A Tailor-made Data Quality Approach for Higher Educational Data

Cinzia Daraio[1][†], Renato Bruni[1], Giuseppe Catalano[1], Alessandro Daraio[1], Giorgio Matteucci[1], Monica Scannapieco[2], Daniel Wagner-Schuster[3], Benedetto Lepori[4]

[1]DIAG, Sapienza University of Rome, Italy
[2]Italian National Institute of Statistics, Italy
[3]JOANNEUM RESEARCH, Institute for Economic and Innovation Research, Austria
[4]Università della Svizzera italiana, Faculty of Communication sciences, Switzerland

Abstract

Purpose: This paper relates the definition of data quality procedures for knowledge organizations such as Higher Education Institutions. The main purpose is to present the flexible approach developed for monitoring the data quality of the European Tertiary Education Register (ETER) database, illustrating its functioning and highlighting the main challenges that still have to be faced in this domain.

Design/methodology/approach: The proposed data quality methodology is based on two kinds of checks, one to assess the consistency of cross-sectional data and the other to evaluate the stability of multiannual data. This methodology has an operational and empirical orientation. This means that the proposed checks do not assume any theoretical distribution for the determination of the threshold parameters that identify potential outliers, inconsistencies, and errors in the data.

Findings: We show that the proposed cross-sectional checks and multiannual checks are helpful to identify outliers, extreme observations and to detect ontological inconsistencies not described in the available meta-data. For this reason, they may be a useful complement to integrate the processing of the available information.

Research limitations: The coverage of the study is limited to European Higher Education Institutions. The cross-sectional and multiannual checks are not yet completely integrated.

Practical implications: The consideration of the quality of the available data and information is important to enhance data quality-aware empirical investigations, highlighting problems, and areas where to invest for improving the coverage and interoperability of data in future data collection initiatives.

Originality/value: The data-driven quality checks proposed in this paper may be useful as a reference for building and monitoring the data quality of new databases or of existing databases

---

† Corresponding author: Cinzia Daraio (E-mail: daraio@diag.uniroma1.it).

Research Paper

available for other countries or systems characterized by high heterogeneity and complexity of the units of analysis without relying on pre-specified theoretical distributions.

1   History of ETER and its political importance for research on higher education

Studies, analyzes and policy investigations about the positioning and the characterization of education and research systems need data to be performed. Whenever we need data, we need a method for the management of data, and in the Big Data era, a crucial role is played by data quality. Therefore, higher education policies and indicators development need data quality techniques to increase the value of data and improve the exploitation of the available data.

The availability of data and information about Higher Education Institutions (HEIs) is then the first requirement for the development of empirical studies. The second relevant aspect is the consideration of the quality of the available data and information. In higher education, we observe a kind of paradox. While we are leaving in the Big data era, in which huge amount of data are produced, stored in non-SQL databases, and analyzed on large scale, in this field still relational databases are used to organize the existing data and information, and cases of "little data or no data" at all (Borgman, 2015) are the normality.

Higher Education Systems are complex systems and their assessment is complex too. The development of models of indicators or metrics for a quantitative assessment requires a comprehensive framework, which should include the specification of the underlying theory, methodology, and data properties. Models of metrics are necessary to assess the meaning, validity, and robustness of metrics (Daraio, 2017). Daraio and Glanzel (2016) identified the following critical issues: i) data quality issues (OECD, 2011) including completeness, validity, accuracy, consistency, availability, and timeliness; ii) comparability problems related to heterogeneous definitions of the variables, data collection practices, and databases; iii) lack of standardization; iv) lack of interoperability; v) lack of modularization; vi) problems of classification; vii) difficulties in the creation of concordance tables among different classification schemes; viii) problems and costs of the extensibility of the system; ix) problems and costs of updating of the system.

The development of the European Tertiary Education Register (ETER) has grown up of the recognition that, beyond aggregated data at the country and regional level provided by EUROSTAT, there is an urgent need to have information on individual

HEIs and their individual profiles. On the one hand, New Public Management approaches to higher education governance (Capano, 2011; Ferlie et al., 1996) focused on "steering at distance" and on transforming HEIs into strategic actors, which are capable to develop their own profile and strategy (Bonaccorsi & Daraio, 2007a). On the other hand, empirical studies have shown that higher education systems are highly heterogeneous as of the type and characteristics of HEIs (Daraio et al., 2011) and, therefore, analyzes based on natural aggregates might lead to incorrect conclusions. Moreover, the emergence of international rankings emphasized the importance of comparing institutions rather than countries; in that respect, while comparative analyzes of publication outputs were available since many years (van Raan, 2013), until recently very few analysis were performed including both inputs and outputs of universities. Daraio, Bonaccorsi, and Simar (2015) proposed a methodological contribution that overcomes four main criticisms of university rankings, including monodimensionality, statistical robustness, dependence on university size and subject-mix and lack of consideration of the input–output structure. They illustrated their method on European university data and pointed out on the importance of investing in the data collection and integration for research and policymaking. Daraio and Bonaccorsi (2017) after summarizing the main criticisms of rankings, and recent trends in indicators development, proposed an approach to overcome rankings based on the integration of multidimensional data in open platforms. More recently, Lepori, Geuna, and Mira (2019) compared European and USA universities.

The development of ETER has grown up from this recognition and from the objective of the European Commission to improve transparency and accountability of higher education in Europe (European Commission, 2011). From the beginning, ETER was entitled with two main functions: first, establishing a register of higher education institutions and, accordingly being able to identify them and locate them in the European space; second, collecting statistical data on relevant dimensions of HEIs as identified by the scholarly literature in the field (Huisman et al., 2015). The first function raised complex issues of delimiting higher education and defining inclusion and exclusion criteria, which turned out to be largely conventional (Lepori & Bonaccorsi, 2013). The second entailed a complex work of addressing comparability problems between national systems (Bonaccorsi et al., 2007); while ETER could build on standardization work by EUROSTAT for what concerns students and graduates (Unesco, OECD, Eurostat–UOE-2013), the project had to work out its own definitions for what concerns finances and staff data, as well as suitable mappings from (heterogeneous) national classifications.

Research Paper

The establishment process of ETER took however more than one decade because of the complexity of the European statistical system, which was by large composed of different national statistical systems with their specificities (Lepori & Bonaccorsi, 2013) and of the lack of a suitable institutional framework, as the option of managing ETER within EUROSTAT was discarded because of practical and legal constraints.

An important role for the success of ETER has been played by a pioneering European research project, called AQUAMETH, that integrated for the first time comparable data on six European countries (Italy, Norway, Portugal, Spain, Switzerland, and the UK), showing the feasibility and interest of this data integration for research analysis (Bonaccorsi & Daraio, 2007b). Williams (2008) in his book review of Bonaccorsi and Daraio (2007c) in the *London Review of Education* highlighted the importance of data for making econometric analysis and in particular comparison of the efficiency of universities across Europe. He wrote: "*the main intention of the book is to use these data to undertake institution-level cross-national econometric analyses of the efficiency of universities* (Williams, 2008)" "*Their analyses will… serve the purpose of showing serious mathematical economists in Europe that their higher education systems are potentially a fruitful subject of study and are beginning to produce data that are worth serious analytical attention*", "*this book and the AQUAMETH studies that underpin it can serve an important proselytising function. It deserves to be widely read by serious higher education researchers* (Williams, 2008)". Daraio (2018, 2019) recently considered the important role of the availability and quality of institutional level data for econometric analysis.

After the AQUAMETH project, the European Commission launched in 2011 a large-scale pilot, called EUMIDA that provided for the first time a complete mapping of European higher education and proved the feasibility of a large-scale data collection (Niederl et al., 2014). From 2013 to 2019, ETER was established as a regular (yearly) data collection on European higher education with the aim of reaching a comparable level of quality and completeness as the US Integrated Postsecondary Data Service IPEDS (https://nces.ed.gov/ipeds/). The development process of ETER entailed the consolidation of methodology and definitions inherited from the pilot (Lepori et al., 2015), the introduction of a set of procedures for the collection of data and the verification of their quality and, finally, the programming of a database to manage data collection and host the data, as well as of a public website so that users can search and download data (www.eter-project.com).

A Tailor-made Data Quality Approach for Higher Educational Data   Cinzia Daraio et al.

Research Paper

2   Aim of the paper

The main objective of this paper is to describe the flexible approach developed for monitoring the data quality of ETER, illustrate its functioning and highlight the main challenges that still have to be faced. More specifically, we will focus on the data quality checks that are helpful to identify outliers, extreme observations, and to detect ontological inconsistencies not described in the available meta-data. We aim also to raise awareness on the users of institutional data about the importance of data quality issues for a correct interpretation of the results and to show the functioning of the proposed approach that can be easily adapted, mutatis mutandis, to other complex institutional databases characterized by a high heterogeneity of their units of analysis.

The paper is organized as follows. Section 3 provides an outline of the European Tertiary Education Register information system. Section 4 introduces the current data quality approach developed for ETER and its management, keeping into account the peculiarity of the ETER data collection. The methodology developed for the multiannual and the cross-sectional checks is then explained in Section 5. After that, in Section 6, we describe the results obtained by applying the proposed approach to the last version of the ETER database available. Finally, in the concluding section, we outline the strength of the proposed approach and its potential applicability to other databases as well as existing challenges and possible extensions.

3   An outline of the European Tertiary Education Register

ETER is a database of microdata on Higher Education Institutions (HEIs) in Europe, concerning their basic characteristics and geographical location, staff, finances, education, and research activities. ETER includes the following main groups of variables:

– Institutional descriptors and geographical information on the included HEIs.
– Data on students and graduates, including breakdowns by International Standard Classification of Education (ISCED-2011) level, gender, citizenship, mobility, and field of education.
– Data on research, including PhD students and graduates, as well as R&D expenditure and participation to European Framework Programmes for Research and Innovation.
– Financial data: expenditures and revenues of the HEI.
– Staff data (academic and non-academic), including some breakdowns by gender, citizenship, and field.

Research Paper

When compared with the data provided by education and R&D statistics at EUROSTAT, ETER includes very similar variables and breakdowns for students and graduates, since ETER readily adopted the definitions from the UOE manual on education statistics. However, HEI-level data are provided rather than national aggregates.

ETER provides substantial additional information concerning the other dimensions: descriptors are of the outmost importance in order to characterize types of HEIs and their history, while geographical information allows for an analysis of the distribution of HEI activities across the European space. ETER also provides more detailed information on expenditures and revenues, including an important breakdown of revenues by core budget and third-party funds, which is not foreseen in education statistics. Additional data have also been collected concerning staff, including the number of full professors and breakdowns by gender and citizenship.

The ETER database is targeted to include 37 countries composed of the 27 EU Member States, plus the UK, plus EFTA countries (CH, IS, LI, NO) and other five EU candidate countries (AL, ME, MK, RS, TR). In principle, ETER data are provided by National Statistical Authorities (NSAs), Higher Education Ministries, or Higher Education Agencies, based on national statistical databases or higher education information systems, with few exceptions. Descriptors and geographical information are mostly collected by the ETER consortium. ETER data have been collected for six years (2011–2016).

The ETER database includes 3,198 unique HEIs over all years. For the academic year 2015/2016, 22.1 million undergraduate and graduate students and around 688 thousand Ph.D. students are accounted in ETER.

Data quality is a relevant issue for any data collection and is a greater challenge for microdata multi-sources data collection processes as is the case of ETER. A basic but very important dimension of the data quality is completeness that evaluates the share of missing values in the considered dataset. The current ETER dataset has an overall completeness index of 63%, meaning that the number of missing and confidential data is around 37%.

The lower completeness observed is due to the inclusion of some countries in which limited data have been collected namely Albania (AL), Iceland (IS), Republic of North Macedonia (MK), Montenegro (ME), and Turkey (TR) or only descriptors and geographical information is available, as is the case for the French part of Belgium (BE) and Romania (RO). In general terms, the level of completeness also varies largely by country (see Table 1). There are 10 countries for which completeness is 90% or more, including Austria (AT), Switzerland (CH), Cyprus (CY), Germany (DE), Ireland (IE), Liechtenstein (LI), Malta (MT), Portugal (PT), Sweden (SE), and the UK. For some countries, such as Italy (IT) and Poland (PL), data are largely

complete for universities, but there are missing information for other institutions (particularly about staff and financial data).

Table 1.   Completeness of data by country in the ETER Database.

Country	Completeness (2011–2016)			
	Average Completeness	Min	Max	Range
High level of completeness				
Switzerland CH	0.99	0.93	1.00	0.06
Liechtenstein LI	0.98	0.98	0.99	0.01
Germany DE	0.97	0.23	0.99	0.76
United Kingdom UK	0.96	0.13	1.00	0.87
Sweden SE	0.95	0.29	1.00	0.71
Portugal PT	0.92	0.23	0.97	0.74
Malta MT	0.92	0.83	0.95	0.13
Cyprus CY	0.91	0.49	0.99	0.51
Ireland IE	0.90	0.80	0.95	0.15
Austria AT	0.90	0.85	0.96	0.11
Medium-High level of completeness				
Spain ES	0.87	0.81	0.94	0.13
Estonia EE	0.86	0.85	0.99	0.14
Finland FI	0.85	0.60	0.95	0.35
Norway NO	0.84	0.13	0.93	0.80
Bulgaria BG	0.84	0.83	0.87	0.04
Slovakia SK	0.83	0.76	0.88	0.12
Lithuania LT	0.81	0.24	0.98	0.74
Italy IT	0.80	0.36	0.92	0.56
Latvia LV	0.79	0.12	0.90	0.79
Czech Republic CZ	0.78	0.23	0.90	0.67
Poland PL	0.77	0.24	0.89	0.65
Medium level of completeness				
Hungary HU	0.74	0.12	0.92	0.80
Netherland NL	0.74	0.35	0.83	0.48
Greece GR	0.74	0.34	0.90	0.55
Croatia HR	0.73	0.26	0.90	0.63
Denmark DK	0.65	0.10	0.94	0.84
North Macedonia MK	0.57	0.11	0.84	0.73
Luxemburg LU	0.53	0.29	0.93	0.63
Low level of completeness				
France FR	0.45	0.06	0.96	0.89
Slovenia SI	0.45	0.11	0.85	0.74
Belgium BE	0.42	0.09	0.97	0.87
Iceland IS	0.41	0.13	0.55	0.42
Serbia RS	0.35	0.09	0.83	0.74
Albania AL	0.33	0.13	0.69	0.56
Turkey TR	0.26	0.11	0.64	0.53
Montenegro ME	0.12	0.11	0.12	0.01
Romania RO	0.10	0.06	0.12	0.06

Research Paper

The level of completeness largely varies by domain and variable. It is higher for data on students and graduates, although some breakdowns by field and by mobility are more problematic. Completeness is lower for financial data on income and expenditure (around 40% on average). The lack of availability of this information is due to the absence of standardized collection procedures on the national level in some countries. R&D expenditure is available in around 33% of cases. Data on staff are in an intermediate position around 50–55% for both Head Count (HC) and Full Time Equivalent (FTE), except for academic staff breakdowns.

4   ETER's current approach and management of data quality

Data quality is a relevant interdisciplinary issue, studied in statistics, management and computer science. Poor data quality greatly reduces data value: inaccuracy, incompleteness, out-of-dateness may cause data to become useless (Batini-Scannapieco, 2016).

Some international standards for defining the data quality concepts and related dimensions have been proposed. ISO 25012 introduces and defines three possible levels (views) of data quality to be considered individually, namely:

– Internal Data Quality, related to values and formats of data (e.g. consistency, completeness);
– External Data Quality, related to characteristics of the software and hardware used to store and access data (e.g. response time, portability);
– Data Quality in Use, related to the final user of data (e.g. effectiveness, level of satisfaction).

Validation and data quality controls are indeed central tasks in ETER, facing challenges rose by the specific nature of ETER data: i) micro-data at institutional level with an high level of heterogeneity, instead of aggregated data, ii) secondary data collection based on data collected nationally largely without a common reference framework. The latter is alternative to a primary data collection, hence implying also a limited control on the overall data collection process.

We hereby summarize the ETER data quality process, which is purposefully defined to address the specificities of ETER data collection. This process combines different methods, including a systematic analysis of internal quality of data (format accuracy, completeness, consistency, and timeliness), and advanced statistical methods for outlier detection and analysis of comparability, based on metadata for checking external validity by comparing ETER data with other data sources.
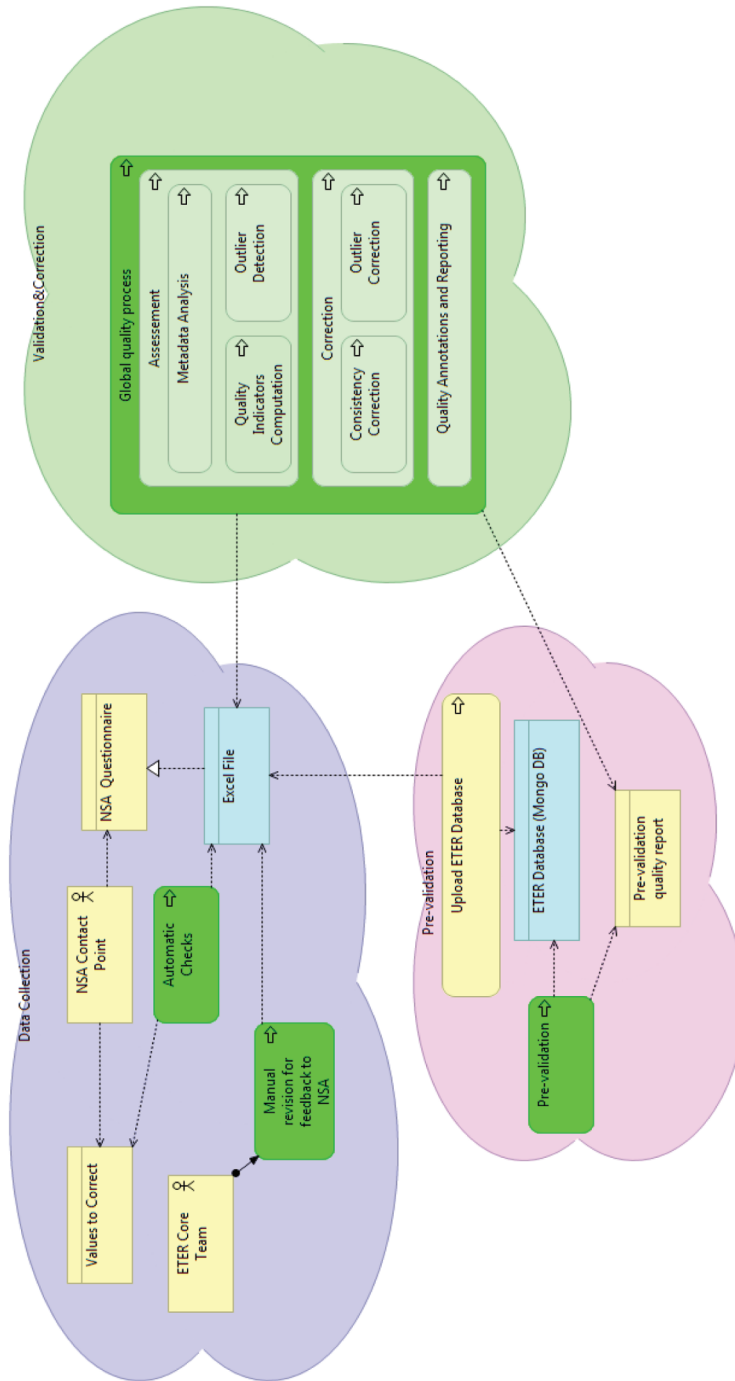
Figure 1.　Overall ETER data quality process.

Research Paper

More specifically, the ETER Quality Validation and Reporting process consists of the following phases:

- Quality Metadata Collection, contextual with data collection.
- Quality indicators calculation and validation checks performed within the data collection phase on a country basis and on the whole dataset. They are described in the following.
- Multi-annual checks.
- Cross-sectional ratios to detect comparability problems.
- Investigation of the comparability dimension on the base of the previous analysis.
- Checks with external data sources, either to assess the overall coverage against official statistics, national aggregates, or explain/correct problems detected in the previous steps.

The overall ETER data quality process is depicted in Figure 1, where the darker green processes are the specific ones introduced to deal with data quality:

- During the Data Collection phase, both manual and automated checks are performed.
- Right after the Data Collection, a Pre-validation phase is carried out.
- A dedicated Quality Review and Correction phase is later performed, based on the methodology illustrated in the next section and a relevant Quality Annotation and Reporting process, also following Eurostat indications (Eurostat, 2014, 2019). See Appendix 1 for the annotation and flags adopted.

The data quality indicators adopted within the ETER project belong to the ISO25012 and relate to the internal data quality, they are:

- *Accuracy* To evaluate the conformity of the provided values to the specified format in the collected data sets.
- *Completeness* To evaluate the number and meaning of missing values that are present in the collected data sets.
- *Consistency* To verify possible violations of semantic rules defined over the involved data, and specifically between different variables.
- *Timeliness* To evaluate the lapse of time between the ETER collection date and the source release date.

Here we describe briefly the checks performed to investigate accuracy, completeness, and consistency (for further details, see Lepori et al., 2018).

Accuracy checks

Accuracy checks verify that data entered have the right format foreseen by the handbook and that no logically impossible values are found. These checks are performed in the data collection sheet and on delivered data. Simple mistakes are corrected directly, whereas unclear cases are reported back to NSAs/NEs for clarification.

Completeness checks

No blank cells are allowed in the dataset, except for remarks. Blanks should be recoded correctly as missing, confidential, not applicable, or "0". This control is extremely important for the final quality of the database. Blank cells are highlighted automatically. Clear cases are recoded directly and ambiguous cases (for example between missing and not applicable) are reported back to national experts and NSAs for clarification.

Consistency checks

a) These checks control for logical consistency between different variables (for example when the highest degree delivered is at ISCED 7 level, all values for students and graduates at ISCED 8 level should be not applicable). See Appendix 2 which reports the list of consistency indicators.

b) Further, these checks control whether the sums of breakdowns by subcategories equals the total and numerical relationships between values (example R&D expenditures lower than total expenditures). Deviant values are identified and checked. In case there are specific reasons, an explanation is added to the metadata for that specific HEI.

Check of missing data

An analysis of missing data is performed (including also issues of breakdowns by subcategories). When it is expected that data should be available, possibly with some limitations, this is requested to NE/NSAs.

Control of metadata completeness

Metadata are systematically controlled for the completeness, taking into account also issues emerging from the checks on the data. When metadata are missing or incomplete, further information is requested.

Expert checks

Expert checks based on knowledge of national systems, as well on information available on the Web and EUMIDA data, are performed in order to ensure that provided data are realistic. Potential problematic cases are notified back to national

Research Paper

experts and NSAs. When these are related to methodological issues, the corresponding remarks are integrated in the metadata.

The data quality management system in the ETER project has been built to meet several challenges of large-scale international data collections. The first challenge was to use a reproducible scalable system, which can be applied to 37 different countries. Secondly, the process needs to keep the workload for data deliverer as small as possible and to reduce the margin of error as much as possible. Thirdly, data from different departments are collected in one template. This demands extensive control mechanisms in order to ensure that no inconsistencies between the data exist. In order to meet these challenges, the following data quality management procedure has been developed within the ETER project (see Figure 2):

1. Because of its widespread usage, Microsoft Excel has been chosen as tool for the perimeter validation and data collection. In the first step, national authorities identify the higher education institutions for the respective data collection year. Any demographic events are tracked and added as variables to the dataset. After confirmation of the perimeter, data collection templates are sent out. These templates are prefilled with information based on previous years, which is not expected to change (e.g. foundation year, geographical information etc.).

2. The data collection files already include a high number of control mechanisms in order to make the data deliverer aware of potential irregularities. These mechanisms screen the template for issues in completeness (for mandatory variables), accuracy (e.g. a NUTS 2 region need to have four digits) and consistency (e.g. sums of breakdowns equal to the aggregate variable). Specifically, consistency checks verify a possible violation of semantic rules defined over the involved data, and specifically between different variables. The list of indicators and involved variables is reported in Appendix 2. In order to prevent any overwriting of automated checks, a macro has been implemented into the data collection file. This macro allows only pasting values and therefore ensures that the preprogrammed automatism cannot be bypassed.

3. Already clean data are then imported into the database, where the data collection is managed henceforward. Data collection files can be produced by the ETER infrastructure if updates or changes are necessary.

4. Automated data validation and data quality checks are run on the imported data on the database. Data validation is an extended form of the control mechanisms already implemented in the data collection templates (completeness, accuracy, and consistency) and produces a pdf report per country and academic year. Then, an extensive automated data quality

procedure is performed on the data. This internal data quality process includes a multiannual analysis as well as cross-sectional outlier detection. Suspect cases are either corrected (in cooperation with NSAs) or flagged. Additionally, the ETER data are subject to external data quality controls, where the data are compared to equivalent data like EUROSTAT national aggregates or data from U-Multirank.

5. The final product of this procedure is a high-quality dataset, which is published in spring of each year on the ETER web interface. Because of continuous work on the dataset, updates in terms of additions or corrections are regularly.
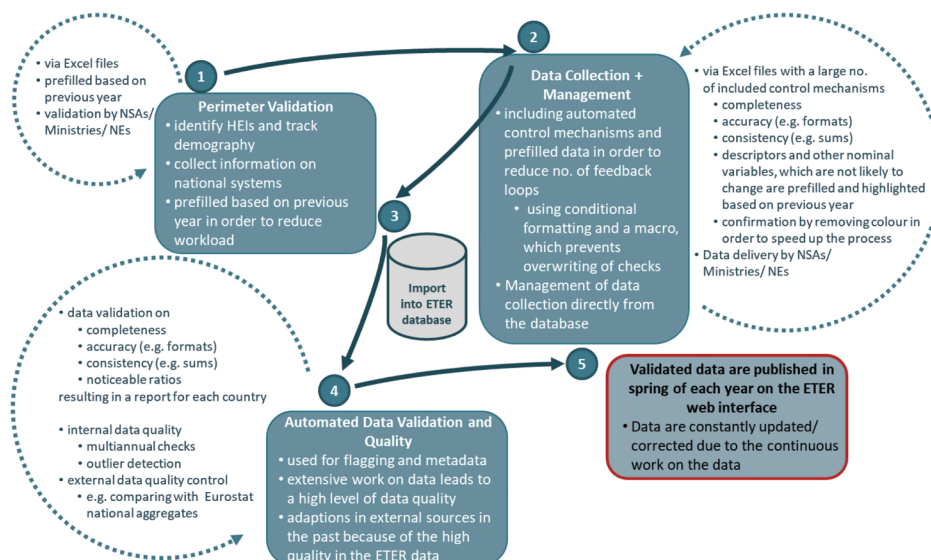
Figure 2.   Current approach in data quality management.

The effectiveness of the presented approach has been validated by proof-of-concept and user experience. At the side of the data deliverer, the implemented process minimized the additional quality control burden focusing the interactions on a limited number of selected cases, identified by statistical analysis and automated controls.

5   Methodology

The overall data quality management process, described in the previous section and based on the ETER Data Quality Report (Daraio et al., 2018), combines different approaches. It includes a systematic analysis of internal quality of data (format accuracy, completeness, consistency, and timeliness), the analysis of comparability

Research Paper

based on metadata and check of external validity by comparing ETER data with other data sources. In this section, we describe the data quality checks developed to identify outliers, extreme observations, and to detect ontological inconsistencies not described in the available meta-data which constitute the main objective of this paper.

Given the specificities of the ETER database, the methodology developed to check the consistency and stability of data over time is based on an empirical-oriented approach, which analyzes the observed distributions of the relevant variables without referring to pre-defined theoretical data distributions. This is different from what was done previously within the ETER project when outliers were identified by means of the approach implemented in the R package "extreme values", which compares the empirical data with theoretical distributions.

In the following, we describe the logic and functioning of the multiannual checks as well as the cross-sectional checks which complement them.

5.1   Multiannual checks

Each institution I_j contains the values v_{ji} of a number of variables, some of which usually change over the time horizon used in the database (the years go from 2011 to 2016). Examples of these variables are "number of students" or "number of graduates". To lighten the notation, when there is no ambiguity, we will denote the values of the generic time series for one of these variables with $v_1, v_2,…, v_t$ (without explicit reference to the index j of the institution), and the set of years indices of the time series simply with 1, 2, …, t = T.

The availability of data across different years raises the issue of longitudinal consistency of the data collected (impact of demographic events; revision of variable's categories and definitions, etc.). On the other hand, the availability of several yearly editions of data offers an additional possibility for quality control. Indeed, multi-annual checks can help to detect suspect cases where the level of variation from year to year is very large or otherwise anomalous when compared with the average changes in the sample. This type of check is particularly useful in detecting and reporting mistakes of respondents and/or changes in the methodology for data collection.

The availability of only six years of data, however, does not allow the use of methods specific for time series analysis, which requires much longer time series. Moreover, ETER dataset contains different typologies of variables (e.g. structural descriptors rather than quantitative variables) with a different propensity to change over time.

For these reasons, the methodological approach developed for the multiannual checks consists of multiple procedures and it is based on the use of different techniques:

A Tailor-made Data Quality Approach for Higher Educational Data
Cinzia Daraio et al.

Research Paper

- manual check of the impact of demographic events (take-over, spin-off) on concerned institutions' figures and respective flagging (the code "b" for breakdown in time series was already foreseen);
- analytic control of descriptors and status variables supposed to be stable over time, i.e. legal status, foundation year, geographical information, lowest/ highest degree awarded, etc.;
- comparison of national aggregates over time for a selected number of quantitative variables already during the validation phase, with an alarm if the variation is over a pre-defined threshold;
- use of measures of statistical dispersion (interquartile range comparison over time) to assess the overall stability of the distribution of quantitative variables;
- statistical analysis to highlight the HEIs with annual growth outstanding from the overall distribution (outlier).

The approach proposed in this work for the checks of the described time series has been developed in order to be flexible and scalable. It is based on thresholds and parameters, which can be tuned tacking into account expert knowledge or may be determined from the empirical distribution of the observed data. This approach is easily implementable and can be executed within the most common software tools used in data management, e.g. R, Matlab or even MS Excel. This approach has been adopted in the data quality assessment of the ETER European research project in replacement of the previous methodology based on outlier analysis.

The proposed approach relies on two types of controls to identify potentially erroneous time series in the HEIs:

1) *Check of the discontinuity*: this control is aimed at identifying large variations in the values of the variable under analysis, and therefore capturing its volatility over time. It is based on the computation of the annual variations, called deltas, and on its possible normalization using a measure of the size of the institution. A scale invariance parameter is defined, in order to choose the desired level of normalization.

2) *Check of the variance of deltas*: this control is aimed at identifying fluctuations in the size of the deltas, i.e. a second order information with respect to the value appearing in the time series. This information allows to identify institutions having an overall moderate range of variation, which are not detected in the previous control, but having anomalous isolated "jumps". Again, normalization is possible using a measure of the size of the institution, setting a scale invariance parameter.

After their identification, the HEIs containing inconsistent values in the time series should be validated or corrected by using subsequent procedures, depending on the specific case (for instance, checking external sources of the same data).

Research Paper

In more detail, the methodology is composed of the following four steps.

i) Exclusion of irrelevant HEIs. This step is performed because very small institutions may exhibit very large percentage fluctuations in the values simply because of their small size, without necessarily revealing errors. For example, the number of students in a very small university may easily double or halve from one year to the next. An analysis of similar cases would be quite complex, and on the other hand, its impact on the global situation would be negligible. Thus, small institutions are generally excluded from data quality checks. To determine where to set the division between relevant and irrelevant HEIs, we compute the geometric mean μ of the whole time series $v_1, v_2, \ldots, v_t$ for the variable under analysis, and consider μ a measure of the size of the institution for that variable.

$$\mu = \left(\prod_{i=1}^{t} v_i \right)^{1/t} \tag{1}$$

Then, we compute a threshold S_1 such that $\mu \geq S_1$ for a predetermined percentage of cases (e.g. 95% or 98%). Now, any institution with $\mu < S_1$ will be considered irrelevant for the variable under analysis. Computation of the *Discontinuity Measure* (DM) and *Jump Variance* (JV) for each HEI. We call *delta* the difference between any two consecutive values of the time series of the variable under analysis in a given institution. To lighten the notation, we will not explicitly write the indices of the variable and of the institution, obtaining so $\delta_1 = v_2 - v_1$, $\delta_2 = v_3 - v_2$, and so on. The set of the delta values of an institution will be denoted as Δ. Within Δ, we consider the sum of the deltas having positive values, denoted as $\Delta+$, and the sum of the deltas having negative values, denoted as $\Delta-$. Then, we compute the *Discontinuity Value* (DV) of the variable under analysis in the given institution as the absolute value of the product of the two mentioned sets, as follows.

$$DV = | \Delta+ \Delta- | \tag{2}$$

In other words, DV measures the amount of the "jumps" in the time series of the variable under analysis in the given institution, and it reduces when all jumps tend to be in the same direction. This evaluates the "discontinuity" in the time series. In order to introduce scale invariance at a controlled intensity, DV is divided by the geometric mean of the same variable of the given institution raised to a power σ, obtaining DM

$$DM = DV / \mu^{\sigma} \tag{3}$$

When σ = 1, DV is fully "normalized" by the size of the institution, obtaining so a *scale invariant* measure. On the other extreme, when σ = 0, the value of DV is fully dependent on the size of the institution. Any value of σ between 0 and 1 can also be selected, and that will constitute the desired *level of scale invariance*.

Another measure computed with the set Δ is the *Jump Variance* (JV), that is the variance of the elements in Δ, computed as follows, where |Δ| denotes the cardinality of the set Δ.

$$JV = \Sigma_i \, (\delta_i - \delta_{mean})^2 \, / \, |\Delta| \qquad (4)$$

The aim of the JV measure is to identify time series having a not excessive value of DV, hence not highlighted by the DV measure, but containing some anomalous jumps, for example, because one isolated value in the time series contains an error. This value can again be normalized by using the power μ^σ of the geometric mean, with a technique very similar to the previous case, obtaining the *Jump Diversification* (JD)

$$JD = JV \, / \, \mu^\sigma \qquad (5)$$

ii) Issue of *alarm flags*. For each HEI in the relevant sample of the variable under analysis (that is, with $\mu \geq S_1$), we determine whether to mark it with alarm flags or not by using the following criteria:

    a. HEIs with the highest values of discontinuity measure DM for the variable under analysis (e.g. the top 5% or 10%) are flagged with Alarm 1. The demarcation value will be called S_2.

    b. HEIs with the highest values of jump diversification JD for the variable under analysis (e.g. the top 5% or 10%) are flagged with Alarm 2. The demarcation value will be called S_3.

iii) *Check* of the alarmed HEIs. Finally, we check the institutions which received alarm flags and therefore appear to contain one or more inconsistent series of values. Note that, due to the nature of the data, the presence of alarm flags does not guarantee the presence of errors, but only that the time series are "uncommon". Now, depending on the specific case, correction or validation can be performed by checking external sources of the same data, by inspection, etc.

To further clarify the described approach, we analyze in detail the case of the variable "Enrolled students ISCED 5-7". The same approach has been applied to the other relevant variables listed in Table 2.

Research Paper

Step i) We compute the geometric mean μ over the years for the number of students enrolled for each single HEI, and we find the threshold S_1=142 to exclude 5% of the smallest HEIs.

Step ii) For each HEI not excluded by the previous step, we compute the set of the yearly variations Δ of the number of students enrolled, from which we compute $DV = |\Delta + \Delta -|$ and $JV = \Sigma_i (\delta_i - \delta_{mean})^2 / |\Delta|$ for each HEI. Then, we select the scale invariance parameter σ = 0.5, thus we choose a partial level of scale invariance. Finally, for each HEI, we compute $DM = DV / \mu^\sigma$ and $JD = JV / \mu^\sigma$.

Step iii) We set alarm flags for the values of students enrolled for all the relevant HEIs in the top 5% of the values of DM and for all the relevant HEIs in the top 5% of the values of JD. Note that an institution can also receive both alarm flags, but one is enough to require check. The threshold identified on DM with this procedure is $S_2 = 2.08$ while that on JD is $S_3 = 1.96$. The total number of flagged HEIs is 285.

Step iv) The 285 HEIs flagged at Step iii) for their values of students enrolled have been checked. In particular, every Country expert examined those belonging to his/her Country and took the adequate actions after consultation with national statistical offices and/or ministries. The actions have been of 3 possible types: (a) confirming suspect data; (b) flagging and explaining, (c) correcting the data.

Table 2. List of variables considered for the multiannual checks.

Variable
Total expenditure (PPP)
Total revenues (PPP)
Total academic staff (FTE)
Total academic staff (HC)
No. of administrative staff (FTE)
Total staff (FTE)
Total staff (HC)
Total students enrolled (by ISCED level)
Total graduates (by ISCED level)

5.2   Cross-sectional checks

The multiannual checks are complemented by a check to control for cross-sectional consistency. The method is based on the analysis of the distribution of the values of a ratio between two interrelated variables, e.g. the amount of personnel expenditure and the number of staff. To take into account specificities linked to country settlement and type of institution, the method has been applied to sub-distributions of HEIs:

– By country
– By institutional category (university, university of applied sciences, other).

The analysis may not include HEIs below a minimum size thresholds since experience tells us that very small institutions show a number of contextual peculiarities and are sort of "outliers" by definitions (in principle very small institutions can be left outside ETER perimeter, but many countries include them). The method is applied as follows:

– For each ratio R_i (the considered ratios are listed in Table 3 and indicated as $R_1,\ldots, R_i, \ldots, R_8$) starts the analysis by computing its values on all the institutions and sorts its values ascending;
– Identify the value of the ratio that leaves out e.g. the 5% of the cases with the lowest value and call it X_l and identify the value of the ratio that cuts out the 5% of the cases with the highest values and call it X_u;
– Create two sets of sub-distributions for the analysis; the first set is defined according to the institutional category which contains three sub-groups of records including respectively universities, universities of applied science, other; in the second set records are grouped by country;
– For each sub-distribution calculate the value of the aggregate ratio "R_s"- obtained dividing the aggregate value of the variable at the numerator by the aggregate value of the variables used as denominator;
– For each sub-distribution calculate the ratio for each record, sort the ratios ascending and identify the records with the value of ratio below $X_{l,Rs}$ and above $X_{u,Rs}$ where X_l and X_u are the parameters calculated on the overall distribution and "Rs" indicates the value of the aggregate ratio of the sub-distribution
– Alarm cases which are either below the lower-bound thresholds ($X_{l,Rs}$) and above the upper-bound one ($X_{u,Rs}$).

This multistep methodology allows treating in a simple way the heterogeneity of the higher education systems in different countries and across them with reference to categories of institutions.

Table 3.　List of cross-sectional ratios for checks.

Code	Name
R_1	Enrolled Students / Academic Staff
R_2	Academic staff / Total staff
R_3	Personnel expenditure / Total staff
R_4	Personnel expend. / Total expenditure
R_5	Total expenditure / Total revenue
R_6	Basic Government funds / Total revenue
R_7	Graduates ISCED 5-7 / Enrolled students ISCED 5-7
R_8	Graduates ISCED 8 / Enrolled students ISCED 8

The described controls can be applied in ETER in two different phases:

1. Preliminary validation checks, performed directly within the data collection phase on a country basis, in order to allow for an easy return from the respondents, to correct the data before online integration;
2. Further in-depth quality checks and validation, to perform more accurate controls that can also provide indications about appropriate data usage and possible quality improvements for future data collections.

6   Results

6.1   Results from the multiannual checks

The methodology for multiannual checks described above has been applied to 19 of the ETER variables considering all reference years (2011–2016), keeping the year 2016 as a base. Around 2,800 cases, spread in 33 countries, have been highlighted and checked in detail (Table 4a and 4b).

The distribution by country, in general, follows the size of the country in terms of the number of institutions in ETER, with the six larger countries (DE, ES, IT, PL, TR, the UK) accounting for around one half of the cases.

The detected cases can be grouped into three categories:

1. Breaks in time series already known and flagged, both as a consequence of demographic events or methodological discontinuities. Examples are the change in the classification of curricula in Spain from 2013 onward, a different method for counting academic staff in Swiss UAS in 2013, etc.;
2. Country systemic issues, involving a large number of HEIs in one country, and therefore pointing to breaks in time series, which have not been notified or flagged before. Examples are a generalized drop of ISCED5 students and a paralle increase of ISCED6 in 2014 in Ireland, the sharp drop of academic staff HC registered in Italy in 2014, etc.;
3. Individual cases, which may be the consequence of errors in data reporting or to peculiarities of the institution (i.e. recently founded HEIs show tremendous growth in the first year).

All cases in categories 2 and 3 have been controlled individually by the consortium interacting with NSAs.

In terms of variables, more than two third of cases concern student population (students and graduates) but volatility emerges in all variables considered.

The multiannual checks beside the identification of individual outliers cases and mistakes in the reporting that were revised with data providers and corrected,

A Tailor-made Data Quality Approach for Higher Educational Data   Cinzia Daraio et al.

Research Paper

Table 4a.   Outcome of the multiannual checks. Number of cases detected by variable and country.

Variable	AT	BE	BG	CZ	DE	ES	GR	HU	IT	NL	PL	PT	RS	SI	TR	UK	Total
Academic Staff FTE	9		6	9	10	10		9			12	9	1			5	80
Academic Staff HC	20	1	8	5	40	8	14	16	35			12	1		3	13	176
Non-Academic Staff FTE	4		10	8	24	4		21		2	29					5	107
Total current expenditures (NC)	1	1		4	14											3	23
Total current revenues (NC)				2	19						13					6	40
Total Graduates ISCED 5								8		1					4	18	31
Total Graduates ISCED 5–7	12		7	5	35	22	2	11	14		48	13	3	5	19	8	204
Total Graduates ISCED 6	7		5	3	21	44	5	13	10		36	5	9	4	10	4	176
Total Graduates ISCED 7	5	1	5	1	18	9	7	17	8		23	17	7	1	12	9	140
Total Graduates ISCED 7 long degree	4		1	1	6	39			4		11	1				5	72
Total Graduates ISCED 8	3		15	4	16	13	6	4	11			6	2		9	15	105
Total Staff FTE	4		4	9	12	4		17		1	16	36				1	103
Total Staff HC	4		3	7	20	2	11	14	11			34				4	110
Total Students ISCED 5	10							9		6					6	19	50
Total Students ISCED 5–7	9	3	10	3	16	8	35	11	26	1	36	13	2		15	17	205
Total Students ISCED 6	2	2	4	3	11	27	32	7	35		25				12	3	163
Total Students ISCED 7	3	2	9	3	11	13	9	22	67	1	22	19	5		14	23	223
Total Students ISCED 7 long degree	2				2	47			3		12					2	68
Total Students ISCED 8	2	2	13	1	4	5	7	1	24	1	2	11	3		16	17	109
All variables	101	12	100	68	279	255	128	180	248	13	285	176	33	10	120	177	2,185

Research Paper

Table 4b. Outcome of the multiannual checks. Number of cases detected by variable and country.

Variable	AL	CH	CY	DK	EE	FI	HR	IE	LI	LT	LV	MK	MT	NO	SE	SK	Total
Academic Staff FTE		1	2			4	2			15				4		13	41
Academic Staff HC		9	1				2	4		5	1	1		5		13	41
Non-Academic Staff FTE		1	3			5				9				3	1		22
Total current expenditures (NC)			2			1		1		5				3		2	14
Total current revenues (NC)										3					1	12	16
Total Graduates ISCED 5				1				15			5				1		24
Total Graduates ISCED 5-7			8	2		3	8	2	1	6	7		2	11	9	5	64
Total Graduates ISCED 6			4	1		1	1	1		1	2		2	7	3	4	24
Total Graduates ISCED 7			5		1	2	1	4		1	4			2	5	4	30
Total Graduates ISCED 7 long degree										1				1	6	3	11
Total Graduates ISCED 8				1			5	2						2	2	5	21
Total Staff FTE			1			3				4	1					10	32
Total Staff HC		7						3		15	1	2		3		9	32
Total Students ISCED 5			3	5		1		14		5	5		2	5	2		27
Total Students ISCED 5-7	18		11	18		1	6				4	9	1	5	14	7	96
Total Students ISCED 6			8	7			2	6		1			1	5	13	4	51
Total Students ISCED 7			9		1		1	2			4	9	1	3	13	7	51
Total Students ISCED 7 long degree						1				1				1	1		2
Total Students ISCED 8			2	1			5	4			3			1		7	24
All variables	18	18	59	36	2	22	33	58	1	72	37	21	9	61	71	105	623

allowed highlighting problems of comparability across waves of data collection. In several cases, changes in curricula with an impact on their classification according to the ISCED system caused a sudden increase/decrease of the number of enrolled students and degrees in specific ISCED level. The changes in the figures therefore did not reflect substantial changes of the pattern of enrollment in higher education, but a simple normative effect. During the years, several countries (e.g. Spain) were affected by these changes due to the still ongoing adaptation to the Bologna process. Other artificial changes due to the administrative rules were found concerning counting methods and rules for reporting staff figures, especially contract and part time staff (e.g. Switzerland and Italy).

6.2 Cross-sectional ratios for consistency analysis

Financial data are gathered according to ETER breakdowns, this fact may induce issues on the quality of such data since ETER itemization could mismatch with the categorization adopted by National Authorities that provide the data funneled into ETER.

Concerning revenues, national data are available with different levels of granularity; in Italy, for example, the content of some revenue breakdowns (i.e. basic core budget and other core budget) differs between state and non-state HEIs because of the different granularity of the national data available. The match with ETER in most cases is made without significant problems of attribution. However, some categories may not perfectly match. For example, referring to non-state universities in the UK, data are available in domestic statistics with lower level of details; also the content of some revenue breakdowns (i.e. basic core budget and other core budget) differs between state and non-state universities because of the different granularity of the national data. Finally, on some occasions, non-recurring revenues are not distinguishable from the others, although regular funding is only a share of total current revenues.

Considering costs and expenses, national authorities provide data in different categories sometimes including depreciation, depending on the accounting system adopted by HEIs. However, depreciation is not included in ETER reporting, since it considers capital expenses according to a cash accounting approach, as in the majority of countries. This fact may create mismatches since the perimeter of capital expenses may differ for those HEIs that adopt the accrual accounting system and consider depreciation of fixed assets in their financial reports instead of registering capital expenses. For example, in Italy, during the time span covered by ETER, there has been the progressive implementation of the reform of the university accounting rules. As a result of these changes, universities have progressively adopted an

Research Paper

accrual accounting system and have moved to the "single budget", consolidating the data of all the centers with managerial and administrative autonomy which make up the organizational structure of the universities. In such cases, without an expert based control on data, inconsistent information may be provided in both cases when comparing HEIs within the same country or across different countries.

Considering the total employees count, some data inconsistencies may arise since HEIs may utilize different methods in counting staff (Head Count vs FTE) or may include in the academic staff different categories of employees, for example, some university hospitals may include doctors in specialist medical training within the academic staff.

Following the previous considerations, it emerges that detecting inconsistent data and comparability issues within the same country and the same year amongst different HEIs is a very important topic.

Heterogeneity in terms of size and span of HEIs (for example, polytechnics vs general purpose universities) may hide some comparability issues. Namely, comparing numbers of different orders of magnitude may not produce useful information while, after a normalization process, comparisons may provide better insights. Thus, the approach of financial/managerial ratios (Woelfel, 1987) has been deemed suitable to compare HEIs with different orders of magnitude and span. Following this approach, it was defined as an *ad hoc* set of ratios. Each ratio has been defined as the relative magnitude of two selected numerical values taken from ETER, in particular ratios were built mainly with the purpose of analyzing financial and staff data.

By comparing the ratios with national standards or relevant (expert based) threshold values, data inconsistencies or comparability issues may be detected. Moreover, although ratios may not be directly comparable between HEIs that adopt different accounting methods, they could help in detecting such differences.

After several experiments and group discussions, a set of eight ratios was defined mainly considering financial and staff data, thresholds have been defined through an expert based approach in order to spot data inconsistencies both within each country and between different countries. As a result, over 2,000 cases have been detected in the first test on the last wave of data collection. Table 5 lists the set of ratios to detect comparability issues that have been proposed and implemented and it also reports the percentage of detected cases for each ratio. It emerges, at a first glance, that the majority of detected cases involved the staff counts, either as the percentage of academic staff on enrolled students (R_1) or as the percentage of total staff (R_2), also the percentage of staff on personnel expenditure (R_3).

Table 5. Cross-sectional ratios for consistency analysis.

Description	Code	%
Enrolled Students / Academic Staff	R1	27.6%
Academic staff / Total staff	R2	18.7%
Personnel expenditure / Total staff	R3	15.1%
Personnel expend. / Total expenditure	R4	5.2%
Total expenditure / Total revenue	R5	8.9%
Basic Government funds / Total revenue	R6	6.1%
Graduates 5-7 / Enrolled students 5-7	R7	15.0%
Graduates 8 / Enrolled students 8	R8	3.2%

Table 6 reports the percentage of detected cases for each ratio and in total for each country involved in ETER.

Table 6. Cross-sectional ratios – a country by country reporting.

	R1	R2	R3	R4	R5	R6	R7	R8	Total
Austria	8.0%	1.9%	0.9%			8.1%	1.7%		3.4%
Belgium	0.5%	0.2%							0.2%
Bulgaria	4.0%	1.1%					0.7%	4.4%	1.6%
Croatia							2.1%		2.0%
Cyprus	1.3%	1.7%	8.9%	10.2%	6.3%		5.0%		3.9%
Czech Republic	6.7%	4.7%	0.2%	2.7%		17.3%	4.0%		4.6%
Estonia				4.1%	2.4%	0.6%	1.2%		31.4%
Finland					0.8%	1.7%			2.3%
Germany	21.3%	26.5%	53.5%	55.8%	70.6%	1.2%	16.1%	22.0%	3.2%
Greece	7.3%	1.1%					6.4%		3.3%
Hungary	1.8%	13.1%					2.1%		2.6%
Ireland	0.1%	0.8%	0.9%		1.6%		0.2%		1.3%
Italy	6.8%	4.2%					4.5%	16.5%	0.8%
Latvia	2.6%	4.2%	11.5%	2.0%	3.6%		1.4%	1.1%	6.4%
Lithuania	0.5%	2.1%	11.0%	1.4%	2.4%	0.6%	0.2%	1.1%	4.4%
Luxembourg	0.1%					0.6%			1.1%
Malta			0.5%	0.7%	0.4%	0.6%			1.2%
Netherlands	0.8%	0.6%	3.8%		0.8%	0.6%		9.9%	2.7%
North Macedonia	0.3%	0.6%							9.2%
Norway	1.4%	1.3%					0.9%		0.5%
Poland	11.2%	3.6%				19.7%	9.5%	1.1%	3.9%
Portugal	2.6%	16.3%	0.2%	2.0%	1.6%	1.2%	1.9%		0.1%
Serbia	2.2%						1.9%	5.5%	3.9%
Slovakia	0.4%		1.4%	2.0%	0.4%	11.6%	6.6%	15.4%	0.2%
Slovenia							5.9%	5.5%	3.7%
Spain	3.6%	1.9%					4.0%	9.9%	0.2%
Sweden	1.0%	0.2%	0.5%	1.4%		8.7%	1.7%		0.6%
Switzerland	2.1%	0.2%	3.5%	11.6%	1.6%	1.2%	0.5%		0.2%
Turkey	7.5%						9.7%	4.4%	0.3%
UK	6.0%	13.8%	3.1%	6.1%	7.5%	26.6%	11.6%	3.3%	1.1%

7   Discussion and conclusions

As recalled, the ETER features lead to specific challenges for the data quality process due to the nature of microdata, the extreme heterogeneity of typologies of rules and HEIs' categories across countries, the lack of control over the complete data collection process.

The methodology developed to account for these specificities combines quantitative statistical checks with an expert-based interaction with the national data providers. This approach can be complemented with the results of imputation procedures to fill in missing data (Bruni, Daraio & Aureli, 2020).

A strength of the developed approach is its empirical-oriented flexibility that allows the user to personalize the quality investigation with respect to the observed distribution of the variables considered instead of using theoretical-based distribution functions for the data analysed.

Given this flexibility, the developed methodology could be extended, mutatis mutandis, to other institutional data of Higher Educational systems of those countries for which there is a lot of information available in documents and public sources but it has not been collected and integrated yet in a unique register for the monitoring of the system over time.

The current methodology, although consolidated and assessed, could be improved in different ways. It could be useful to invest in better combining multiannual and cross-sectional checks to further reduce the number of cases to inspect manually and to pre-identify the problems or possible explanations, going towards the implementation of a fully automatized control. This ambitious goal would require the revision of the current architecture of the data and of the overall data quality management process.

Another limit of the current approach to data quality relies on the reporting of data quality information. Although flags are incorporated in the dataset for each variable, the information about the explanation of the problems and the way they should be treated (i.e. if the impact on the comparability of the data is high, low or null) is fragmented in different sections of the dataset including the notes available for groups of variables, metadata at the variable level and additional more in-depth information. This fragmentation of the relevant information and the difficulties to read together with the data and the metadata hamper a full data quality aware use of the data, especially by policy makers or analysts who are not specialists in the field.

The main challenges in ETER data collection that remain open are:

– Dealing with the heterogeneity of data sources through a formal and unambiguous way of representing metadata. Computational ontologies can be a possible solution in this direction, being able to provide an harmonized view on concepts expressed in a machine-readable way.

– Dealing with "advanced" quality controls. In several cases, quality checks can go beyond syntactic representation and, instead, be based on the semantics of the concepts. For instance, "total expenditures" should be properly specified in terms of mandatory and optional components, as, e.g. "R&D expenditures" are available only for a subset of countries.

Daraio et al. (2016a, 2016b) introduced the ontology based data management approach to coordinate, integrate, and maintain the data needed for science, technology and innovation policy and illustrate its potentials for specifying Science, Technology and Innovation indicators and developing science of science policies. They outline the main advantages of OBDM that are conceptual access to the data, re-usability, documentation and standardization, flexibility, extensibility, openness, interoperability, and data quality.

In the future, a possible development is inherent to addressing the above cited challenges and having a "modernized" ETER data collection that could benefit to ontologies and Semantic Web models and languages.

In particular, the use of such an ontology based approach could allow to achieve (i) a harmonized data collection, overcoming sources heterogeneities and (ii) richer quality controls, which could be specified in a declarative way and be designed on the basis of an explicit semantic representation of concepts involved in the ETER data collection.

Acknowledgments

Author contributions

Proposing the research problems and designing the research framework: Cinzia Daraio (daraio@diag.uniroma1.it). Performing the research: Renato Bruni (bruni@diag.uniroma1.it), Giuseppe Catalano (catalano@diag.uniroma1.it), Alessandro Daraio (a.daraio@gmail.com), CD, Giorgio Matteucci (matteucci@diag.uniroma1.it), Monica Scannapieco (monica.scannapieco@istat.it), Daniel Wagner-Schuster (Daniel.Wagner-Schuster@joanneum.at), and Benedetto Lepori (benedetto.lepori@usi.ch). Collecting and analyzing the data: RB, AD, CD, GM, DW-S, and BL. Writing and revising the manuscript: CD, RB, GC, AD, GM, MS, DW-S, and BL.

References

Batini, C., & Scannapieco, M. (2016). Data and information quality. Springer. doi: 10.1007/978-3-319-24106-7

Research Paper

Bonaccorsi, A., & Daraio, C. (2007a). Theoretical perspectives on university strategy. In A. Bonaccorsi & C. Daraio (Eds.) Universities and Strategic Knowledge Creation. Specialization and Performance in Europe (pp. 3–30). Cheltenham: Edwar Elgar.

Bonaccorsi, A., & Daraio, C. (2007b). Universities as strategic knowledge creators: Some preliminary evidence. In A. Bonaccorsi & C. Daraio (Eds.) Universities and Strategic Knowledge Creation. Specialization and Performance in Europe (pp. 31–81). Cheltenham: Edwar Elgar.

Bonaccorsi, A., & Daraio, C. (2007c). Universities and Strategic Knowledge Creation. Specialization and Performance in Europe. Cheltenham: Edwar Elgar.

Bonaccorsi, A., Daraio, C., Lepori, B., & Slipersaeter, S. (2007). Indicators on individual higher education institutions: Addressing data problems and comparability issues. Research Evaluation, 16(2), 66–78.

Borgman, C.L. (2015). Big data, little data, no data: Scholarship in the networked world. MIT press.

Bruni, R., Daraio, C., & Aureli, D. (2020). Imputation Techniques for the Reconstruction of Educational Institutions Data, Technical Report DIAG, Sapienza University of Rome.

Capano, G. (2011). Government continues to do its job. A comparative study of governance shifts in the Higher Education Sector. Public Administration, 89(4), 1622–1642.

Daraio, C. (2017). A framework for the assessment of research and its impacts. Journal of Data and Information Science, 2(4), 7–42.

Daraio, C. (2018). Nonparametric Methods and Higher Education. In: Teixeira P., Shin J. (Eds) Encyclopedia of International Higher Education Systems and Institutions. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-9553-1_82-1

Daraio, C. (2019). Econometric approaches to the measurement of research productivity, in Springer Handbook of Science and Technology Indicators edited by Glänzel W., Moed H.F., Schmoch H. and Thelwall M., 633–666.

Daraio, C., & Bonaccorsi, A. (2017). Beyond university rankings? Generating new indicators on universities by linking data in open platforms. Journal of the Association for Information Science and Technology, 68(2), 508–529.

Daraio, C., Bonaccorsi, A., & Simar, L. (2015). Rankings and university performance: A conditional multidimensional approach. European Journal of Operational Research, 244(3), 918–930.

Daraio, C., Bruni, R., Catalano, G., Matteucci, G., Daraio, A., Scannapieco, M., Wagner-Schuster, D., & Lepori, B. (2019). European Tertiary Education Register (ETER): Evolution of the Data Quality Approach, in Proceedings of the 17th International Conference on Scientometrics & Informetrics, 2–5 September 2019, pp. 2766–2767.

Daraio, C., & Glänzel, W. (2016). Grand challenges in data integration. State of the art and future perspectives: An introduction. Scientometrics, 108(1), 391–400.

Daraio, C., Lenzerini, M., Leporelli, C., Moed, H.F., Naggar, P., Bonaccorsi, A., & Bartolucci, A. (2016). Data integration for research and innovation policy: An ontology-based data management approach. Scientometrics, 106(2), 857–871.

Daraio, C., Lenzerini, M., Leporelli, C., Naggar, P., Bonaccorsi, A., & Bartolucci, A. (2016). The advantages of an ontology-based data management approach: Openness, interoperability and data quality. Scientometrics, 108(1), 441–455.

Daraio, C., Scannapieco, M., Catarci, T., & Simar, L. (2018). ETER Data Quality Report.

Daraio, C., Bonaccorsi, A., Geuna, A., Lepori, B., Bach, L., Bogetoft, P., Cardoso, M.F., Castro-Martinez, E., Crespi, G., & De Lucio, I.F. (2011). The European university landscape: A micro

characterization based on evidence from the Aquameth project. Research Policy, 40(1), 148–164.

European Commission. (2011). Supporting growth and jobs—An agenda for the modernisation of Europe's higher education systems Brussels: European Commission, SEC(2011) 1063. doi: 10.2766/17689

Eurostat ESS handbook for quality reports. (2014). Available at https://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf

Eurostat Quality Assurance Framework of the European Statistical System. (2019). Available at https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646

Ferlie, E., Ashburner, L., Fitzgerald, L., & Pettigrew, A. (1996). The New Public Management in Action. Oxford: Oxford University Press.

Huisman, J., Lepori, B., Seeber, M., Frølich, N., & Scordato, L. (2015). Measuring institutional diversity across higher education systems. Research Evaluation, 24(4), 369–379.

Lepori, B., & Bonaccorsi, A. (2013). The socio-political construction of a European census of higher education institutions. Minerva, 51(3), 271–293.

Lepori, B., Bonaccorsi, A., Daraio, A., Daraio, C., Gunnes, H., Hovdhaugen, E., Ploder, M., Scannapieco, M., & Wagner-Schuster, D. (2018). Implementing and Disseminating the European Tertiary Education Register – Handbook for data collection. Brussels.

Lepori, B., Bonaccorsi, A., Daraio, A., Daraio, C., Gunnes, H., Hovdhaugen, E., Ploder, M., Scannapieco, M., & Wagner-Schuster, D. (2015). Establishing a European Tertiary Education Register. Final Report Brussels: European Commission.

Lepori, B., Geuna, A., & Mira, A. (2019). Scientific output scales with resources. A comparison of US and European universities. PloS One, 14(10): e0223415.

Niederl, A., Bonaccorsi, A., Lepori, B., Brandt, T., De Filippo, D., Schmoch, U., Schubert, T., & Slipersaeter, S. (2014). Chaper 1. Mapping the European higher education landscape: New insights from the EUMIDA project. Knowledge, Diversity and Performance in European Higher Education: A Changing Landscape. doi: 10.4337/9781783472000

OECD. (2011). Quality Framework and Guidelines for OECD Statistical Activities. OECD Publishing, Paris.

UOE. (2013). UOE data collection on education systems. Volume 1. Manual. Concepts, definitions, classifications Montreal, Paris, Luxembourg: UNESCO, OECD, Eurostat.

van Raan, A.F. (2013). Universities scale like cities. PloS One, 8(3), e59384.

Williams, G. (2008). Universities and strategic knowledge creation: Specialization and performance in Europe. London Review of Education, 6(2), 191–192.

Woelfel, C.J. (1987). Financial statement analysis for colleges and universities. Journal of Education Finance, 13(1), 86–98.

Appendix 1. Annotation, record level metadata and flags

The information reported in this Appendix is taken from Lepori et al. (2018).

Record level metadata (completeness metadata and notes)

When data are not available for any variable/HEI, in order to avoid blank cells, a specific level of metadata should be inserted substituting the missing figure. A list of standardised completeness metadata is the following:

Metadata Code	Description
a	refers to the fact that the variable is not applicable to the unit of observation (for example number of PhD students for a HEI which does not have the right to award doctorates).
m	refers to the fact that the data in question is missing.
x	should be applied when a specific breakdown is not available, but the data are included in the total.
xc	should be used when the value is included in another subcategory (e.g. private funding, which are included in third party funding but cannot be singled out).
xr	should be used for data that are included in other rows, which can occur when an institution is part of another institution.
nc	should be used for data that have not been collected in the reference year (e.g. the gender breakdown of full professors was not collected for the academic year 2011/2012, but was introduced in the next data collection).
c	is used in the public database only for data with restricted access (in the full dataset the data are available, but the same flag "c" is used).
s	is used in the public database only for data below 3 to keep anonymity of individuals (in the full dataset the data are available).

In addition, to better highlight quality problems applying at the level of single HEI or groups of them, it is required to add specific remarks in dedicated columns for "notes" referring to a variable or to a group of them (e.g. notes on enrolled students at ISCED level 6).

Data Flags and Remarks

One of the results of the data quality process is a documentation of the quality evaluation of the data set through the provision of specific flags and notations accompanying the data. Flags signal problems or specificities of data both relating to format accuracy, consistency, completeness, and comparability. Flags can be attributed to i) individual cells; ii) one dimension or group of variables in a country (i.e. all variables concerning revenues); iii) all variables for one or more HEIs in a specific country (i.e. all private HEIs).

In general, flags are also accompanied by a specific remark providing more explanation on the issue highlighted by the flag.

The ETER flag system builds on a simplified and reduced version of the one adopted by EUROSTAT, but introduces a few additional codes for cases relevant at the level of individual HEIs.

A Tailor-made Data Quality Approach for Higher Educational Data
Cinzia Daraio et al.

Research Paper

Aside from the completeness of the metadata, which "substitutes" a figure that is not available or has to be hidden, in order to avoid blank cells which are ambiguous (special codes: "a", "m", "x", "xc", "xr", "nc", "c", "s"), data flags "accompany" an existing figure and act as a warning or an explanation. These are provided for each record in a column next to the corresponding variable, where multiple flags should be separated by commas (","). When relevant short explanations are included in the corresponding "Notes" columns of the data set in order to quickly identify the reasons for the flags.

Detailed explanations of the flags are included in the metadata sheets (particularly for the country-level flags).

Flags are introduced from the following sources:

– Analysis of the metadata in particular to identify important cases of departures from definitions, which should be put to the knowledge of the users.
– Information from checks and data validation processes concerning deviant ratios and inconsistencies.
– Results of the data quality analysis.

Flag Code	Description	Definition
b	break in time series	When changes in definitions or data collection procedures imply that the data are not comparable across years. The change is explained in the remark section
de	break in time series due to a demographic event	When changes in the perimeter due to demographic events (the same ETER ID, but institution changed, i.e. spin-outs and take-overs) imply that data are not comparable across years.
d	definition differs	Differences in definitions adopted for data collection imply that figures significantly differ from those complying with the ETER methodology and are not comparable across countries.
i	see metadata	There are specific conditions that imply that the value of a cell should be interpreted in a different way or not directly compared with others.
ic	inconsistent	Either when the sum of the break down differs from the total or if another semantic rule is violated.
rd	rounded	When data have been rounded by the data provider and thus are included in this format in the database.
c	confidential	When data are available, but restricted to public access (this flag is relevant only for user with unrestricted access).
ms	missing subcategory	This flag is applied to totals in order to warn users that the total does not include one relevant subcategory (for example total expenditures not including capital expenditures).
p	provisional	Data quality checks highlight some anomalies, like abnormal ratios or large fluctuations between years. Either these anomalies are not explained or a generated by data issues that could not be resolved. The corresponding data may be revised in the future.
r	remark	While the data are methodologically correct, some special event generates data anomalies, like a very large number of graduations in a single year. The remark field explains the source of anomaly.

Research Paper

Appendix 2. Consistency indicators

The information reported in this Appendix is taken from Lepori et al. (2018).

Consistency indicator
1   Total Expenditure=SUM(personnel expenditure, non-personnel expenditure, capital expenditure, unclassified expenditures)
2   Total expenditure>0
3   Total Income=SUM(core budget, third party funding, tuition fees, revenues unclassified)
4   Total Income>0
5   Staff Total (HC and FTE)=SUM(academic staff, non-academic staff)
6   Staff Total>0
7   Academic staff total=SUM(female academic staff, male academic staff, unclassified)
8   Academic staff total=SUM(national academic staff, foreign academic staff, unclassified)
9   Academic staff total=SUM(academic staff by field of education)
10   Academic staff total-full professors>0
11   Full professors=SUM(female full professors, male full professors, unclassified)
12   *If* lowest degree delivered=ISCED 8 *then* Enrolled Students, Graduates ISCED 5-7 ="a"   *If* lowest degree delivered=ISCED 7 *then* Enrolled Students, Graduates ISCED 5-6 ="a"   *If* lowest degree delivered=ISCED 6 *then* Enrolled Students, Graduates ISCED 5 ="a"
13   *If* highest degree delivered=ISCED 5 *then* Enrolled Students, Graduates ISCED 6-8 ="a"   *If* highest degree delivered=ISCED 6 *then* Enrolled Students, Graduates ISCED 7-8 ="a"   *If* highest degree delivered=ISCED 7 *then* Enrolled Students, Graduates ISCED 8 ="a"
14   Student Total=SUM(female students, male students, unclassified) (for each ISCED level)
15   Student Total=SUM(national students, foreigner students, unclassified) (for each ISCED level)
16   Student Total=SUM(resident students, mobile students, unclassified) (for each ISCED level)
17   Student Total=SUM(students by fields of education) (for each ISCED level)
18   SUM(Total students enrolled ISCED 5-7, Total students ISCED 8)>0
19   Graduates Total=SUM(female graduates, male graduates, unclassified) (for each ISCED level)
20   Graduates Total=SUM(national graduates, foreigner graduates, unclassified) (for each ISCED level)
21   Graduates Total=SUM(resident graduates, mobile graduates, unclassified) (for each ISCED level)
22   SUM(Total graduates ISCED 5-7, Graduates ISCED 8)>0
23   *If* Number of students=0 *then* number of graduates=0 (for each ISCED level)
24   *If* Non research active *then* R&D expenditure "a"
25   Total expenditure-R&D expenditure>0
26   Ancestor year $\leq$ foundation year $\leq$ legal status year