in Social sciences falls in the category of observational and experimental studies. Data are intensive, contextual and time-dependent (Curty 2016), often requiring an extra effort to make the dataset human and machine readable. Moreover, data variety depends on the different research approaches (quantitative, qualitative and quali-quantitative) as well as research techniques (surveys, questionnaires, interviews, focus groups, etc.). Primary research data in Social sciences can also include video, audio and photos.

3.1.3 Types of dataset providers

A specific characteristic in this field is that a significant portion of data is produced for purposes other than research (Borgman 2007). These are data created by governmental bodies that have to comply with transparency regulations (such as the UK Freedom of information Act) and make data they collect publicly available. Examples of these data comprise census figures, cohort and longitudinal studies, cross national surveys, economic indicators, etc. Among governmental bodies it is worth mentioning the data produced by national statistical offices that apply standard procedures to collect and process data, provide detailed supplementary documentation to describe the datasets, and also guarantee long-term preservation. This data collection constitutes trustworthy information, on which many other studies are based representing an important data source not only for social scientists.

While these sources of information can be compared to the big data produced by STEM, long-tail data are produced by social scientists to investigate local phenomena in small collaborative groups often within interdisciplinary projects or individually. They are usually facing privacy issues that make the dataset sharing more complex.

3.1.4 Modes of sharing/publishing datasets

There is a general consensus on the difference between published and Published data (with capital letter) distinguishing between datasets available for instance on a personal website and data Published "as permanently available as possible on the Internet" (Lawrence et al. 2011) and undergone "processes that add value to the users, such as metadata creation and peer review" (Mayernik et al. 2015).

As mentioned above, currently, data validation is more accurately performed in data journals and data repositories. Therefore, validation of data is directly connected to quality measures applied by data publishers, either journals or repositories. However, analyses by

Candela et al. (2015), Carpenter (2017) show that there is room for improvement, as peer review in data journals varies a lot and is mostly focused on metadata, rather than data themselves, aiming at assessing to assess the documentation and metadata description that facilitate data reuse. Assante et al. (2016), in the analysis of generalist repositories (Zenodo, Dryad, Figshare etc.), also come to the conclusion that different criteria and quality control mechanisms are implemented, based on varied policies and/or guidelines.

In Social sciences, data publication model is mostly related to dataset submission in a data repository. In fact, there is only one data journal covering Humanities and Social sciences: "Research Data Journal (RDJ)"[6] It was created by DANS[7] in 2016 with the aim to increase the visibility of data stored in the archive and to provide more extensive and detailed documentation. This journal conforms to well established data journals in other disciplines such as Earth System Science Data, Geoscience Data Journal, and Scientific Data. It assigns a DOI to each article and provides the related DOI assigned to the dataset stored in the DANS archive, but it does not provide a standard description to cite the article. Currently eight data papers have been published and two papers refer to the field of Social sciences.

Considering trusted data repositories in Social sciences, their main feature is that of data centres that act at national level as main information sources in this field. Worth mentioning are the UK Data Archive[8], GESIS[9] and DANS. The majority of these national centralized data centres are also part of two consortia, CESSDA at a European level and ICPSR[10] at international level. These consortia provide a single access to international and national data and also develop and coordinate initiatives on standards, protocols and best practices to support data management and dissemination. Most of them provide access to data produced by governmental bodies and by research groups.

3.1.5 Modes of validating datasets

The above-mentioned data repositories are certified by the Data Seal of Approval[11] that has identified 16 requirements based on 5 criteria: data availability on the Internet, accessibility (clear rights and licenses), usability (format), reliability and identification of dataset through a persistent identifier. Note that these also correspond to the criteria used to evaluate the data themselves.

Given that data validation represents an iterative process that encompasses the entire research lifecycle, these trusted repositories provide guidelines on how to develop a data management plan at the very beginning of a research to assure data quality. Moreover, they require data producers to establish copyright and appropriate licenses, to use proper data formats and metadata schemas to facilitate access and reuse.

Trusted data repositories provide guidelines and/or templates for a correct data ingestion according to the metadata schema of DDI[12], a standard supported by the Social sciences community that facilitate data replication and/or reproduction. They assure long-term data

---

[6] Research Data Journal: https://brill.com/view/journals/rdj/rdj-overview.xml

[7] Data Archiving and Networked Services – DANS: https://dans.knaw.nl/en

[8] UK Data Archive: https://www.data-archive.ac.uk/

[9] Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen – GESIS: https://www.gesis.org/en/home/

[10] Interuniversity Consortium for Political and Social Research – ICPRS: https://www.icpsr.umich.edu/icpsrweb/

[11] Data Seal of Approval https://www.datasealofapproval.org/en/

[12] Data Documentation Initiative http://www.dcc.ac.uk/resources/metadata-standards/ddi-data-documentation-initiative

preservation and curation, develop data discovery tools (such as landing pages; Callaghan 2015), suggest users a data citation format that acknowledge data provenance. The suggestion of a data citation format represents an important feature to support data citations that are an indirect appraisal of the quality of the dataset in the post-publication phase.

Some trusted repositories have adopted tools to track data use. For instance, DANS provides data users with a validation template to rank data set available in EASY: users can provide the rating (up to five stars) to data quality, quality of documentation, completeness of the data, consistency, structure and usefulness of the file format[13].

3.2 Interviews with the Human Mortality Database managers

HMD is a well-known data source that provides detailed mortality and population data providing a detailed documentation on the methods used to analyse the raw data obtained by national statistical offices of 39 countries. HMD has numerous data users worldwide belonging to different scientific communities as well as to the business sector. Table 2 provides a brief description of HMD main characteristics focusing on particular to types of data available and related documentation.

Table 2: The Human Mortality Database in a nutshell

- The Human Mortality Database (HMD) is an open database that provides detailed, consistent and high quality data to researchers, students, journalists, policy analysts, and others interested in the history of human longevity and its prospects for the future (https://www.mortality.org/).
- HMD is a joint project of the Department of Demography at the University of California, Berkeley (UCB), and the Max Planck Institute for Demographic Research (MPIDR). Recently the HMD project is also supported by the French Institute for Demographic Studies (INED).
- Along with raw data, coming mostly from national statistical offices, HMD provides uniform death rates and complete and abridged period life tables. In addition, cohort life tables are provided when the observation period is sufficiently long to include at least one cohort observed from birth until extinction. All data are provided with the highest level of detail and include some unique information on old age mortality up to age 110.
- *Documentation available*: Methods Protocol - Country specific documentation - Guidelines for citation - User agreement - Citation report
- Country specific documentation describes in depth all necessary information to understand the population dynamics as well as the issues related to the estimation of the raw data. It also discusses any data quality issues that might arrive from the original statistics. This report is updated each time new data are analysed.
- *Type of data*: Original input data - Unsmoothed death - Population estimates - Death counts Period and cohort life tables - Life expectancy at birth and all other ages

The interviews with HMD community were conducted on the 31st of January and 1st of February 2018 at the Max Planck Institute for Demographic Research in Rostock, Germany. They were performed according to interviewees' role in HMD. The two directors, two researchers in their role of country responsible (in charge of analysing data for specific countries) were interviewed. These interviews covered the majority of HMD staff (4 out of 7).

[13] http://datareviews.dans.knaw.nl/details.php?l=en&pid=urn:nbn:nl:ui:13-0an-1ei

The following paragraphs summarize the key points of the interviews carried out with the two directors and two country specialists (CSs). Interviews of HMD staff were planned aiming at exploring the following topics:

- Origin, motivations and organizational features
- Goal and main features of the database
- Data quality assessment process
- Opinion on Open access of data

The summary content presented was revised, commented and approved by the interviewees.

3.2.1 Origin, Motivations and Organizational Features

Two previous relevant experiences guided the development of the database: the Kannisto-Thatcher Database on Old Age Mortality (KTD) at the MPIDR and the Berkeley Mortality Database (BMD), founded by John Wilmoth at UCB. Both experiences were concerned with what was at that time an emerging phenomenon of low mortality at young and adult ages, falling mortality at old ages, and greater survival to an advanced age, leading to a potential increase in the number of people exposed to degenerative diseases, which are difficult to treat or prevent. To understand this phenomenon, it was necessary to analyse and model longevity and survival of humans with a special emphasis on advanced age over a long period of time. This research needed reliable data at international level providing long-term and continuous series without gaps, running up to the highest ages, providing information on age, time, and cohort dimensions, ensuring sufficient quality and comparability across time and populations. HMD was therefore developed to answer this scientific question providing a methodology based on the previous mentioned experiences as well as freely available high-quality data [26]. The two HMD directors explained that “the collaboration was originally, (and still is), based on a small, very well-established group of internationally based demographers who were willing to serve the scientific community interested in demographic studies”. The workload is equally distributed among the team that comprises CSs who have high-level competences on demographic development of a set of specific countries and are responsible for collecting and analysing data from the related national statistical offices. Other tasks comprise the development of computer codes, which are also made freely available to the end user who wants to reproduce the analysis, as well as the management of the website. Strong collaboration pertains to the data quality process performed before data are publicly available, which constitutes a form of internal pre-publishing peer review process. During the interview the two directors agreed that “trust among the team and scientific curiosity are the drivers of this successful cooperation”, that only recently was formalized by a Memorandum of understanding that reports the common lines of action of the cooperating partners.

3.2.2 Goals and Main Features of the Database

The main goal of HMD is to support research on human mortality and longevity providing open data on 39 countries and some sub-areas and sub-populations with series starting as early as 1751 (i.e. Sweden) and covering more than 100 years for 16 populations. Birth and death counts are generally based on data from national vital registration systems, while data on population are based on the national census and estimates between censuses. However, differences may exist among countries in the periodicity of census, methods and definition used as well as in data format. Moreover, some countries have experienced changes in their territorial boundaries, have suffered substantial loss during war periods and/or faced substantial consistent migration over the period covered by HMD. For these reasons, as underlined by the two directors, HMD has developed a methodology to produce detailed death counts and population estimates, to correct mortality estimates at old ages, and to build high quality life tables (as described in detail in the Methods protocol). "All HMD data are prepared using this standard methodology. This assures comparability in time and across countries". The two Country specialists explained, that "when special methods are needed to accommodate issues in data availability, this is documented in the country-specific documentation as well as reported in summary tables"[14] Country-specific details related to the data quality and statistical system in each country are therefore documented in the country-specific Background and Documentation file accessible from each country webpage. The application of these thorough procedures, "the punctual explanation of the estimations and refinements of data sources make this database different from other sources providing mortality rates". These procedures guarantee a uniform analysis of raw data, facilitating the comparability across time and space, while the detailed documentation and the availability of source data allow end user to reproduce the analysis. The HMD team has also developed software code that guide them in the evaluation of data quality as well as software packages that facilitate end users to import and working with HMD data. These tools are freely available to end users along with technical reports explaining how to use these scripts[15]. This is another value-adding feature of HMD.

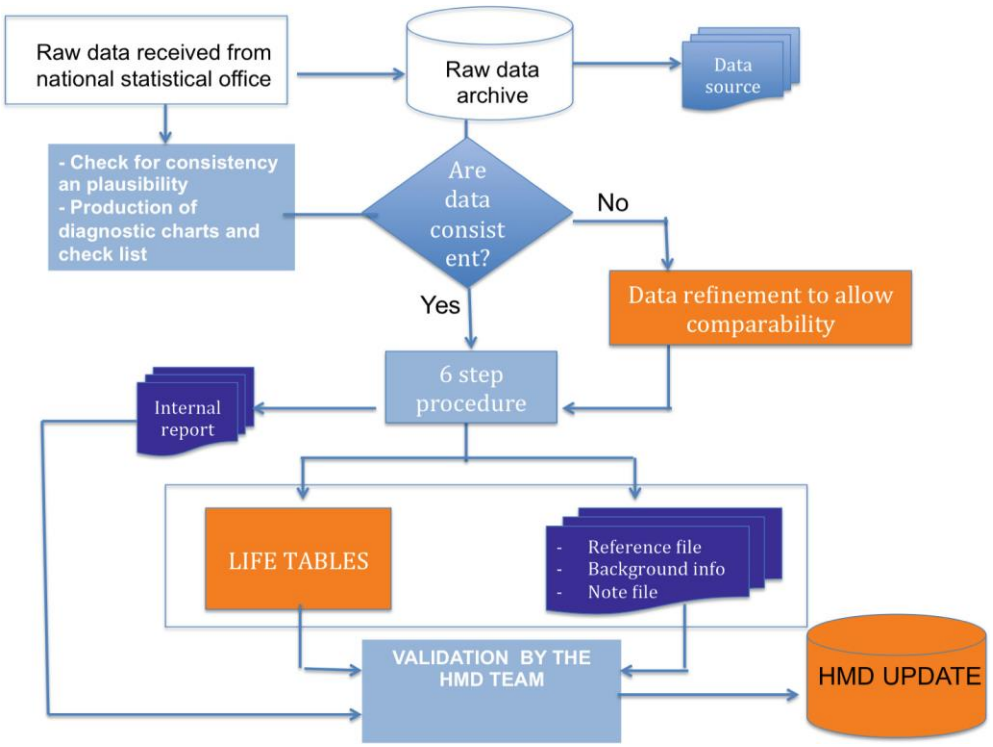
3.2.3 The HMD Data Quality Assessment Process

The HMD team has developed a set of procedural steps to ensure data quality. This important topic was addressed in the interviews with the two directors and particularly explored in the interviews with the CSs. An activity diagram that reconstructed the workflow of the activities performed before data publication was presented to the CSs and discussed to have further insights in the procedures adopted to assess data quality. This intended to explore whether collaborative activities resembling a peer review process could be tracked in HMD data quality assessment. A high-level description resulting from the interviews is provided in Figure 1.

[14] https://www.mortality.org/Public/Docs/SpecialMethods.pdf

[15] https://www.demogr.mpg.de/en/projects_publications/publications_1904/mpidr_technical_reports/all.htm

Figure 1. Data quality assessment process



Source: CNR-IRPPS elaboration.

During the interviews the CSs explained that each country or area is assigned to an individual researcher, a CS, who maintains a close relationship with a local expert generally at national statistical offices, and has an extensive knowledge of the population dynamics as well as how data are collected at national level. A CS is responsible for the first quality checks that evaluate consistency and plausibility of input data, prepares pre-calculation file (Lexis file) and analyses the results on the basis of a predefined data quality checklist and diagnostic charts that help him/her to explore unusual fluctuations and/or any other issues in data sources. The results of this analysis are shared within the HMD community via an internal report and are the basis for the application of the six-step procedure to produce the complete data series (exposures to risk, death rates, life expectancy and other life tables). Before data are published, the HMD team performs an additional phase of validation. These activities are crucial especially when a new country has to be included in HMD. However, they constitute a routine procedure every time data are updated. "In cases of unexpected changes in national statistical systems or in regimes of national statistical registration, the updating procedures are non-trivial". All steps in the computing of data analysis are documented in detail and made available to end users in the different files (Background and documentation, Data source and Explanatory Notes). According to the CSs interviewed, this is the distinctive feature of HMD: "Data refinements and harmonization that allows comparison across countries are documented in detail so that researchers in this field are aware of possible problems in the data and know how these issues have been solved".

3.2.4 Opinion on Open Access of Data and Peer Review

HMD management team declared that open access and open data in particular are very important for the development of demographic studies. Although they have no official statements on open policy, since its beginning, HMD provided open access data, based on a user agreement indicating that the data in the HMD are provided free of charge to all individuals who request access to the database"[16]. Moreover, users are required to cite the database in their publications, following the citation guidelines provided by HMD[17]. Citations tracked through Google scholar are also reported in the website, and further steps to improve their collection are planned in the next future.

When asked about long preservation of data, it emerged that the two HMD directors are dependent on funds. At the moment MPIDR support their activities ("MPIDR researchers are allowed to spend half of their work time on HMD"), while the UCB team has to provide its own funds. A clear commitment of the organization would therefore be very important and would also mean a clear recognition of their activities.

Between the lines, it emerged that publication of scientific papers are generally considered more important than managing a database. In their opinion, "the analysis of data, their quality check is not only a service for the community of reference but is a research activity in itself." The majority of the interviewees has heard about open review of journals but has little knowledge on all its traits. If they see a similarity with peer review of data, this is associated in particular with transparency as a means of reconstructing the methods and procedures used for the data analysis.

3.2.5 Conclusion

Some important indications emerged from the analysis of the interviews that can drive the adoption of data quality assessment, and hence peer review, as well as some principles that can incentivize other scientific communities to share their research data. As stated by the HMD interviewees, the guiding principles to create an open access database were: comparability, flexibility, accessibility and reproducibility. Comparability was reached using a uniform, scientific methodology to calculate the various statistics of the 39 countries included in the database. Flexibility was achieved in the analysis of results using a uniform set of procedures for each population, but at the same time giving significant attention to each population in terms of its history and socio-political development. This have direct implications on the formats availability of output data series. This is achieved thanks to the experiences and knowledge of country specialists, that are persons in charge of collecting data from a specific number of countries, who interact with statistical offices, check data consistency and provide population statistics together with a country report that explains specificity and motivation of analysis. Accessibility was guaranteed from the beginning by free of charge access of data, as well as by the provision of data in an open, no-proprietary format. Reproducibility is provided by the reconstruction of the data lifecycle that includes the availability of raw data, the method applied, the related results as well as the explanatory documentation. One of the main successful features of HMD is its transparent way of data managing and sharing that has two central phases of data validation. The first one is carried out by the CSs, who analyse the raw data according to a common predefined checklist that verifies consistency and plausibility of

[16] https://www.mortality.org/Public/UserAgreement.php
[17] https://www.mortality.org/Public/CitationGuidelines.php

data. The second one is carried out in a collaborative way within the HMD team that validate the statistics before their publication, each time the database is updated.

Moreover, another successful component of HMD was its collaborative approach that is based on a strong scientific interest in the field as well as on the trust among the involved community that only recently has formally signed a Memorandum of understanding.

The interviews also highlighted some indications that confirm some concerns already mentioned by other surveys. Interviewees stressed the importance of having a strong commitment of the organization in supporting the development of data infrastructures. This pertains to different aspects: a long-term financial support (beyond the project duration), a policy endorsement on open data as well as a formal recognition of scientists for the efforts in data curation and quality assurance.

3.3 Human Mortality Database users' survey

3.3.1 The sample

The survey was open to all HMD registered users. The questionnaire was sent to the users' mail address provided during registration. The number of registered users may be biased by multiple accounts and changes in mail address.

More than 35500 invitations have been sent, 1049 came back for incorrect address, 1553 completed the questionnaire. The response rate was 4.5%.

The survey includes two filter questions (q. 9 and q. 12). The first one allows us to single out registered users from the ones regularly accessing the database (i.e. 1408 active users), while the second one distinguishes between users, who only visualize data (i.e. 170) from those that download/copy them to make further analysis. Specific questions on the practices on data use were therefore asked to the remaining 1238 respondents.

3.3.2 Respondent's demographic profile

An overview of the respondent's demographic profile is given in Table 3. The majority of respondents are male (68.8%) and they fall mainly into two age groups (43.7% of 20-39 years old and 38.6% of 40-59 age range). Most female respondents (29.5%) fall in the same age groups (respectively 16.9% and 10.5%). Respondents residing in Europe are 59.9% of total responses, followed by America (25.3%), Asia (10.6%), Oceania (2.9%), and Africa (1.2%).

Table 3: Respondents' demographic profile

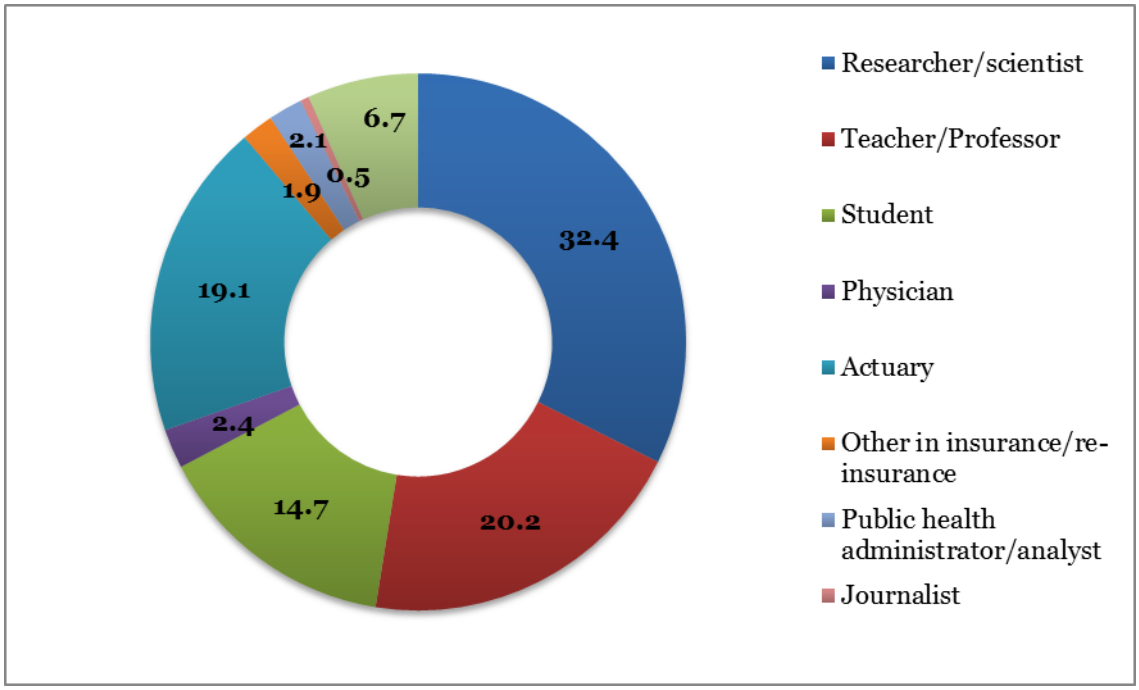
	no.	**%**
Sex		
Male	1069	68.8
Female	458	29.5
Prefer not to answer	26	1.7
Total	1553	100.0
Age		
<20	4	0.3
21-39	679	43.7
40-59	600	38.6
60+	270	17.4
Total	1553	100.0

Country of residence		
Africa	19	1.2
America	393	25.3
Asia	165	10.6
Europe	931	59.9
Oceania	45	2.9
Total	1553	100.0

Source: CNR-IRPPS elaboration.

Concerning respondents' occupation (Figure 2), the majority of them are researchers/scientists (32.4%,) teachers/professors (20.2%), students (14.7%,) and actuaries (19.1%) outlining a consistent, well defined type of users that all together covers the 86.4% of respondents. 6.7% of *Other* comprises a high variety of occupation such as data analysts, employees, citizens and retired people.

Figure 2: Respondents' occupation (%)



Source: CNR-IRPPS elaboration

On the basis of these results and in line with the scope of the survey, results are stratified by occupation in the following categories:

- *Scientist* comprises teachers/professors and researchers;
- *Actuary* includes respondents reporting to be actuary as well as those belonging to "other in insurance and re-insurance;
- *Student*
- *Other* includes the answers of the pre-defined questionnaire categories: Physicians, Journalists, Public health administrators/analysists as well as the respondents that specified different types of occupation in the free text variable.

Table 2 shows the users' demographic profile by the above-mentioned category of occupation. An almost gender balanced composition is present among students (54.1% male, 45.4% female), whose 90.4% falls in the age range 20-39 years old. In the other categories the majority of users are male. Concerning age group, the category of scientists and others show a similar distribution in the ranges 40-59 and >60 years (respectively 45.2% and 22.4; 47.5% and 26%), while actuaries are almost totally concentrated in the two ranges 20-39 and 40-59 years (49.2% and 40.4%).

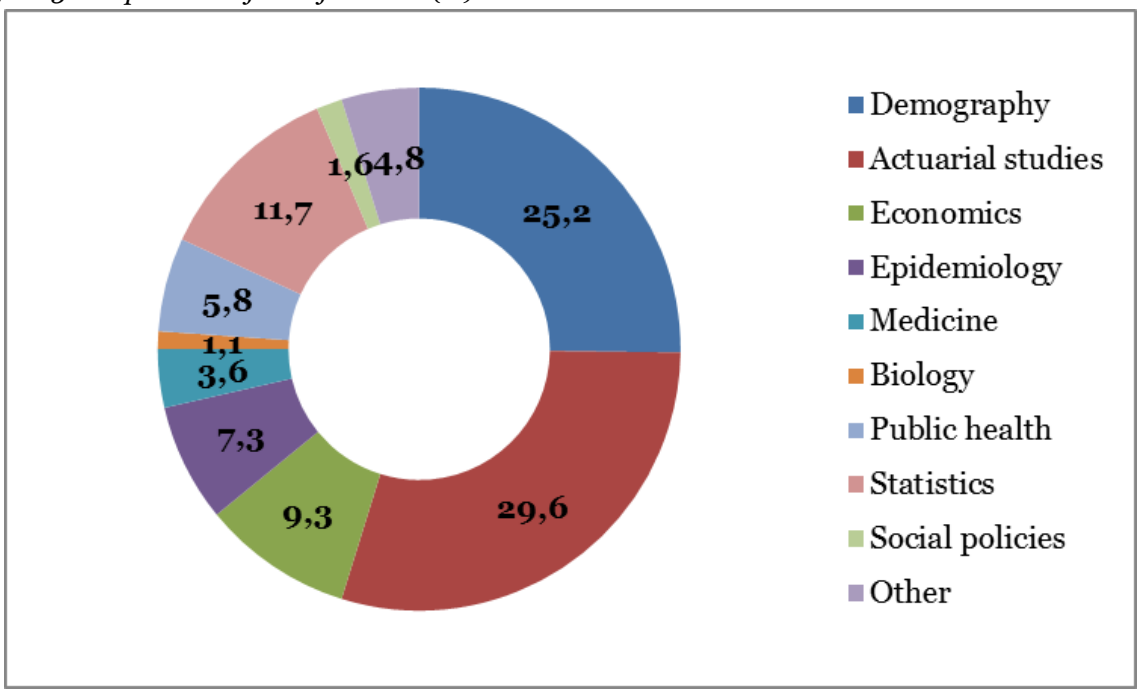
Table 4: Users' demographic profile by category of occupation

	Scientist		Actuary		Student		Other	
	no.	%	no.	%	no.	%	no.	%
Sex								
Male	571	70	238	72.8	124	54.1	136	75.1
Female	231	28.3	82	25.1	104	45.4	41	22.7
Prefer not to answer	14	1.7	7	2.1	1	0.4	4	2.2
	816	100.0	327	100.0	229	100.0	181	100.0
Age								
<20	-	-	-	-	3	1.3	1	0.6
21-39	264	32.4	161	49.2	207	90.4	47	26
40-59	369	45.2	132	40.4	13	5.7	86	47.5
>60	183	22.4	34	10.4	6	2.6	47	26
Total	816	100.0	327	100.0	229	100.0	181	100.0

Source: CNR-IRPPS elaboration.

Responses by discipline (Figure 3) show that the main fields of interest are actuarial studies (29.6%.), demography (25,2%) and statistics (11.7%.).

Figure 3: Respondents' field of interest (%)



Source: CNR-IRPPS elaboration.

In term of disciplines (Table 5), within the four categories of occupation, the major field of interest for scientists, students and others is Demography (respectively 33.1%, 30.1% and 23.8%), while as expected almost all actuaries are interested in actuarial studies (92.4%).

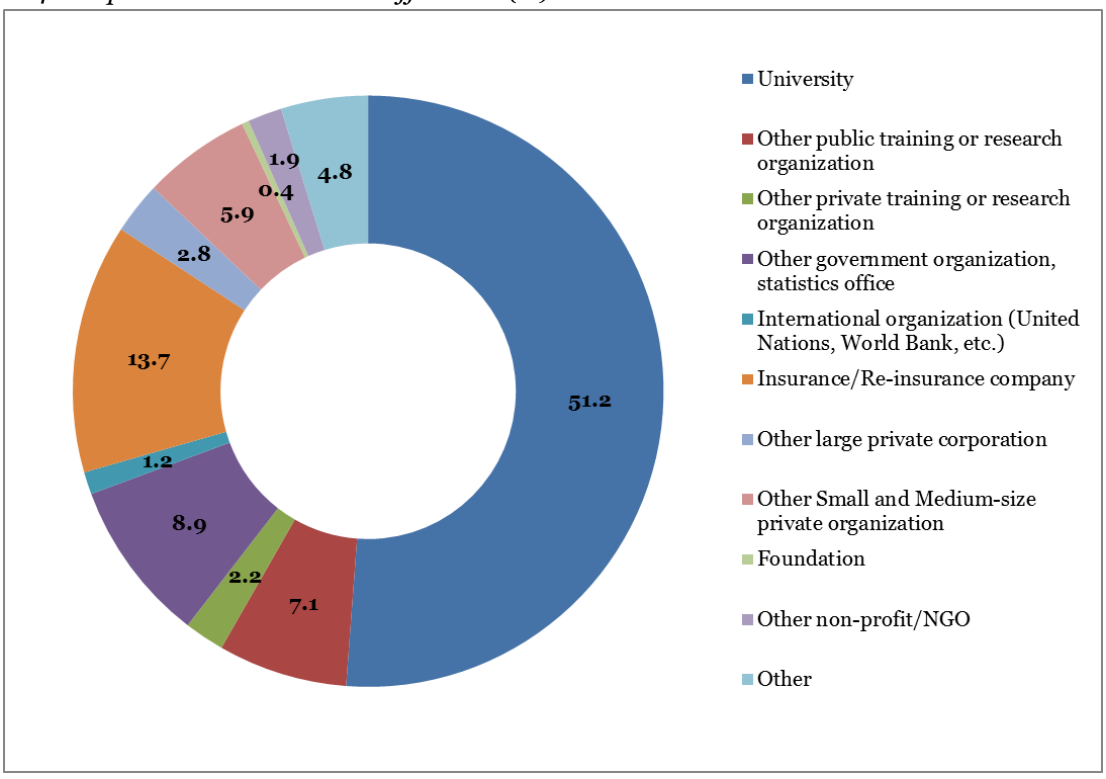
Table 5: Field of interest by category of occupation

	Scientist		Actuary		Student		Other	
	no.	%	no.	%	no.	%	no.	%
Demography	270	33.1	9	2.8	69	30.1	43	23.8
Actuarial studies	94	11.5	302	92.4	52	22.7	12	6.6
Economics	111	13.6	2	0.6	20	8.7	11	6.1
Epidemiology	87	10.7	-	-	12	5.2	14	7.7
Medicine	28	3.4	-	-	2	0.9	26	14.4
Biology	14	1.7	-	-	1	0.4	2	1.1
Public health	54	6.6	1	0.3	13	5.7	22	12.2
Statistics	105	12.9	12	3.7	38	16.6	27	14.9
Social policies	14	1.7	1	0.3	4	1.7	6	3.3
Other	39	4.8	-	-	18	7.9	18	9.9
Total	816	100.0	327	100.0	229	100.0	181	100.0

Source: CNR-IRPPS elaboration.

Moreover, 60.2% of respondents work at research institutions, in detail by University (51.2%), Other public training or research organization (7.1%) and Other private training or research organization (2.2%), while 13.7% are employed in Insurance/Re-insurance companies. Among the female respondents, most of them work at University (16.0%),

Figure 4: Respondents' institutional affiliation (%)



Source: CNR-IRPPS elaboration.

The distribution by occupation category (Table 6) confirms the results mentioned above.

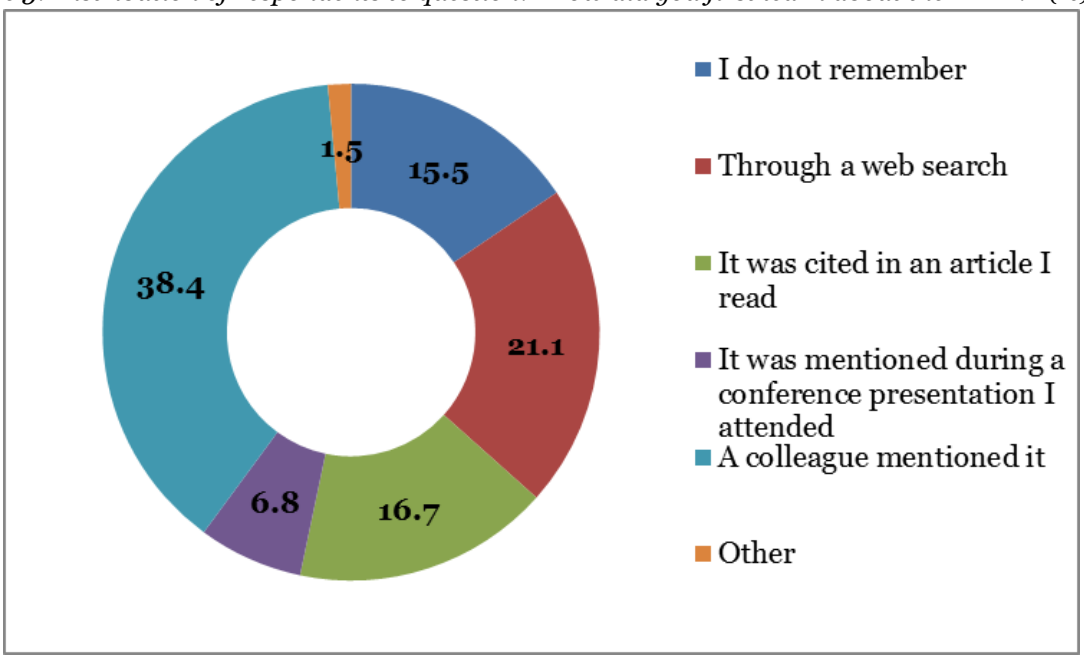
Table 6: Type of institution by category of occupation

	Scientist		Actuary		Student		Other	
	no.	%	no.	%	no.	%	no.	%
University	538	65.9	19	5.8	215	93.9	23	12.7
Other public training or research organization	97	11.9	-	-	4	1.7	9	5
Other private training or research organization	22	2.7	7	2.1	-	-	5	2.8
Other government organization, Statistics office	63	7.7	24	7.3	2	0.9	49	27.1
International organization (United Nations, World Bank, etc.)	14	1.7	1	0.3	-	-	4	2.2
Insurance/Re-insurance company	11	1.3	194	59.3	2	0.9	5	2.8
Other large private corporation	7	0.9	25	7.6	1	0.4	11	6.1
Other Small and Medium-size private organization	28	3.4	42	12.8	1	0.4	21	11.6
Foundation	5	0.6	-	-	-	-	1	0.6
Other non-profit/NGO	13	1.6	4	1.2	2	0.9	10	5.5
Other	18	2.2	11	3.4	2	0.9	43	23.8
Total	816	100.0	327	100.0	229	100.0	181	100.0

Source: CNR-IRPPS elaboration.

If we consider how respondents have learnt about HMD database, a consistent number of the answers (38.4%) indicate that it was mentioned by a colleague/teacher/professor, or found in a web search (21.1%) or cited in an article (16.7%).

Figure 5: Distribution of respondents to question: "How did you first learn about the HMD?" (%)



Source: CNR-IRPPS elaboration.

The tendency of learning about HMD through word of mouth is particularly diffused in the first three occupation categories, while the set of people belonging to *Other* usually get to know HMD via a web search (Table 7).

Table 7: Becoming aware of HMD by category of occupation

	Scientist		Actuary		Student		Other	
	no.	%	no.	%	no.	%	no.	%
A colleague/professor mentioned it	281	34.4	149	45.6	129	56.3	37	20.4
I do not remember	138	16.9	51	15.6	16	7	36	19.9
It was cited in an article I read	150	18.4	51	15.6	32	14	27	14.9
It was mentioned during a conference presentation I attended	58	7.1	31	9.5	10	4.4	6	3.3
Through a web search	175	21.4	42	12.8	39	17.0	72	39.8
Other	14	1.7	3	0.9	3	1.3	3	1.7
Total	816	100.0	327	100.0	229	100.0	181	100.0

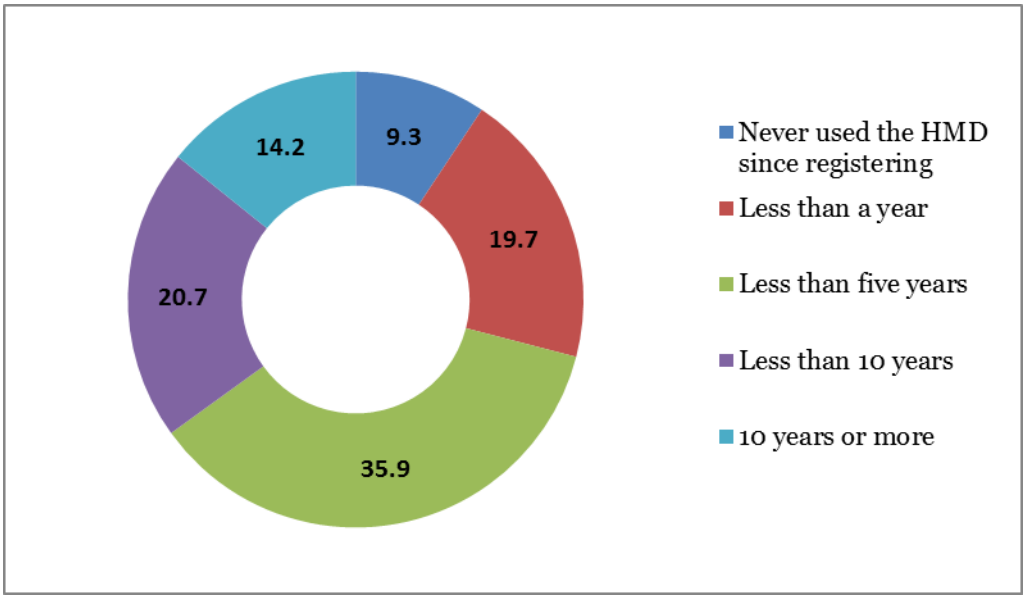
Source: CNR-IRPPS elaboration.

3.3.3 General information on access

This paragraph summarizes length and frequency of use as well as countries that attract more interest by HMD users.

Considering length of HMD use (Q.9), the majority of responses (55.6 %) registered less than 5 years and less than a year, while 34.9% are long-standing users (less than 10 years and 10 years or more). 9.3% of respondents declare that they never used HMD after registration, therefore they have not completed the remaining questions. Thus, the analysis of further questions is based on the sample of current users, that is 1408 respondents.

Figure 6: Distribution of respondents to question: "For how long have you been an HMD user?" (%)



Source: CNR-IRPPS elaboration.

Scientists and actuaries represent the category that prevalently use HMD for a longer time, while the other two categories tend to be recent users of the database.

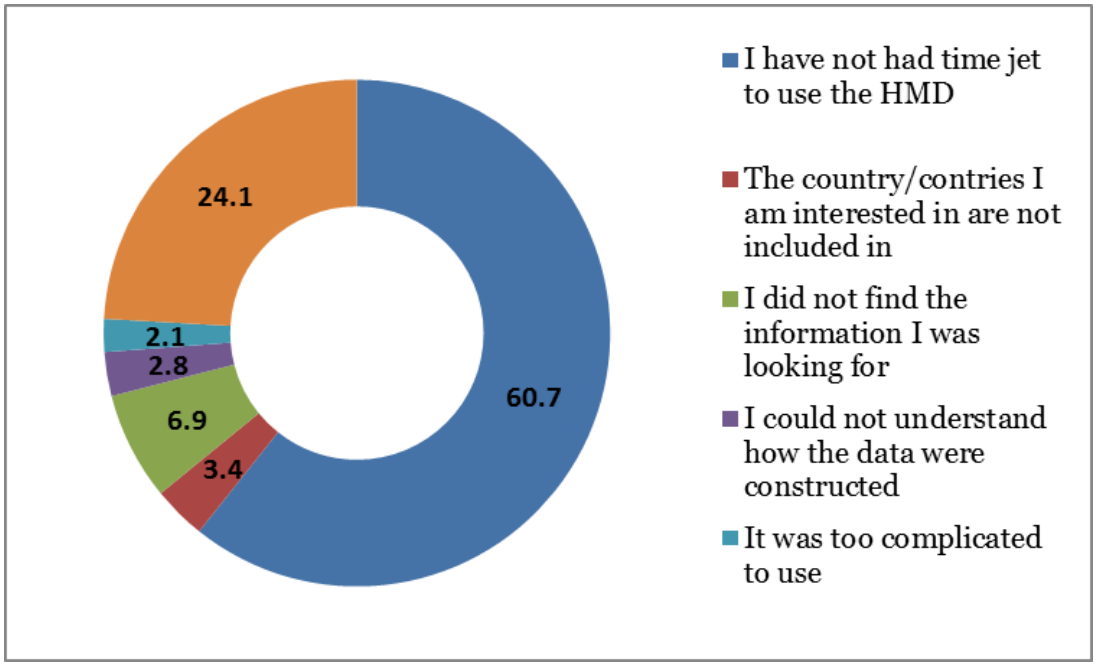
Table 8: Length of HMD use by category of occupation

	Scientist		Actuary		Student		Other	
	no.	%	no.	%	no.	%	no.	%
Never used the HMD since registering	80	9.8	17	5.2	20	8.7	28	15.5
Less than a year	114	14	40	12.2	110	48	42	23.2
Less than five years	264	32.4	146	44.6	87	38	62	34.3
Less than 10 years	198	24.3	82	25.1	11	4.8	31	17.1
10 years or more	160	19.6	42	12.8	1	0.4	18	9.9
Total	816	100.0	327	100.0	229	100.0	181	100.0

Source: CNR-IRPPS elaboration.

Considering users who registered but have never used HMD, the majority of them reported that they have not had time to do it yet (60.7%) and only 6.9% indicate that they have not found the information they wanted. The few respondents that added some comments did not report any problem in using the database and did generally mentioned a decline in interest in this topic.

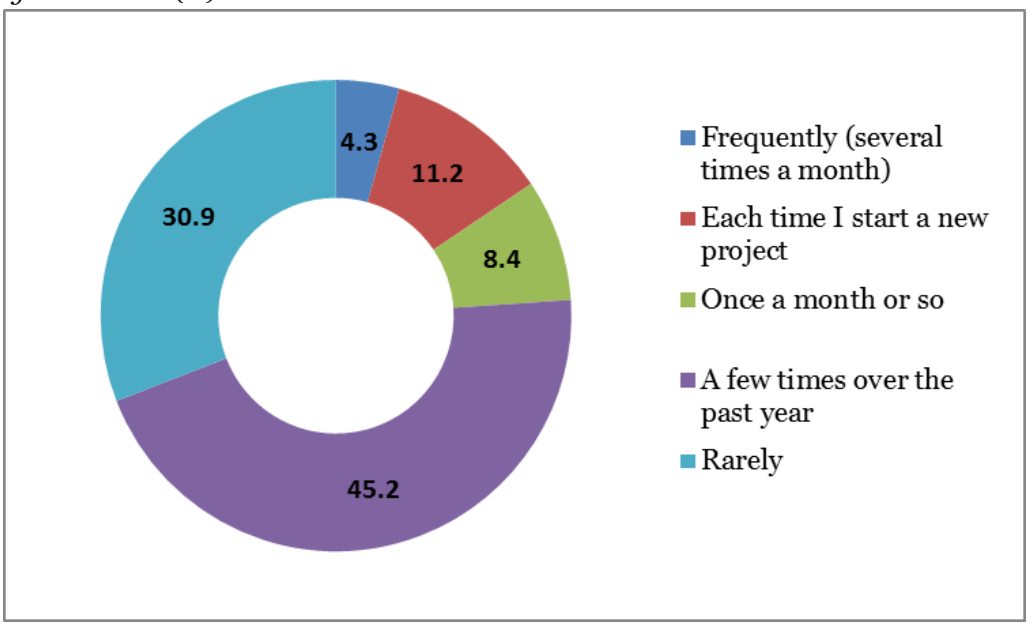
Figure 7: Distribution of respondents to question: "Please tell us more about why you have never used the HMD" (%)



Source: CNR-IRPPS elaboration.

Related to the frequency of accessing the database (Q.10), the majority of respondents (76.1%) consult it a few times over the past (45.2%) and rarely (30.9%). This can depend on the types of data and but also by their updating that it is generally done at 2- to 3-year intervals. (Barbieri 2015).

Figure 8: Distribution of respondents to question: "How frequently do you access the Human Mortality Database?" (%)



Source: CNR-IRPPS elaboration.

No relevant differences can be detected, if considering frequency of use distributed by occupation category (Table 9).

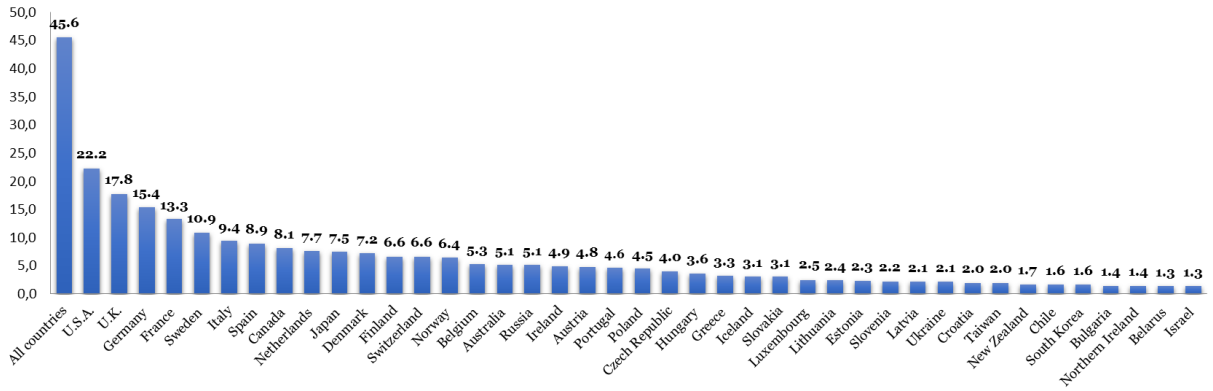
Table 9: Frequency of use by category of occupation

	Scientist		Actuary		Student		Other	
	no.	%	no.	%	no.	%	no.	%
Frequently (several times a month)	40	5.4	5	1.6	14	6.7	2	1.3
Each time I start a new project	95	12.9	26	8.4	30	14.4	7	4.6
Once a month or so	67	9.1	20	6.5	19	9.1	12	7.8
A few times over the past year	341	46.3	156	50.3	74	35.4	65	42.5
Rarely	193	26.2	103	33.2	72	34.4	67	43.8
Total	736	100.0	310	100.0	209	100.0	153	100.0

Source: CNR-IRPPS elaboration.

When asked about country of interest (Figure 9), a high number of respondents (45.6%) report that they access data related to all countries available in HMD. The countries whose data are the most accessed ones are: the U.S.A (22.2%), followed by the U.K (17.8%), Germany (15.4%), France (13.3%), and Sweden (10.9%).

Figure 9: Distribution of respondents by question: "Which HMD countries/regions are you most interested in?" Multiple responses (%)



Source: CNR-IRPPS elaboration.

Table 10 points out that HMD users tend to be interested in data covering all countries (45.6%) or *vice versa* are specifically focused on single country (25.7%). Between these two extremes, data analysed by deciles, confirm that the rest of users tend to get information on a limited number of countries, at most between 2 and 11 countries.

Table 10: Number of countries accessed

Number of accessed countries	**Users no.**	**%**
1	362	25.7
2-11	334	23.6
12-21	41	3
22-31	27	1.9
>32	2	0.2
All countries	642	45.6
Total	1408	100.0

Source: CNR-IRPPS elaboration

The distribution by occupation category further details specific features. Differently from the other three categories, actuaries tend to be interested in specific countries (in particular the UK 32.6% and the USA 31.3%) and less focused on the data comparison of all countries

Table 11: Countries accessed by category of occupation

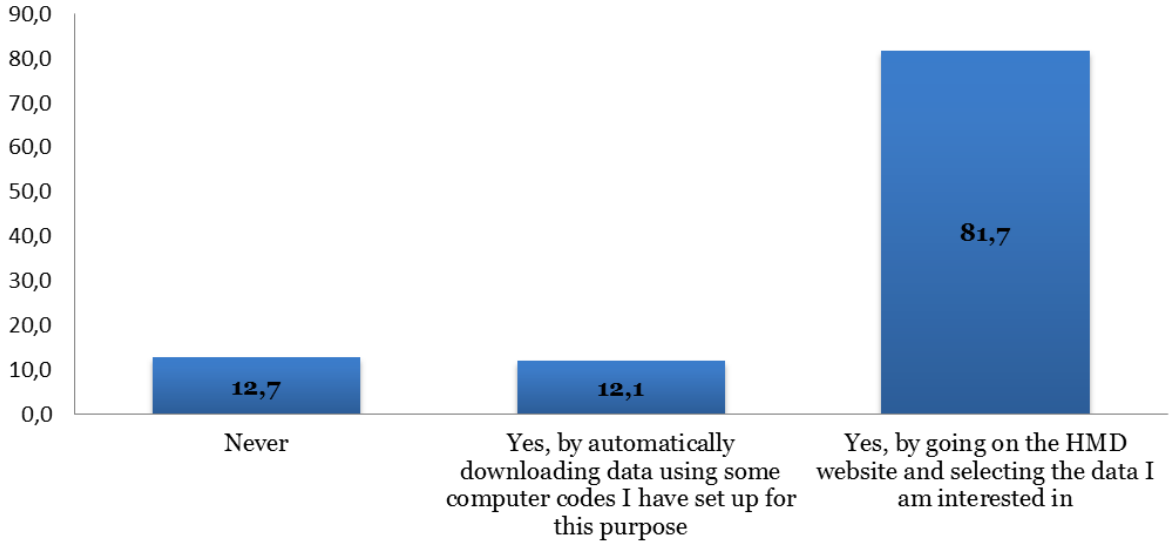
	Scientist		Actuary		Student		Other	
	no.	%	no.	%	no.	%	no.	%
All countries	399	54.2	81	26.1	100	47.8	62	40.5
Australia	28	3.8	25	8.1	10	4.8	9	5.9
Austria	36	4.9	19	6.1	10	4.8	3	2
Belarus	14	1.9	1	0.3	3	1.4	1	0.7
Belgium	38	5.2	23	7.4	10	4.8	4	2.6
Bulgaria	13	1.8	4	1.3	2	1	1	0.7
Canada	41	5.6	50	16.1	12	5.7	11	7.2
Chile	7	1	10	3.2	3	1.4	3	2
Croatia	18	2.4	5	1.6	2	1	3	2
Czech Republic	34	4.6	8	2.6	10	4.8	5	3.3
Denmark	56	7.6	23	7.4	17	8.1	5	3.3
Estonia	23	3.1	3	1	5	2.4	2	1.3
Finland	52	7.1	21	6.8	13	6.2	7	4.6
France	87	11.8	57	18.4	30	14.4	13	8.5
Germany	111	15.1	63	20.3	29	13.9	14	9.2
Greece	29	3.9	8	2.6	5	2.4	4	2.6
Hungary	29	3.9	13	4.2	6	2.9	3	2
Iceland	27	3.7	10	3.2	3	1.4	3	2
Ireland	30	4.1	30	9.7	6	2.9	3	2
Israel	7	1	8	2.6	3	1.4	1	0.7
Italy	66	9	38	12.3	20	9.6	8	5.2
Japan	48	6.5	33	10.6	16	7.7	9	5.9
Latvia	19	2.6	6	1.9	3	1.4	2	1.3
Lithuania	22	3	6	1.9	4	1.9	2	1.3
Luxembourg	20	2.7	7	2.3	6	2.9	2	1.3
Netherlands	51	6.9	37	11.9	15	7.2	5	3.3
New Zealand	12	1.6	7	2.3	4	1.9	1	0.7
Northern Ireland	13	1.8	4	1.3	2	1	1	0.7
Norway	50	6.8	21	6.8	13	6.2	6	3.9
Poland	39	5.3	14	4.5	6	2.9	4	2.6
Portugal	31	4.2	20	6.5	10	4.8	4	2.6
Russia	38	5.2	13	4.2	14	6.7	7	4.6
Slovakia	26	3.5	9	2.9	6	2.9	2	1.3
Slovenia	19	2.6	8	2.6	3	1.4	1	0.7
South Korea	54	7.3	45	14.5	15	7.2	12	7.8
Spain	12	1.6	7	2.3	1	0.5	3	2
Sweden	88	12	35	11.3	19	9.1	11	7.2
Switzerland	36	4.9	41	13.2	14	6.7	2	1.3
Taiwan	17	2.3	7	2.3	3	1.4	1	0.7
U.K.	109	14.8	101	32.6	22	10.5	18	11.8
U.S.A.	147	20	97	31.3	30	14.4	39	25.5
Ukraine	16	2.2	5	1.6	8	3.8	1	0.7

Source: CNR-IRPPS elaboration.

3.3.4 Modes of dataset acquisition

The second section investigated the types and acquisition mode of HMD dataset (Q.12, Q.16). A specific question (Q.12) about data acquisition mode leads to the distinction between registered users who usually only consult HMD data and those who download and/or copy files from HMD website. This is a first indicator of data usage. 12.1% of respondents affirm that they never download/copy files, while 81.7% declare that they only select data from HMD website and 12.7% automatically download data using some computer codes.

Figure 10: Distribution of respondents by question: "Have you ever downloaded/copied files from the HMD website?" Multiple responses (%)



Source: CNR-IRPPS elaboration.

This is slightly different if we consider the distribution by occupation category, in which scientists (14.7%) and students (15.4%) tend to retrieve data using some computer codes more than the actuaries (10.6%). However, the general tendency of HMD users is to select the data they are interested in on the website.

Table 12. Data acquisition by category of occupation

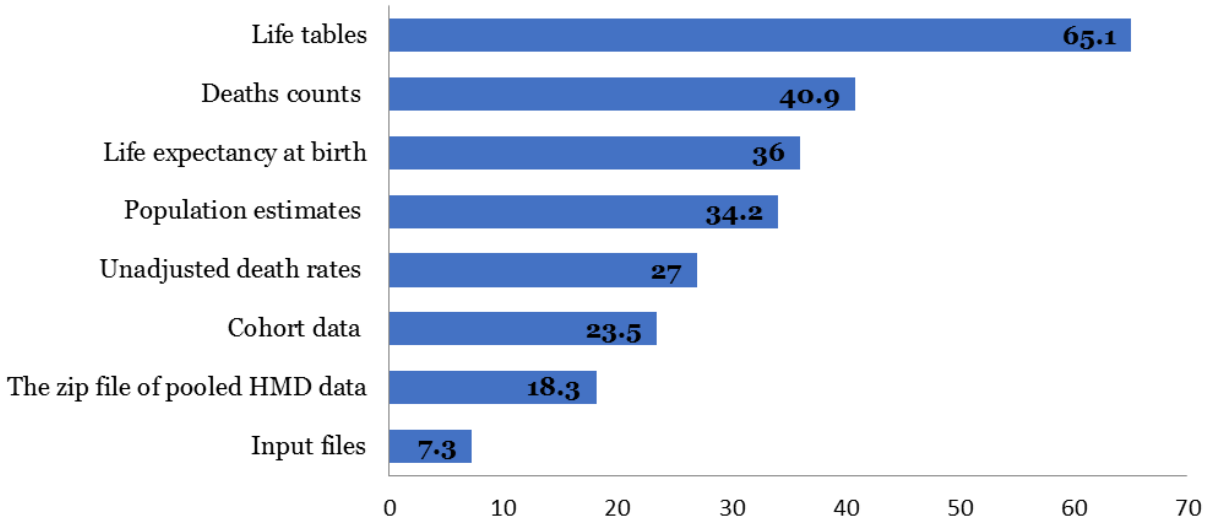
	Scientist		Actuary		Student		Other	
	no.	%	no.	%	no.	%	no.	%
Never	81	11	32	10.3	21	10	36	23.5
By going on the HMD website and selecting the data I am interested in	602	81.8	266	85.8	168	80.4	14	74.5
By automatically downloading data using some computer codes I have set up for this purpose	108	14.7	33	10.6	32	15.3	6	3.9

Source: CNR-IRPPS elaboration.

To explore which types of dataset are the most downloaded, respondents were asked about their preferences (multiple choice allowed, respondents are equal to 1238, thus excluding respondents who answered *never* in the previous question). Figure 10 shows that 65.1% of all respondents indicate life table, followed by death counts (40.9%), life expectancy at birth (36%), population estimates (34.2%), unadjusted death rates (27%), cohort data (23.5%) and

zip file of pooled HMD data (18.3%). It is interesting to note that only 7.3% of respondents usually access input files. As mentioned above, these are the baseline data on which HMD results are computed. This is a probable indicator of the reliability of HMD data, as users usually do not have the need to access input data to reproduce the analysis.

Figure 11. Distribution of respondents to question: "Which type of HMD files have you downloaded over the past 12 months (for any country)?" Multiple responses (%)



Source: CNR-IRPPS elaboration.

Table 13 analyses whether users download more than one type of file. Only a minority of respondents use more the three types of file.

Table 13: Number of downloaded files

Number of type of HMD files downloaded	**Users no.**	**%**
1	443	35.8
2	265	21.4
3	238	19.2
4	144	11.6
5	73	5.9
6	45	3.6
7	16	1.3
8	14	1.1
Total	1238	100.0

Source: CNR-IRPPS elaboration.

No remarkable differences are detected, when considering the type of downloaded files distributed by occupational category. The only exception is that actuaries use life expectancy data and cohort data less than the other three categories.

Table 14: Files downloaded by category of occupation

	Scientist		Actuary		Student		Other	
	no.	%	no.	%	no.	%	no.	%
The zip file of pooled HMD data	124	18.9	45	16.2	36	19.1	21	17.9
Input files	49	7.5	19	6.8	14	7.4	8	6.8
Life tables	420	64.1	182	65.5	126	67.0	78	66.7
Unadjusted death rates	174	26.6	90	32.4	42	22.3	28	23.9
Population estimates	245	37.4	97	34.9	51	27.1	30	25.6
Deaths counts	268	40.9	126	45.3	83	44.1	29	24.8
Life expectancy at birth	259	39.5	68	24.5	72	38.3	47	40.2
Cohort data	182	27.8	41	14.7	40	21.3	28	23.9

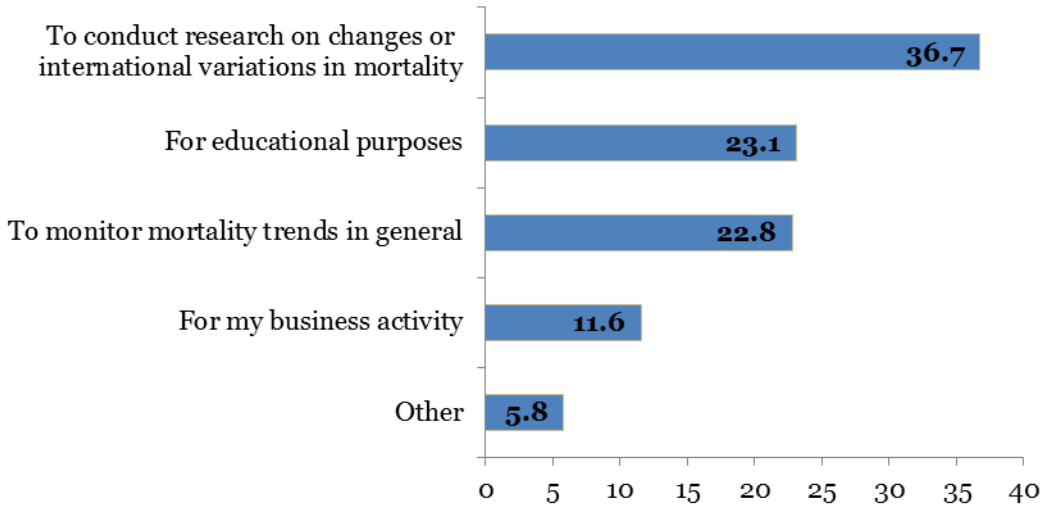
Source: CNR-IRPPS elaboration.

3.3.5 Use of dataset

The third section of the questionnaire analyses why and how HMD data are used (Q.14, Q.15, Q.18 and Q.19).

Considering the purposes of accessing datasets (multiple choice allowed), 36.7% of respondents use HMD for research in mortality, 23.1% for educational purposes, 22.8% to monitor mortality trends, while 11.6% indicate that they use data for business activity.

Figure 12: Distribution of respondents to question: "Your main purpose in using the HMD is?"(%)



Source: CNR-IRPPS elaboration.

Table 15 provides results that are coherent with HMD occupational categories. The majority of scientists use the database to conduct research, actuaries have the aim of both monitoring mortality trends and carry out their business activities, while students consult HMD for educational purposes.

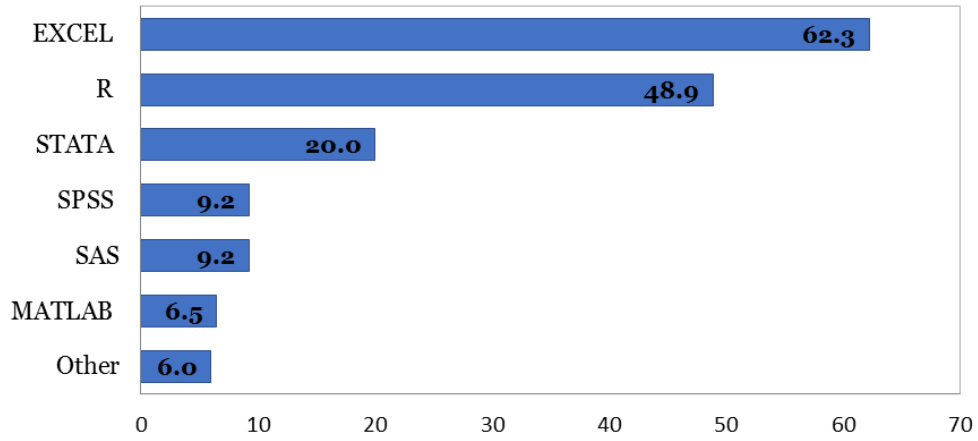
Table 15. Respondents' purposes by category of occupation

	Scientist		Actuary		Student		Other	
	no.	%	no.	%	no.	%	no.	%
To monitor mortality trends in general	142	19.3	117	37.7	23	11.0	39	25.5
To conduct research on changes or international variations in mortality	389	52.9	41	13.2	52	24.9	35	22.9
For educational purposes	140	19,0	25	8.1	123	58.9	37	24.2
For my business activity	21	2.9	116	37.4	3	1.4	23	15.0
Other	44	6.0	11	3.5	8	3.8	19	12.4

Source: CNR-IRPPS elaboration

HMD database share data contents in ASCII text files, imported into Excel tables, or into a statistical package (e.g., R, SAS, Stata, SPSS, etc). Figure 13 shows that when respondents were asked on the type of software used to process HMD data (multiple choice allowed), the most frequent answers are Excel (62.3%) and R (48.9%).

Figure 13: Distribution of respondents to question: "Which software do you use to process HMD data?" Multiple responses (%)



Source: CNR-IRPPS elaboration.

Table 16 shows whether respondents use more than one type of software. Apart from a majority of respondents that use only one software, there is a consistent number of them that use two or three different software.

Table 16: Number of software used

Number software used	Users no	%
1	692	55.9
2	365	29.5
3	145	11.7
4	32	2.6
5	3	0.2
6	1	0.1
Total	1238	100.0

Source: CNR-IRPPS elaboration.

The distribution by occupation category shows a clear preference in the use of Excel, which increases in particular in the case of actuaries and others. R is generally the second mostly used software. The other types of software are prevalently used especially by scientists and students, even if with different percentages.

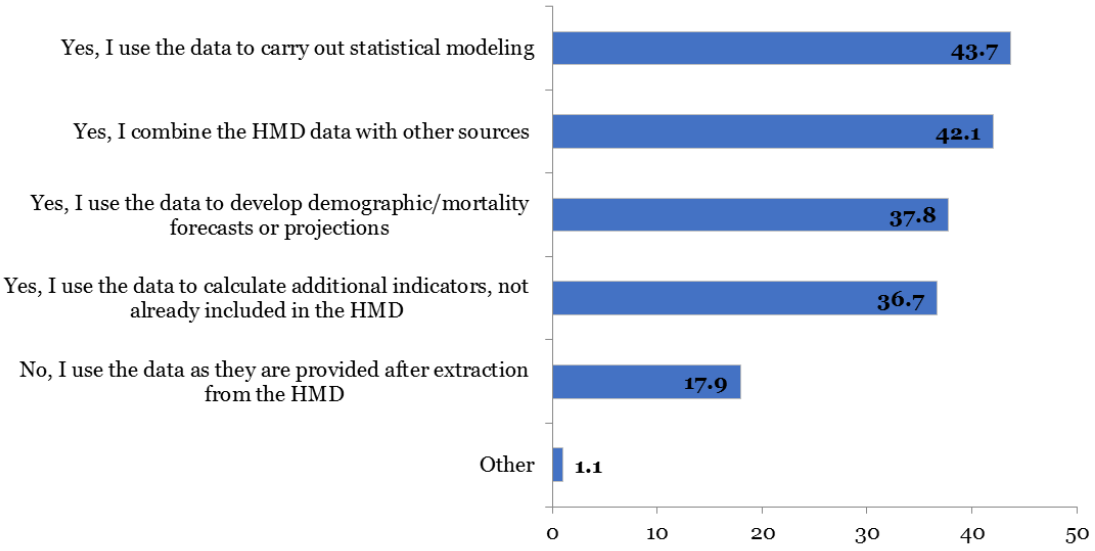
Table 17: Software used by category of occupation

	Scientist		Actuary		Student		Other	
	no.	%	no.	%	no.	%	no.	%
R	322	49.2	130	46.8	105	55.9	48	41.0
SAS	67	10.2	30	10.8	4	2.1	13	11.1
STATA	190	29.0	1	0.4	43	22.9	14	12.0
SPSS	74	11.3	3	1.1	20	10.6	17	14.5
MATLAB	59	9.0	10	3.6	11	5.9	-	-
EXCEL	359	54.8	225	80.9	104	55.3	83	70.9
Other	40	6.1	17	6.1	7	3.7	10	8.5

Source: CNR-IRPPS elaboration

Additional questions were focused on respondent's practices in using HMD data. The first one asked whether and how they elaborated HMD data to conduct their further analysis (Figure 14). Multiple answers were allowed. As most frequently reported, HMD is the basis for statistical modelling (43.7%), demographic forecasts (37.8%) or for the identification of additional indicators (36.7%). Some respondents also combine HMD data with other sources (42.1%).

Figure 14: Distribution of respondents to question: "Do you process the HMD data beyond simply downloading what you need?" Multiple responses (%)



Source: CNR-IRPPS elaboration.

The distribution by occupational category generally reflects the users' aims and confirms that HMD data are usually the reference point to make further research. Scientists are the category that use HMD data much more to carry out a wider range of different analysis,

actuaries are generally more focused on two types of analysis, while the category of students and others use HMD data in a more homogeneously distributed way.

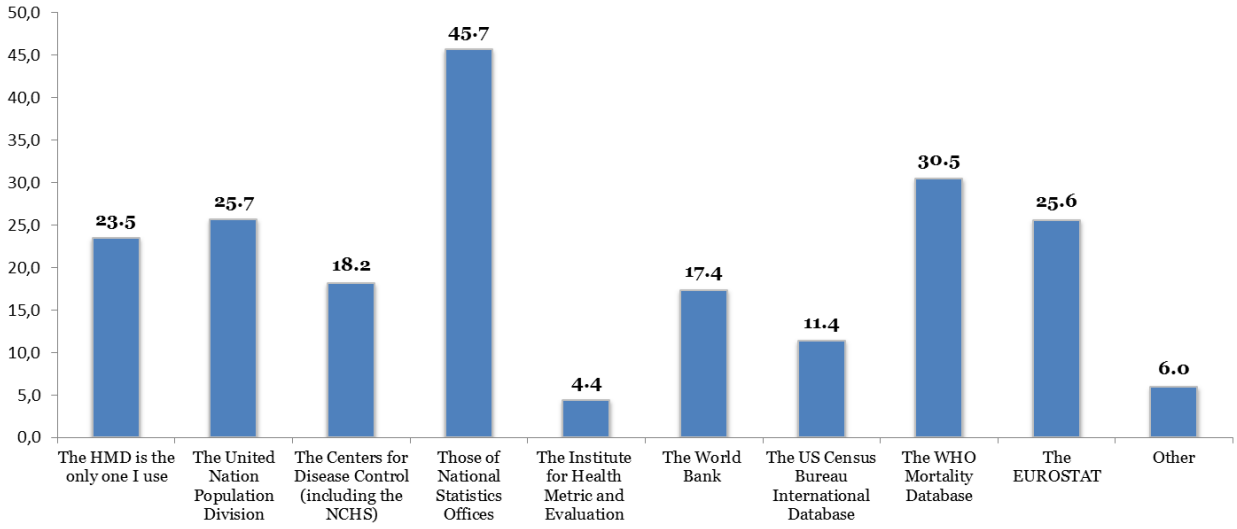
Table 18: Data processing by category of occupation

	Scientist		Actuary		Student		Other	
	no.	%	no.	%	no.	%	no.	%
No, I use the data as they are provided after extraction from the HMD	103	15.7	51	18.3	35	18.6	33	28.2
Yes, I combine the HMD data with other sources	317	48.4	100	36,0	65	34.6	39	33.3
Yes, I use the data to calculate additional indicators, not already included in the HMD	278	42.4	83	29.9	63	33.5	30	25.6
Yes, I use the data to carry out statistical modeling	331	50.5	114	41,0	61	32.4	35	29.9
Yes, I use the data to develop demographic/mortality forecasts or projections	224	34.2	149	53.6	57	30.3	38	32.5
Other	8	1.2	1	0.4	1	0.5	3	2.6

Source: CNR-IRPPS elaboration.

When asked which other information sources are consulted, respondents report that they also access on regular basis data provided by National Statistical Offices 44.5% as well as by International organizations such as the WHO Mortality Database (30.5%) the United Nation Population Division (25.7%), the Eurostat (25.6%), the Centers for Disease Control (including the NCHS) (18.2%), the World Bank (17.4%), the US Census Bureau International Database (11.4%) and the Institute for Health Metric and Evaluation (4.4%).

Figure 15: Distribution of respondents to question: "Which other website or databases do you consult on a regular basis to collect information on national mortality levels?" Multiple responses (%)



Source: CNR-IRPPS elaboration.

With the exception of students that preferably use HMD as a single source of information (39.7%) the other categories reflect the general trend consulting in particular also data from statistical offices and/or data contained in the WHO mortality database.

Table 19: Databases consulted by category of occupation

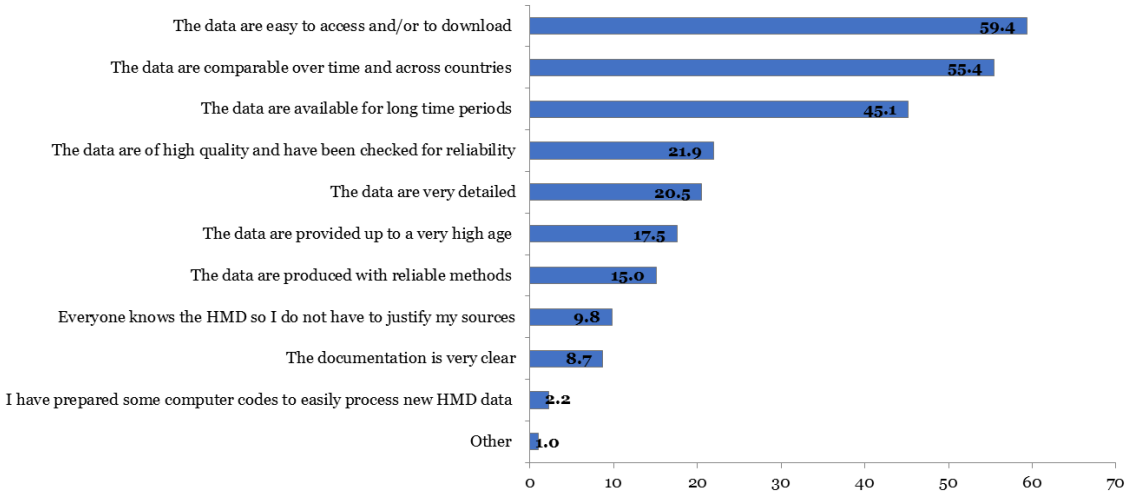
	Scientist		Actuary		Student		Other	
	no.	%	no.	%	no.	%	no.	%
The HMD is the only one I use	146	19.8	63	20.3	83	39.7	26	17,0
The United Nation Population Division	242	32.9	28	9.0	44	21.1	42	27.5
The Centers for Disease Control (including the NCHS)	147	20,0	46	14.8	32	15.3	22	14.4
Those of National Statistics Offices	352	47.8	162	52.3	66	31.6	52	34,0
The Institute for Health Metric and Evaluation	38	5.2	6	1.9	4	1.9	12	7.8
The World Bank	158	21.5	22	7.1	44	21.1	16	10.5
The US Census Bureau International Database	106	14.4	17	5.5	21	10,0	13	8.5
The WHO Mortality Database	266	36.1	69	22.3	52	24.9	34	22.2
The EUROSTAT	219	29.8	71	22.9	35	16.7	26	17,0
Other	35	4.8	27	8.7	12	5.7	10	6.5

Source: CNR-IRPPS elaboration.

3.3.6 User's perception of HMD

The last section of the questionnaire (Q.17 and Q.20) intended to explore users' perception on the advantage in using HMD data. Respondents reported most often the following options (Figure 16): the easily accessible data (59.4%), the comparability over time and across-countries (55.5%), and the long time periods of the data available (45.2%).

Figure 16: Distribution of respondents to question: "What is/are the advantage(s) of using the HMD?" Multiple responses (%)



Source: CNR-IRPPS elaboration.

The appraisal distributed by occupation category mirrors the similar results. Remarkable differences are related to the easy to access, which is especially appreciated by students (67.9%) and actuaries (60%), while scientists value in particular the comparability over time and across

countries (57.9%). Among the other options, the category of students (28.7%) and others (26.8%) also appraise that data are very detailed.

Moreover, looking at the final overall comments provided by some respondents (16.9%), important indications on users' needs and expectations can be drawn. Among the suggestions of improvements, there is the request of providing a more detailed geographic distribution of data (e.g. provinces, municipalities), expanding the number of countries to be included in HMD, including data on causes of death. Different needs emerged on the file formats. The most frequent demand concerns the improvement of tools for data automatic imports into statistical packages, such R and Stata. Even if some tools have been developed by the HMD team, it is clear that some users are not aware of them. Therefore, there is room for improvements in the HMD web interface, making the link with these scripts more evident. Moreover, respondents mention the need of a more frequent timely update of datasets.

Table 20: HMD appraisal by category of occupation

	Scientist		Actuary		Student		Other	
	no.	**%**	**no.**	**%**	**no.**	**%**	**no.**	**%**
The data are easy to access and/or to download	396	53.8	186	60,0	142	67.9	86	56.2
The data are comparable over time and across countries	426	57.9	167	53.9	109	52.2	79	51.6
The data are very detailed	142	19.3	46	14.8	60	28.7	41	26.8
The data are provided up to a very high age	133	18.1	52	16.8	31	14.8	31	20.3
The data are available for long time periods	353	48,0	146	47.1	83	39.7	53	34.6
The data are produced with reliable methods	116	15.8	54	17.4	25	12.0	17	11.1
The documentation is very clear	57	7.7	29	9.4	23	11,0	13	8.5
The data are of high quality and have been checked for reliability	175	23.8	69	22.3	39	18.7	26	17.0
Everyone knows the HMD so I do not have to justify my sources	71	9.6	41	13.2	11	5.3	15	9.8
I have prepared some computer codes to easily process new HMD data	15	2.0	8	2.6	3	1.4	5	3.3
Other	7	1.0	4	1.3	2	1.0	1	0.7

Source: CNR-IRPPS elaboration.

Among the many appreciations reported by respondents, some of them summarize well the characteristics of HMD. This pertains to the availability of the data free of charge, the transparency of in the data processing procedures, the detailed documentation and the trustworthiness of the database. As a respondent reported "It is also easy to use and reference, and a trustworthy source, so I don't have the need to look elsewhere for data". Another respondent expressed his/her appreciation: "You are the gold standard in the field and an example of the good work that can be done, but we need more like you to have the rest of the world a la HMD".

3.3.7 Conclusions

Although HMD attracts the interest of a wide range of professionals from different disciplinary fields as well as citizens, the majority of the users belong to academia/research and actuarial settings. This outlines a user profile that is expert both in demographic issues and data analysis, is accustomed to accessing HMD for quite a long period (more than 10 and 5 years) and does it at a regular basis (a few times over the past year). Moreover, both the academia/research user category and the actuarial one consult other specialized databases along with HMD confirming expertise in the data analysis. Differences between these two categories concern the retrieval of country-related data. While the academia/research users tend to use data related to all countries, benefitting from their comparability based on a common methodology to monitor demographic trends, the actuarial users tend to concentrate on specific countries. The student user profile shares similar characteristics with the academia/research one. Interesting to note that they learned about HMD at university showing that HMD is a popular/well-established source also for teaching purposes.

Looking at responses that can outline data re-use, similar behaviour can be detected especially among the academia/research users and the actuarial ones. The general tendency of selecting the data directly on the web site and using excel for the analysis represent a common feature, even if the use of computer codes is slightly higher in academic/research setting and so is the contemporary use of more than one specialized statistical packages. HMD data are generally re-used to perform further research. The academic/research users tend to re-use them for several types of analysis, combining them with different data sources, for the creation of additional indicators and for statistical modelling, while actuarial users tend to privilege statistical modelling and forecasts or projections. These different types of analysis are done by all the identified user categories retrieving in particular life tables and death counts, and in the case of academic/researchers and students also analysing life expectancy data at birth.

The general agreement across user categories on the appraisal of HMD for its data comparability over time and countries, for the availability of the long time series and the ease of access confirm the appreciation of the scientific efforts made by the HMD community. For these reasons the requests reported by some respondents of expanding the database to other countries and regions is in line with the positive evaluation of HMD.

4. Final remarks

The OpenUP pilot on research data sharing in Social sciences is part of the seven pilots related to the three pillars of the project (Peer Review, Impact Assessment and Innovative Dissemination) that aimed to implement, test, and verify the outputs and results obtained in OpenUP. The pilots were carried out in close cooperation with selected, devoted research communities from four scientific areas: Arts and humanities, Social sciences, Life sciences, and Energy. They contributed to raising awareness and increasing skills related to the tested open science approaches among the involved communities, and generated lessons learned and an evidence based knowledge on various aspects of the tested approaches.

The key findings and lessons learned from all the pilots were included in OpenUP final recommendations (OpenUP final recommendations 2018) for policy makers and research organizations. These recommendations identified strategies and policies to promote Open Science on the basis of the analysis carried out during the projects as well as of issues,

challenges and existing best practices captured from the communities involved in the different pilots.

In particular, the pilot described in this paper provided evidences to support policies and actions for the development of successful data repositories and improve the availability of reliable and quality open data.

- A best practice coming from the HMD community showed that effective and transparent procedures related to the provision of raw data as well as the methods used to calculate country statistics have positive impact on data reproducibility. Therefore, successful data repository requires a quality assessment that is transparent for the users, making it possible to reproduce the research results.
- Interviews with the HMD community showed the importance of introducing alternative methods to recognize and reward data management activities. HMD community stated that "the analysis of data, their quality check is not only a service for the community of reference but is a research activity in itself." Therefore, researchers' career evaluation should be expanded to include different skills, such as data curation and management and outreach activities to communicate results also to the general public.
- HMD is a joint project that involved researchers of the Department of Demography at the University of California Berkeley (UCB), the Max Planck Institute for Demographic Research (MPIDR) and the French institute of demographic studies (INED). A successful feature of HMD is the strong collaboration among the team. The two HMD directors explained that "the collaboration was originally, (and still is), based on a small, very well- established group of internationally based demographers who were willing to serve the scientific community interested in demographic studies". Therefore, it is necessary to incentivize networking and community building to strength the cohesion of existing or emerging groups of scientists.

These high-level principles can be also valid to research communities outside Social sciences. Nevertheless, the complexity and variability of data management in specific research fields require the development of further pilots and analyses. Moreover, ad hoc incentives to share open data and actions to support infrastructures that facilitate long-term preservation of high-quality data should be put in place to sustain an open science-oriented culture.

Appendix: Questionnaire's resulting frequencies and percentages

1. Sex:	**No.**	**%**
Male	1069	68.8%
Female	458	29.5%
Prefer not to answer	26	1.7%
Total	1.553	100%

2. Age group

>20	4	0.3%
21-39	679	43.7%
40-59	600	38.6%
60+	270	17.4%
Total	1.553	100%

3. Country of residence

Andorra		
Afghanistan	1	0.1%
Antigua and Barbuda		
Anguilla		
Albania		
Armenia	1	0.1%
Angola		
Antarctica		
Argentina	3	0.2%
American Samoa		
Australia	38	2.4%
Aruba		
Åland Islands		
Azerbaijan		
Austria	21	1.4%
Algeria	1	0.1%
Bahamas		
Bahrain		
Bangladesh	2	0.1%
Barbados		
Belarus	2	0.1%
Belgium	29	1.9%
Belize		
Benin		
Bermuda		
Bhutan		
Plurinational State of Bolivia		
Sint Eustatius and Saba Bonaire		
Bosnia and Herzegowina	1	0.1%
Botswana		
Bouvet Island		
Brazil	15	1.0%
British Indian Ocean Territory		
Brunei Darussalam		
Bulgaria	7	0.5%

Burkina Faso		
Burundi		
Cambodia		
Cameroon		
Canada	73	4.7%
Cape Verde		
Cayman Islands		
Central African Republic		
Chad		
Chile	10	0.6%
China	38	2.4%
Christmas Island		
Cocos Islands		
Colombia	6	0.4%
Comoros		
Congo		
The Democratic Republic of The Congo		
Cook Islands	1	0.1%
Costa Rica	2	0.1%
Côte d'Ivoire		
Croatia	6	0.4%
Cuba		
Curaçao		
Cyprus		
Czech Republic	33	2.1%
Denmark	38	2.4%
Djibouti		
Dominica		
Dominican Republic		
Ecuador	3	0.2%
Egypt		
El Salvador		
Equatorial Guinea		
Eritrea		
Estonia	3	0.2%
Ethiopia	1	0.1%
Falkland Islands		
Faroe Islands	1	0.1%
Fiji		
Finland	17	1.1%
France	55	3.5%
French Guiana		
French Polynesia		
French Southern Territories		
Gabon	1	0.1%
Gambia		
Georgia		
Germany	135	8.7%
Ghana	1	0.1%
Gibraltar		
Greece	13	0.8%
Greenland		
Grenada		

Guadeloupe		
Guam		
Guatemala	1	0.1%
Guernsey		
Guinea		
Guinea-bissau		
Guyana		
Haiti	1	0.1%
Heard and McDonald Islands		
Holy See		
Honduras		
Hong Kong	2	0.1%
Hungary	14	0.9%
Iceland	2	0.1%
India	9	0.6%
Indonesia	5	0.3%
Iran	3	0.2%
Iraq	3	0.2%
Ireland	14	0.9%
Isle of Man		
Israel	8	0.5%
Italy	74	4.8%
Jamaica	2	0.1%
Japan	17	1.1%
Jersey		
Jordan	1	0.1%
Kazakhstan		
Kenya	5	0.3%
Kiribati		
Democratic People's Republic of Korea		
Republic of Korea	6	0.4%
Kuwait		
Kyrgyzstan		
Lao People's Democratic Republic		
Latvia	2	0.1%
Lebanon	1	0.1%
Lesotho		
Liberia		
Libya		
Liechtenstein		
Lithuania	9	0.6%
Luxembourg	3	0.2%
Macao		
The Former Yugoslav Republic of Macedonia	1	0.1%
Madagascar		
Malawi		
Malaysia	6	0.4%
Maldives		
Mali		
Malta		
Marshall Islands		
Martinique		
Mauritania		

Mauritius		
Mayotte		
Mexico	21	1.4%
Federated States of Micronesia		
Republic of Moldova	2	0.1%
Monaco	1	0.1%
Mongolia		
Montenegro		
Montserrat		
Morocco	2	0.1%
Mozambique		
Myanmar	1	0.1%
Namibia		
Nauru		
Nepal		
Netherlands	57	3.7%
New Caledonia		
New Zealand	4	0.3%
Nicaragua	1	0.1%
Niger		
Nigeria		
Niue		
Norfolk Island		
Northern Mariana Islands		
Norway	16	1.0%
Oman		
Pakistan	1	0.1%
Palau		
State of Palestine		
Panama		
Papua New Guinea	1	0.1%
Paraguay		
Peru	2	0.1%
Philippines	2	0.1%
Pitcairn		
Poland	17	1.1%
Portugal	26	1.7%
Puerto Rico		
Qatar		
Réunion		
Romania	3	0.2%
Russian Federation	43	2.8%
Rwanda	1	0.1%
Ascension and Tristan Da Cunha Saint Helena		
Saint Barthélemy		
Saint Kitts and Nevis		
Saint Lucia		
Saint Pierre and Miquelon		
Saint Vincent and The Grenadines		
Samoa	1	0.1%
San Marino		
Sao Tome and Principe		
Saudi Arabia		

Senegal		
Serbia	1	0.1%
Seychelles		
Sierra Leone		
Singapore	3	0.2%
Sint Maarten		
Slovakia	8	0.5%
Slovenia	1	0.1%
Solomon Islands		
Somalia		
South Africa	4	0.3%
South Georgia and The South Sandwich Islands		
South Sudan		
Spain	77	5.0%
Sri Lanka	1	0.1%
Sudan		
Suriname		
Svalbard and Jan Mayen Islands		
Swaziland		
Sweden	39	2.5%
Switzerland	47	3.0%
Syrian Arab Republic		
Province of China Taiwan	8	0.5%
Tajikistan	1	0.1%
United Republic of Tanzania	1	0.1%
Thailand	1	0.1%
Timor-leste		
Togo		
Tokelau		
Tonga		
Trinidad and Tobago		
Tunisia	1	0.1%
Turkey	3	0.2%
Turkmenistan		
Turks and Caicos Islands		
Tuvalu		
Uganda		
Ukraine	5	0.3%
United Arab Emirates	1	0.1%
United Kingdom	148	9.5%
United States	249	16.0%
United States Minor Outlying Islands		
Uruguay	4	0.3%
Uzbekistan		
Vanuatu		
Bolivarian Republic of Venezuela		
Vietnam		
Virgin Islands (British)		
Virgin Islands (US)		
Wallis and Futuna Islands		
Western Sahara		
Yemen		
Zambia		

Zimbabwe	1	0.1%
Total	1.553	100%

4. City

Plain text answers

5. Main field of interest

Demography	391	25.2%
Actuarial studies	460	29.6%
Economics	144	9.3%
Epidemiology	113	7.3%
Medicine	56	3.6%
Biology	17	1.1%
Public health	90	5.8%
Statistics	182	11.7%
Social policies	25	1.6%
Other	75	4.8%
Total	1.553	100.0%

6. Types of Institution

University	795	51.2%
Other public training or research organization	110	7.1%
Other private training or research organization	34	2.2%
Other government organization, Statistics office	138	8.9%
International organization (United Nations, World Bank, etc.)	19	1.2%
Insurance/Re-insurance company	212	13.7%
Other large private corporation	44	2.8%
Other Small and Medium-size private organization	92	5.9%
Foundation	6	0.4%
Other non-profit/NGO	29	1.9%
Other	74	4.8%
Total	1.553	100.0%

7. Occupation

Researcher/scientist	503	32.4%
Teacher/Professor	313	20.2%
Student	229	14.7%
Physician	37	2.4%
Actuary	297	19.1%
Other in insurance/re-insurance	30	1.9%
Public health administrator/analyst	32	2.1%
Journalist	8	0.5%
Other	104	6.7%
Total	1.553	100.0%

8. How did you first learn about the HMD?

I do not remember	241	15.5%
Through a web search	328	21.1%
It was cited in an article I read	260	16.7%
It was mentioned during a conference presentation I attended	105	6.8%
A colleague mentioned it	596	38.4%
Other	23	1.5%
Total	1.553	100.0%

9. For how long have you been an HMD user?		
Never used the HMD since registering	145	9.3%
Less than a years	306	19.7%
Less than five year	559	36.0%
Less than 10 years	322	20.7%
10 years or more	221	14.2%
Total	1.553	100.0%
9bis. Please tell us more about why you have never used the HMD		
I have not had time jet to use the HMD	88	60.7%
The country/contries I am interested in are not included in	5	3.4%
I did not find the information I was looking for	10	6.9%
I could not understand how the data were constructed	4	2.8%
It was too complicated to use	3	2.1%
Other	35	24.1%
Total	1.408	100.0%
Missing	145	
10. How frequently do you access the Human Mortality Database?		
Frequently (several times a month)	61	4.3%
Each time I start a new project	158	11.2%
Once a month or so	118	8.4%
A few times over the past year	636	45.2%
Rarely	435	30.9%
Total	1.408	100.0%
Missing	145	
11 Which HMD countries/regions are you most interested in?		
All countries	642	61.3%
Australia	72	6.9%
Austria	68	6.5%
Belarus	19	1.8%
Belgium	75	7.2%
Bulgaria	20	1.95
Canada	114	10.9%
Chile	23	2.2%
Croatia	28	2.7%
Czech Republic	57	5.4%
Denmark	101	9.6%
Estonia	33	3.1%
Finland	93	8.9%
France	187	17.8%
Germany	217	20.7%
Greece	46	4.4%
Hungary	51	4.9%
Iceland	43	4.1%
Ireland	69	6.6%
Israel	19	1.8%
Italy	132	12.6%
Japan	106	10.1%
Latvia	30	2.9%
Lithuania	34	3.2%
Luxembourg	35	3.3%

Netherlands	108	10.3%
New Zealand	24	2.3%
Northern Ireland	20	1.9%
Norway	90	8.6%
Poland	63	6.0%
Portugal	65	6.2%
Russia	72	6.9%
Slovakia	43	4.1%
Slovenia	31	3.0%
South Korea	23	2.2%
Spain	126	12.0%
Sweden	153	14.6%
Switzerland	93	8.9%
Taiwan	28	2.7%
U.K.	250	23.9%
U.S.A.	313	29.9%
Ukraine	30	2.9%

12. Have you ever downloaded/copied files from the HMD website?

Never	170	12.1%
Yes. by going on the HMD website and selecting the data I am interested in	1150	81.7%
Yes. by automatically downloading data using some computer codes I have set up for this purpose	179	12.7%
Missing	145	

13. Are you satisfied with the format in which HMD data are provided?

Yes	1350	95.9%
No	58	4.1%
Missing	145	

13bis Why?

Plain text answers

14 Which software do you use to process HMD data?

R	605	48.9%
SAS	114	9.2%
STATA	248	20%
SPSS	114	9.2%
MATLAB	80	6.5%
EXCEL	771	62.3%
Other	74	6%
Missing	315	

15. Do you process the HMD data beyond simply downloading what you need?

No. I use the data as they are provided after extraction from the HMD	222	17.9%
Yes. I use the data to calculate additional indicators. not already included in the HMD	454	36.7%
Yes. I use the data to develop demographic/mortality forecasts or projections	468	37.8%
Yes. I combine the HMD data with other sources	521	42.1%
Yes. I use the data to carry out statistical modeling	541	43.7%
Other	13	1.1%
Missing	315	

16. Which type of HMD files have you downloaded over the past 12 months (for any country)?

The zip file of pooled HMD data	226	18.3%
Input files	90	7.3%
Life tables	806	65.1%
Unadjusted death rates	334	27.0%
Population estimates	423	34.2%
Deaths counts	506	40.9%
Life expectancy at birth	446	36.0%
Cohort data	291	23.5%
Missing	315	

17. What is/are the advantage(s) of using the HMD? (Please select at most 3 answers)

The data are easy to access and/or to download	837	59.4%
The data are comparable over time and across countries	781	55.4%
The data are very detailed	289	20.5%
The data are provided up to a very high age	247	17.5%
The data are available for long time periods	636	45.1%
The data are produced with reliable methods	212	15.0%
The documentation is very clear	122	8.7%
The data are of high quality and have been checked for reliability	309	21.9%
Everyone knows the HMD so I do not have to justify my sources	138	9.8%
I have prepared some computer codes to easily process new HMD data	31	2.2%
Other	14	1.0%
Missing	145	

18. Your main purpose in using the HMD is?

To monitor mortality trends in general	321	22.8%
To conduct research on changes or international variations in mortality	517	36.7%
For educational purposes	325	23.1%
For my business activity	163	11.6%
Other	82	5.8%
Missing	145	

19. Which other websites or databases do you consult on a regular basis to collect information on national mortality levels?

The HMD is the only one I use	331	23.5%
The United Nation Population Division	362	25.7%
The Centers for Disease Control (including the NCHS)	256	18.2%
Those of National Statistics Offices	644	45.7%
The Institute for Health Metric and Evaluation	62	4.4%
The World Bank	245	17.4%
The US Census Bureau International Database	161	11.4%
The WHO Mortality Database	430	30.5%
The EUROSTAT	361	25.6%
Other	84	6.0%
Missing	145	

20. Is there any comment. suggestion or feedback you would care to provide about the Human Mortality Database (regarding its content. utilities. tools. website or any other aspect of the database)?

Plain text answers

References

Assante, M., Candela, L., Castelli, D., & Tani, A. (2016). Are scientific data repositories coping with research data publishing? *Data Science Journal*, 15: 1–24. doi:10.5334/dsj-2016-006.

Barbieri, M., Wilmoth, J.R., Shkolnikov, V.M., et al. (2015). Data resource profile: the human mortality database (HMD). *Int. J. Epidemiol.*, 44(5), 1549–1556 https://doi.org/10.1093/ije/dyv105.

Blümel C., et al. (2018). Project OpenUP-Deliverable D6.3 – Final Use Case Evaluation Report, 14 September 2018. http://doi.org/10.5281/zenodo.2557435.

Borgman, C.L. (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. MIT Press, Cambridge.

Callaghan, S., Steve D., Sam P., et al. (2012). Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres. *Int. J. Digital Curation,* 7 (1): 107–13. doi:10.2218/ijdc.v7i1.218.

Callaghan, S., Murphy F., Tedds J., et al. (2013). Processes and Procedures for Data Publication: A Case Study in the Geosciences. *Int. J. Digital Curation,* 8 (1): 193–203. doi:10.2218/ijdc.v8i1.253.

Callaghan, S., Tedds J., Kunze J., et al. (2014). Guidelines on Recommending Data Repositories as Partners in Publishing Research Data. *Int. J. Digital Curation,* 9 (1): 152–63. doi:10.2218/ijdc.v9i1.309.

Callaghan, S., Tedds J., Lawrence R., et al. (2014). Cross-Linking Between Journal Publications and Data Repositories: A Selection of Examples. *Int. J. Digital Curation,* 9 (1): 164–75. doi:10.2218/ijdc.v9i1.310.

Callaghan, S. (2015). Data without peer: examples of data peer review in the earth sciences. *D-Lib Mag.,* 21(1/2). https://doi.org/10.1045/january2015-callaghan.

Candela, L., Castelli, D., Manghi, P., Tani, A. (2015). Data journals: a survey. *J. Assoc. Inf. Sci. Technol.,* 66(9), 1747–1762. https://doi.org/10.1002/asi.23358.

Carpenter, T.A. (2017). What Constitutes Peer Review of Data: A Survey of Published Peer Review Guidelines, April. http://arxiv.org/abs/1704.02236.

Curty, R.G. (2016). Factors Influencing Research Data Reuse in the Social Sciences: An Exploratory Study. *Int. J. Digital Curation,* 11 (1): 96–117. doi:10.2218/ijdc.v11i1.401.

Faniel, I.M., Kriesberg, A., Yakel, E. (2015). Social scientists' satisfaction with data reuse. J. Assoc. *Inf. Sci. Technol.*, 67(6), 1404–1416. https://doi.org/10.1002/asi.23480.

Hellauer, T.R. (2017). What is open peer review? A systematic review. *F1000Research*, 6:588 https://doi.org/10.12688/f1000research.11369.2

Kim, Y., Adler, M. (2015). Social scientists' data sharing behaviours: investigating the roles of individual motivations, institutional pressures, and data repositories. *Int. J. Inf. Manage.*, 35, 408–418. https://doi.org/10.1016/j.ijinfomgt.2015.04.007.

Kratz, J.E., Strasser, C. (2015). Researcher perspectives on publication and peer review of data. *PLoS ONE,* 10(2), e0117619. https://doi.org/10.1371/journal.pone.0117619.

Lawrence, B., Jones, C., Matthews, B., Pepler, S., Callaghan, S. (2011). Citation and peer review of data: moving towards formal data publication. *Int. J. Digital Curation,* 6(2), 4–37 https://doi.org/10.2218/ijdc.v6i2.205.

Luzi, D., Ruggieri, R., Pisacane, L., & Di Cesare, R. (2017). Verso una (open) peer review dei dati: uno studio pilota nelle scienze sociali. In *Scienza aperta e integrità della ricerca*. III Convegno AISA, 9-10 Novembre 2017, Milano. https://archiviomarini.sp.unipi.it/id/eprint/744

Luzi, D., Ruggieri, R., & Pisacane, L. (2019). The OpenUP Pilot on Research Data Sharing, Validation and Dissemination in Social Sciences. In Manghi, P., In Candela, L., & In Silvello, G. (2019). *Digital Libraries: Supporting Open Science*: 15th Italian Research Conference on Digital Libraries, IRCDL 2019, Pisa, Italy, January 31 – February 1, 2019 Proceedings. Communications in Computer and Information Science 988, Springer

Mayernik, M.S., Callaghan, S., Leigh, R., Tedds, J., Worley, S. (2015). Peer review of datasets: when, why, and how. *Bull. Am. Meteorol. Soc.*, 96(2), 191–201 https://doi.org/10.1175/BAMS-D-13-00083.1.

Priem, J., Taraborelli, D., Groth, P., Neylon, C. (2010). Altmetrics: A manifesto, 26 October 2010.http://altmetrics.org/manifesto.

Stančiauskas, V., Banelytė, V. (2017). OpenUP survey on researchers' current perceptions and practices in peer review, impact measurement and dissemination of research results survey, 19 April 2017. https://doi.org/10.5281/zenodo.556157.

Tenopir, C., et al. (2011). Data sharing by scientists: practices and perceptions. *PLoS One,* 6(6), e21101. https://doi.org/10.1371/journal.pone.0021101.

Vignoli M. (2017). Project OpenUP-Deliverable D6.1 – Use Cases and Pilots Definition of Methodology. DOI:10.5281/ZENODO.2557426.

Vignoli M. (2018). Project OpenUP-Deliverable D6.2 – Interim Use Case Evaluation Report, 30 November 2017. http://doi.org/10.5281/zenodo.2557428.