Implementing and Disseminating the European Tertiary Education Register

Implementing and Disseminating the European Tertiary Education Register

Contract EAC 2015-0280

Handbook for data collection

Benedetto Lepori, Andrea Bonaccorsi, Alessandro Daraio, Cinzia Daraio, Hebe Gunnes, Elisabeth Hovdhaugen, Michael Ploder, Monica Scannapieco, Daniel Wagner-Schuster

1.1 Disclaimer

This document has been prepared for the European Commission; however, it reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Document summary

This document provides an in-depth description of the data collection methodology of the ETER project, including basic definitions, the perimeter for data collection, classification schemes for data, definition of variables and guidelines for data collection and quality control.

This version is updated to the status of July 2017: it includes methodological improvements, particularly concerning staff, finances and metadata, as well the description of the new release of the ETER database and of the user interfaces.

Table of Contents

Table 1. List of abbreviations

Abbreviation	Full Name
DG EAC	Directorate General Education and Culture
DG RTD	Directorate General for Research and Innovation
EC	European Commission
EEA	European Economic Space
EFTA	European Free Trade Agreement
ERA	European Research Area
ETER	European Tertiary Education Register
EU	European Union
EUMIDA	European Microdata Project
EUROSTAT	European Statistical Office
FOE	Fields of Education
FTE	Full Time Equivalents
FYROM	The Former Yugoslav Republic of Macedonia
HC	Head Count
HEI	Higher Education Institutions
ISCED	International Standard Classification of Educational Degrees
NE	National Experts
NIFU	Nordic Institute for Studies in Innovation, Research and Education
NSA	National Statistical Authority
OCED	Organisation for Economic Cooperation and Development in Europe
UAS	Universities of applied sciences
UOE	UNESCO OECD EUROSTAT handbook on education statistics
USI	Università della Svizzera italiana
a.a.	Academic year

1 Introduction

The European Tertiary Education Register (ETER) is a study contracted by the European Commission, Directorate general of Education and Culture (contract EAC-2015-0280). The study's aim is to manage a regular annual European Tertiary Education Register (ETER) for two years, which takes into account ETER's results to date.

The study is a joint undertaking of five partners - USI – Università della Svizzera Italiana, Lugano, Center for Organizational Research, Joanneum Research, Graz, NIFU – Nordic Institute for Studies in Innovation, Research and Education, Oslo, University of Rome La Sapienza, Department of Computer, Control and Management Engineering Antonio Ruberti, Rome, Department of Electrical Engineering, University of Pisa – together with a network of experts in the concerned countries. It is supervised by the Directorate General Education and Culture of the European Commission, in cooperation with DG Research and Innovation and with EUROSTAT.

It follows the European MIcroData project (EUMIDA) a large-scale study supported by the European Commission from 2009 to 2011, which demonstrated the feasibility of a European-level data collection on individual HEIs, and the first phase of the ETER study (2013-2015), where the European Tertiary Education Register was successfully implemented.

The study began on August 1st, 2015 and lasted until July, 31st, 2017. Two waves of data collection have been performed for the baseline years 2013 (academic year 2013/2014) and 2014 (academic year 2014/2015). The countries covered are the EU-28 countries, EEA/EFTA countries (Iceland, Liechtenstein, Switzerland and Norway), as well as candidate countries to the European Union (Montenegro, Serbia, Turkey, Former Yugoslav Republic of Macedonia).

This handbook, which largely builds on the EUMIDA handbook, provides a detailed description of the methodology for the data collection in ETER, covering the following main items:

- The basic conceptual principles for ETER data collection (chapter 2).
- The definition of the perimeter of Higher Education Institutions to be included (chapter 3).
- The list of classifications to be used for different types of data, including students, graduates and staff (chapter 4).
- The definitions of the variables to be collected (chapter 5).
- The definition and procedures for calculating indicators on HEIs (chapter 6).
- The procedures for data validation, quality control and production of metadata (chapter 7).
- The procedures and tools for data collection in ETER (chapter 8).
- The description of the on-line tool for data publication (chapter 9).

This version of the handbook refers to the data collection for the year 2014 (academic year 2014/2015) and includes the changes and updated in the methodology, specifically for what concerns staff data, financial data and metadata. It also includes the description of the new database and on-line interface published in June 2016.

1.1 Project organization and roles in data collection

Figure 1 provides an overview of the ETER project structure and of the respective roles of different actors in the data collection and publication process.

Figure 1. Organizational structure of ETER

European Commission and NSA Task Force
- Project supervision
- Decision on main methodological choices
- Ownership of the dataset

ETER project team
Managing the data collection
Data validation and quality control
Managing the dataset and publishing data

ETER national expert
Linking with NSAs
Providing support for data collection

National Statistical Authorities
Providing data
Ownership of the raw data

1) At the highest level, the overall project supervision is by the European Commission, Directorate General Education and Culture. It is seconded by a task force composed of representatives of the National Statistical Authorities from the participating countries, which are responsible for validating the project's methodological choices (for example concerning the list of variables, data publication, etc.).

2) The ETER project team is composed by members responsible for each of the four involved institutions, namely for USI, Benedetto Lepori who is the coordinator of the project; for Joanneum Michael Ploder, for Sapienza Cinzia Daraio, for NIFU Elisabeth Hovdhaugen, for university of Pisa, Andrea Bonaccorsi. The main functions of the project team are as follows:
- Coordinating the entire process and developing its methodology, as stipulated in this handbook.
- Preparing perimeter and data collection templates and sending them to national statistical authorities.
- Performing data validation and quality analysis on collected data and reporting any problems to the NSAs.
- Integrating the data in the ETER dataset and publishing them.

3) For each country, the ETER consortium has indicated a National Expert (NE) with the following tasks:
- Maintaining contact between the ETER consortium and the NSA.
- Assisting the NSA in the delivery of data. In particular, the NE might provide information on guidelines, help the NSA for the collection of non-statistical data

(descriptors, geographical information), and assist the NSA in putting the data into the ETER format.

- Perform a preliminary check of the data and clarify any relevant issues with NSAs.

Most national experts are members of the ETER projects, but for a few countries experts directly in the country have been hired by the project (for example because of linguistic reasons).

4) National statistical authorities are the main provider of data in ETER, possibly with the exception of descriptors and geographical information.

2 Background and conceptual foundations

To remain consistent with EUMIDA, the ETER study is built around following main principles (Bonaccorsi and Daraio, 2007):

- a focus on Higher Education Institutions (HEIs) as the main building blocks of higher education and thus as the basic statistical unit in ETER.
- An understanding of HEIs as multi-input and multi-output organizations, where the separation between education and research at the level of inputs and activities is in principle problematic.
- A focus on characterizing the diversity of HEIs and constructing profiles across their main dimensions of activity rather than the attempt to rank them on a mono-dimensional scale.
- A methodological approach that builds on existing statistical frameworks (specifically the UOE data collection on education), but introduces new data when required, while adapting definitions to the specific conditions of HEIs.

2.1 HEIs as the main statistical units

ETER comes with a focus on Higher Education Institutions (HEI) as the object of inquiry and the main statistical units for which data are collected. This draws on an understanding of *Higher Education Institutions* (HEIs) as organizations that have at least some level of autonomy for the State and are able to strategically decide their focus concerning activities (for example teaching vs. research, respective of the subject specialization). Of course, we recognize that the level of autonomy and strategy of HEIs strongly depends on specific national settings and the type of HEIs considered and, that, in most cases, the State keeps a substantial level of control over a number of HEI decisions.

Substantive implications of this principle are: first, the construction of the perimeter in terms of a list of Higher Education Institutions to be included or excluded based on a set of criteria (section 3.1). This also implies that this perimeter is conceptually different from the one defined by educational statistics, where educational programs are the main statistical unit. Second, the data is aggregated at the level of the institution and no internal differentiation is made, for example concerning individual establishments (multiple campuses).

2.2 HEIs as multi-input and multi-output organizations

Second, ETER comes with an understanding of higher education institutions as multi-input multi-output organisations, which use sets of inputs – financial resources, human resources, infrastructure – to produce multiple sets of outputs, including research output, educational outputs, transfer activities towards society and economy (Bonaccorsi and Daraio, 2007). We consider that the production processes for these outputs cannot be easily separated, which is why the choice was made to characterize the whole set of inputs and outputs at the level of the whole institution instead of looking at individual activities. The ETER framework then overcomes the traditional distinction between educational and R&D statistics and does not attempt to separate these activities, but rather provides for distinct indicators concerning educational and research production.

Thus, the basic conceptual scheme of ETER considers a higher education institution as the unit of analysis, characterized by a set of inputs and outputs, as well as internal processes embedded in a broader environment, related to state regulation, funding sources, potential students; we acknowledge that this environment widely differs from country to country, as shown by comparative studies of higher education systems, and this needs to be taken into account in the analysis when comparing HEIs in different countries.

2.3 Diversity and profiling

In recent years, under the pressure from the massification of tertiary education, national higher education systems have undergone deep transformations and a fast expansion, both in terms of numbers of students and of institutions (Kyvik, 2004). In this process, completely new subsectors of higher education have emerged, like universities of applied sciences in so-called binary countries (Kyvik and Lepori, 2010), while at the same time governments have tried to promote differentiation of HEIs by introducing competition in the allocation of resources (Jongbloed, 2008). This debate comes with an increasing focus on the diversity of HEIs as a key characteristic of national systems, which responds to increasingly differentiated societal requests – both in education and research (Daraio *et al*, 2011; Huisman, 2000).

In analytical terms, these insights have led to the development of methodological approaches which allows for the systematic characterization of the different activity dimensions of HEIs, like those developed in the AQUAMETH and EUMIDA projects (Daraio et al, 2011, Bonaccorsi 2014) and, complementarily, in U-MAP and MULTIRANK projects (Van Vught *et al*, 2008).

ETER (and, previously, EUMIDA) builds on this approach for the selection of variables and the construction of indicators, even if not all dimensions can be adequately covered through data from official statistics.

More specifically, following EUMIDA, the relevant dimensions for the characterization of HEIs can be identified in: educational profile (subject domains offered, importance of different degree level), student's profile (the composition of the student's body), research activities, involvement in knowledge exchange activities (technology transfer), international orientation of HEI activities, and engagement of HEIs in regional development.

2.4 Developing a methodology tailored specifically to HEI

International statistical agencies like EUROSTAT and OECD, as well as National Statistical Authorities, already perform data collections on higher education institutions based on existing methodological manuals. This includes particularly:

- UNESCO-UIS/OECD/EUROSTAT data collection on education statistics (UOE) provides internationally comparable data on key aspects of education systems, specifically on the participation and completion of education programmes, as well as the cost and type of resources dedicated to education. Thus, UOE provides for data collection concerning students, degrees, educational staff, finances and educational expenditures; data is disaggregated by level of education (using the International Standard Classification of Education, ISCED) and by fields of education. The UOE manual is also largely the reference for the ETER data collection and whenever possible, definitions and classifications are adopted; however, there are also differences related to the fact that the basic unit in UOE are educational programs, while in ETER they are higher education institutions – this aspect will be made clear in a subsequent section of the handbook. Data collection is managed in most countries by national statistical institutes (NSI), which deliver summary tables of national aggregates to EUROSTAT.

- OECD/EUROSTAT joint data collection on research and development (R&D) statistics provides data on R&D expenditures and R&D staff, based on the definitions, classifications and procedures for collecting R&D data provided in the OECD Frascati manual (OECD, 2002) and, for EU Member States, following the requirements of Community Regulation CR/753/2004. R&D statistics are based on regular surveys of research performers, even if in some countries higher education R&D expenditures are derived from higher education statistics. The Frascati manual is relevant when looking at research expenditures of higher education institutions. The Fields of Science (FOS) classification of scientific fields might be also relevant for ETER.

The basic approach of ETER will be that, whenever the data collected through these methodologies reasonably corresponds to the ETER framework, they will be used with as few changes as possible. This most specifically concerns data regarding students, graduates, expenditures and revenues of HEIs. However, in individual cases, some slight departures have been decided to take into account specific issues which emerged from the EUMIDA study: for instance, ISCED 7 long degrees (masters without an intermediate bachelor) will be singled out; nominal variables have been introduced to characterize how capital expenditures are accounted, etc.

Further, ETER introduces a number of variables not present in the current statistical data, in two separate areas (as they lack a definition of HEIs as such):
• First, non-statistical descriptors that characterize the legal status and the history of HEIs, like the name, the official type of HEI, the foundation year.
• Second, geographical information, such as the region of establishment, city, geographical coordinates. This allows for interlinking with context data (like those from regional statistics), as well as to take into account the impact of geography (for example distance between HEIs).

Finally, in the specific case of staff data, based on the EUMIDA experience, it is considered that existing definitions in educational statistics do not specifically consider HEI staff structures and issues like the specific status of PhD students – constituting a large part of the researcher's workforce in some countries - as well as the status and role of permanent professors among the academic staff. In this area, ETER introduces a set of ad hoc definitions and will perform a specific metadata collection in order to come to a robust methodology for the analysis of HEI staff.

While being largely a data collection study, a major goal of ETER is to lead to a further development of statistical foundations concerning Higher Education Institutions, taking into account their deep differences with other levels of schools.

3 Definition of the perimeter

3.1 Introduction

A central issue both for data collection and for the handbook is the perimeter to be considered for the data collection. The foundations for this perimeter originate from the EUMIDA feasibility study and the experiences made during the project. These experiences are now used to address issues in the ETER data collection and improve this process. The aim of this section is therefore to provide definitions and guidelines for identifying the perimeter for the ETER data collection.

A sensible choice for the perimeter needs to take into account at least the following criteria:

- The coherency with the overall conceptual framework of the ETER data collection, namely to define higher education institutions (HEI) as the basic unit of data collection, considered as nationally recognized HEI-units with a sufficiently well-defined perimeter and some unity in terms of operations and strategy. This is a central criterion to ensure the conceptual coherency of the proposed set of definitions and the related data collection.
- The compatibility with existing manuals and data collection and, especially, with UOE and R&D statistics. This is relevant in order to reduce the workload for data collection, to allow for comparisons between institution-level and national-level data and to avoid communication problems.
- The compatibility and acceptance with national-level systems since quite a number of countries already have national systems of data collection at the individual level. Wherever it is possible, differences should be avoided to make the proposed data collection feasible and accepted by national statistical offices. Moreover, broader acceptance of the perimeter of higher education and research-active institutions is required at the national level.
- Issues of feasibility and workload for data collection, since the choice of the perimeter is likely to strongly influence the number of units, but also the extent to which additional data collection is required in respect to already existing data. This is relevant since contacts with national statistical offices showed that the issue of workload and resources will be critical for them and is likely to influence their willingness to put the system in place.

Unlike EUMIDA, which distinguished between a broader and a restricted perimeter, all indicators in ETER are collected for all higher education institutions (HEIs).

3.2 Perimeter for core set data collection: higher education institutions (HEIs)

Statistical units in ETER are Higher Education Institutions: this means that, in principle, all data will be collected for individual HEIs considered as a unique identity (identified by a specific ID code; see chapter 3.3) and no disaggregation will be provided for internal subunits. Further, while data could be further aggregated for analytical purposes – for example at the country and regional level - HEI-level data are the main focus of the project and it is intended that these will be published as such (and not in an aggregated level).

Importantly, there is currently no statistical definition of HEIs since educational statistics consider programs as the main statistical units. Hence, in the EUMIDA project, a specific definition has been provided and successfully tested. The definition and criteria employed in ETER basically follows the ones employed in EUMIDA.

3.2.1 Basic definition

ETER collects data on *higher education institutions*, defined as entities:

- which are recognisable as distinct organisations,

- which are nationally recognised as HEIs, and
- whose major activity is providing education at the tertiary level (ISCED 2011 level 5, 6, 7 and/or 8). R&D activities might be present, but are not a necessary condition for inclusion in the perimeter.

A HEI is nationally recognised if it is officially accredited as such by a legitimate organization in a country. Recognisable as a distinct organisation means that the perimeter of these institutions can be identified rather unambiguously, they have an internal organisational structure and, at least in principle, their own budget.

3.2.2 Delimitation criteria

To operationalize this definition, the following three main criteria should be used:
- *Main activity*. Institutions for which education at ISCED 2011 level 5, 6, 7 and/or 8 (former ISCED 97 level 5 and/or 6) is a major activity and constitutive part of their mission. These institutions can have a large share of research activity, but it should not be their only purpose. In practical terms, this means that a substantial share of staff time is devoted to education.
- *Graduation at ISCED 2011 level 5, 6, 7 and/or 8.* In principle, institutions delivering courses in curricula, where other institutions attribute the degree, are excluded.
- *National recognition as higher education institutions*, for example in national laws, is a relevant criterion of inclusion. This might include the specific mention of the institution in the law ruling higher education, accreditation by a specific body, or mention in the list of officially recognized HEIs.

Further, for unclear cases, the following two additional criteria might be used:
- *Size and visibility* is a further criterion, also for practical purposes. Size should be measured by staff data and number of students enrolled as a reference. As a general rule, institutions with less than 30 FTEs of academic staff and less than 200 students should only be included in exceptional cases, specifically for institutions mostly awarding degrees at ISCED level 8.
- *Continuity*, hence the intention of being a stable organization, is another delimitation criteria. Thus, structures that are in principle transitory, for example by offering just one training cycle, should not be included in the perimeter.

3.2.3 Examples of higher education institutions

Examples of higher education institutions to be included are universities (PhD awarding), as well as universities of applied sciences (Fachhochschulen, Polytechnics). Other examples are Colleges of Arts and Music, theological schools, schools of pedagogy, and distance education universities.
Military academies should be included as separate institutions, unless other institutions in the same country already account for their expenditures.

3.2.4 Exclusion cases

Institutions offering only services for education (but no curricula) are not included in the ETER data collection despite their inclusion in the UOE data collection (non-instructional institutions; see UOE handbook 4.3.2).
Research institutions, like public research institutes and Academies of Sciences, whose principal mandate is performing R&D, are excluded (on the basis of the main activity criterion) even if they are delivering some educational activities, except they deliver a large number of doctoral degrees.
The same applies for organizations providing other types of professional services, which offer training courses as a side activity.

Figure 2. Decision tree for identifying institutions as HEIs
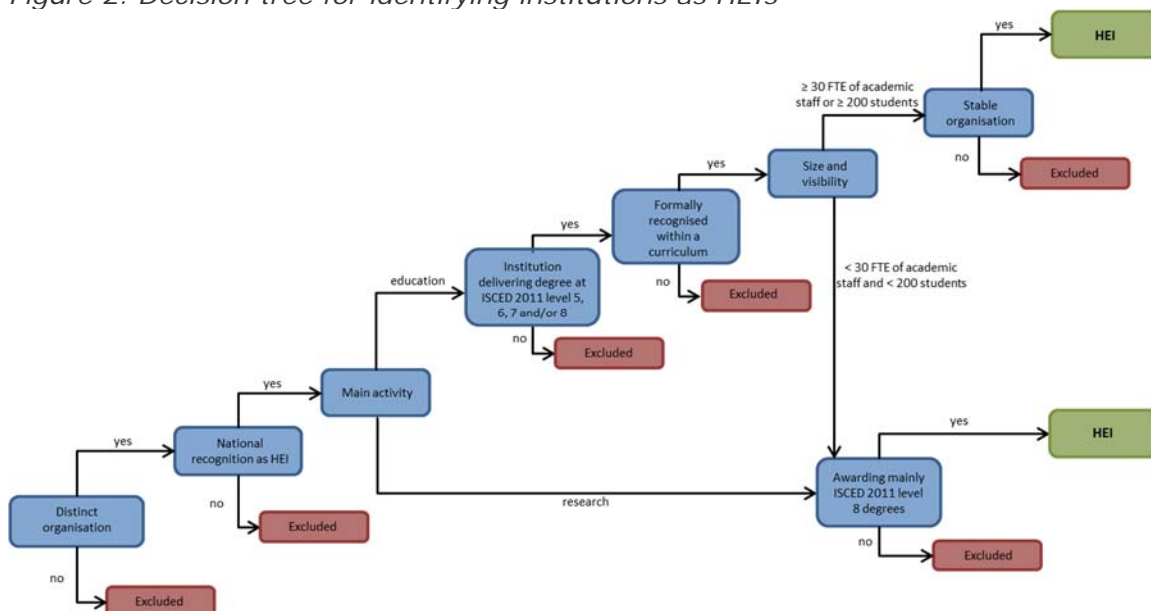
3.2.5 Coverage of HEI activities

As a principle, when an institution is included in the ETER perimeter, all its activities should be considered in the data collection. This is especially relevant for contract research, consultancies and services to the public sector and private economy (so-called third-mission activities), which are included in the data collection perimeter. The perimeter should then include all non-instructional services considered in the UOE handbook (2.9.1), as well as expenditures, funding and staff devoted to these activities. This is a fundamental reason why in the ETER perimeter, only institutions for which education (and R&D) is the main activity should be included.

Additionally, research centres should be included to the extent they are organisationally part of the HEI. That means that they are subject to the same basic regulations, and to the organisational hierarchy or juridical entity and/or they are integrated into accounting.

Specific national cases where the inclusion of associated centres might be unclear are discussed in detail in the Annex to this handbook.

3.2.6 Relationships with the definition of tertiary education

As introduced, the ETER perimeter is strongly related, but not identical to the perimeter of tertiary education in educational statistics (which includes all programs at the ISCED levels 5, 6, 7 and 8).

A first difference is related to differences between the program-oriented approach of UOE and the institution-oriented approach of ETER. While in most cases the two broadly coincide, like for example in universities, there are cases (especially in professional education) of institutions whose main missions and activities are outside the educational sector – like professional associations – but to a limited extent they also deliver tertiary education diplomas. These educational providers are not included in ETER.

A second difference is related to feasibility issues and the rationale to include small-scale educational units or parts of the public administration, when their size is quite small. This is relevant since in ETER, data have to be provided and published at the level of individual HEIs, while for educational statistics data are first aggregated by NSAs (and, in some cases, regional authorities or private providers) and then transmitted in aggregated form to EUROSTAT. Thus, in practical terms, including too small units would strongly increase the burden for data collection, validation, and

management without significantly adding to the data collection. Concerning the stability of the ETER-perimeter this would increase the longitudinal vulnerability against minor institutional changes. As a matter of fact, 20% of the smallest HEIs in the EUMIDA data set (about 480 HEIs) include just 0.6% of the undergraduate students.

If HEIs exist and fulfil the perimeter requirements, but data are missing or cannot be delivered because of confidentiality issues, they should nonetheless be included in the perimeter. Information about the missing data will be included in the metadata.

In terms of the number of students, the coverage of ETER is expected to be slightly smaller than the whole tertiary education for undergraduate students (ISCED 5-7), whereas it should be almost identical for doctoral students given their stronger concentration in large universities. Concerning staff, the perimeter of ETER might be larger than the tertiary education perimeter, owing to the inclusion in university staff data of research centres.

As a general principle, coverage in ETER of at least 90% of tertiary education at the country level and of 95% for the whole data set should be envisaged. Experience from data collection shows that, with the chosen definition, coverage of HEIs graduating at least at level ISCED 6 (bachelor) is fairly complete, whereas this is less the case for HEIs graduating only at level ISCED 5 (professional diplomas of less than three years). This is largely explained by the different structure of professional education and the fact that it is usually provided by a very large number of small institutions.

3.2.7 Perimeter description

Since ETER does not necessarily cover the whole of tertiary education, an accurate description of the coverage is important in order to match data with EUROSTAT national aggregates, but also in order to ascertain whether individual cases of HEIs are excluded.

This description provides a list of the main institutional sectors of tertiary education in the concerned country and for each sector list the following information

- A general description of the perimeter.
- Sector description.
- Curricula delivered by the sector (with reference to the ISCED mapping) and their coverage in ETER.
- Important excluded cases of HEIs in ETER.
- A reference, for example a link to the official list of HEIs in the country.

Information is partially provided by NSAs, partially derived from the UOE ISCED mappings and form other sources like Euryidice.

An example of mapping is provided on the following page.

Table 2. Perimeter description for Switzerland (example)

| Perimeter description | The perimeter includes all institutions which are officially recognized as part of the Higher Education System by law, including universities, universities of applied sciences and teacher-trainign institutions; are excluded a few very small-scale HEIs. The perimeter does not include professional schools outside the higher education sector, which enrolled in 2011 about 20% of total students at the tertiary level (however with a strongly decreasing share of tertiary education in the last ten years). About half of these students are enrolled in public higher professional schools (largely full time), the rest in part-time professional courses. There are about 150 higher professional schools with an average number of students below 200; no disaggregated data are available at national level. | | | | | | |

Switzerland	ISCED level	N. of Institutions (estimated)		ETER coverage by students' numbers	Sector description	Programme (ISCED mapping) included in ETER	Programme (ISCED mapping) not included in ETER	Most relevant exclusion cases in ETER (institutions)
		All	Included in ETER					
University sector	6,7,8	12	12	Nearly 100%	This sector includes the 10 Cantonal universities, two federal institutes of technology as well as a small number of HEIs recognized by the law.	7.1 University diploma / 6.3 University/UAS bachelor / 7.02 University master /8.01 University doctorate	7.03 university post-graduate	Facoltà di Teologia di Lugano, Franklin College Switzerland, Theologische Hochschule Chur, Institut Universitaire Kurt Bösch excluded because of size threshold
Universities of applied sciences	6,7,8	9	8	Nearly 100%	This sector includes the seven public and the two private universities of applied sciences currently recognised by the Confederation.	6.2 university of applied science diploma / 6.04 Fachhochschule, post-graduate		UAS Les-Roches-Gruyere excluded since it is below the size threshold
Teacher training universities	6,7,8	15	15	Nearly 100%	This sector includes the 15 teacher training universities not integrated in the UAS.	6.1 Pedagogical university diploma (bachelor included in figures for the UAS sector)		Hochschule für Logopädie Rorschach excluded since it is below the size threshold
Federal PET Diploma examination / higher vocational education, stage I	5,6	some hundreds (estimate)	0	Not covered	Exams regulated at the federal level, no mandatory curricula, but preparatory courses offered by a large number of mostly private providers.		6.05 Federal PET Diploma examination / higher vocational education, stage I	No reliable data on students as there are not mandatory and structured curricula.
higher vocational education, stage I (no regulation on the federal level)	5,6	some hundreds (estimate)	0	Not covered	Exams regulated at the federal level, no mandatory curricula, but preparatory courses offered by a large number of mostly private providers.		5.15 Higher vocational education, stage I (no regulation on the federal level)	No reliable data on students as there are not mandatory and structured curricula.
PET College /technical school	6	150	0	Not covered	Cantonal schools (Höhere Fachschulen) delivering professional tertiary education degrees (duration 2-3 years). Older ISCED-1997 5B, classified at ISCED-2011 level 6.		6.06 PET College /technical school / 6.07 Postgraduate course PET college	This sector is not included in ETER since it is too fragmented and only aggregated data at regional level are available.
Advanced Federal PET diploma examination / higher vocational education, stage II	7	some hundreds (estimate)	0	Not covered	Exams regulated at the federal level, no mandatory curricula, but preparatory courses offered by a large number of mostly private providers.		7.04 Advanced Federal PET diploma examination / 5.16 higher vocational education, stage II	No reliable data on students as there are not mandatory and structured curricula.
Source								
Official list of recognised institutions by the swiss law from the Swiss rectors conference (last consulted 24.06.2016)								
https://www.swissuniversities.ch/en/higher-education-area/recognised-swiss-higher-education-institutions/								

3.3 Identifiers and demographic events affecting the perimeter

This data is meant to identify unambiguously the institution. They include:
- a unique identifier (ETER ID),
- a national ID used in official registers,
- the official institutional name both in the national language and translated in English (only the official translation).

3.3.1 Institutional code/register identifier (ID)

The ETER study team establishes a register of identifiers (ID) for each HEI in the perimeter. These institutional codes are administrated centrally.
Identifier: country code (ISO 3166) + integer four digits code (ex. AT0001).
Identifiers are pre-assigned in the ETER data set and are unique through the whole data set and multi-annual data collection, i.e. identifiers used in the past for HEIs that do not exist anymore are not reused.
Type of variable: ISO code (two-digit country code and four digits numeric).
Data source: assigned by ETER study team.
In principle, ETER provides a yearly data collection. Depending on the variable, there are different possible references: the perimeter definition and thus the identifier should refer to the last day of the calendar year (31.12.XXXX).
Even if most higher education institutions are rather stable, and thus tracking institutional history does not have the same relevance as it does for the business enterprise register, this procedure is likely to be very relevant in a long-term perspective and for multi-year analysis. Hence, the ETER handbook provides for a specific notation for institutional changes, adapting the categories used in the business enterprise register to the specificities of the higher education sector. According to this, IDs of HEIs that have been included in the EUMIDA perimeter will be adapted for the ETER perimeter as possible.

3.3.2 Demographic events

In ETER, for simplicity and to avoid double counting, demographic events of institutions are handled annually. This means that the reference year N in the perimeter includes only the institutions existing on the last day of the year, while demographic information links the perimeter between consecutive years (N-1 to N).
Precise rules to decide when to record a demographic event to be inserted
To this aim, a specific register of demographic events is included in the database and each registration will take the following form:
- country code,
- ETER ID,
- status in reference year (included in perimeter of reference year or not),
- demographic event in reference year,
- affected institutions of demographic events (in case of merger, split, spin-off or take-over, and
- remarks, which include a short description of demographic events.

When the exact date of the event is not available, an approximate date (for example calendar year) can be inserted.
This approach with a separate register is preferred since we assume that demographic events are rather rare for most HEIs and thus this solution is more desirable than recording the event directly on the individual HEI records. In the case of a merger or split of institutions, the new HEI is assigned a new ID. Original IDs are documented on an annual basis. Further, IDs are unique throughout the whole data set and across all years, which means that IDs are not reused if an institution doesn't exist anymore.
Institutions involved in demographic events that are not in the HEI register will be recorded only in the demographic event register in order to keep track of them. Of

course, the register will cover only the demographic events that involve at least one institution in the HEI registry.

Events refer to real-world changes, which are relevant for the data collection; implications in terms of the register of institutions in the perimeter are discussed in the data collection section. The categories of demographic events are listed in Table 3 below.

Demographic events are directly integrated in the database as specific HEI variable for the reference year (see section 5.2.6).

3.3.3 Exemplary cases of demographic changes

- *Merger*. Aalto University in Helsinki (Finland) has been founded in 2010 as a merger of three pre-existing institutions, Helsinki University of Technology, Helsinki School of Economics and University of Arts and Design Helsinki. In EUMIDA (reference year 2009), the three schools are included in the perimeter with their respective IDs. These three IDs will not be active any more in the 2011 perimeter (but they will be reserved), whereas Aalto University will receive a new ID.
- *Take-over*. The teacher-training university of Southern Switzerland was merged in 2010 with the University of Applied Sciences in the same region (SUPSI). It is therefore included in the EUMIDA 2009 perimeter with an individual ID. In the 2011 perimeter, the ID is not used any more. SUPSI keeps its 2009 ID.
- *Spin-out*. The teacher training university of the Swiss Canton Zug was part of the teacher training university of central Switzerland until 01.01.2013. It is not recorded as an individual institution in the 2009 and 2011 perimeter, but be attributed a new individual ID in the 2013 perimeter. The parent HEI keeps its ID.

Table 3. Demographic events

Demographic Event		Change of identifier (ID)
Birth or foundation. The creation of a new institution. (Defining the name, place). The change of status is not handled as a birth or a foundation.	**Death**: The complete closure of activities of an existing institution.	New independent identifier (ID) introduced in the case of births. Identifiers of disappearing HEIs are reserved and not reused.
Entry. The entry of an existing institution in the data collection perimeter, for example because of a change of status, which includes the legal status, new activities, etc.	**Exit** from the data collection perimeter.	New independent identifier (ID) introduced in the case of entries. Identifiers of exiting HEIs are reserved and not reused.
Merger between two institutions into a new institution (death of the merged institutions, new name, legal status, accreditation etc.). The (date of the) merger corresponds to the (year of) birth (foundation) or entry.	**Split** of an existing institution into two or more independent institutions (two or more new institutions, new name, accreditation, maybe legal status).	New independent identifier (ID) introduced analogously to births. Identifiers of antecedent HEIs are reserved and not reused. Metadata of the register allow the connection of interlinked IDs.
Take-over of one institution by another one (death of the institution which was taken over, the taking over institution continues to exist). The new institution retains the identity of the institution taking over, if name and location are not changed. In order to show that data are not comparable across years after a take-over, data for the institution taking-over have to be flagged with "de".	**Spin-out (Spin-off)**. Splitting of a section of an institution to become a separate institute (old institution exists unchanged, new institution has new name, legal status, accreditation, governance, etc.). The remaining unit keeps name, legal status and location of the previous institution. As required for a take-over the flag "de" has to be added for the parent institution (the HEI losing a part) in order to show that data are not comparable across years after a spin-out.	In the case of take-overs, the identifier of the dominant HEI taking over another will be held (even if this HEI is smaller than the HEI taken over). The identity of the HEI taken over is reserved and not reused. In the case of spin-outs, the ID of the original HEI will continue. Analogous to a birth, a new ID is created for the new HEI spinning out. Metadata of the register allow the connection of interlinked IDs.

3.4 Multi-site institution

Even if most higher education institutions are basically one-sited, with most of their staff and activities concentrated on a single location, there are in fact a few cases of truly multi-site institutions. The ETER-perimeter has to face two types of multisite:

- National ones (usually different establishments in the same region) – in this case they are treated as a unique HEI and no disaggregated data are collected, but ETER includes a dummy for multisite and the possibility of subordinate NUTS codes
- Foreign campuses – consistently with UOE these are treated as self-standing HEIs in the country where they are established. This should be restricted to larger units meeting the delimitation criteria sketched above. According to the information gathered in the EUMIDA project, there are relatively few cases where this distinction is relevant, but this issue might become increasingly relevant with the internationalisation and Europeanization of higher education.

In general, ETER does not collect data on national establishments except the region (NUTS level 3). Aggregated data is provided for the whole institution, except for foreign campuses when data is provided by the local establishment separately (coherently with the current UOE practice). However, ETER provides data in a way that foreign campuses are now related to their legal institution in the parent country.

Foreign campuses should be treated as self-standing HEIs in the country where they are established and included if they comply with the definitions for the perimeter of the ETER data collection. Decisions in this respect should be taken by the national statistical authority of the country where the delimitation criteria are identical to those sketched above. Conforming UOE practice, data for foreign campuses are collected by the national statistical authority of the country where the campus is located. The time and institutional history is documented according to the definitions of HEI – demography.

4 Basic categories and classifications

This chapter introduces the basic conventions and classifications that will be used for the ETER data collection and production of indicators. It covers the following items:

- Classification by educational fields (UOE fields).
- Classification by educational level (ISCED classification).
- Geographical classification (NUTS regions).
- Conventions concerning the reference period for variables.
- Conventions concerning the use of monetary units.

While the chapter has been written so that these classifications can be used directly for the purposes of data collection, for a detailed description of classification schemes, references are made to the official descriptions provided by Eurostat/OECD/UNESCO.

As a general rule, in the ETER data collection, when data are divided by subcategories, totals should also be provided in order to account for unclassified cases and to check for consistency. The unclassified category is included in all classification schemes.

4.1 Subject classification

Classifying HEI activities by field is of prime importance for different reasons: first, HEIs display very different mixes of subjects and this represents an important dimension of HEI diversity, which needs to be investigated carefully (Lepori *et al*, 2010). Second, different subject domains might display different forms of organization and mixes of activities, owing to underlying differences in disciplinary structures, and hence using data disaggregated by field is of primary importance.

In principle, all HEI activities covered in ETER could and should be disaggregated by subject. However, on the one hand availability of disaggregated data is considered as problematic for a number of variables. On the other hand, different activities are classified using different principles – for example students are classified based on the subject of the curriculum to which they are enrolled and scientific publications based on the scientific domain of the journal where they are published.

The general principles of ETER are to keep the original subject classification of the specific type of data; correspondence tables can then be constructed for analytical purposes.

Table 4. Main classification schemes by type of data

Type of data	Classification scheme	Use in ETER	Remarks
Students and graduates	ISCED 2011 Fields of Education and training	Yes, at the broad field level (10 fields).	ISCED-F classification has been adopted to avoid changes in definitions in the next years.
Staff	ISCED 2011 Fields of Education and training Fields of Science	Yes, at the broad field level (10 fields).	
Funding	No classification	No	
Expenditures	No classification	No	
R&D expenditures	Fields of Science (FOS)	No	Foreseen by the Frascati manual, but not used in ETER.

4.1.1 ISCED Fields of Education and Training classification

For the classification of students and degrees, the Fields of Education and Training 2013 classification (ISCED-F) should be used at the first level (broad fields). This classification is consistent with the previous Fields of Education classification 1997 and 2011, as well as with Eurostat Fields of Education and Training 1999 (FoE 1999), except for including a specific code for Business, Administration and Law (included in social sciences in the previous schemes) and for ICT (included in natural sciences in FOE-1997).

Remark: when data are available based on FOE-1997, which does not break down fields 04 and 06 of ISCED-F, data for field 04 (Business, administration and law) should be set to "included in other columns" (xc) and the following remark should be added: "included in field 03". Similarly, data for field 06 (ICT) should be set to "included in other columns" (xc) and the following remark to be added: "included in field 05".

In ETER, first-level classifications of ISCED-F 2013 are used. The subfields and ISCED-1997 FOE are only provided for mapping purposes.

References:

UNESCO, International Standard Classification of Education: Fields of Education and Training 2013, draft, May 2013.

http://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx

Table 5. Fields of Education and training classification

Code	Name	Subfields	ISCED 1997 FOE
00	General programmes and qualifications	001 Basic programmes and qualifications 002 Literacy and numeracy 003 Personal skills	01 Basic programmes 08 Literacy and numeracy 09 Personal development
01	Education	011 Education	14 Teacher training and education science
02	Humanities and Arts	021 Arts 022 Humanities 023 Languages	21 Arts 22 Humanities
03	Social sciences	031 Social and behavioral science 032 Journalism and information	31 Social and behavioral science 32 Journalism and information
04	Business and law	041 Business and administration 042 Law	34 Business and administration 38 Law
05	Natural Science, mathematics and statistics	051 Biological and related sciences 052 Environment 053 Physical sciences 054 Mathematics and statistics	42 Life sciences Part of 62 (natural parks and wildlife) 44 Physical sciences 46 Mathematics and statistics
06	Information and communication technologies	061 Information & Communication Technologies	48 Computing
07	Engineering, manufacturing and construction	071 Engineering and engineering trades 072 Manufacturing and processing 073 Architecture and construction	52 Engineering and engineering trades (plus most of 85 environmental protection) 54 Manufacturing and processing 58 Architecture and building
08	Agriculture, forestry, fisheries and veterinary	081 Agriculture 082 Forestry 083 Fisheries 084 Veterinary	62 Agriculture, forestry and fishery (minus natural parks and wildlife) 64 Veterinary
09	Health and welfare	091 Health 092 Welfare	72 Health 76 Social services
10	Services	101 Personal services 102 Safety services 103 Security services 104 Transport services	81 Personal services Part of 85 environmental protection (community sanitation and labor protection and security) 86 Security services 84 Transport services

4.1.2 Fields of science

The classification by fields of science and technology (FOS) has been introduced in the Frascati manual in the 1960s and slightly revised since then (Frascati manual, 3.6.2). OECD conducted the last revision of the FOS classification in 2007 (FOS-2007). Classification should be based on the scientific fields in which most of the activities of an institutional unit (for example a department or a research centre) are undertaken. The Frascati manual recommends performing this classification at the level of subunits where activities are sufficiently homogeneous.

FOS is not directly used in ETER and is based on different principles than FOE, thus the two classifications should not be directly compared. However, at the level of broad fields, the following correspondence table might be used when required (for example for matching staff data classified based on FOS).

Table 6. Correspondence table FOE - FOS

ISCED-F 2013	Fields of Science FOS - 2007
00 General programmes and qualifications	-
01 Education	5.3 Educational sciences
02 Humanities and Arts	6. Humanities
03 Social sciences	5. Social sciences without 5.2, 5.3 and 5.5
04 Business and law	5.2 Economics and Business 5.5 Law
05 Natural Science, mathematics and statistics	1. Natural sciences without 1.2
06 Information and communication technologies	1.2 Computer and information sciences
07 Engineering, manufacturing and construction	2. Engineering and technology
08 Agriculture, forestry, fisheries and veterinary	4. Agricultural sciences
09 Health and welfare	3. Medical sciences
10 Services	-

* Includes urban planning, which is in the FOS classification included in 5. Social sciences.

4.2 Levels of education (ISCED)

Data on students and graduations will be divided by the level of education of the program to which they are enrolled, using the International Standard Classification of Education (ISCED) in its 2011 version. This version is particularly well suited for ETER purposes, as it includes the distinction between "Bologna" levels of education (Bachelor, Master and Doctorate). Relevant ISCED levels for ETER are ISCED 5 (short-cycle tertiary), ISCED 6 (bachelor), ISCED 7 (Master) and ISCED 8 (Doctoral level).

The table below provides the main categories and description for the relevant levels for ETER data collection. For full details the reader should refer to the ISCED-2011 Handbook.

References
http://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx

Table 7. Levels of education

ISCED-2011 level	Definition	Criteria
ISCED 5 short-cycle tertiary education	Programmes at ISCED level 5, or short-cycle tertiary education, are often designed to provide participants with professional knowledge, skills and competencies. Typically, they are practically based, occupationally specific and prepare students to enter the labour market. However, these programmes may also provide a pathway to other tertiary education programmes. Academic tertiary education programmes below the level of a Bachelor's programme or equivalent are also classified as ISCED level 5.	Duration: 2-3 years Entry requirements: ISCED 3 or 4
ISCED 6 Bachelor's or equivalent levels	Programmes at ISCED level 6, or Bachelor's or equivalent level, are often designed to provide participants with intermediate academic and/or professional knowledge, skills and competencies, leading to a first degree or equivalent qualification. Programmes at this level are typically theoretically based but may include practical components and are informed by state of the art research and/or best professional practice. They are traditionally offered by universities and equivalent tertiary educational institutions.	Duration: 2-3 years Entry requirements: ISCED 3 or 4 Usually: first degree at tertiary level
ISCED 7 Master of equivalent level	Programmes at ISCED level 7, or Master's or equivalent level, are often designed to provide participants with advanced academic and/or professional knowledge, skills and competencies, leading to a second degree or equivalent qualification. Programmes at this level may have a substantial research component but do not yet lead to the award of a doctoral qualification. Typically, programmes at this level are theoretically based but may include practical components and are informed by state of the art research and/or best professional practice. They are traditionally offered by universities and other tertiary educational institutions.	Duration: 2-3 years Entry requirements: ISCED 6 Usually: second degree at the tertiary level Direct access to ISCED 8 level
ISCED 7X6 Master or equivalent level long degrees	Long first-degree programme at a Master's or equivalent level with a cumulative theoretical duration (at the tertiary level) of at least five years (that does not require prior tertiary education). These programs will be singled out in ETER when possible given their different characteristics and their impact on the number of diplomas.	Duration: at least 5 years Entry requirements: ISCED 3 or ISCED 4 Usually: first degree at the tertiary level Direct access to ISCED 8 level
ISCED 8 Doctoral or Equivalent level	Programmes at ISCED level 8, or doctoral or equivalent level, are designed primarily to lead to an advanced research qualification. Programmes at this ISCED level are devoted to advanced study and original research and are typically offered only by research-oriented tertiary educational institutions such as universities. Doctoral programmes exist in both academic and professional fields.	Duration: at least 3 years Entry requirements: ISCED 7 Research-based programs (not only courses).

4.3 Geographical classification

4.3.1 NUTS regions

For the geographical localization of HEIs, the NUTS classification (Nomenclature of Territorial Units for Statistics) is adopted.

The NUTS classification comprises three main levels (NUTS-0 codes being the National States):

- NUTS1 regions are major socio-economic regions, usually corresponding with a major administrative division like Länder in Germany and groups of regions in Italy.
- NUTS2 regions are basic regions for the application of regional policies like Regierungsbezirke in DE and regions in IT.
- NUTS 3 regions are smaller regions for more specific analyses, like Kreise in Germany and Province in Italy.

For a correspondence table with national structures refer to EUROSTAT website, a correspondence table with Postcodes is also available. This correspondence table should be used to crosscheck the consistency of the NUTS classification with postcode information in ETER.

Main reference for NUTS classification:

http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction

Eurostat, Regions in the European Union. Nomenclature of Territorial Units for Statistics, 2011 Edition, Brussels (including detailed country maps).

4.3.2 Postcodes

National standards for postcodes should be used for the ETER database. In most countries the postcodes are purely numeric. The following table provides information on the format of postcodes in ETER countries.

Table 8. Format of postal codes in ETER countries

Countries	Format
Iceland	3-digits numerical (XXX)
Austria, Belgium, Bulgaria, Denmark, Hungary, Latvia, Liechtenstein, Luxembourg, FYRM, Norway, Portugal, Slovenia, Switzerland	4-digits numerical (XXXX)
Croatia, Czech Republic, Estonia, Finland, France, Germany, Greece, Italy, Lithuania, Montenegro, Poland, Serbia, Slovakia, Spain, Sweden, Turkey	5-digits numerical (XXXXX)
Romania	6-digits numerical (XXXXXX)
Ireland, Malta, Netherlands, UK	Alfa-numeric postcodes. Ireland: 7 characters, mixes letters and numbers. Netherlands: 4 digits and 2 letters Malta: 3 letters and 4 digits. UK: different formats by area.

4.4 Reference period for data collection

In principle, ETER provides a yearly data collection.

As a general principle, all data should be collected every year, including those data that generally are not expected to change in almost all cases (name of the HEI, foundation year, etc.). This choice has several advantages:

- First, yearly data sets are complete and the management of the dataset can be simplified.

- Second, mistakes in the data can be corrected in the following years.
- Third, rare cases of changes (like a HEI changing its name) can be tracked.

However, for HEI-level descriptors, questionnaires will be pre-filled with data for the previous year, while national correspondents will be asked to mark any changes made in these descriptors.

Depending on the nature of the data and the practices of data collection, individual data refer to slightly different periods as detailed in Table 9. Specific national departures from these practices should be detailed in the metadata.

Table 9. Units of measure and reference periods

Variable	Reference period/date
Descriptors and geographical information	Last day of calendar year (31st of December)
Expenditures	Calendar year (1st January – 31st of December)
Revenues	Calendar year (1st January – 31st of December)
Staff	Calendar year based on person-years for FTEs. Last day of the calendar year for HC.
Students	End of first month of begin of academic year.
Degrees (including PhD degrees)	Academic year or calendar year (to be specified)

Therefore, descriptors for the reference year "2011" should refer to the end of the civil year, financial data to the civil year 2011 and student's data to the beginning of the academic year 2011/2012.

Departures from these standard reference periods should be noted in the metadata.

When data is not collected annually, like in the case of R&D expenditures for some countries, data for missing years should be coded as missing ("m"), but a remark should be inserted that they are missing because data are not collected every year and are available for other years (for example 2011 instead of 2012).

It is important to remark that these practices slightly different from those adopted by EUROSTAT, as in EUROSTAT year '2012' refers to the academic year 2011/2012 for students and graduates, whereas in ETER to the academic year 2012/2013. In both cases, year '2012' refers to the calendar year for finances, staff and graduates.

In order to avoid confusion, data should be labelled correctly by using the calendar or academic year when applicable (for example academic year 2012/2013 for students).

4.5 Special codes

For all variables in the ETER data collection, the following special codes apply, which conform to guidelines for UOE data collection:
- Code "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).
- Code "m" refers to the fact that the data in question is missing.
- Code "x" should be applied when a specific breakdown is not available, but the data are included in the total. If a specific breakdown is not available for a country at all, "x" should be also applied in cases where the total is "0".
- Code "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).
- Code "xr" should be used for data that is included in other rows, which can occur when an institution is part of another institution.

- Code "nc" should be used for data that hasn't 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)
- Code "c" is used in the public dataset for confidential data (see section 9.1).
- Code "s" is used in the public dataset cell values below 3 to keep the anonymity of individuals (see section 9.1).

In general, no blank cells are allowed in the data collection, except for "Notes"-fields. "0" should only be used for null data (e.g. no students in that category).

An unclassified category is provided for all breakdowns (students, staff, revenues, expenditures). This should be handled as follows:
- When all cases are classified in a subcategory, this column should be set to "0" (not to "a").
- When some cases cannot be classified, this column should provide the exact number of unclassified cases.
- When all data for a breakdowns are missing, this column should be "m" or "x" (not the total of all cases) depending on whether the total is available or not.

4.6 Monetary units

All monetary data should be expressed in the national currency (euro for countries in the euro-zone) in use for the reference year at the current prices of the reference year.

This also implies that if countries change currency from year to year, data should be recorded with different currencies.

ISO_4217 codes for the currency are included in a separate field in the data collection (see Table 12).

For purposes of comparative analysis, the ETER data set includes two types of conversion rates to EUR for each year:
- Average annual exchange rates from national currency to euro for non-euros countries. Source: EUROSTAT (http://ec.europa.eu/eurostat/web/exchange-rates), download date 14th June 2016.

Table 10. Average annual exchange rates 2011-2014

ISO currency code	Currency	2011	2012	2013	2014
EUR	Euro	1	1	1	1
BGN	Bulgarian lev	1.9558	1.9558	1.9558	1.9558
CHF	Swiss Franc	1.2326	1.2053	1.2311	1.2146
CZK	Czech koruna	24.59	25.149	25.98	27.536
DKK	Danish krone	7.4506	7.4437	7.4579	7.4548
HRK	Croatian kuna	7.439	7.5217	7.5786	7.6344
HUF	Hungarian forint	279.37	289.25	296.87	308.71
GBP	Pound sterling	0.86788	0.81087	0.84926	0.80612
ISK	Islandic króna	161.42	160.73	162.38	154.86
LTL	Lithuanian litas	3.4528	3.4528	3.4528	3.4528
LVL	Latvian lats	0.7063	0.6973	0.701500	na
MKD	Macedonian denar	61.48	61.5235	61.585	61.6244
NOK	Norwegian krone	7.7934	7.4751	7.8067	8.3544
PLN	Polish zloty	4.1206	4.1847	4.1975	4.1843
RON	Romanian new leu	4.2391	4.4593	4.419	4.4437
RSD	Serbian dinar	101.9502	113.1277	113.1369	113.1369
SEK	Swedish krona	9.0298	8.7041	8.6515	9.0985
TRY	Turkish lira	2.3378	2.3135	2.5335	2.9065

Purchasing power parities to euro (average of EU-28) for each country and by year. This allows for the conversion of monetary data while taking into account the average price level in each country; importantly, PPPs might be different than 1 also for EUR countries. Source: EUROSTAT Source: EUROSTAT (http://ec.europa.eu/eurostat/web/purchasing-power-parities), download date 14th June 2016.

Table 11. Purchasing power parities 2011 – 2014

	2011	2012	2013	2014
AT	1.10745	1.08438	1.08813	1.08600
BE	1.11012	1.09786	1.10502	1.10371
BG	0.926067	0.916728	0.922444	0.903934
CH	1.89896	1.80335	1.79170	1.77043
CY	0.920421	0.930649	0.931213	0.910928
CZ	17.7602	17.6973	17.4887	17.4448
DE	1.03947	1.03874	1.05322	1.04290
DK	10.0725	10.0611	10.0557	10.0572
EE	0.694821	0.694382	0.719507	0.725238
ES	0.933167	0.914517	0.912642	0.895232
FI	1.20325	1.20662	1.23699	1.24120
FR	1.11924	1.12178	1.10948	1.10044
GR	0.928133	0.881012	0.833208	0.817580
HR	5.00663	4.87163	4.89387	4.81332
HU	165.461	168.074	171.309	174.900
IE	1.10301	1.09414	1.11089	1.11435
IS	178.625	181.726	184.072	185.612
IT	1.01907	0.997928	1.01080	1.00607
LI	1.89896	1.80335	1.79170	1.77043
LT	0.607129	0.602721	0.604162	0.603272
LU	1.18573	1.19520	1.20986	1.19968
LV	0.669830	0.671114	0.679320	0.674978
ME	0.490966	0.490714	0.501535	0.496415
MK	25.5855	25.0768	25.6116	25.1765
MT	0.756277	0.768500	0.787234	0.802796
NL	1.10059	1.09608	1.09438	1.09448
NO	11.9103	11.9451	12.2607	12.5582
PL	2.42259	2.39784	2.40872	2.40593
PT	0.821281	0.780645	0.789918	0.779218
RO	2.10956	2.07534	2.20831	2.21197
RS	49.6669	50.6366	53.6405	53.7735
SE	11.7358	11.5182	11.8063	11.9887
SI	0.835573	0.812118	0.811720	0.799669
SK	0.687298	0.678660	0.674144	0.661563
TR	1.31481	1.35655	1.46306	1.56287
UK	0.927463	0.919998	0.937283	0.939808

4.7 List of ISO country codes for countries covered in ETER

For full reference on ISO country codes see:
http://www.iso.org/iso/country_codes
For Greece, use of the code GR is adopted; for United Kingdom the code UK is adopted.

Table 12. ISO country and currency codes

ISO country code	Country	Currency	ISO currency code
AT	Austria	Euro	EUR
BE	Belgium	Euro	EUR
BG	Bulgaria	lev (pl. leva)	BGN
CH	Switzerland	Swiss Franc	CHF
CY	Cyprus	Euro	EUR
CZ	Czech Republic	Czech koruna (pl. koruny)	CZK
DE	Germany	Euro	EUR
DK	Denmark	Danish krone (pl. kroner)	DKK
EE	Estonia	Euro [1]	EUR
GR	Greece	Euro	EUR
ES	Spain	Euro	EUR
FI	Finland	Euro	EUR
FR	France	Euro	EUR
HR	Croatia	kuna (inv.)	HRK
HU	Hungary	forint (inv.)	HUF
IE	Ireland	Euro	EUR
IS	Iceland	króna (pl. krónur)	ISK
IT	Italy	Euro	EUR
LI	Liechtenstein	Swiss franc	CHF
LT	Lithuania	litas (pl. litai) [3]	LTL
LU	Luxembourg	Euro	EUR
LV	Latvia	lats (pl. lati) [3]	LVL
ME	Montenegro	Euro	EUR
MK	Former Republic of Macedonia	denar (pl. denars)	MKD
MT	Malta	Euro	EUR
NL	Netherlands	Euro	EUR
NO	Norway	Norwegian Krone	NOK
PL	Poland	zloty (pl. zlotys)	PLN
PT	Portugal	Euro	EUR
RO	Romania	Romanian leu (pl. lei)	RON
RS	Serbia	Serbian dinar	RSD
SE	Sweden	Swedish krona (pl. kronor)	SEK
SI	Slovenia	Euro	EUR
SK	Slovakia	Euro [2]	EUR
TR	Turkey	Turkish lira (inv.)	TRY
UK	United Kingdom	pound sterling (pl. pounds)	GBP

(1) From the year 2011. For 2008 EUMIDA data national currency Estonian Krone (EEK) was used.
(2) From the year 2009. For 2008 EUMIDA data national currency Slovak Koruna (SKK) was used.
(3) Euro has been adopted 01.01.2014 (Latvia), respectively 01.01.2015 (Lithuania).

Information on the currency used in the data collection is included in the metadata.

5 Definition of variables

This chapter outlines the variables collected in the ETER data set. Variables are here defined as data collected on a specific issue that later can be used to calculate indicators. Data refer to a specific year and are collected every year. In cases where the variables are not expected to change across time, the value will be pre-filled in the data sheet, based on the previous years data.

For a detailed description of the classification schemes used in collecting these data, please refer to chapter 4.

5.1 Overview: variables in ETER

The following table summarizes the proposed set of variables in ETER, while the following sections discuss each dimension in a more in-depth way by looking at the choices made. Detailed definitions are provided in the following chapter of this report.

We notice that in the technical implementation of the dataset all variables are in text format in the database, since letters are used for special codes.

Table 5.1 List of variables in data set

Dimension	Variables	Format*
Identifiers	ETER ID	Text
	National identifier (optional)	Text
	Institution name (in own language)	Text
	English institution name (if available)	Text
	Acronym	Text
	Year	Integer
	Demographic event (past)	Nominal
	Affected HEIs (past)	Text
	Remarks (past)	Text
	Demographic event (future)	Nominal
	Affected HEIs (future)	Text
	Remarks (future)	Text
Basic institutional descriptors	Country Code	ISO code
	Legal status	Nominal
	Institution category, national definition (in own language)	Text
	Institution category, national definition (in English, if available)	Text
	Institution category standardized	Nominal
	Foreign campus	Binary
	Foundation year	Integer
	Legal status year	Integer
	Ancestor year	Integer
	University hospital	Nominal
	Institutional website	Text
Geographic information	Region of establishment, NUTS2 code	NUTS code
	Region of establishment, NUTS3 code	NUTS code
	Name of the city	Text
	Postcode	Text
	Multi-site institution	Binary
	Geographical coordinates	Numeric

Educational activities	Highest degree delivered	Nominal
	Lowest degree delivered	Nominal
	Number of enrolled students at ISCED levels 5, 6, 7, by fields of education, gender, citizenship and mobility	Integer
	Total number of students enrolled at ISCED 5-7	Integer Integer
	Number of graduates at ISCED levels 5, 6, 7, by fields of education, gender, citizenship and mobility	Integer
	Total number of graduates at ISCED 5-7	Binary
	Distance education institution	
Research activities	Research active institution	Binary
	Number of enrolled students at ISCED levels 8, by fields of education, gender, citizenship and mobility	Integer
	Number of graduates at ISCED levels 8 (doctorates), by fields of education, gender, citizenship and mobility	Integer Numeric
Expenditure	Currrent expenditure	Numeric
	Personnel expenditure	Numeric
	Non-personnel expenditure	Numeric
	Expenditure unclassified	Numeric
	Capital expenditure	Numeric
	Accounting system of capital expenditures	Nominal
	R&D expenditure	Numeric
Revenue	Current revenue	Numeric
	Core funding	Numeric
	Basic government allocation	Numeric
	Other core funding	Nominal
	Third party funding	Numeric
	Public third-party funding	Numeric
	Private third-party funding	Numeric
	Third-party funding from abroad	Numeric
	Third party funding unclassified	Numeric
	Tuition fees	Nominal
	Student fees funding	Nominal
	Revenues unclassified	Nominal
	Non-recurring revenues	Nominal
Staff	Number of academic staff in FTEs and headcounts	Numeric/Integer
	Number of academic staff by fields of education, gender and citizenship in headcounts	Integer
	Number of non academic staff in FTEs and headcounts	Numeric/Integer Integer
	Number of professors by gender	Binary
	Inclusion of PhD students	Numeric/Integer
	Number of total staff in FTE and HC	
Erasmus students	Number of incoming Erasmus students	Integer
	Number of outgoing Erasmus students	Integer

*In the ETER database, all variables are coded as text since numeric variables also include non-numeric special codes ("m", "a", etc.).

5.2 Identifiers

This data is meant to unambiguously identify the institution. They include:
- a unique identifier (ETER ID),
- a national identifier, and
- the official institutional name both in the national language and translated in English (only the official translation).

For a description of which institutions are to be included in ETER, see chapter 3.2.

5.2.1 Institutional code/register identifier

Identifier: country code (ISO 3166) + integer (four digits code ex. AT0001).
Identifiers are pre-assigned in the ETER data set and are unique through the whole data set and multi-annual data collection, i.e. identifiers used in the past for HEIs that do not exist anymore are not reused.
Type of variable: ISO code (two-digit country code and four digits numeric).
Data source: assigned by ETER.

5.2.2 National ID

This variable includes the ID used in national official registers, in order to simplify matching between the ETER database and national data sources. When there is no such an official register it should be "not applicable".
Type of variable: text.
Data source: NSA.

5.2.3 Legal name of the institution

The legal name of the institution as stipulated in the legal registry or founding act. Importantly, the name of the institution should be in the reference year for data collection, changes of names across years without a change of IDs is possible when they simply imply a relabeling of the same institution (keeping its location, most of activities, staff, etc.). This also includes name changes related to changes in status (Hogeschoole > University).
Type of variable: text
Data source: NSA (preferred)/National Experts (optional).

5.2.4 Official English translation

Official English translation of the name of the institution when available or commonly used translation.
Type of variable: text.
Data source: NSA (preferred)/National Experts (optional).

5.2.5 Institution Acronym

The acronym of the institution as it is frequently used for example in official communication, on the website or on institutional affiliations. This notion of acronym is broader than the official ones, it can refer also to widely used short names, for example which can be found on institutional websites.
When no acronym can be found, this variable should be "m".
Type of variable: text.
Data source: NSA (preferred)/National Experts (optional).

5.2.6 Demographic events

In ETER, demographic events are directly integrated in the data for each individual year; this approach allows users to recognize immediately demographic events which might affect comparisons across years.
To this aim, ETER includes the following variables.

Demographic event (past). This variable identifies whether the considered HEI in a specific year was affected by a demographic event between the previous year data collection and the current year.

Type of variable: nominal. The following list of codes is applied.

Table 13. Codes for demographic events

Demographic event	
0	no demographic event
1	entry
2	exit
3	birth
4	death
5	merger
6	split
7	take-over
8	spin-out (spin-off)

Affected HEIs (past). The ETER ID of the affected HEIs in the previous year. When no demographic event happened or when the event is "entry" or "birth", the code "a" should be applied. Multiple HEIs should be separated by a semicolon (;).

Demographic event (future). This variable identifies whether the considered HEI in a specific year is affected by a demographic event between the current year of data collection and the next year

Type of variable: nominal. The same list of codes as for past events is adopted.

Affected HEIs (future). The ETER ID of the affected HEIs in the following year. When no demographic event happened or the event it "exit" or "merger", the code "a" should be applied. Multiple HEIs should be separated by a semicolon (;).

Example. In Austria, the two HEIs AT001 and AT002 were merged to AT0033 between data collection 2011 and 2012. At the same time, AT0003 gave rise to a spin-out HEI AT0034. This lead to the following demographic notations in the ETER dataset.

Table 14. Demographic Events

ETER ID	Year	Demographic event (past)	Affected HEIs (past)	Remarks (past)	Demographic event (future)	Affected HEIs (future)	Remarks (future)
AT0001	2011	0	a		5	AT0033	
AT0002	2011	0	a		5	AT0033	
AT0033	2012	5	AT001; AT002		0	a	
AT0003	2011	0	a		8	AT0003; AT00034	
AT0003	2012	8	AT0003		0	a	
AT0034	2012	8	AT0003		0	a	

Remarks and specific cases
- Demographic events are repeated for the past and future years, but codes are managed differently in the two cases.
- Future year demographic events are added only retrospectively, i.e. those for reference year 2012 will be added during data collection 2013. In the meantime, they will be coded as missing ("m").

- Past demographic codes for the ETER first year (2011) refer to events since the EUMIDA data collection for 2008.

5.3 Basic institutional descriptors

These descriptors provide some general information on the HEI considered. Proposed descriptors are the following:
- country of establishment,
- legal status,
- type of institution (national),
- type of institution (standardized),
- foreign campus,
- university hospital,
- institution's website,
- legal status year,
- foundation year, and
- ancestor year.

5.3.1 Country

The country where the institution is established, using official ISO 3166 country codes (http://www.iso.org/iso/english_country_names_and_code_elements). By country of establishment it is meant where the institution develops most of its activities, for example where the largest part of the staff is located, even if this is not the legal seat of the institution.

Foreign campuses (branch campuses) should be included in the country where they are located and distance institutions in the country (and region) where their headquarters are located (independent from the location of the students).

Type of variable: ISO code (two digits).

Data source: pre-assigned in ETER.

5.3.2 Legal status (public and private)

The classification between **public** and **private** is made according to whether a public agency or a private entity has <u>ultimate control</u> over the institution. *Ultimate control* is decided with reference to who has the power to determine the general policies and activities of the institution and to appoint the officers managing the school. Ultimate control will usually also extend to the decision to open or close the institution. As many institutions are under the operational control of a governing body, the constitution of that body will also have a bearing on the classification.

Private institutions should be further divided between *government dependent* – which either receives more than 50% of their core funding from government agencies or whose teaching staff are paid by a government agency – and *independent private*.

Please refer to UOE manual for details on how to classify institutions and for a discussion of special cases (UOE 4.3.3).

Type of variable: nominal (public=0, private=1, private government-dependent=2).

Data source: NSA.

5.3.3 Type of institution, national definition

Specification of the type of institution, using the national types of institutions. It should be provided in the national language and in English if a translation is available.

References might be, for example, the categories used in the national higher education act or commonly used national categories.

Type of variable: text.

Data source: NSA (preferred)/National Experts (optional).

5.3.4 Institution category standardized

This variable specifies a European-level standardized classification of Higher Education Institutions, built on top of the national categories. It is relevant in order to provide comparative analysis of higher education systems and analyse subgroups.

The following categories are used:

- UNI (university). These HEIs display a largely academic orientation (without excluding some focus on applied research), they have the right to award the doctorate and can bear the full name of "University" (including variants like technological university, etc.). In general, awarding doctorates should be the main criterion to classify HEIs in this category, even if a few doctoral-awarding HEIs might be included in the two following categories.
- UAS (university of applied sciences/college). These institutions are officially recognized as a part of higher education, though not as universities (see definition above). Commonly these institutions have a focus on professional education. In most cases they do not have the right to award a doctorate (exceptions are possible). National names are for example Fachhochschule (Austria, Germany), Hogescholen (Netherlands), colleges (Norway), Polytechnics (Finland). In most cases, such institutions constitute a second sector of higher education, but exceptions are possible where the college sector is now merged into a unitary HE system, but colleges still have a distinct mission and constitute a clearly distinct sector, even when ruled by the same law (like in the case of Sweden).
- Other. All institutions that do not fit the description of university/university of applied science will be categorized as "other". This may apply to institutions like art academies, military schools; also technological and professional schools in countries without a binary system (like the UK or France) should be classified in this way.

Remark: the other category will be analysed more in-depth by the consortium.

The distinction between universities and universities of applied sciences are usually stipulated in legal documents like national higher education acts. The categorization might be refined in a future release of this handbook.

Type of variable: nominal (0=other, 1=UNI, 2=UAS).

Data source: NSA (preferred)/National Experts (optional).

5.3.5 Foreign campus

If the institution is a foreign campus of another institution, for instance Maltese Campus of Middlesex University, please fill in the following variables.

Type of variable: binary (0=no, 1=yes).

Data source: NSA/pre-assigned in ETER.

If foreign campus=1, then please fill in a text field with the name and country of the mother institution.

Type of variable: text.

Data source: NSA/pre-assigned in ETER.

5.3.6 University hospital

This variable should be 1 if data from university hospital(s) are included and 0 if they are not. Following the Frascati manual, university hospitals are included in higher education (and then covered by ETER as part of the parent HEI) when they have a teaching component (training students in medicine) or are research units associated with HEIs that performing research as well.

Type of variable: binary (0=no, 1=yes).

Data source: NSA (preferred)/National Experts (optional).

5.3.7 Institutional website

Registration of the official website.

Type of variable: text – format www.website_name.
Data source: NSA (preferred)/National Experts (optional).

5.3.8 Year of Establishment

The foundation year is highly relevant in order to understand the context of the considered institutions (for example it might be assumed that older institutions are more prestigious or well-known). However, since many HEIs underwent significant changes during their history, the identification of the foundation year might prove difficult in a number of cases. Hence, ETER will collect data for a main variable:

- **foundation year**, defined as the year when the institution was established (corresponding to a birth in demography).

Additionally, two optional variables will be collected if needed:

- **legal status year**, i.e. the year when the institution received their legal status. Examples of legal status events include accreditation as an HEI, accreditation as university, etc.
- **ancestor year**, i.e. the foundation year of the oldest ancestor which can be identified. This variable is relevant when an HEI claims to descend from earlier institutions, but the current HEI is quite different from its ancestor. If there is no ancestor prior to the foundation year of the current HEI, this variable should be set to "not applicable".

Remark. In some cases, it might be difficult to decide whether some kind of change gave rise to birth of a new institution or rather was a gradual change. Criteria for decisions include a completely different name, a different scope in terms of the subject covered or a completely new internal organization.

Examples:

- University of Lucerne (CH) – current status and foundation year 2000, when the institution was officially founded as a comprehensive university. First ancestor: Faculty of Theology year 1600.
- University of Stavanger (NO) – current status year 2005, when the institution was accredited with university status. Foundation year: 1994 when the State University College of Stavanger was established. Ancestor year: 1968 with the establishment of the District University College of Stavanger.
- University of Aalto (FI) – current status and foundation year 2010, when the merger was undertaken – ancestor year 1849 when the oldest of the three merged schools (Helsinki University of Technology) was founded.

The exact calendar year should be used. A short description of the change of status should be included in the metadata worksheet. All three variables should refer to the legal event (for example the accreditation decision) and not to the effective 'beginning of activities' (even if, for older institutions, it might be difficult to trace back to this distinction).

If legal status year is the same as foundation year, the same date should be inserted in the data collection sheet. If they are missing because this information is not available, the code "m" should be used.

Type of variable: integer (four digits).
Data source: NSA (preferred)/National Experts (optional).

5.4 Geographical information

This dimension will register where the institution is located, and includes the following variables:

- Region of establishment (NUTS2 and NUTS3),
- Postal code of main seat,
- Name of city/town of main seat,
- Geographical coordinates, and
- Multi-site institutions.

5.4.1 Region of main seat - NUTS code

This variable requires the NUTS2 and NUTS3 region where the institution's main seat (see perimeter chapter 3.4) is established. For the NUTS2 and NUTS3 codes, see classification chapter 4.4.1.

When HEIs activities are distributed in more than one NUTS3 region, meaning no main seat can be identified, this variable should be "not applicable" and the NUTS3 codes entered for regional campuses.

Examples:

- The University of Applied Sciences of Western Switzerland (HES-SO) has its activities distributed among six NUTS3 regions without a central location can be identified.
- The University of Bologna has regional campuses in other cities, but the legal seat and most activities are concentrated in Bologna, which is then considered as the main seat.
- The University of Modena and Reggio Emilia has two locations, which are largely equivalent in terms of their size, activities, etc., and thus it has no main seat.

Distance education institutions have to be classified according to the location of their main seat, independent of where the students live (and of the existence of decentralized student support services).

Type of variable: text (4 characters for NUTS 2 region, 5 characters for NUTS 3 region).

Data source: NSA (preferred)/National Experts (optional).

5.4.2 Postal code

The postal code of the official address of the HEI. Different postal code systems throughout Europe are described in chapter 4.4.2.

This variable is meant to roughly identify the geographical location of the HEI. A remark should be added when the legal address is located far away from the place where the main activities are (for example in a different city).

Type of variable: text.

Data source: NSA (preferred)/National Experts (optional).

5.4.3 Name of city/town

The city/town where the main seat and most of the activities are located. This variable should be "not applicable" when no central city can be found, as the HEI is too dispersed. The name should be in English.

Type of variable: text.

Data source: NSA (preferred)/National Experts (optional).

5.4.4 Multi-site institutions

Multi-sited institutions are defined as institutions with local establishments in NUTS3 region(s) that are different from the main seat.

Type of variable: binary (0=not multi-sited, 1=multi-sited).

Data source: NSA (preferred)/National Experts (optional).

5.4.5 Regional campuses

If MULTI-SITED = 1, then please include the NUTS3 codes of the other campuses. If more than one regional campus, please separate the NUTS3 numbers with a comma. The location of the campuses should be entered in the remarks section.

Type of variable: text.

Data source: NSA.

A remark with the exact location of the other campuses should be inserted.

5.4.6 Geographic coordinates

Geographic coordinates, i.e. longitudes and latitudes, based on the postcode of the official address.

Type of variable: numeric

Data source: Will be calculated by ETER from postcodes. Geographic coordinates have been calculated automatically from Google Maps through the website http://www.doogal.co.uk/BatchGeocoding.php. HEIs, which couldn't be localized correctly by the tool, have been corrected manually by the core team using Google Maps.

This variable is not meant to allow for micro-localization within a single city, but rather for a broader localization at the national and European level. When the HEI activities are divided between different sites in different locations and no central place can be identified, the localization of the legal seat should be used, but the variable should be flagged and a remark should be added.

5.5 Educational activities

The following variables are collected for this dimension:
- lowest degree delivered,
- highest degree delivered,
- the number of enrolled students at ISCED levels 5, 6, and 7 (singling out ISCED 7 long degrees) by fields of education, gender, citizenship and mobility, and the total number of enrolled students at ISCED levels 5-7,
- the number of graduates at ISCED levels 5, 6 and 7 (singling out ISCED 7 long degrees) by fields of education, gender, citizenship and mobility, and the total number of graduates at ISCED levels 5-7, and
- distance education institution: yes/no.

These variables provide a rather simple but complete view of the main dimensions of educational activities, without a particular burden for data collection.

Description of the ISCED-2011 classification of educational degrees is found in chapter 4.2.1.

Remark: when a HEI does not have the right to award degrees at some ISCED level, the corresponding fields for students and degrees should be set to "not applicable". Accordingly, cases where the number of students and degrees for a specific ISCED level are set to "0", should be checked carefully. More specifically, this applies to the following cases:

- When a HEI does not have the right to award PhD degrees, all data for students and graduates at ISCED 8 level should be set to "a" (highest degree: ISCED 5, 6 or 7).
- When a HEI does not have the right to offer ISCED 5 (respectively ISCED 7 long degree) degrees, all data for students and graduates at ISCED 5 (respectively ISCED 7 long degree) level should be set to "a".
- When a HEI does not adopt the Bologna system and offers only long degrees at level 7, data for students and graduates at levels 6 and 7 should be set to "a".

When students and degrees are available only by ISCED1997 and no reliable mapping can be made to ISCED 2011 based on national qualifications, the data for ISCED 5-1997 should be entered under total students at ISCED 5-7. Fields for ISCED 5, 6 and 7, should be labelled with the code "xc" (data included in other categories

5.5.1 Lowest and Highest degree delivered

The lowest and highest degrees delivered by the institutions should be defined using the following classification:
- ISCED 5 diplomas with duration of less of three years,
- ISCED 6 bachelor (3 or 4 years),
- ISCED 7 Master or equivalent diploma in the pre-Bologna system (for example 4/5 years license), and
- ISCED 8 qualification equals doctorate. Type of variable: nominal.

Data source: NSA (preferred)/National Experts (optional).

5.5.2 Number of students

Numbers of students enrolled per institution at ISCED level 5, 6, 7 levels (with separate figures for long ISCED 7 students) should comply with the UOE manual for definitions and data collection procedures (UOE manual, 3.1). Data should reflect the number of students enrolled at the beginning of the academic year (last day of the first month of the winter semester academic year) and be based on the count of students. Exceptions to this rule should conform to UOE data collection practices.

In general, ETER data should reflect the number of students enrolled at a specific HEI and, therefore, double counting of students enrolled in joint degrees is possible. The preferred method is apportioning of students (so that the sum of proportions gives 1), but simple double counting or pro-rating (and thus fractional numbers) are allowed. However, the exact method should be explained in the metadata.

Breakdowns by gender, fields of education, citizenship and mobility will be provided separately by level of education, but not combined in order to reduce the burden for data collection.

In addition, a column is included in the dataset for the total number of students at levels 5-7, which can also serve as a control to ensure that the number of students at ISCED level 5, 6 and 7 are totalled correctly.

When long ISCED 7 students and degrees cannot be singled out, they should be included together with the other ISCED 7 students and degrees (and the code "xc" used for these cells).

Gender
Breakdown between men and women. If this breakdown is not possible, please use the appropriate special code (see chapter 4.5).
Type of variable: integer.
Data source: NSA.

Citizenship
Breakdown between national and foreign students. Students are considered to be foreign if they don't have the citizenship of the country for which data is collected (see UOE handbook, chapter 4.6.2).
Foreign students are defined as non-citizens of the country in which they study. If this breakdown is not possible, please use the appropriate special code (see chapter 4.5).
Type of variable: integer.
Data source: NSA.

Internationalisation
International/Mobile students are defined as foreign students who have physically crossed a national border and moved to another country with the objective of studying (see UOE handbook, chapter 4.6.2). In other words, the student has moved from what we in this context call the country of origin to the reporting country of study (also called country of destination). The country of origin is defined as the country of prior education i.e. country where upper secondary diploma was obtained. The status of mobile student is maintained throughout the whole education at tertiary level (i.e. students who entered at bachelor level are still considered as mobile at the PhD level).
If this breakdown is not possible, please use the appropriate special code (see chapter 4.5).
Type of variable: integer.
Data source: NSA.

Fields of Education

Students should be divided by fields of education. The categories used in this data set are listed in chapter 4.2.1. If this breakdown is not possible, please use the appropriate special code (see chapter 4.5).
Type of variable: integer.
Data source: NSA.
Total number of students ISCED 5-7
The sum of students at levels ISCED 5-7.
Type of variable: integer.
Data source: NSA.

5.5.3 Number of graduates
Numbers of graduates per institution at ISCED level 5, 6, 7 level by fields of education (with separate figures for long ISCED 7 degrees), gender, mobility and citizenship.
Breakdowns by gender, fields of education and citizenship/mobility will be provided separately by the level of education, but not combined in order to reduce the burden for data collection.

Gender
Breakdown between men and women. If this breakdown is not possible, please use appropriate special code (see chapter 4.5).
Type of variable: integer.
Data source: NSA.

Citizenship
Breakdown between national and foreign students. Students are considered to be foreign if they are not citizens of the country for which data is collected (see UOE handbook, chapter 4.6.2).
Foreign students are defined as non-citizens of the country in which they study. If this breakdown is not possible, please use appropriate special code (see chapter 4.5).
Type of variable: integer.
Data source: NSA.

Internationalisation
International/Mobile students are defined as foreign students who have crossed a national border and moved to another country with the objective of studying (see UOE handbook, chapter 4.6.2). In other words, the student has moved from what we in this context call the country of origin to the reporting country of study (also called country of destination). To the extent possible, definitions based on country of prior education (graduation at ISCED 3A level) should be adopted. Information on the exact criterion adopted should be provided in the metadata.
If this breakdown is not possible, please use the appropriate special code (see chapter 4.5).
Type of variable: integer.
Data source: NSA.

Fields of Education
Graduates should be divided by fields of education, using the ISCED-F 2013 classification. The categories used are listed in chapter 4.2.1. If this breakdown is not possible, please use the appropriate special code (see chapter 4.5).
Type of variable: integer.
Data source: NSA.

Total number of graduates
The sum of graduates at levels ISCED 5-7.
Type of variable: integer.

Data source: NSA.

5.5.4 Distance education institution

This discrete variable (yes/no) is introduced to identify those institutions whose main mission is to deliver distance education and where almost all programs are offered at distance to students that are not regularly coming to the HEI campus, except for workshops or short seminars. HEIs with at least 80 % of total students attributed to distance education should be classified as distance education institutions. Typical examples are Fernuniversität Hagen or the UK Open University.

Institutions offering some degrees at distance or part-time, blended learning curricula should not be included in this category.

Type of variable: binary (no=0, yes=1).

Data source: NSA (preferred)/National Experts (optional).

5.6 Research activities

Four types of data are collected for this dimension:
- the status of the research active institution,
- the number of enrolled students at ISCED level 8 by fields of education, gender, citizenship and mobility,
- the number of graduates at ISCED level 8 by fields of education, gender, citizenship and mobility, and
- R&D expenditures.

5.6.1 Research-active institutions

Research-active institutions are those who have institutionalised research activities.

Criteria for inclusion are then the following:
- The existence of an official research mandate.
- The existence of research units institutionally recognised (for example on the institutional website).
- The inclusion in R&D statistics (availability of R&D expenditure data), as sign of institutionalised research activity.
- Awarding doctorates or ISCED 8 degrees.
- Consideration of research in an institution's strategic objectives and plans.
- Regular funding for research projects either from public agencies or from private companies.

Institutions fulfilling at least three of these criteria should be included. It is generally expected that non-research active HEIs have no or very low numbers of ISCED 8 students and graduates.

Type of variable: binary (0=non-research active; 1=research active).

Data source: NSA (preferred)/National Experts (optional).

5.6.2 ISCED 8 students

Numbers of students enrolled per institution at ISCED 8 level should comply with the UOE manual definitions and data collection procedures (UOE manual, 3.1). Data should reflect the number of students enrolled at the beginning of the academic year (last day of the first month of the winter semester academic year) and can be provided in full-time equivalents. Exceptions to this rule should conform to UOE data collection practices. Note that in ISCED-2011 research masters are included in the level ISCED 7.

Note that ISCED 8 students are attributed to the HEI where they are enrolled, which does not necessarily correspond to the one where they are under an employment contract. If a different criterion than enrolment is adopted, this should be noted in the metadata and the data should be flagged.

Breakdowns by gender, fields of education, citizenship and mobility will not be combined in order to reduce the burden for data collection.

Remark: When a HEI does not have the right to award a doctorate, the cells on the number of ISCED 8 students and degrees should be set to "not applicable". If it has the right, but no students or graduates at this level, they should be set to "0".

Gender
Breakdown between men and women. If this breakdown is not possible, please use the appropriate special code (see chapter 4.5).
Type of variable: integer.
Data source: NSA.

Citizenship
Breakdown between national and foreign students. Students are considered to be foreign if they are not citizens of the country for which data is collected (see UOE handbook, chapter 4.6.2).
Foreign students are defined as non-citizens of the country in which they study. If this breakdown is not possible, please use the appropriate special code (see chapter 4.5).
Type of variable: integer.
Data source: NSA.

Internationalisation
International/Mobile students are defined as foreign students who have crossed a national border and moved to another country with the objective of studying (see UOE handbook, chapter 4.6.2). In other words, the student has moved from what we in this context call the country of origin to the reporting country of study (also called country of destination). If this breakdown is not possible, please use the appropriate special code (see chapter 4.5).
Type of variable: integer.
Data source: NSA.

Fields of Education
Students and awarded degrees should be divided by fields of education. The categories used in this data set are listed in chapter 4.2.1.
Type of variable: integer.
Data source: NSA.

5.6.3 ISCED 8 degrees

This variable asks for the number of degrees delivered at ISCED 8 qualification. It excludes intermediate stage ISCED 8 programs like the DEA in France and ISCED 8 post-doctorate qualifications (like the German habilitation). Definitions and data collection should comply with the UOE manual (UOE manual 17.1). Following the UOE manual, graduations should be based on the academic year.
Breakdowns by gender, fields of education, citizenship and mobility will not be combined in order to reduce the burden for data collection.
Type of variable: integer.
Remark: When a HEI does not have the right to award the doctorate, the cells on number of ISCED 8 degrees should be set to "not applicable".
Data source: NSA.

Gender
Breakdown between men and women.
Type of variable: integer.
Data source: NSA.

Citizenship

Breakdown between national and foreign students. Students are considered to be foreign if they don't have the citizenship of the country for which data is collected.

Foreign students are defined as non-citizens of the country in which they study. If this breakdown is not possible, please use appropriate special code (see chapter 4.5).

Type of variable: integer.

Data source: NSA.

Internationalisation

International/Mobile students are defined as foreign students who have crossed a national border and moved to another country with the objective to studying. In other words, the student has moved from what we in this context call the country of origin to the reporting country of study (also called country of destination).

Type of variable: integer.

Data source: NSA.

Fields of Education

Students and awarded degrees should be divided by fields of education. The categories used in this data set are listed in chapter 4.2.1.

Type of variable: integer.

Data source: NSA.

5.7 Revenue and expenditure

5.7.1 General definitions and principles

Financial data in ETER deal with two distinct aggregates:

- *Revenue*, i.e. the amount of money received by the HEI, either as a general allocation from the state/public authorities or other funding for research, education and other services (which could be coming from public and/or private sources).
- *Expenditure*, i.e. the amount of money spent by an HEI in order to manage its activities, for example for paying its employees, external services, general expenditures and maintenance costs, as well as capital costs.

It is therefore generally assumed that the HEI has its own budget and that there might be a difference between the amount of revenues and expenditures received during the financial year; some differences might be due to different schedules in revenues and expenditures (for example delayed payment for services), other differences might be due to the fact that HEIs incur profit or loss. Finally, differences might be due to the treatment of capital costs, as explained below. However, very large and systematic differences over many years will be checked as they might signal methodological problems.

This distinction is a major difference with UOE data collection, which has a single perimeter for expenditures and provides information on sources of funds (thereby implicitly assuming revenues = expenditures). This also implies that adjustment to fund balances reported in UOE data collection are not included in ETER (UOE manual, section 5.3.3.3, Finance 2).

Accounting systems. ETER acknowledges the variety of the accounting systems of HEIs in Europe, some having the traditional cash-based approach of public administration (where all revenues and expenditures are accounted in the current year and the HEI cannot build reserves and make profits and losses), others having a commercial accounting system based on the accrual principle. The major difference between the two methods concerns how capital expenditures are accounted (see below).

Currency. All financial data should be provided in the national currency at current prices All data are in currency units (not thousands). In the ETER database, they will be converted in euros (at current exchange rate) and in Purchasing Power Parities.

Reference periods. It is generally assumed that all financial variables refer to the calendar year (01.01-31.12). Other definitions of the reference period should be noted in the metadata.

5.7.2 Structure of the ETER financial information

The structure of the ETER financial information, both on the revenues and expenditures side, is based on the distinction between operating accounts and non-operating accounts. The former includes revenues and expenditures which are related to the normal operations of the university, while the latter include extraordinary costs and revenues, like large investments and non-recurrent donations as endowment capital. This distinction is meant to make data on the current revenues and expenditures more comparable over time and across HEIs and to limit comparability problems due to differences in accounting systems.

Table 15. Classification of revenues and expenditures

Expenditure			
Current expenditure			Non-current expenditure
Personnel expenditures	Non-personnel expenditure	Expenditure unclassified including depreciation (accrual basis) or	Capital expenditure (cash basis)
Revenues			
Current revenues			Non-recurring revenues
Core budget	Third-party funding	Students fees	Revenues unclassified	Contributions for investments, donation (capital)

The main aim of this classification is to improve the comparability between HEIs having different accounting systems, particularly by distinguishing current expenditures and revenues on the one side, non-current expenditures and revenues on the other side. This takes into account lasting differences in this respect between those universities having an accrual accounting system and those having a cash accounting system:

- In an accrual accounting system, capital costs are not part of the operating statement, which includes however depreciations of past investments.
- In a cash accounting system, capital costs are part of the operating statement in the year they have been disbursed; revenues might also include large investment contribution from the State.

The table below provides a comparisons of the expenditures, respectively revenues categories for the two accounting systems for two HEIs with the same level of current expenditures and which both undertook an investment of 500 units in the first year, one having an accrual accounting system and one a cash accounting system.

	Accrual basis					
	Current expenditures	Capital costs/depreciation	Total expenditures	Current revenues	Contributions for investments	Total revenues
2010	1000	100	1100	1100	0	1100
2011	1000	100	1100	1100	0	1100
2012	1000	100	1100	1100	0	1100
2013	1000	100	1100	1100	0	1100
2014	1000	100	1100	1100	0	1100
	Cash basis					
	Current expenditures	Capital costs/depreciation	Total expenditures	Current revenues	Contributions for investments	Total revenues
2010	1000	500	1500	1000	500	1500
2011	1000	0	1000	1000	0	1000
2012	1000	0	1000	1000	0	1000
2013	1000	0	1000	1000	0	1000
2014	1000	0	1000	1000	0	1000

In both cases, total expenditures and total revenues over the five-year period are the same, but the repartition by year is different. Moreover, the levels of current revenues will differ since the first HEIs is financing its investments from the current revenues, whereas the second one receives them as a specific endowment.

Therefore excluding capital costs and depreciations, this reporting model ensures comparability between current expenditures and removes the largest differences in revenues which are due to large one-time contributions for investments.

5.7.3 Expenditure

5.7.3.1 Current expenditure

This variable includes all current expenditures of the considered HEI for its activities, including salaries, payment for services, repairs and maintenance expenditures, electricity, financial costs, etc. . It excludes depreciations and capital costs even if they are included in the operational account (cash accounting principle).

This variable measures the total level of expenditure of higher education institutions in the reference period. The perimeter corresponds to expenditures inside institutions as defined in the UOE manual (UOE manual section 2.9) covering all types of goods and services, It includes both expenditures for educational core goods and services, R&D expenditures and other non-instructional goods and services, like meals and transportation, when directly made by HEI institutions.

Special case. Expenditure for hospitals. Consistently with the UOE manual, expenditure by or on teaching hospitals (sometimes referred to as academic hospitals oruniversity hospitals) is excluded from educational expenditure, particularly all costs of patient care and other general expenses of academic hospitals, even if such expenses are paid by the education authorities. However, expenditure by or on teaching hospital that it is directly and specifically related to the training of medical students, expenditure on R&D at teaching hospitals to the extent that it is included in the OECD/DSTI data collection on R&D are included.

Remark: the sum of personnel, non-personnel and unclassified expenditure should equal operating expenditure.

Variable: operating expenditure.
Type of variable: integer, national currency.
Data source: NSA.

5.7.3.2   Personnel expenditure

This variable includes salaries and social expenses such as payroll tax (employer's contribution), insurance, pensions etc. It requires having employees (UOE manual p. 76).
Type of variable: integer, national currency.
Data source: NSA.

5.7.3.3   Non-personnel expenditure

Includes other current expenditure, i.e. all current expenditure on goods and services other than compensation of staff (UOE manual p. 76/77). This includes electricity, costs for equipment which are included in the current expenditures (for example small equipment), services paid outside the HEIs, financial costs, rents, repairs and maintenance of infrastructure.
Type of variable: integer, national currency.
Data source: NSA.

5.7.3.4   Expenditure unclassified

In case there are any current expenditures, which are not assigned to personnel, non-personnel current expenditure, they should be indicated as unclassified. When all expenditures can be assigned to the provided categories, this cell should contain "0". When a breakdown is not available at all, this column should be "x" (included in current expenditure) and only total expenditures be filled in.
Type of variable: integer, national currency.

5.7.3.5   Capital expenditure

Capital expenditure is expenditure on assets that last longer than one year, they include spending on construction, renovation and major repair of buildings, and expenditures on new or replacement equipment. It is understood that most countries report small outlays for equipment below a certain cost threshold, as current rather than capital spending. This variable should refer to large investments like new facilities and buildings.
Accounting of capital expenditure is based on different principles depending on the accounting system:
- In accrual accounting, it should correspond to the depreciation as recorded in the operating statement.
- In cash accounting, it should correspond to the investment cost as recorded in the operating statement.

Type of variable: integer, national currency.
Data source: NSA.

Remark. If capital expenditure cannot be separated from current expenditures, this variable should be "x" (included in current expenditure) and the current expenditure variable flagged with "d".

5.7.3.6   Accounting system of capital expenditure

This variable identifies between three basic methods of computing capital expenditures:
- not included in university account, when the universities facilities are owned and financed by the State directly (or by third parties).
- cash accounting, meaning that capital expenditures are accounted for in the year the expenses are incurred (following provisions of UOE manual, section 2.9.2)
- Accrual account basis, meaning that capital expenditures are depreciated across years in the HEI accounts and only these are included in the operating accounts.

Type of variable: nominal (not included in university account=0, cash accounting=1, accrual accounting=2). This variable is collected at the HEI level since the method might different between HEIs.

5.7.3.7 R&D Expenditure

Data on R&D expenditure can be obtained either by surveys or in form of administrative data as described in detail in chapter 9 of the Frascati manual (OECD, 2015). The calculation of research expenditure of higher education institutions is particularly complex since in most cases research and educational activities are not clearly separated (both in real activities and in accounting).

The corresponding methods are based on a breakdown of worked hours of HEI staff, which are then converted into breakdown of cost by use of data on salaries. Data can be collected from time-survey of HEI staff or through administrative information, frequently by using R&D coefficients to convert data to R&D costs.

If data are not available at the HEI level – for example since average percentage of R&D activities at national level, this variable should be missing. Metadata on the methodology of the R&D statistics should be included in the metadata work sheet.

For non-research-active institutions, this variable should be "not applicable" (and not "m" for "missing").

Type of variable: integer, national currency.

Data source: NSA.

5.7.4 Revenue

5.7.4.1 Current Revenue

This variable measures the amount of money received by the considered HEI in the reference period, excluding non-recurring revenues which are recorded in a separate variable.

Two different types of breakdowns of revenues are possible. UOE provides a breakdown by source of funds, by distinguishing between public, private and households. A second type of breakdown is the distinction, *which is not used in UOE*, between core funding (resources for the general operations of the HEI) and third-party funds, i.e. resources devoted to specific activities (usually at the level of researchers or organizational subunit). The two definitions do not match each other and, therefore, care needs to be taken to achieve consistency.

Table 16. Revenues classification

	Government	Private	Households	Abroad
Core funding	Basic government allocation	Other revenues (donations, interests, etc.)		
Third-party funding	Public third-party funds	Private funds (contracts and payments for education)		Third-party funds from abroad
Student's fees			Fees (student only)	

Variable. Current revenues.
Type of variable: integer, national currency.
Data source: NSA.

5.7.4.2 Core funding

Core funding is defined as funding available for the operations of the whole institution, which are not earmarked to specific activities and whose internal allocation can be decided fairly freely by the institution itself. Thus, the main criterion to separate core funding and third-party funds is that the former are managed at the level of the whole HEI and there is discretion to which activities to devote them, while third-party funds are usually attributed and managed directly by organizational subunits.

In most institutions, the main component of core budget is the government-based allocation (either from a national or regional government). Other components are financial revenues, donations at the institutional level, as well as income from the rent of premises and other income like sales from commercial activities and paid services to students. When possible, ETER asks for a distinction between the two.

Variable: core funding.

Type of variable: integer, national currency.

Data source: NSA.

Core funding is divided in two distinct and non-overlapping categories, i.e. basic government allocation and other core funding. If the breakdown is not available, total should be entered in the aggregated variable and the subcategories should receive the code "x".

5.7.4.3   Basic government allocation

Core government allocation is defined as the amount of money transferred to the HEI by the government for its basic operations; in most cases, the use of this money is at the discretion of the HEI. Both contributions from national and federal government should be included.

Variable: basic government allocation.

Type of variable: integer, national currency.

Data source: NSA.

5.7.4.4   Other core funding

This variable includes all other components of the HEI core budget, like interests and endowments, donations at the institutional level, as well as income from the rent of premises and other non-identified income

When only basic government allocation is available and is equal to total core budget, other core budget should be coded as "xc" (assuming that it is included for example in unclassified revenues).

Variable: other core funding.

Type of variable: integer, national currency.

Data source: NSA.

5.7.4.5   Third party funding

Third party funding is funding earmarked for specific activities and institutional units, in most cases it is also limited in time. Most of these funds come from public sources, but they constitute a different stream than core funding. Third party funding specifically includes:

- Grants from national and international funding agencies for research activities, such as research councils. This includes for example grants from agencies like the Norwegian Research Council or DFG in Germany, European Union framework programmes, international programmes like Eureka or COST.
- Funds from charities and non-profit organizations for specific research and educational purposes (like the Wellcome Trust of the Bill Gates foundation).
- Contracts from public bodies, non-profit organizations and private companies for specific research and services.
- Fees/payments from companies for educational services and research and service grants from companies.

In practical terms, third-party funding variable will be the sum of the categories above. Detailed explanations on which items have been included should be provided in the metadata.

Variable: total third-party funding.

Type of variable: integer, national currency.

Data source: NSA.

Third party funding is divided in two distinct and non-overlapping categories depending on whether funds come from public and private sources. If the breakdown is not

available, total should be entered in the aggregated variable and the subcategories should receive the code "x".

5.7.4.6 Public third-party funding

Public third-party funding includes third-party funding provided by public entities, like government bodies, research funding organizations, etc. It includes for example funding from national research councils, public administration contracts, grants from national sources.
Variable: public third-party funding
Type of variable: integer, national currency.
Data source: NSA.

5.7.4.7 Private third-party funding

It includes funding by private entities on contract research and contract education, including private businesses and non-profit organisations, religious organisations, charitable organisations, business and labour associations, as well as households.
Variable: private third-party funding
Type of variable: integer, national currency.
Data source: NSA.

5.7.4.8 third-party funding from abroad

It includes funding from abroad, like funding from international research programs and from companies abroad. This amount should therefore correspond to funding from abroad in the EUROSTAT definition.
Variable: third-party funding from abroad.
Type of variable: integer, national currency.
Data source: NSA.

5.7.4.9 third-party funding undivided

This variable includes third-party funding which cannot be separated between the above categories, particularly revenues items where the distinction between public and private is not possible (revenue items like "other contracts for education"). When the whole third-party funding cannot be broken down, this variable should be 'xc' (included in total third-party funding).
Variable: third-party funding undivided.
Type of variable: integer, national currency.
Data source: NSA.

5.7.4.10 Student fees funding

The amount of money the institutions get from student fees paid by households and students to higher education institutions for participation in educational programmes (UOE manual, 5.2.4.1). Conforming to the manual, these include:
• tuition fees,
• other fees charged for educational services and other services, and
• fees paid for other welfare services furnished to students by the institution.
Payments are reported as net amounts, after subtracting any scholarships or other forms of financial aid (such as reductions in tuition fees or waivers of fees) provided to students by the educational institutions themselves. Other financial aid to students from governments or privates are not netted out.

Variable: student fees funding
Type of variable: integer, national currency.
Data source: NSA.

5.7.4.11 Tuition fees

This is a variable that registers if the institution charges tuition fees for all, some or none of the students. It takes three values:
• no fees,

- fees for some students (e.g.: Sweden has tuition fees for foreign students but not for national), and
- fees for all students.

Variable: tuition fees.

Type of variable: nominal (0=no fees, 1= partial fees, 2= fees for all students).

Data source: NSA (preferred)/National Experts (optional).

5.7.4.12 Revenue unclassified

All revenues, which were not assigned to the above revenues, should be indicated as unclassified. When all revenues can be assigned to the provided categories, this cell should contain a "0", when a breakdown is not available at all, this column should be "x".

Type of variable: integer, national currency.

5.7.4.13 Non-recurrent revenue

This variable includes all revenues which have an extraordinary and non-repeating character. Therefore, in an accrual accounting system this kind of revenues is not included in the operating statement. Singling out extraordinary revenues is therefore meant to avoid large changes from year to year and to limit comparability problems between accounting systems (as extraordinary revenues are treated in different ways in cash accounting and accrual accounting systems). The most relevant cases of extraordinary revenues are:

- specific contributions for investment, when for example the State provides a large amount of money for a facility, which is then capitalized and spent over the year when the construction progresses, are excluded. This should be the case only in cash accounting systems (as in accrual system such contribution will be included in the revenues).
- Large capital endowments or large non-recurrent donations (for example a charity providing a large endowment for the HEI).

Variable: extraordinary revenues.

Type of variable: integer, national currency.

Remark. If non-recurrent revenues cannot be separated from current revenues, this variable should be "x" (included in current revenues) and the current revenues variable flagged with "d". If the HEI has accrual accounting, this variable should be "a" (not applicable).

5.8 Staff

These variables provide information on the staff employed by higher education institutions.

As a basic definition, HEI staff comprises all persons employed by the HEI and whose activities are required for the HEI operations (UOE, section 3.5).

This definition is extensive as it includes all persons:

- Involved in student's instructions and R&D activities.
- Providing professional support to students, including both academic support and counselling and social support.
- Involved in the management and administration of the HEI.
- Providing services to the HEI like social services, food and maintenance.

Inclusion criteria are not dependent on the level or duration of employment; therefore, also temporary staff should be included as a general rule. Exception might be made (particularly for data on headcount) for staff employed with a very small percentage. Temporary replacement staff should be included as well.

Exclusions and special cases. As a general rule:

- Persons working for HEI subcontractors are included *only if the staff working for the HEI can be distinguished from staff devoted to other services* This might include for

example staff working for outsourced services like the cantina or for a technology transfer office, serving only the HEI. Other subcontracted work is excluded.

- Student aids (for example master students or PhD students) are included to the extent they have a contractual relationships with the HEI. This applies particularly for PhD students.
- Staff temporarily not at work (e.g. for reasons of illness or maternity) is included.
- Retired staff is not included.

5.8.1 Academic and non-academic staff

The basic distinction for HEI staff is between academic and non-academic staff.
Academic staff includes:

- staff whose primary assignment is instruction, research or public service,
- staff who hold an academic rank, like professor, assistant professor, lecturer or an equivalent title.
- staff with other titles (like dean, head of department, etc.) if their principal activity is instruction or research, and
- PhD students employed for teaching assistance or research.

Non-academic staff includes:

- HEI-level management staff, like director, administrative director, head of service, etc.
- HEI-level administrative staff, including both central-level and department-level staff.
- Staff engaged in maintenance and operations, including special services like IT.
- Undergraduate student employed for teaching assistance or research.

Remark: given that it is a source of comparability differences between countries, a metadata is provided on the inclusion of medical staff, particularly whether staff of the university hospital is included in the data.

The sum of academic staff and non-academic staff should be equal to total staff.

5.8.2 Measurement units

Two measurement units are adopted for HEI staff:

- Numbers of staff in headcounts (HC).

There are three different ways of computing headcounts.

One approach is based on *apportioning*: staff having different functions or active in different fields is distributed between functions proportionally to their engagement, so that the total HC equals the number of individuals working in the institution. This is the preferred methods. It implies that some breakdowns will be fractional even if based on HC.

A second method is based on *functions*: individuals are counted more than once when they have different functions. This method implies that the HC figures will exceed the number of individuals working in the institutions and, therefore, there will be inconsistencies between breakdowns and totals.

A third method is based on *single attribution*: individuals are attributed only once to the function or field where they are mostly engaged. With this methods, the total HC equals the number of individuals working in the institution.

The choice made has to be indicated in the metadata.

- Staff activities in Full Time Equivalents (FTE). Full Time Equivalents are defined as the number of actual working hours of HEI staff during a reference period (*usually the calendar year*) divided by the total number of hours conventionally worked in the same period by a full time employed individual.

For most purposes, FTE data are the most important measure of HEI staff, as they measure the total effort performed by staff in HEI activities. HC data are relevant when observing personnel characteristics referring to individuals, like their gender, nationality and field.

5.8.3 Staff breakdown

The following breakdown is required:

- **citizenship**: breakdown of academic staff between national and foreign staff using citizenship as a criterion. This breakdown is relevant to analyze the internationalization of HEI.
- **gender**: breakdown of academic staff and of professors between men and women. This breakdown is relevant to analyze the gender balance of HEI.
- **fields of education**: breakdown of academic staff by fields of education. This breakdown provides the opportunity to analyze differences in the composition of fields of education between institutions, and may also contribute to explain observed differences.

Data should be collected both in full time equivalents (FTE) and headcounts (HC). However, breakdowns by citizenship, gender and fields of education are required only for headcounts.

Additionally, data are collected for a specific subcategory of academic staff, i.e. full professors.

Table 17. List of variables and breakdowns

Variable	Unit	Breakdown
Total staff	FTE and HC	
Total non-academic staff	FTE and HC	
Total academic staff	FTE	
Total academic staff	HC	Field of Education, Nationality, gender
Number of full professors	HC	Gender

5.8.4 List of variables

5.8.4.1 Total number of academic staff (FTEs).

The total number of academic staff employed by the HEI during the reference period in Full-Time Equivalents.
Type of variable: integer.
Data source: NSA.

5.8.4.2 Academic staff by gender (HC)

Breakdown between men and women in HC. If this breakdown is not possible, please use appropriate special code.
Type of variable: integer.
Data source: NSA.

5.8.4.3 Academic staff by citizenship (HC)

Breakdown between national and foreign staff in HC. Employees are considered to be foreign if they don't have the citizenship of the country for which data is collected.
If this breakdown is not possible, please use appropriate special code.
Type of variable: integer.
Data source: NSA.

5.8.4.4 Academic staff by Field of education (HC)

Academic staff in HC should be divided by the fields of education they work in, by using the Field of Education and Training classification (FET_2013; see chapter 4.2.1). If this breakdown is not possible, please use appropriate special code.
Type of variable: integer.
Data source: NSA.

5.8.4.5 Number of full professors (HC)

Full professorship requires a main position as tenured professor (minimum 75%). The following categories developed in the She figures (European Commission, 2012), only staff in category A are included, i.e. the single highest grade/post at which research is normally conducted, corresponding usually to national classifications like ordinary professor, full professor etc. (not including associate professors and assistant professors). Competence level: PhD degree, tutoring PhD-students, participating in decision-making bodies. Reference can be made to detailed correspondence tables in the She figures methodological annex (year 2015; https://ec.europa.eu/research/swafs/index.cfm?pg=library&lib=gender_equality).
This category is a subcategory of academic staff.
Type of variable: integer (headcount).
Data source: NSA.

5.8.4.6 Full professors by gender (HC)

Breakdown between men and women. If this breakdown is not possible, please use appropriate special code.
Type of variable: integer

5.8.4.7 Total number of non-academic staff (FTEs).

The total number of non-academic staff employed by the HEI during the reference period in Full-Time Equivalents.
Type of variable: integer.
Data source: NSA.

5.8.4.8 Total number of staff (FTEs).

The total number of staff employed by the HEI during the reference period in Full-Time Equivalents.
Type of variable: integer.
Data source: NSA.

5.8.4.9 Total number of staff (HC).

The total number of staff employed by the HEI during the reference period in Headcounts.
Type of variable: integer.
Data source: NSA.

5.8.4.10 Inclusion of PhD-students

This variable characterizes whether PhD students are included in the reported data of academic staff or excluded, be it because they don't have a regular work contract or they are not covered in the statistics. The selection should refer to the prevailing practices.
For HEIs that do not award PhD degrees, this variable should be "0" (independent from the fact that PhD students enrolled in other institutions are employed as researchers).
Type of variable: binary (0=most PhD students are not included in staff data, exceptions possible, 1=most PhD students are included in staff data, exceptions possible).
Details on the degree of inclusion should be provided in the remark text to this variable, respectively in the metadata.
Data source: NSA.

5.9 Erasmus students data

These variables provide information on the number of incoming and outgoing students supported by the Erasmus program. They are based on data from the Education, Audiovisual and Culture Executive Agency of DG EAC (https://eacea.ec.europa.eu/homepage_en).

The data source for the years 2011-2013 (a.a. 2011/2012 to 2013/2014) are micro-data published on the open data portal of the European Commission: http://data.europa.eu/euodp/en/data/dataset/erasmus-mobility-statistics-2013-14

Data for 2014/2015 are currently under preparation.

5.9.1 Erasmus charter code

The Erasmus charter code of the ETER HEI, based on matched with the Erasmus list of codes. The codes for the program 2007-2013 are currently adopted, they might be updated to the new Erasmus+ 2014-2020 codes at later stage.
For the ETER HEIs that could not be identified in the Erasmus program, the code "a" is used since probably they did not participate.
Type of variable: text.
Data source: Erasmus programme.

5.9.2 Erasmus incoming students

The number of Erasmus students hosted by the HEI during the respective academic year. This includes students' mobility at ISCED levels 5 to 8. It is planned to disaggregate these data by ISCED level in a future release of ETER.
When the HEI has no Erasmus code, the code "a" is used.
Type of variable: Integer.
Data source: Erasmus programme.

5.9.3 Erasmus outgoing students

The number of Erasmus students sent out by the HEI during the respective academic year. This includes students' mobility at ISCED levels 5 to 8. It is planned to disaggregate these data by ISCED level in a future release of ETER.
When the HEI has no Erasmus code, the code "a" is used.
Type of variable: Integer.
Data source: Erasmus programme.

6 Indicators in ETER

In addition to the variables described in chapter 5, the ETER database includes a number of indicators which characterize HEIs across a number of important dimensions, like gender, mobility and internationalization, as well as research orientation.

These indicators can be seen as an added value in the project. The indicators highlight traits and characteristics in HEIs that are in many cases not available through other sources. Hence, it is likely that indicators will draw a lot of attention from final users, and will possibly inspire research and further use of the ETER database.

The choice of directly integrating indicators in the database is motivated by robustness considerations: the indicators include those that the ETER consortium considers sufficiently robust and which are available for a sufficient number of countries.

This chapter presents the rationale for the selection of indicators, the current list of indicators and rules for their calculation.

6.1 General approach

There are several reasons for selecting the specific set of indicators used in ETER. Firstly, the indicators presented contribute valuable information on HEIs, as they can be used to characterize relevant dimensions of HEIs. In addition to this is the set of indicators also selected based on the availability of data in the database and the robustness of the indicator itself.

All indicators are available on the institutional level, and it is possible to compare different types of institutions within a country or between countries. It is further possible to select a certain type of institutions, either by size or by type (for example private/public or university/university of applied science) and compare this subgroup. This is what makes the ETER indicators unique, compared to official statistics from Eurostat, which are conducted on a national level.

The indicators discussed in this document are all derived from the variables present in the ETER database or the variables that will be included in the second wave of data collection in ETER.

Some of the variables stand alone as indicators, without having to be calculated. For instance the total number of students or the total number of graduates etc. at an institution give an indication of the size of this HEI compared to others in this country. It also makes it possible to compare institutions of approximately the same size throughout Europe.

The indicators used are calculated using variables available in the data set. These are indictors such as gender balance (share of women), share of foreigners, share of internationally mobile individuals and subject mix at institution/among students/graduates/staff and degree orientation of the institution. In addition, indicators such as full professor's share of total academic staff and share of researchers among staff, as well as core budget share of total budget will be included among indicators.

Currently, ETER provides data on HEIs in 31 countries, but not all countries delivered all variables and breakdowns. When a variable or breakdown is missing for a country, this generally applies to all HEIs in this country. It is important that the indicator is available for a meaningful number of HEIs and countries in order to provide meaningful results.

Figure 3. Availability of variables in ETER by number of countries (year 2012)
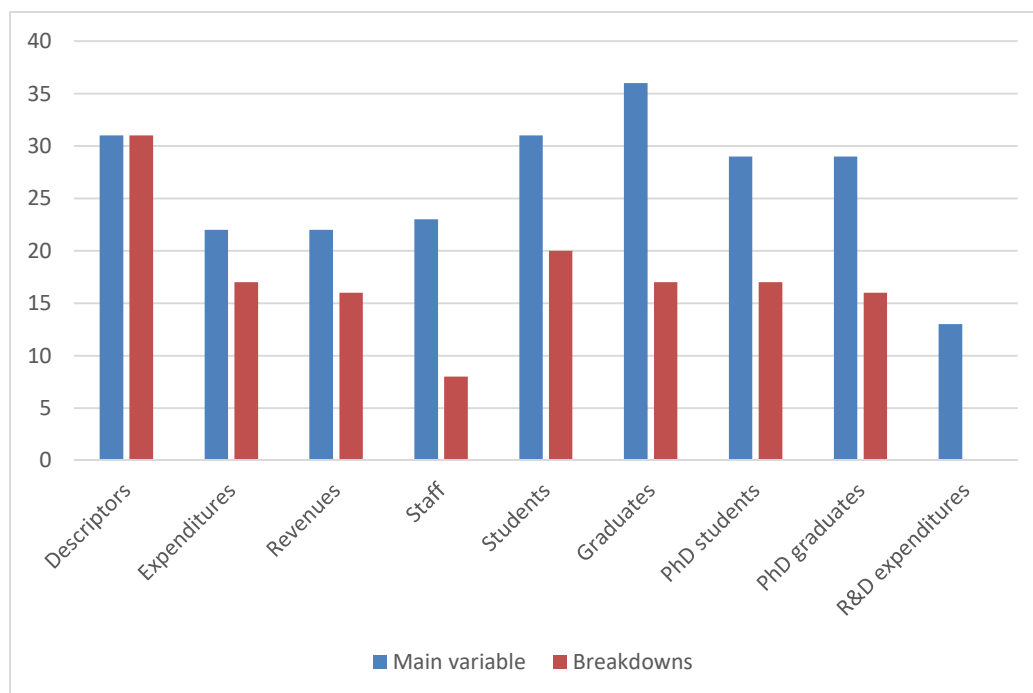
Figure 3 indicates that ETER has good coverage of variables on students (ISCED level 6-8), graduates (ISCED level 6-8) and staff (with the exception of the breakdown of staff by field). Coverage of ISCED level 5 is quite limited and, therefore, no indicators are provided at this level.

Approximately two-thirds of the countries have delivered data on revenues, expenditure and half of them for R&D expenditure. In the initial phase of ETER we suggest only using data from the areas where the coverage is good for the calculation of indicators. This implies that we currently have a few indicators related to research, and more indicators related to gender balance and teaching.

Economic indicators, like costs per student, are problematic methodologically for a few different reasons: first, costs per student vary strongly by subject domain and HEI mission and, therefore, comparing aggregated HEI-level values might lead to misleading interpretations in terms of efficiency. Second, even when financial data are compared across countries using Purchasing Power Parities, these do not fully account for the different cost structures in higher education.

Figure 4. Availability of breakdowns as a percentage of the HEIs
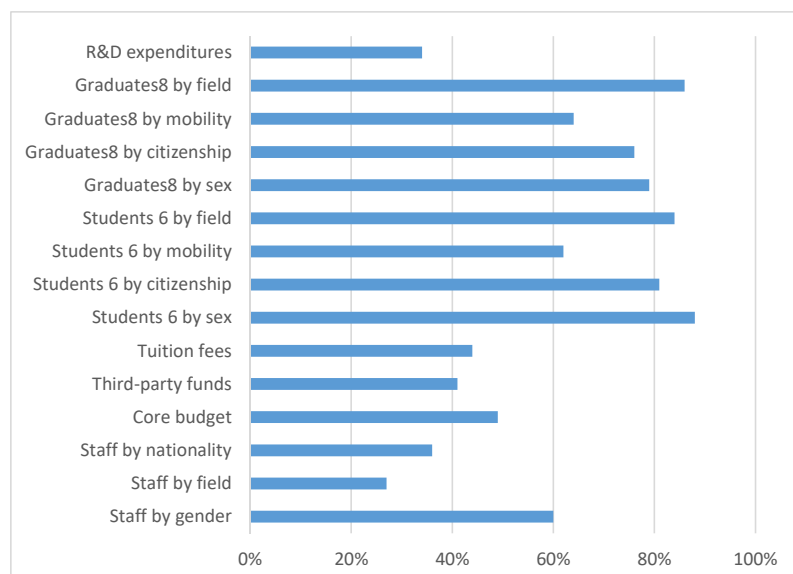
Figure 5, when counting individual HEIs rather than countries, the situation does not change substantially: breakdowns of students by sex, citizenships and field of study are available in most cases, slightly less so for mobility. The situation is more problematic when considering breakdowns of staff data (even when aggregates are available in most cases) and for revenues and R&D expenditures.

The direct integration of indicators in the ETER database presents advantages in terms of usability, but also of robustness, since the ETER consortium holds the needed competence to deal with problematic cases. Drawing on this competence, the consortium has, and will continue to discuss, which indicators should and should not be included in the ETER database.

6.2 List of indicators

Table 18 provides the list of indicators currently included in ETER. Most of these indicators are related to students, graduates, PhD-students, PhD-graduates and staff in HEIs, since these are the types of variables where most countries have delivered data. The majority of these indicators are also non-controversial, as they only provide further descriptions of the institutions/countries, to facilitate comparisons.

Table 18. List of indicators in the ETER database by dimension

Dimension	Variable	Note	N.
Gender in HEIs	Students (ISCED 6 and 7)		2
(share of women)	Graduates (ISCED 6 and 7)		2
	PhD-students		1
	PhD-graduates		1
	Academic staff		1
	Full professors	Included in 2012	1
Nationality in HEIs	Students (ISCED 6 and 7)		2
(share of	Graduates (ISCED 6 and 7)		2
foreigners)	PhD-students		1
	PhD-graduates		1
	Academic staff		1
Mobility in HEIs	Students (ISCED 6 and 7)		2
Share of	Graduates (ISCED 6 and 7)		2
mobile individuals)	PhD-students		1
	PhD-graduates		1
Subject mix	Students (ISCED 5-7)		1
(Herfindahl index)	PhD-graduates		1
	Staff		1
Degree orientation	PhD-intensity		1
Staff	Full professors, as share of academic staff (HC)		1
	Academic staff (FTE), as share of total staff		1
Revenues	Total core budget as share of total current revenues		1
	Total third party funding as share of total current revenues		1
	Student fees funding as share of total current revenues		1

The numbers for students and graduates are somewhat incomplete for ISCED 5, as only 436 out of 2249 HEIs in the current ETER database graduate at level ISCED 5. Thus, we suggest only creating the indicators for students and graduates on ISCED levels 6 to 8. All indicators for the ISCED level 7 include both regular master degrees and integrated (or long) master degrees.

6.3 Calculation of indicators

6.3.1 Overall rules and special cases

While specific formulas for each indicator are provided below, a number of general rules apply.
First, values that are unclassified will not be included in the calculation of indicators. This applies for example if the gender is unknown, share of women will be calculated as the share of women among men and women, where any unclassified person (missing) is

excluded. A similar case is when the nationality is unknown; in these cases the share of foreigners will only be calculated as the share of foreigners among foreigners and nationals. However, indicators computed from breakdowns where the share of unclassified cases exceeds 25% will be flagged with "ic".

Cases where the indicator is clearly meaningless because of underlying data problems – for example negative shares of budget or share of foreigners greater than one – will be deleted and replaced with "m".

Second, for countries that have reported restricted data, the indicator using that kind of data will be set to restricted as well (special code "c") and coded as "c" in the public databases. Additionally, indicators that are calculated using small values coded to "s" are also coded to "s" in order to prevent recalculation.

Flags related to data will also be applied to the indicators using that data as follows:

- When data used for calculating an indicator is missing (codes "m", "x", "xr" or "xc"), the indicator will be set to missing.
- When data used for calculating an indicator is not applicable ("a"), the indicator will be set to not applicable as well ("a").
- When a total is "0", the value for the indicator referring to a breakdown will be set to "a" (not applicable). For example, this applies when a HEI has "0" students at a specific ISCED level.

All indicators are calculated for individual years and HEIs.

6.3.2 Gender

Indicators concerning gender are computed as follows:

$$share\ of\ women = \frac{number\ of\ women}{number\ of\ women + number\ of\ men}$$

"Unclassified" is not included in the denominator.

These indicators are computed for the following variables in the ETER database:
- WomenStudentsISCED 6. Students at level ISCED 6.
- WomenStudentsISCED 7. Students at level ISCED 7 (including ISCED 7 long degrees).
- WomenStudentsISCED 8. Students at level ISCED 8. For HEIs not delivering the PhD, this indicator is set to "a" (not applicable).
- WomenGradISCED 6. Graduates at level ISCED 6.
- WomenGradISCED 7. Graduates at level ISCED 7 (including ISCED 7 long degrees).
- WomenGradISCED 8. Graduates at level ISCED 8. For HEIs not delivering the PhD, this indicator is set to "a" (not applicable).
- WomenAcad. Academic staff in Headcounts.
- WomenProf. Professors in Headcounts.

These indicators are highly relevant to analyse gender balance and, specifically, how it differs by activity and level, respectively by type of HEIs.

6.3.3 Share of foreigners

Indicators concerning citizenship are computed as follows:

$$share\ of\ women = \frac{number\ of\ foreign\ citizens}{number\ of\ foreign\ citizens + number\ of\ nationals}$$

"Unclassified" is not included in the denominator.

These indicators are computed for the following variables in the ETER database:

- ForeignStudentsISCED6. Students at level ISCED 6.
- ForeignStudentsISCED7. Students at level ISCED 7 (including ISCED 7 long degrees).
- ForeignStudentsISCED8. Students at level ISCED 8. For HEIs not delivering the PhD, this indicator is set to "a" (not applicable).
- ForeignGradISCED6. Graduates at level ISCED 6.
- ForeignGradISCED7. Graduates at level ISCED 7 (including ISCED 7 long degrees).
- ForeignGradISCED8. Graduates at level ISCED 8. For HEIs not delivering the PhD, this indicator is set to "a" (not applicable).
- ForeignAcad. Academic staff in Headcounts.

The availability of indicators on foreign academic staff is rather limited.
These indicators are relevant to analyse the internationalization process of HEIs and its determinants depending on level and activity.

6.3.4 Share of mobile students and graduates

Indicators concerning mobility are based on the EUROSTAT definition of place of prior education (providing access to tertiary education). They provide a more precise measure of student's mobility, which allows excluding foreigners born in the country.

$$share\ of\ women = \frac{number\ of\ mobile\ students}{number\ of\ mobile\ students + number\ of\ resident\ students}$$

"Unclassified" is not included in the denominator.

These indicators are computed for the following variables in the ETER database:
- MobStudentsISCED6. Students at level ISCED 6.
- MobStudentsISCED7. Students at level ISCED 7 (including ISCED 7 long degrees).
- MobStudentsISCED8. Students at level ISCED 8. For HEIs not delivering the PhD, this indicator is set to "a" (not applicable).
- MobGradISCED6. Graduates at level ISCED 6.
- MobGradISCED7. Graduates at level ISCED 7 (including ISCED 7 long degrees).
- MobGradISCED8. Graduates at level ISCED 8. For HEIs not delivering the PhD, this indicator is set to "a" (not applicable).

These indicators are relevant to analyse the internationalization process of HEIs and its determinants depending on level and activity.

6.3.5 Subject concentration in education and graduate education

These indicators characterize the level of specialization of HEIs education by field. The indicator is calculated as a Herfindahl index of the distribution of students and graduates by field of education.

$$Herfindahl = \frac{1}{n^2} * \sum_{1}^{11} n_j^2$$

Where n_j^2 is the number of students or graduates in field j and n is the total number of students or graduates for that level within the HEI. The sum runs over the 11 fields of education. The index runs from 1, when all students are in the same field, to 0.1 when the students are equally distributed across fields in an HEI.
- SubConcEducation. The indicator is calculated using the distribution of the total number of students at levels ISCED 5-7 by field of education (ISCED-F).

- SubConcPhD. The indicator is computed using the distribution of PhD graduates (ISCED 8) by field of education (ISCED-F). Graduates are preferred to students, as data are considered more reliable. Since PhD education is closely associated with research, this can also be considered as an indicator of subject concentration in research.

When data are computed using ISCED-1997 FOET (8 fields), a flag is added ("d") and a remark is included.

6.3.6 Research intensity

This is an indicator frequently used to characterise the level of orientation to research of an HEI, in respect to the volume of educational activities at the undergraduate level.
It is computed as follows:

$$research\ intensity = \frac{number\ of\ graduates\ at\ level\ ISCED\ 8}{number\ of\ graduates\ at\ ISCED\ levels\ 5,6,7\ and\ 7\ long}$$

When the HEI does not deliver the doctorate or when an HEI awards only PhD degrees, this indicator is set to "a".

6.3.7 Indicators related to staff composition

These indicators provide information on staff structure of HEIs, which is relevant to analyse their internal organization.

Two indicators are included in ETER.

$$ShareAcadStaff = \frac{number\ of\ academic\ staff\ (FTE)}{number\ of\ academic\ staff\ (FTE) + number\ of\ administrative\ staff\ (FTE)}$$

Is a relevant indicator to analyse the weight of administration for HEIs.

$$ShareProf = \frac{number\ of\ professors\ (HC)}{number\ of\ academic\ staff\ (HC)}$$

It is relevant to analyse the academic staff structure, also depending on the country and type of HEI.

6.3.8 Indicators related to revenues

Three complementary indicators are provided for the share of revenue sources of HEIs. Cases where the share of unclassified revenues exceeds 25% will be flagged.

$$ShareCore = \frac{Total\ core\ budget}{Total\ current\ revenues}$$

$$ShareThirdParty = \frac{Total\ third\ party\ funding}{Total\ current\ revenues}$$

$$ShareFees = \frac{Student\ fees\ funding}{Total\ current\ revenues}$$

7 Metadata, data validation, quality control

7.1 General introduction and relevance of data quality

Data quality is a relevant issue in any data collection, but often underestimated in practical data collection procedures with the exception of the Official Statistics field, in the context of which data undergo a well-defined correction and control procedure.

Data quality is well recognized as a multidimensional concept including several distinct dimensions (Batini and Scannapieco, 2006) and a number of different definitions have been proposed in various contexts (Scannapieco and Catarci, 2002). It can be intended as "the distance between the data views presented by an information system and the same data in the real world"; such a definition can be seen as an "operational definition", although evaluating data quality on the basis of comparison with the real world is a very difficult task.

A crucial dimension of data quality is data accuracy: it measures the closeness between a value v and a value v', considered as the correct representation of the real-life phenomenon that v is intended to represent. However, quality is more than simply data accuracy. Other significant dimensions such as completeness, consistency, and actuality (i.e. degree of up-to-datedness), just to cite some significant ones, play a role in the definition of the Data Quality concept.

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).

Internal quality control is a major issue for HEI-level international data collection, as, on the one hand, the lower level of disaggregation of data makes them more sensible and increases the chances of mistakes and, on the other hand, there are many different sources of comparability problems between HEIs and countries.

In addition, ETER data collection is performed by integrating data already collected by statistical institutions by means of different surveys or administrative data. This means that the "design survey" phase of a classical statistical production process cannot be applied in this case to the "overall" process, and, consequently, the proposed quality indicators cannot be reused in a direct way in ETER quality validation and reporting.

For these reasons, quality validation and reporting on data collected by ETER is performed as follows:

☐ A dedicated quality validation activity is performed on collected data, independent of the survey-specific quality characterization, but instead based on a newly proposed quality characterization described in Section "Quality Indicators".

☐ Where data available at National Statistical Offices can be accompanied by statistical process quality reporting (like prescribed by (EUROSTAT, 2007), such quality reporting is part of a detailed quality characterization of the "source data" to the ETER process.

The overall process is depicted in Figure 5.

Figure 5. Overall ETER data collection and quality management
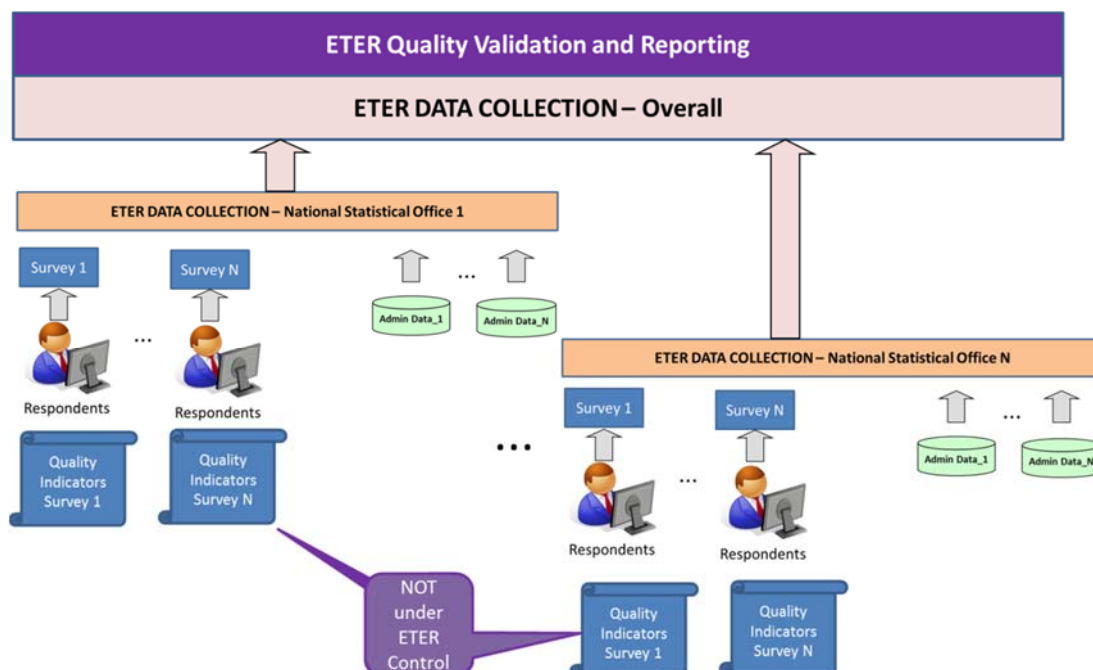
This chapter describes in detail the ETER Quality Validation and Reporting process. On one side, a set of operational steps is detailed in order to make this process feasible, on the other side the process itself is described in terms of design choices and quality methods used.

The ETER Quality Validation and Reporting process consists of the following phases:
- Quality Metadata Collection, contextual with data collection, in which NSAs provide the set of quality metadata as specified in the next section.
- Preliminary level quality and validation checks performed within the data collection phase on a country basis and on the whole dataset. These checks are described in detail in chapter 8
- Quality control and validation consisting of:
  o Computation of Quality Indicators, mainly exploiting quality metadata as defined in Section 7.3 "Quality Indicators";
  o Outlier detection, consisting of an in-depth data analysis performed according to advanced statistical methods described in Section 7.4 "Outlier Detection";
  o Checks with external data sources, either to assess the overall coverage against official statistics, national aggregates, or explain/correct problems detected with previous methods.

7.2 Metadata

In addition to the variables described in chapter 5, a set of metadata is collected in order to characterize the quality of ETER data. According to the ISO standard, metadata is defined as data that defines and describes other data and processes. The proposed metadata will be one of the inputs of the data quality and validation process. Although metadata are part of the data collection, the relevant cells should be filled in only when the required information is relevant according to the nature of the variable.

These metadata can exported directly with the ETER data from the on-line search tool (see chapter 9 of this handbook).

We distinguish between metadata collected for each country at the variable level and metadata collected at the single record level, which will be filled-in within the data collection sheet and are directly available in the flags and remarks field in the database.
The table summarizes the proposed set of metadata in ETER, specifying the coverage of collection (only quantitative variables / all variables, including descriptors). Detailed definitions are provided in the following subsections.

Table 19. List of Meta Data

Level of collection	Metadata	Format	Coverage
Variable level	Content and departure from definition Reference period/date Source release date Reference to data source Type of providing data source In-depth information	Text Mixed Date Text Nominal Text	All Quantitative variables All All Quantitative variables Selected variables
Record level	Completeness metadata Data flags Notes	Nominal Nominal Text	All All All

7.2.1 General metadata

These metadata are collected for all variables in the dataset by country and year and provide a general characterization of the data source. Since most of this information does not change across years, it is stored in the database including a start and end data of validity.

7.2.1.1 Content and departures from ETER definition

This is the exact national definition of the data filled-in and the remark of differences with ETER variable definitions, when relevant. This metadata is very useful for the creation of a common data dictionary and to address the issue of cross-country comparability that is very relevant in ETER. Therefore the most detailed description is required, as the information will be standardised during the data quality and validation phase.

In order to minimize the burden on respondents, the variable definition is required only when readily available from national statistical sources and only for quantitative variables. For descriptors, it is only required to add comments or underline departures from handbook definitions.

Type of variable: Text
Data source: NSA/NE

7.2.1.2 Reference period/date

This metadata defines the exact date or period of time to which data refer. It is particularly important when the reference date or period is different from the international standard as specified in paragraph 4.4 of the handbook. For example, for enrolled students the international standard refers to "end of first month and beginning of academic year"; each respondent is required to specify the exact date in her own country and, even more important, to highlight any deviations from the standard (i.e. enrolled students reported at the end of the academic year).

 Type of variable: Mixed (Date; Text)
Data source: NSA

7.2.1.3 Source release date

This is the date when data were officially made available by the providing source at the national level and that is necessarily prior to the date of ETER data collection compilation (and subsequent to reference period/date). When the concerning data are already published, the source release date refers to the date of publication at the national level; in other cases it refers to the date when data are available from the NSA (i.e. survey completed, administrative data acquired in national statistical database, etc.). For example, in country X financial data are collected with reference to 31/12/2011 (reference date) and are published at the national level on 30/09/2012 (source release date).

For descriptors and when data are collected directly from the institution's website by national experts, the date of data collection should be reported instead of the source release date (which would not apply).

Type of variable: Date

Data source: NSA, NE

7.2.1.4 Reference to data source

The exact name of the data source used and link to the relevant website when publicly available online.

Type of variable: Text

Data source: NSA/NE

7.2.1.5 Type of providing data source

We distinguish between three types of data sources: 0=survey, 1=administrative, 2=other

Type of variable: Nominal

Data source: NSA

7.2.1.6 Date of update

This variable identifies the year in which the metadata changed.

Type of variable: numeric, four digit (year).

Data source: NSA.

7.2.1.7 Change for previous year

This variable describes the change in the metadata, for example the recourse to a new data source..

Type of variable:text.

Data source: NSA.

7.2.2 In-depth information

The EUMIDA feasibility study confirmed that there are a few variables which are particularly problematic in terms of comparability among countries (Bonaccorsi *et al*, 2007). For these variables, a more specific collection of information on national practices/situations is envisaged, as an input to analysis, report and recommendations.

This paragraph contains the list of concerned variables and the issues to be clarified for each.

The structure of these variables are as follows:

- Reference year.
- Country.
- Variable category (for example 'expenditures').
- ETER variable (for example 'personnel expenditures').
- Metadata description.
- Metadata content.

This standardized structure allows extracting metadata by country, year or variable from the database for the purposes of comparative analysis and quality assessment.

7.2.2.1 Currency and exchange rates

This information is provided by the consortium and details the currency used in the data for that specific country and year. It includes three variables:

- National currency used.
- Exchange rate nominal for conversion of financial data to euros.
- Purchasing power parities for conversion of financial data to PPPs in euros.

See section 4.6 in this handbook for details.

7.2.2.2 Expenditures

Following metadata are collected.

Expenditures perimeter, i.e. a general description of what is included and excluded in the total expenditures of the HEI.

Method for inclusion or exclusion of ancillary services. Different institutional settlements or national practices may lead to different perimeters of items included in the HEIs balance sheet, with special reference to ancillary services (for example meals or transportation). A clear distinction of services provided for by the HEI and accounted for in its balance sheet by services provided by other institutions (i.e. Government, local authorities, etc.) should be provided here.

Coverage of expenditures in clinical medicine. Since the delimitation between HEIs and their hospital is sometimes problematic and might lead to comparability problems given the different rules by country, NSA are requested to provide information on the extent to which clinical medicine expenditures are included in the ETER perimeter.

Coverage and accounting system for capital expenditures. This metadata asks NSAs to provide information on whether capital expenditures data are available and on how they are accounted, i.e. by using accrual or cash accounting (see section 5.7.3.5 for further information).

Classification of expenditures. For ETER data collection, only a very broad breakdown of expenditures between current (personnel and non-personnel) and capital expenditures is required. However, to have a better understanding of expenditures breakdown and country mapping, details on this breakdown is requested to NSAs.

7.2.2.3 Revenues

Following metadata are collected for what concerns revenues.

Distinctions where the institution has partial fees. Rules on access to higher education and relative costs for students and households widely differ among European countries, and eventually within each country among different HEIs categories (i.e. state vs. private institutions) or even within each institution among cycles or curricula. A description of the fees regime is required here when an institution has partial fees, for example limited to post graduate curricula or to a specific segment of the student population (e.g. mature students).

Revenues perimeter. Details on what items are included among revenues, including for example revenues for services, endowments, etc.

Classification of revenues. This metadata asks for each ETER revenues category to indicate the corresponding revenue items in the national classification. A general remark can be added. Following subcategories are covered by distinct metadata items:

- Core funding - Basic government allocation.
- Core funding - Other core funding.
- Third party funding - Public third-party funding.
- Third party funding - Private third-party funding.
- Third party funding - Private third-party funding.
- Third party funding - Private third-party funding.
- Third party funding - Third-party funding from abroad.
- Third party funding - Third party funding undivided.
- Tuition fees funding.
- Revenues unclassified.
- Non-recurrent revenues.

7.2.2.4 Staff

For what concerns staff, the following metadata are collected.

Staff perimeter and coverage. This metadata provides information on the coverage of staff in the ETER data, for example for what concerns ancillary staff.

Counting method of staff FTEs. This metadata explains how FTEs of staff are calculated, including the reference period, possible exclusions 8for example staff below some percentage of employment, etc.).

Counting method of staff HC. This metadata explains how HC of staff are calculated, including the reference period, possible exclusions 8for example staff below some percentage of employment, etc.).

Inclusion of staff categories. These metadata inquire specifically about the inclusion of some staff categories for which comparability problems might occur (for example inclusion in some countries and exclusion in others). Specifically, following metadata are collected:
- *Inclusion of PhD students*. The inclusion of PhD students among academic staff, which depends on the contractual position of PhD students in the respective country (a binary specific variable is included in the data collection).
- *Inclusion of ancillary and support staff,* for example staff in the support services, transport, cantina, etc., since it is known that this varies by country. In some case there could be autonomous entities connected with the HEI (spin-off, foundations, etc.) employing staff that collaborate with academic staff and contribute to the production of HEI output but is not accounted.
- *Coverage of medical staff,* as depending on the position of the hospital in respect to the university there are large variation in Europe in the extent medical staff is accounted in the HEI perimeter.

National classification system for academic staff by field. Here it is required to describe the national classification system of staff by scientific or academic domains applying in each country, and whenever possible to match with the Frascati Manual "fields of science" at the broadest level respectively with ISCED-F fields of education.
The exact national categories for each ISCED-F field are requested as distinct metadata.

Number of professors. This metadata collects information on the exact national categories included under the professor category in ETER, as comparability problems might be generated by different academic staff structures in Europe.

7.2.2.5 Education

Typologies of degree awarded and levels of education. Detailed breakdown of the number of students by level of education is required for the data collection. Here it is required to describe the actual levels of education provided in each country, and eventually distinguishing among sub-groups of HEIs and typologies of degree awarded. When special rules on the length of curricula apply only to specific subjects (i.e. medicine, engineering, etc.) they also should be underlined here if they can have an impact on international comparability on the number of graduates.

7.2.2.6 R&D expenditure

Method for calculating R&D expenditure. R&D expenditures provided for the data collection should be computed according to international practices in the R&D statistics based on the Frascati manual. It is recognised that only a minority of the countries has been able to devise a statistical system following the guidelines of the Frascati manual. Here it is required to describe the concrete method for calculating R&D expenditures in each country, eventually describing reasons for the non-availability of data at the HEI level.

7.2.3 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 defined by ETER in the form of "special codes" (see section 4.5 for details):

- Code "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).
- Code "m" refers to the fact that the data in question is missing.
- Code "x" should be applied when a specific breakdown is not available, but the data are included in the total.
- Code "x" should be applied when a specific breakdown is not available, but the data are included in the total.
- Code "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).
- Code "xr" should be used for data that are included in other rows, which can occur when an institution is part of another institution.
- Code "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).
- Code "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).
- Code "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 (i.e. notes on enrolled students at ISCED level 6).

7.3 Quality indicators

In order to characterize the quality of data collected in ETER, we decide to evaluate a set of dimensions listed in Table 2. These dimensions belong to the ISO 25012 view related to Internal Data Quality and are a subset of those proposed by EUROSTAT, 2007 resulting from the considerations described above related to the specificity of the ETER collection.

Table 20. Quality Dimensions

Dimension	Purposes
Accuracy (Format)	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.

Quality indicators allow for a quantitative assessment of the quality dimensions listed in table 7.2. They will be compiled on the basis of quality metadata collected from the NSA/NE and described in the previous section of the handbook.

Each quality dimension can be evaluated according to different granularity levels, namely:
- Variables, e.g., Personnel expenditures, Third party funding, etc.
- Dimensions, e.g. Expenditures, Income, etc.
- Data set, e.g. related to the whole bunch of data provided by a single National Statistical Office.

Hence, for each dimension, the following indicators have been defined.

7.3.1 Format Accuracy

Definition: For each variable, *format accuracy* evaluates the compliance of the value to the requested format, as defined in the data chapter of the ETER Handbook.
Variable Format accuracy Indicator:
Provided Format=Requested Format
Value: TRUE, FALSE

7.3.2 Completeness

Definition: For each variable, dimension and data set, *completeness* evaluates the number of missing values (with the meaning relevant to completeness, i.e. unavailable or temporarily unavailable) that are present.

Completeness Variable Indicator:
Number of missing variable values/total number of variable values
Completeness Variable Value Range: [0, 1]

Completeness Dimension Indicator:
Number of missing variable values for all the variables of the dimension/total number of values for all the variables of the dimension.
Completeness Dimension Value Range: [0, 1]
Completeness Data set Indicator:
Number of missing variable values for all the variables of the data set/total number of values for all the variables of the data set.
Completeness Data set Value Range: [0, 1]

7.3.3 Consistency

Definition: Consistency verifies a possible violation of semantic rules defined over the involved data, and specifically between different variables. A first round of consistency checks is performed automatically within the data collection tool: whenever one of the following rules is violated, the respondent is required to provide an explanation. Under the column "notes" the respondent is asked to describe the cause of departure.

Table 21. Consistency indicators

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

Additionally, an annual comparison of national aggregates is implemented in the set of internal data quality mechanisms. In this process, all quantitative variables are aggregated within a country and each case, where a variable changes from the previous year to the current year by more than 20 % will be detected by an automated script. Although aggregation within a country leads to a bias of detected cases in smaller countries, the comparison of national aggregates across years presents a further useful consistency check.

7.3.4 Timeliness

Definition: For each variable, timeliness evaluates the time lag between the ETER publication date and the Source Release date.

Timeliness Variable Indicator: ETER Collection Date- Source Release Date< Tolerance Threshold
Value:
- TRUE, if ETER Collection Date- Source Release Date lower than Tolerance Threshold
- FALSE, otherwise

7.4 Comparability dimensions

The peculiarity of ETER data collection, which is performed by integrating data already collected by statistical institutions by means of different surveys or administrative data in 36 countries, also requires the evaluation of another dimension related to the comparability of data across countries. The EUMIDA feasibility study widely discussed this issue and the impossibility to completely overcome it.

On the basis of the specific metadata collected for each variable ("Content of the variable and departure from definition") and of relevant notes filled in at a record level for sub-categories of HEIs, a common data dictionary will be produced at the end of the data collection wave. For a subset of variables, specific questions will be addressed to NSAs in order to highlight several issues that emerged as relevant from the EUMIDA feasibility study.

Therefore, even if the challenge of comparability would require longer time and cannot be completely addressed within this tender, ETER will contribute:

- to highlight in a transparent way the comparability problems and differences among different countries statistics definition;
- to provide a common definition in the global data dictionary, which NSA could eventually follow for future collection;
- to assess the magnitude of departures and data distortion and decide when integration is possible with specific flags and when it is not possible at all.

7.4.1 Outlier Detection

An outlier can indicate an observation, a processing error, or a special element of the observed population that needs to be treated differently from the bulk in the subsequent processes.

In ETER, the outlier detection phase is part of the quality control and validation process, and has been performed with two different objectives: (i) identification of possible errors in data collected; (ii) better understanding of the collected data in order to perform a proper analysis on them. The ETER outlier detection phase consists of a *basic phase* and an *advanced phase*.

In the basic phase, some checks are performed on a limited number of ratios (see section 7.2 for details) comparing their value with expected values or the observed distribution at the national or European level. For example, the ratio between total expenditure and total income is expected be around 1. Relevant outliers should be

flagged and accompanied by a specific explanation in the notes and eventually rechecked.

The list of ratios used for outlier detection is contained in the table below.

In the advanced phase, a method to *discover* outliers (without a previous knowledge domain injected into the model) is applied and the distribution of the bulk of observations is robustly estimated according to a suitable model distribution. The method is applied on an extended set of ratios between variables either pertaining to the same dimension (i.e. fees funding over total income) or to different dimensions (i.e. personnel expenditure on total staff). The list of ratios is reported in Table 22.

Table 22. Ratios for outlier detection

Ratio	Breakdowns
Total expenditure / Total income	
Core budget / Total income	
Third party funding / Total income	
Student fees funding / Total income	
Tot expenditure / Total enrolled students ISCED 5-8	
Personnel expenditure / Total staff (FTE)	
Academic staff (FTE) / Number of administrative staff (FTE)	
Total staff (FTE) / Total staff (HC)	Academic staff, Administrative staff
Total enrolled students (ISCED 5-8) / Total academic staff	
Foreign enrolled students / Total enrolled students	by ISCED level
Mobile enrolled students / Total enrolled students	by ISCED level
Total students enrolled ISCED 5-7 / Total graduates ISCED 5-7	
Enrolled students ISCED 8 / Number of doctoral degrees	
Number of foreign students / Number of foreign graduates	by ISCED level
Number of mobile students / Number of mobile graduates	by ISCED level
R&D exp / Tot exp	

In the advanced phase, a method to *discover* outliers (without a previous domain knowledge injected into the model) is applied and the distribution of the bulk of observations is robustly estimated according to a suitable model distribution.

The method we use performs parameter estimation by regression on QQ plot positions[1].

For the purpose of outlier detection, the method assumes that the observations y_i are generated by a model probability density, with cumulative density function (cdf) F(Y |θ) where θ is a vector of parameters specifying F. The value of θ can be estimated robustly from the bulk of the observations by minimizing the sum of squares:

$$\theta = \arg\min \sum_{i \in \Lambda} \left[g(y_i) - g(F^{-1}(\hat{F}|\theta)) \right]^2 \quad [EQ\ 1]$$

Where:

---

[1] A Q–Q (Quantile-Quantile) is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

- Λ indexes a subset of the observations y_i , i.e. all the observations that will be used by the fitting phase
- g is a monotonic function, differentiable on the range of Y.
- $\widehat{F}_i\widehat{F}_i$ are plot positions as used in QQ plots.

The method can be applied with different distributions. We used the implementation of the method proposed by the R package "extremevalues"[2]. In particular we found that most of the empirical data show their best fit with the lognormal distribution, a widely empirically validated distribution suitable for neither non-normal nor symmetric variables, as the micro university data are (Ruocco and Daraio, 2013). In addition, the lognormal distribution solving [EQ1] with a suitable transformation g yields linear regression equations in the form:

$$b = (A'A)^{-1}A'x \;\; [EQ2]$$

Where:

- b is a 2-dimensional vector containing functions of the distribution parameters
- A is a $|\Lambda| \times 2$ matrix containing functions of $\widehat{F}_i\widehat{F}_i$ and
- x is a $|\Lambda|$dimensional vector containing functions of yi.

By solving EQ2, is it possible to estimate if the observed values are likely to be generated by the lognormal distribution or not.

Hence the method allows us to compare the real observations with the estimated ones, and thus to obtain a robust estimate for the bulk of the distribution.

7.4.2 Multiannual checks

The availability of data across different years (3 years after the first wave of data collection; 4 years after the second wave) raises the issue of longitudinal consistency of 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 (as compared with what is expected, respectively with the average change in the whole 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 3 or 4 years of data, however, does not enable to apply specific methods for time series analysis which require much more data. In addition the ETER dataset is composed by different typologies of variables (i.e. 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 is based on the use of different techniques:

- *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 alarm is 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;

---

[2]See URL: http://cran.r-project.org/web/packages/extremevalues/extremevalues.pdf

- *statistical analysis* to highlight the HEIs with annual growth outstanding from the overall distribution (outlier) with an adaptation of the methodology already used for yearly outlier detection.

The statistical analysis allows us to robustly estimate[3] the distribution that generates our observations and to identify "outliers" as the observations that are "unlikely" to be generated from the empirically (robustly) estimated distribution. We analyse the distribution of the growth rate over a two year interval calculated "à la Malmquist" and weighted to take into account the absolute size of variation.

We are comparing the value of variable x over two points in time (year0 and year1). Since there is no reason to assume which one is the reference year (i.e. the correct value) the rate is calculated as the geometric average of the absolute values of the growth rates calculated as year1 over year0 and year0 over year1 respectively. In formula:

(1) Growth rate = $\sqrt{|(x_1 - x_0)/x_0 \times 100| \times |(x_0 - x_1)/x_1 \times 100|}$

This methodological approach is inspired by the work of Malmquist (1953) widely used and accepted in literature. In ETER we face another specific challenge due to the large heterogeneity of higher education institutions in the sample ranging from micro institutions (also below the size thresholds set for the ETER database) to huge institutions. In this context it is expected that small institutions register tremendous growth rates when the absolute numbers are small (i.e. enrolled students growing from 10 to 20 at ISCED level 6 lead to a growth rate of 100%). To correct for this possible source of distortion, the growth rate formula (1) is weighted for the difference of the levels of variable x in absolute terms.

(2) Weighted growth rate = $\sqrt{|(x_1 - x_0)/x_0 \times 100| \times |(x_0 - x_1)/x_1 \times 100| \times |x_1 - x_0|}$

The method has been applied by comparing the actual distribution of the growth rates with different distributions. We follow a data driven approach to the selection of the fitting distribution, selecting the one with highest correlation coefficient and most suitable to the specific aim of the analysis. In the ETER data quality process indeed the focus of multiannual checks is not on observations with small or null growth rate (stability of levels over time) but rather on those with the highest variations. In other words we are more interested in upper bound outliers.

The statistical multiannual outlier detection is applied to the 17 variables listed in the table below, covering all ETER dimensions. The comparison is always made between current year against the previous one, but whenever an outlier is detected the whole longitudinal set (starting from 2011) is manually analysed to better interpret its nature. The goal of the statistical method is therefore essentially related to the identification of cases to be controlled. Further an expert based evaluation is performed by the consortium working team.

---

[3] The analysis has been done by using the R package "extremevalues".

Table 9. list of multiannual checks

Variable	Variable
Total expenditure (PPP)	Total staff (FTE)
Total revenues (PPP)	Total staff (HC)
Total academic staff (FTE)	Total students enrolled (by ISCED level)
Total academic staff (HC)	Total graduates (by ISCED level)
Number of administrative staff (FTE)	

7.5 Quality Reporting

Quality reporting has the purpose of summarizing the data quality validation procedures and to report the quality metadata calculated in each wave of data collection.

The overall data quality status of each wave of data collected has been summarised in a technical data quality report, prepared by the ETER data quality managers, which accompanies the data. The Final Quality Report is available at: https://www.eter-project.com/assets/pdf/ETER_quality_report_2017.pdf

Quality problems will be also highlighted at a record level with a system of flags and notes that will be integrated in the dataset by the project team.

7.5.1 Data flags

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:

* individual cells.
* one dimension or group of variables in a country (i.e. all variables concerning revenues).
* all variables for one or more HEIs in a specific country (i.e. all private HEIs).

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.

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.

The following table presents the list of accompanying flags.

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.
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 might be revised at later stage.

8 Data collection and preliminary validation

8.1 General information

The ETER project provides an infrastructure for data collection, which allows for standardization and systematization of the process. This infrastructure includes:

- Templates for data collection including documentation (e.g. flags and special values as commonly used for EUROSTAT-statistics), which guide national data sources (statistical offices, national authorities, other sources) and country experts addressing and supporting national data sources.
- A master database including the upload interface and documentation of the database.

A web application, which allows individual data exports (downloads of all variables for all countries are also possible, for example staff data for a specific country). The specific requirements of the ETER project suggest a centralized web server based collection tool, which will be specified for the needs of the two data collection rounds in ETER. The advantage of such a solution is a closer linkage between data collection, feedback and revision (identification of problems and coordinated support) and finally integration to a raw data set ready for advanced quality, respective of consistency checks and data analysis.

The following figure presents an overview of the entire process starting from data collection and control to data export for statistical analysis.

Figure 6: Data collection interfaces
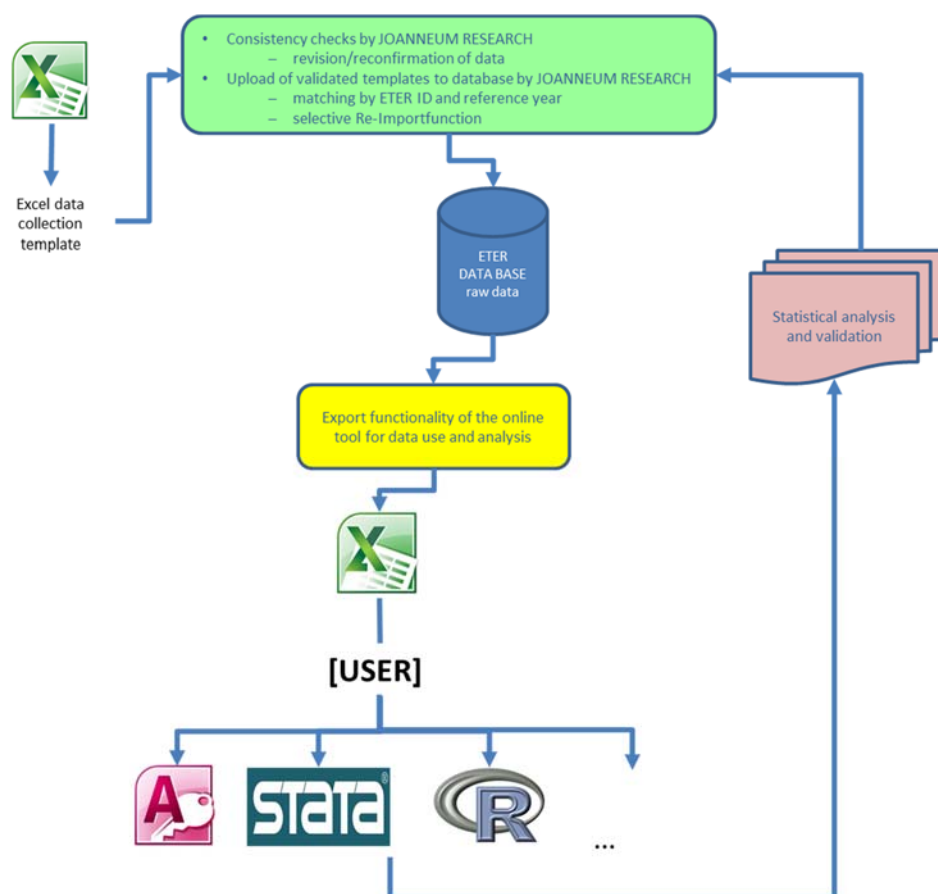
8.2 Structure of the database

The ETER database is based on a MongoDB architecture, which uses an object-oriented approach in order to deal with different structured datasets, so called "documents". Different structured documents lead to a dynamic schema of the database itself, which leads to an iterative and agile approach for the design and development of the ETER database. This implies that the structure of the database will change slightly by every added or withdrawn field (bits and pieces of data). Thus, if new fields need to be added to a document, the field can be created without affecting all other documents in the system, without updating a central system catalogue and without taking the system offline Also, if developers add more features, MongoDB continues to store the updated objects without the need for performing costly alteration operations, or worse - having to re design the schema from scratch.

Figure 7: Current structure of the ETER database (extract)
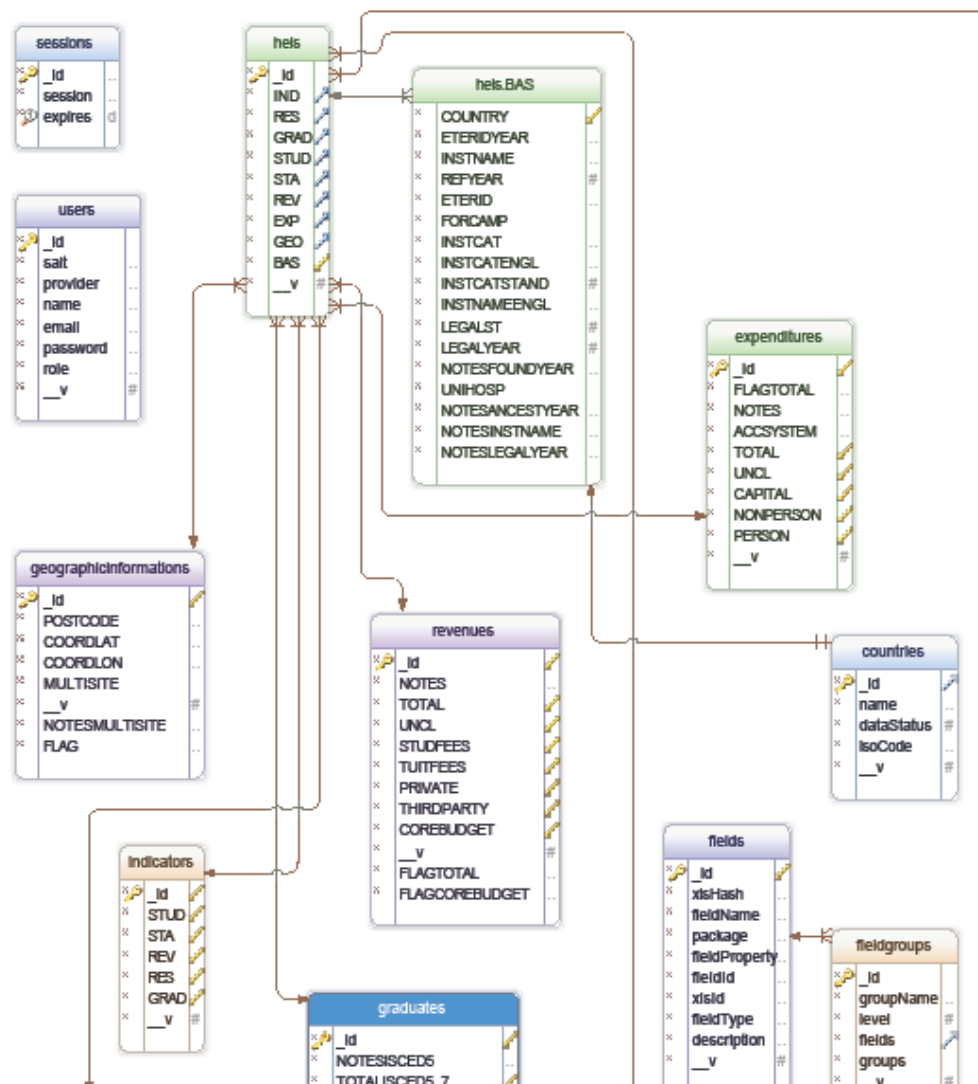
The ETER database consists of some master data and the dynamic data collected via the Excel files. Important parts of the master data are the "field" documents. A field document is a meta-description of a column within the HEI data collection Excel files. The field description defines:

- The type of data that is stored in this field.
- The display labels for the represented column.
- The possible flags available for this field.
- A list of display and export formats.
- The link of the Excel-Column ID and the data path within the database object.

The dynamic data are split into subdocuments in order to improve performance and data handling. Each HEI base document is linked to its subdocuments via a 1:1 connection.

The management of access via the online tool is described in chapter 9.

8.3 Data collection and preliminary validation

The target of the data collection and first phase of the preliminary validation is to ensure a high data quality level while using a standardized and reproducible process. The whole process is shown in Figure 8.

Figure 8. Data collection, validation and integration
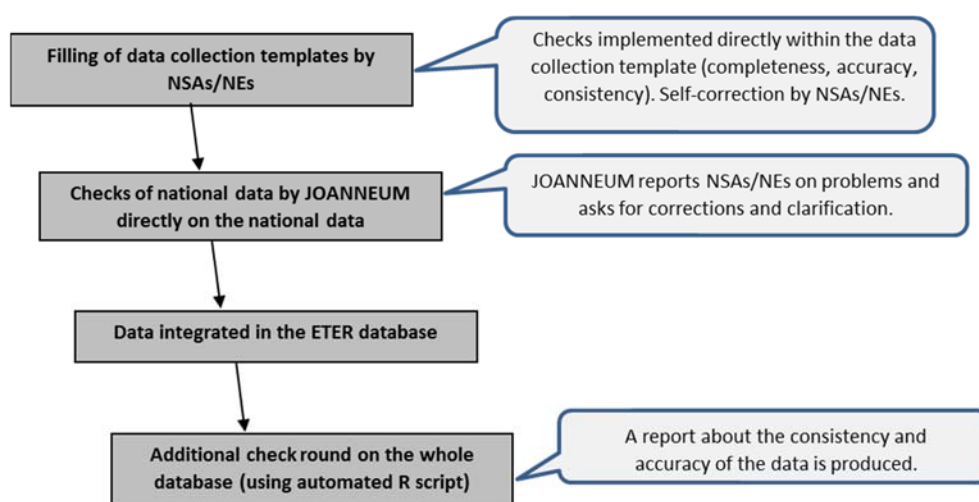
Data collection starts with a revision of each country's perimeter. The process looks as follows:

- The filled perimeter files for the last reference year are adapted to serve as templates. Any changes in the perimeter and demographic changes are recorded by NSAs/NEs in these files.
- Returned files are used to update the information about demographic events and the database. Also, the updated perimeters are transformed into the data collection files in order to start data collection immediately after the validation of the perimeter files.

The data collection is performed in Excel and can be started after revising the perimeter, since Excel files are a common and practical way to share data for the ETER partners. The bases of the data collection templates for the reference year are the files from the preceding year.

An already high level of data quality should be ensured by the implementation of a first phase of quick checks already in the data collection files in order to call attention to NSAs/NEs and thus enable them to correct the data immediately.

The implemented checks alert the deliverer of data about:

- incomplete data and blank cells,
- format accuracy problems, i.e. wrong data formats or irregularities, and
- mistakes in sums and inconsistencies between variables.

For detailed information see "Accuracy checks", "Completeness checks" and "Consistency checks" in Table 24.

The data collection files include prefilled cells with the previous year's values in order to support NSAs/NEs and reduce the burden for data collection. These include the information that is not expected to change systematically from year to year. Prefilled data include:
- descriptors (basic institutional variables, geographic variables),
- nominal variables which are not likely to change and the resulting not applicable values (e.g. no ISCED 5 level in a country, then "a" is prefilled in the respective categories),
- flags (except inconsistence "ic") and notes.
- metadata at the country level.

Prefilled nominal variables, flags and metadata are labelled in order to draw the attention of NSAs/NEs to them. Data are seen as confirmed if the responsible persons remove the labelling. This simultaneously ensures that the persons completing the data get some help in the process, but also seriously deal with these types of data.

The incoming data collection files are screened for any mistakes detected by the implemented automatic checks in the Excel files in order to provide reasonably clean data for the first upload into the database. Problematic cases are analysed by the core team of experts who are responsible for the supervised countries under their responsibilities and decide whether:
- To correct the data (when the source of problems is clear).
- To flag the data, when for example the reason of the problem is already known or explained in the metadata. Flags are already introduced in the data collection phase in order to simplify their management (this will also avoid detecting again problems which have been already flagged).
- To report back to NEs/NSAs when the explanation of problems is not straightforward.

The resulting data sets will be uploaded to the database in order to enable for the tracking of all changes. The imported data will in a further step separately be checked country by country using an R script, which produces a report about the consistency and format accuracy of the data. At the end of the data collection, a report about the consistency and format accuracy of the whole data set will be produced. The R script will detect:
- missing values,
- inconsistencies regarding format accuracy (data formats, irregularities),
- mistakes in sums, and
- noticeable ratios.

A list of ratios to be checked can also be found in chapter 10.

Using the report about the consistency and accuracy of the data, obvious mistakes are corrected by the ETER project team, while all other issues have to be clarified in cooperation with NSAs/NEs. Additionally, the report about the consistency and accuracy of the data forms the basis for the first flagging of data. On that basis, values which do not meet the requirements of quick checks but could be approved (to be flagged on the basis of a manual check). After corrections of data or new data delivery, the concerning data will again be uploaded to the database and checked with the R script. This step should ensure that the detected inconsistencies are corrected and the data include no

new inconsistencies. The goal of this phase of checks is to spot problems which were not detected by the previous phase and thus to systematize checks and validation in order to ensure the best possible data quality.

Reports about the consistency and accuracy of the data (pdf) are time-stamped and available on the web application. This approach ensures that:

- all checks are easily reproducible running the R script,
- all reasons for changes can be traced by the quality report, and
- all changes in the data can be tracked directly in the database by
  - the user who changes the value,
  - the new value, and
  - date and time of change.

The final version of the data is uploaded and published when the pre-validation phase by the R script shows no more unflagged inconsistencies. The coding of confidential and small values for publicly available data is completed by JOANNEUM RESEARCH.

The pre-validated data can then be exported in order to perform quality checks and add additional flags according to the outcome of the quality analysis.

9 Data publication and on-line tool

9.1 Data publication and conditions for access

Most data in the ETER database are publicly available, once the national statistical authorities have given publication permission. Thus, two different types of access are provided in ETER:

- An open public access, where small numbers and all data for which a national statistical authority restricted public access, are coded with "c". These data can be freely downloaded from the ETER website and used for different purposes (under the condition that the source is mentioned).
- A restricted access ("Member Login"), where accredited users receive access to the entire data set for research purposes under the condition that individual data are not disclosed. In order to receive access to all data, interested users have to sign a non-disclosure agreement. This agreement allows for the use of restricted data for research purposes, under the condition that no individual data points are disclosed publicly.

Due to the fact that some data are delivered, but restricted for public access, data have to be recoded for the publicly available data set. The same is valid for values larger than 0 and below or equal to 3, which otherwise could allow the identification of individuals. The following codes are used in this respect:

- When data are available, but restricted to public access, they are recoded to "c" in the publicly available database.
- Values larger than 0 and below or equal to 3 will be recoded to "s". To prevent the reconstruction of coded values, some additional rules have to be fulfilled:
  - o If a total is larger than 0 and below or equal to 3, all breakdowns are coded to "s".
  - o If one or more breakdowns are larger than 0 and below or equal to 3, the unclassified category is also recoded to "s".
  - o If the value of the "unclassified" category is larger than 0 and below or equal to 3, it is not coded to "s" as long as the corresponding values of this breakdown are larger than 3.
    - If the corresponding values are larger than 0 and below or equal to 3, they will be recoded to "s" together with the "unclassified" category.
    - If corresponding values are either larger than 3 or 0, one additional 0 in the breakdowns has to be recoded to "s" in order to prevent recalculation.
  - o If one of the categories "Total academic staff" or "Total administrative staff" include values larger than 0 and below or equal to 3, both categories including breakdowns have to be recoded to "s" in order to preserve the value of total staff.
  - o FTE values in staff data are not recoded as long as the corresponding headcounts are larger than 3.

The ETER website (www.eter-project.com) includes a registration system, which allows potential users to register by indicating the required data, the research purpose etc. and will be managed by JOANNEUM RESEARCH. If access to the restricted data set is approved and they agree to a non-disclosure agreement, users will be activated in order to be able to access the regularly updated required data with their email address and password.

9.2 On-line tool

The new version of the database provides a web interface, which is dedicated to address a broad range of interested user. Therefore, the website does not only host a very detailed database, but also has a focus on dissemination by providing the possibility of

performing visualizations directly online. Particular attention in the reconstruction of the interface has been devoted to improve the usability of the website.

9.2.1 The starting page

The starting page of the website is designed in order to inform the user about news and the progress of the project itself and new analytical results from data exploration. This will be presented as a sliding menu, which automatically scrolls through the content and in this way is dedicated to dissemination of the ETER content. The web application will also provide further resources for information and dissemination of analytical results in order to raise awareness for ETER. This content will be updated regularly with short evidence-based ETER briefs. Thus, the new version of the web interface is strongly user and content oriented, with particular attention paid to the visibility of policy relevant results. The current version of the starting page also allows the user to register for unrestricted access to the database by signing up.
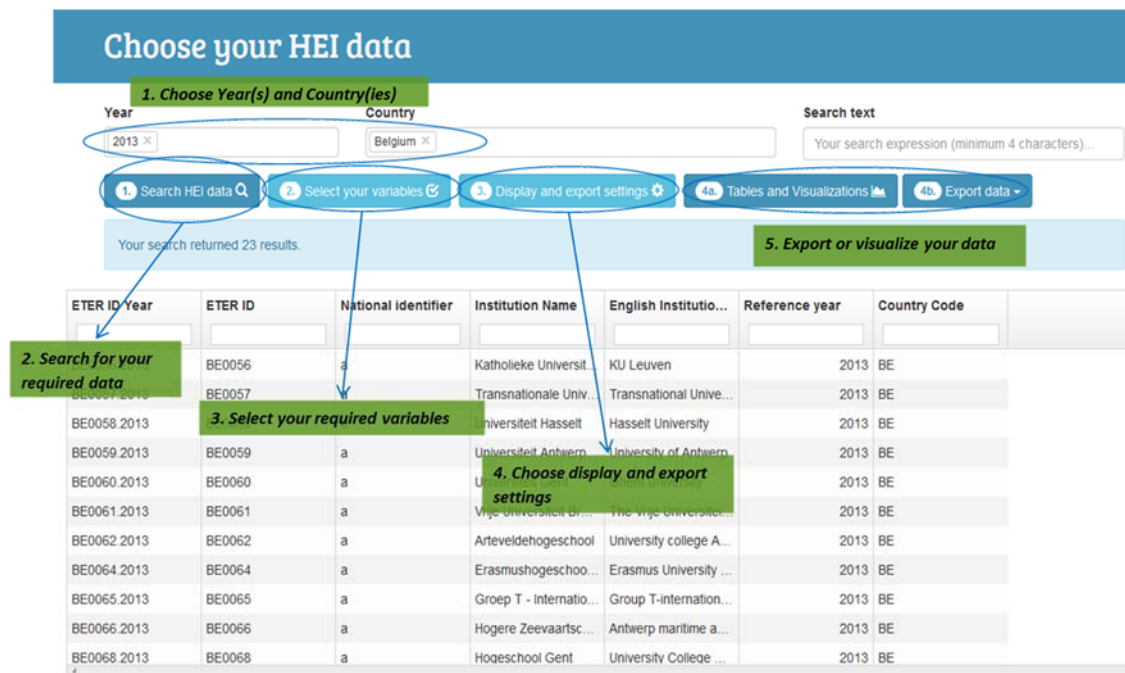
Figure 9: Starting page of the ETER website
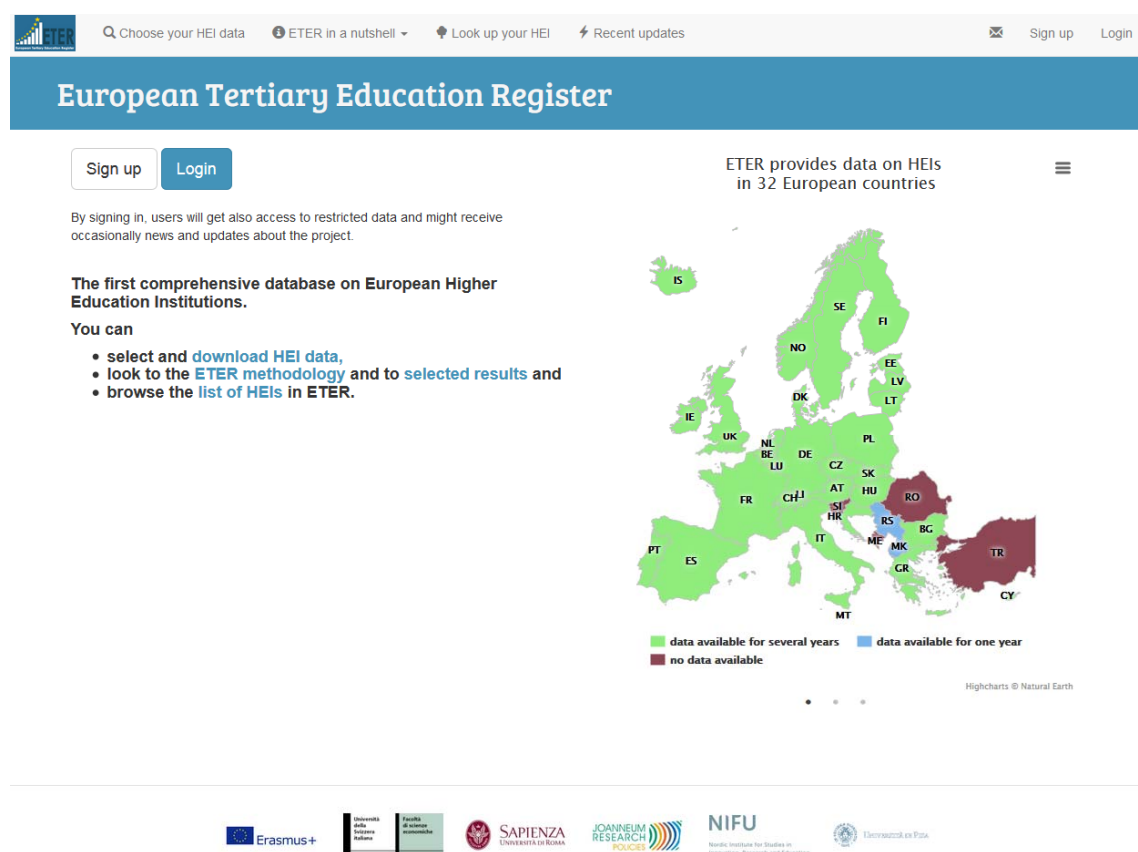
9.2.2 Choose your HEI data

The heart of the website is a search function within the menu *Choose your HEI data*, which allows the user in a first step to choose the required higher education institutions by year and country. Additional possibilities allow the user to refine the selection (*Select your variables*) and filter them. With the group of chosen institutions, the user can go a step forward and decide to download the requested data (*Export data*) or arrange and visualize them directly on the website (*Tables and Visualizations*).

Figure 10: Choose your data

The default results mask includes some basic variables in order to get a first overview. Option *Select your variables* enables the user to select the required variables for export (an exporting of all variables at once is also possible without selecting them – see export function), tables and visualizations.

Figure 11: Choose your HEI data
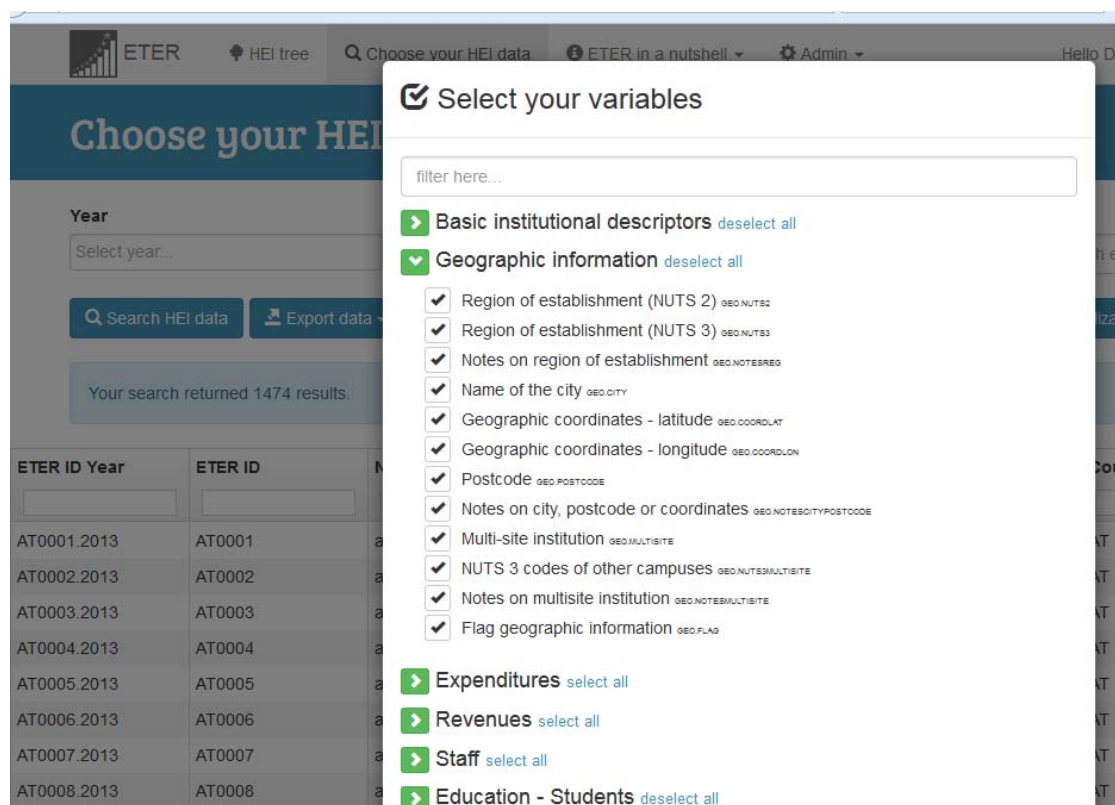
After selecting variables, it is possible to change to the interactive *Tables and Visualizations* space. Using a table component, users are able to customize their data sets directly on the website. Users can drag and drop columns and arrange the data in order to retrieve the required data format. In interactive tables, different variables can be combined, displayed, and adapted to the specific user's needs. The advantage of this possibility is that users don't have to download all data every time, but can choose and also rearrange the data the way they want. The figure below illustrates the practical use of the table function and shows one possible example of practical usage, namely the summarization of Belgium undergraduate students (2013) by NUTS 2-region of institutions.

Figure 12: How to Organize data in the pivot grid
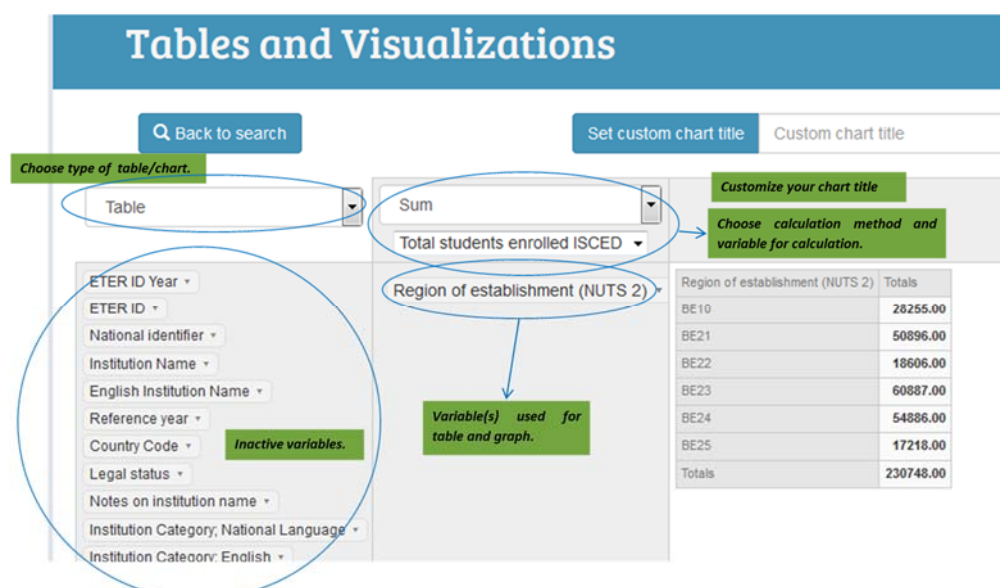
Users can additionally apply filter to all variables and use the table grid in order to calculate shares. The figure below shows the shares of undergraduate students by country and legal status for Austria, Belgium and Germany in the year 2013.

Figure 13: Undergraduate students by country and legal status for AT, BE, DE 2013
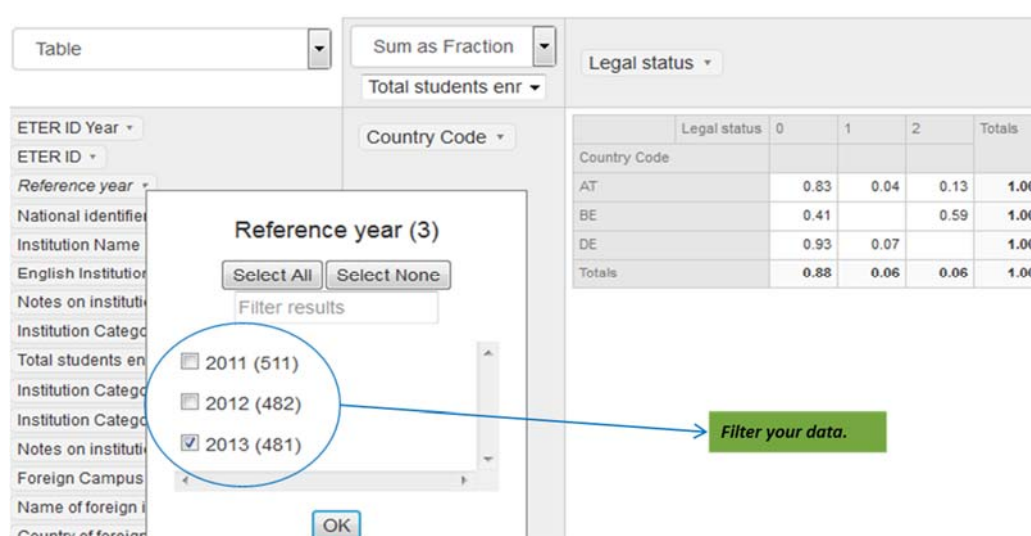
Additionally, graphical options for the selected data are available on the web interface. This is the case for selected indicators and allows the comparison of single institutions in a broader context. All modifications and new features are dedicated to usability and the support of the dissemination objectives of ETER and the contract. The visually oriented user will be able to learn more about structural differences and dynamics. The web interface also allows graphs in a printable quality on that basis. The following figures show some examples of already implemented visualization possibilities. Using exactly the same selections as in the example above, users can now change from 'Table' to 'Column Chart' (or alternatively to 'Bar Chart') in order to display the shares calculated above. The result would look the following way.

Figure 14: Column Chart of share of Undergraduate students by legal status for AT, BE, DE 2013
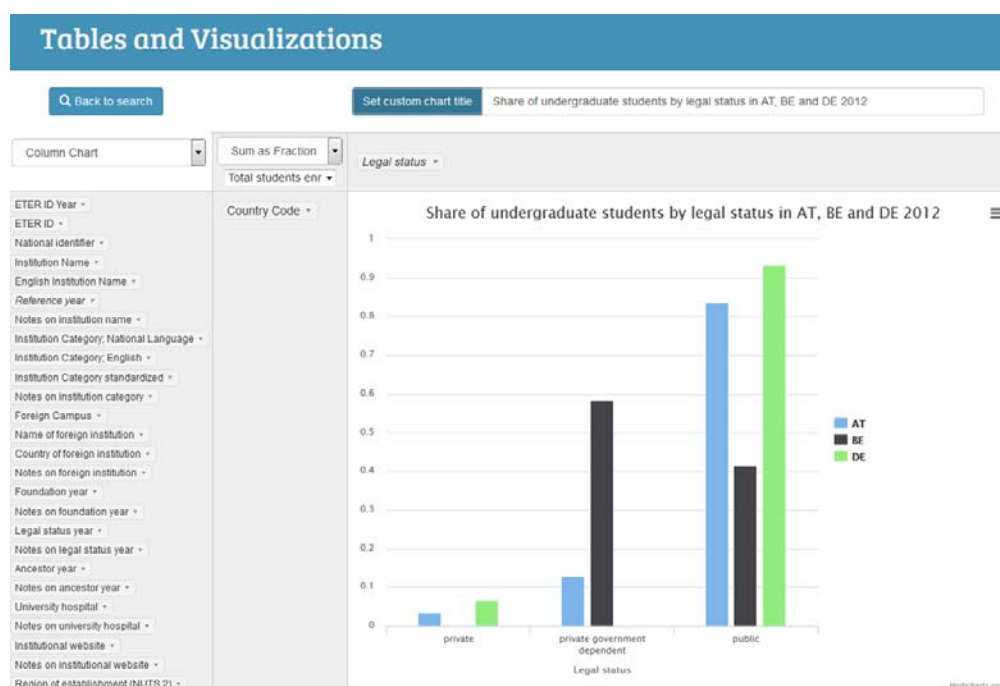
The most important feature for data analysts is the export function, which is dedicated to usability in order to support data analysis and the dissemination objectives of ETER

Figure 15: Choose your export type

Data export will be possible in .xlsx and .csv format, where users have several options:
- choose the data to be exported by applying filter and selecting variables or export all variables at once.
- export the corresponding metadata.
- Choose the export header in order to use either variable labels (full name of a column) or variable names (systematic variable naming).

- replace special codes with fixed values in order to allow the respective statistical software to recognize missing values and therefore enable users to start data analysis immediately after exporting the data.
- replace variable codes, special codes and flags with their full labels in order to support statistical analyses.

Figure 16: Choose your display and export settings

9.2.3 ETER in a Nutshell

Following the approach of the first ETER version, the new web interface also provides a large amount of valuable information in order to enable the user to:
- obtain a general overview about the project itself, its targets and contents.
- gather detailed information on the data collected, the origin of the data and special codes and flags used in the data collection.
- retrieve all relevant documents accompanying the project. These include a report about the main conceptual and methodological choices of ETER, the data collection handbook with an in-depth description of the data collection methodology, a technical report with detailed information on methodology, data collection, data management and data quality, etc.
- access results of analyses (ETER briefs) based on ETER data. These analyses were performed by the project team and individual analysis as well as figures, which can be downloaded separately.
- acquire information on publications based on the ETER data set.
- obtain answers to key questions by having a look at the frequently asked questions (FAQs) section on the website.

Figure 17: ETER in a Nutshell

10 Annex

Table 24: Types of checks performed before the final quality analysis

Type of checks	Description	Procedure
Accuracy checks	Accuracy checks verify that data entered have the right format foreseen by the handbook and that no logically impossible values are found.	Accuracy 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). Rules in this respect are stipulated in the handbook. 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.
Noticeable cases	Standard ratios are calculated (for example students to graduates) and compared to the national averages. See the handbook for the list of ratios.	Noticeable 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. Quality of metadata is critical for the exploitation of the

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

Table 25: Full list of accuracy, completeness and consistency checks

Variable	Check
Institutional Code	prefilled, cell locked, check for missing variables
National identifier	prefilled, cell locked, check for missing variables
Institution Name	prefilled, cell locked, check for missing variables
English Institution Name	prefilled and locked if available, unlocked if empty
Reference year	prefilled, cell locked, check for missing variables
Country Code	prefilled, cell locked, check for missing variables
Institution Acronym	check for missing variables
Legal status	only 0, 1, 2 or "m" possible, check for missing variables
Institution Category; National Language	check for missing variables
Institution Category; English	check for missing variables
Institution Category standardized	only 0, 1, 2 or "m" possible, check for missing variables
Foreign Campus	only 0, 1 or "m" possible, check for missing variables
Name of foreign institution	alarm if "Foreign Campus" = 0 and the value in this cell is not "a" OR if cell is empty OR if "Foreign Campus" = 1 and the value in this cell is "a" OR if "Foreign Campus" = "m" and value in this cell is not "m"
Country of foreign institution	alarm if "Foreign Campus" = 0 and the value in this cell is not "a" OR if cell is empty OR if "Foreign Campus" = 1 and the value in this cell is "a" OR if "Foreign Campus" = "m" and value in this cell is not "m"
Foundation year	check for missing variables
Legal status year	check for missing variables
Ancestor year	check for missing variables
University hospital	only 0, 1 or "m" possible, check for missing variables
Institutional website	check for missing variables
Region of establishment (NUTS 2)	alarm if number of characters is not 4, if value is not "a" or if value is not "m" OR if cell is empty
Region of establishment	alarm if number of characters is not 5, if value is not "a" or if value is not "m" OR if cell is empty

(NUTS 3)	
Name of the city	check for missing variables
Geographic coordinates - latitude	Filled by project team, cell locked, check for missing variables
Geographic coordinates - longitude	Filled by project team, cell locked, check for missing variables
Postcode	check for missing variables
Multi-site institution	only 0, 1, "a", "m" possible, check for missing variables
NUTS 3 codes of other campuses	alarm if "Multi-site institution" = 0 and the value in this cell is not "a" OR if cell is empty OR if "Multi-site institution" = 1 and the value in this cell is "a" OR if "Multi-site institution" = "m" and value in this cell is not "m"
Personnel expenditure	check for missing variables, non-negative
Non-personnel expenditure	check for missing variables, non-negative
Expenditure unclassified	Check for missing variables, non-negative
Total current expenditure	alarm if sum of personnel, non-personnel and unclassified expenditure is not equal to total current expenditure OR if cell is empty OR if value is negative OR if breakdowns are numeric and total is not OR if breakdowns are numeric and unclassified is not
Capital expenditure	check for missing variables, non-negative
Accounting system of capital expenditure	only 0, 1, 2, "a" or "m" possible
Basic government allocation	check for missing variables, non-negative
Other core budget	check for missing variables, non-negative
Total core budget	alarm if sum of basic government allocation and other core budget is not equal to total core budget OR if cell is empty OR if value is negative OR if breakdowns are numeric and total is not
Public third party funding	check for missing variables, non-negative
Private third party funding	check for missing variables, non-negative
Third party funding from abroad	check for missing variables, non-negative
Third party funding unclassified	check for missing variables, non-negative
Total third party funding	alarm if sum of public third party funding, private third party funding, third party funding from abroad and third party funding unclassified is not equal to total third party funding OR if cell is empty OR if value is negative OR if breakdowns are numeric and total is not
Tuition fees	only 0, 1, 2, "a" or "m" possible

Student fees funding	alarm if "Tuition fees" = 0 and the value in this cell is not "a" OR if cell is empty OR if "Tuition fees" is larger number than 0 and the value in this cell is "a" OR if value is negative
Revenue unclassified	check for missing variables, non-negative
Total current revenues	alarm if sum of total core budget, total third party funding, student fees funding and revenues unclassified is not equal to total current revenues OR if cell is empty OR if value is negative OR if breakdowns are numeric and total is not OR if breakdowns are numeric and unclassified is not
Non-recurring revenues	check for missing variables, non-negative
Total academic staff (FTE)	check for missing variables, non-negative
Academic staff - men (HC)	check for missing variables, non-negative
Academic staff - women (HC)	check for missing variables, non-negative
Academic staff - gender unclassified (HC)	check for missing variables, non-negative
Academic staff - national (HC)	check for missing variables, non-negative
Academic staff - foreigner (HC)	check for missing variables, non-negative
Academic staff - citizenship unclassified	check for missing variables, non-negative
Academic staff - ISCED-F 00 (HC)	check for missing variables, non-negative
Academic staff - ISCED-F 01 (HC)	check for missing variables, non-negative
Academic staff - ISCED-F 02 (HC)	check for missing variables, non-negative
Academic staff - ISCED-F 03 (HC)	check for missing variables, non-negative
Academic staff - ISCED-F 04 (HC)	check for missing variables, non-negative
Academic staff - ISCED-F 05 (HC)	check for missing variables, non-negative
Academic staff - ISCED-F 06 (HC)	check for missing variables, non-negative
Academic staff - ISCED-F 07 (HC)	check for missing variables, non-negative
Academic staff - ISCED-F 08 (HC)	check for missing variables, non-negative
Academic staff - ISCED-F 09 (HC)	check for missing variables, non-negative
Academic staff - ISCED-F 10 (HC)	check for missing variables, non-negative
Academic staff - ISCED-F unclassified	check for missing variables, non-negative

Total academic staff (HC)	alarm if sum of men, women and gender unclassified (respectively sum of national, foreigner and citizenship unclassified and sum of ISCED-F) is not equal to total staff OR if cell is empty OR if cell of men and women is numeric and gender unclassified is not (respectively national, foreigner and citizenship unclassified and ISCED-F) OR if men and women are text and gender unclassified is not (respectively national, foreigner and citizenship unclassified and ISCED-F) OR if breakdowns are numbers and total is text OR if value is negative
Number of full professors – men	check for missing variables, non-negative
Number of full professors – women	check for missing variables, non-negative
Number of full professors – unclassified	check for missing variables, non-negative
Total number of full professors	alarm if sum of men, women and gender unclassified is not equal to total OR if cell is empty OR if cell of men and women is numeric and gender unclassified is not OR if men and women are text and gender unclassified is not OR if number is negative OR if breakdowns are numbers and total is text
Inclusion of PhD students	Alarm if value is not 0, 1 or "m" OR if value of total students at ISCED 8 is "a" and value in this cell is not 0 and staff data are not missing
Number of administrative staff (FTE)	check for missing variables, non-negative
Number of administrative staff (HC)	check for missing variables, non-negative
Total staff (FTE)	alarm if sum of "Total academic staff (FTE)" and "Number of administrative staff (FTE)" is not equal to total staff OR if cell is empty OR if breakdowns are numbers and total is text OR if value is negative
Total staff (HC)	alarm if sum of "Total academic staff (HC)" and "Number of administrative staff (HC)" is not equal to total staff OR if cell is empty OR if breakdowns are numbers and total is text or if value is negative
Lowest degree delivered	alarm if value is not 0, 1, 2, 3, "a" or "m" OR if cell is empty OR if "Lowest degree delivered" is higher than "Highest degree delivered"
Highest degree delivered	alarm if value is not 0, 1, 2, 3, "a" or "m" OR if cell is empty OR if "Lowest degree delivered" is higher than "Highest degree delivered"
Students enrolled at ISCED 5 - men	check for missing variables, non-negative
Students enrolled at ISCED 5 - women	check for missing variables, non-negative
Students enrolled at ISCED 5 - gender	check for missing variables, non-negative

unclassified	
Students enrolled at ISCED 5 - national	check for missing variables, non-negative
Students enrolled at ISCED 5 - foreigner	check for missing variables, non-negative
Students at ISCED 5-citizenship unclassified	check for missing variables, non-negative
Students enrolled at ISCED 5 - resident	check for missing variables, non-negative
Students enrolled at ISCED 5 - mobile	check for missing variables, non-negative
Students enrolled at ISCED 5- mobility unclassified	check for missing variables, non-negative
Students enrolled at ISCED 5 - ISCED-F 00	check for missing variables, non-negative
Students enrolled at ISCED 5 - ISCED-F 01	check for missing variables, non-negative
Students enrolled at ISCED 5 - ISCED-F 02	check for missing variables, non-negative
Students enrolled at ISCED 5 - ISCED-F 03	check for missing variables, non-negative
Students enrolled at ISCED 5 - ISCED-F 04	check for missing variables, non-negative
Students enrolled at ISCED 5 - ISCED-F 05	check for missing variables, non-negative
Students enrolled at ISCED 5 - ISCED-F 06	check for missing variables, non-negative
Students enrolled at ISCED 5 - ISCED-F 07	check for missing variables, non-negative
Students enrolled at ISCED 5 - ISCED-F 08	check for missing variables, non-negative
Students enrolled at ISCED 5 - ISCED-F 09	check for missing variables, non-negative
Students enrolled at ISCED 5 - ISCED-F 10	check for missing variables, non-negative
Students enrolled	check for missing variables, non-negative

at ISCED 5-ISCED-F unclassified	
Total students enrolled at ISCED 5	alarm if "Highest degree delivered" = 0 and the value in this cell = "a" OR if cell is empty OR if sum of men, women and gender unclassified (respectively citizenship, mobility and fields of education) is not equal to total students enrolled in ISCED 5 OR if men and women are text and gender unclassified is not (respectively citizenship, mobility and fields of education) OR if cell of men and women is numeric and gender unclassified is not (respectively national, foreigner and citizenship unclassified) OR if value is negative OR if "Lowest degree delivered" = 1,2 or 3 and total is not "a" OR if "Lowest degree delivered" = 0 and total = "a" OR if breakdowns are numbers and total is text
Students enrolled at ISCED 6 - men	check for missing variables, non-negative
Students enrolled at ISCED 6 - women	check for missing variables, non-negative
Students enrolled at ISCED 6 - gender unclassified	check for missing variables, non-negative
Students enrolled at ISCED 6 - national	check for missing variables, non-negative
Students enrolled at ISCED 6 - foreigner	check for missing variables, non-negative
Students at ISCED 6-citizenship unclassified	check for missing variables, non-negative
Students enrolled at ISCED 6 - resident	check for missing variables, non-negative
Students enrolled at ISCED 6 - mobile	check for missing variables, non-negative
Students at ISCED 6 - mobility unclassified	check for missing variables, non-negative
Students enrolled at ISCED 6 - ISCED-F 00	check for missing variables, non-negative
Students enrolled at ISCED 6 - ISCED-F 01	check for missing variables, non-negative
Students enrolled at ISCED 6 - ISCED-F 02	check for missing variables, non-negative
Students enrolled at ISCED 6 - ISCED-F 03	check for missing variables, non-negative

Students enrolled at ISCED 6 - ISCED-F 04	check for missing variables, non-negative
Students enrolled at ISCED 6 - ISCED-F 05	check for missing variables, non-negative
Students enrolled at ISCED 6 - ISCED-F 06	check for missing variables, non-negative
Students enrolled at ISCED 6 - ISCED-F 07	check for missing variables, non-negative
Students enrolled at ISCED 6 - ISCED-F 08	check for missing variables, non-negative
Students enrolled at ISCED 6 - ISCED-F 09	check for missing variables, non-negative
Students enrolled at ISCED 6 - ISCED-F 10	check for missing variables, non-negative
Students at ISCED 6 - ISCED-F unclassified	check for missing variables, non-negative
Total students enrolled at ISCED 6	alarm if "Highest degree delivered" = 1 and the value in this cell = "a" OR if cell is empty OR if "Highest degree delivered" = 0 and the value in this cell is not "a" OR if sum of men, women and gender unclassified (respectively citizenship, mobility and fields of education) is not equal to total students enrolled in ISCED 6 OR if men and women are text and gender unclassified is not (respectively citizenship, mobility and fields of education) OR if cell of men and women is numeric and gender unclassified is not (respectively national, foreigner and citizenship unclassified) OR if value is negative OR if "Lowest degree delivered" = 1 and total = "a" OR if "Lowest degree delivered" = 2 or 3 and total is not "a" OR if breakdowns are numbers and total is text
Students enrolled at ISCED 7 - men	check for missing variables, non-negative
Students enrolled at ISCED 7 - women	check for missing variables, non-negative
Students enrolled at ISCED 7 - gender unclassified	check for missing variables, non-negative
Students enrolled at ISCED 7 - national	check for missing variables, non-negative
Students enrolled at ISCED 7 - foreigner	check for missing variables, non-negative
Students at ISCED 7-citizenship	check for missing variables, non-negative

unclassified	
Students enrolled at ISCED 7 - resident	check for missing variables, non-negative
Students enrolled at ISCED 7 - mobile	check for missing variables, non-negative
Students at ISCED 7 - mobility unclassified	check for missing variables, non-negative
Students enrolled at ISCED 7 - ISCED-F 00	check for missing variables, non-negative
Students enrolled at ISCED 7 - ISCED-F 01	check for missing variables, non-negative
Students enrolled at ISCED 7 - ISCED-F 02	check for missing variables, non-negative
Students enrolled at ISCED 7 - ISCED-F 03	check for missing variables, non-negative
Students enrolled at ISCED 7 - ISCED-F 04	check for missing variables, non-negative
Students enrolled at ISCED 7 - ISCED-F 05	check for missing variables, non-negative
Students enrolled at ISCED 7 - ISCED-F 06	check for missing variables, non-negative
Students enrolled at ISCED 7 - ISCED-F 07	check for missing variables, non-negative
Students enrolled at ISCED 7 - ISCED-F 08	check for missing variables, non-negative
Students enrolled at ISCED 7 - ISCED-F 09	check for missing variables, non-negative
Students enrolled at ISCED 7 - ISCED-F 10	check for missing variables, non-negative
Students at ISCED 7 - ISCED-F unclassified	check for missing variables, non-negative
Total students enrolled at ISCED 7	alarm if "Highest degree delivered" = 2 and the value in this cell or in sum of ISCED 7 long degree students = "a" OR if cell is empty OR if "Highest degree delivered" < 2 and the value in this cell is not "a" OR if sum of men, women and gender unclassified (respectively citizenship, mobility and fields of education) is not equal to total students enrolled in ISCED 7 OR if men and women are text and gender unclassified is not (respectively citizenship, mobility and fields of education) OR

	if cell of men and women is numeric and gender unclassified is not (respectively national, foreigner and citizenship unclassified) OR if value is negative OR if "Lowest degree delivered" = 2 and total in this cell or in ISCED 7 long degrees is "a" OR if "Lowest degree delivered" = 3 and total is not "a" OR if breakdowns are numbers and total is text
Students enrolled ISCED 7 long degree - men	check for missing variables, non-negative
Students enrolled ISCED 7 long degree - women	check for missing variables, non-negative
Students ISCED 7 long degree - gender unclassified	check for missing variables, non-negative
Students enrolled ISCED 7 long degree - national	check for missing variables, non-negative
Students enrolled ISCED 7 long degree - foreigner	check for missing variables, non-negative
Students ISCED 7 long degree- nation. unclassified	check for missing variables, non-negative
Students enrolled ISCED 7 long degree - resident	check for missing variables, non-negative
Students enrolled ISCED 7 long degree - mobile	check for missing variables, non-negative
Students ISCED 7 long degree- mobility unclassified	check for missing variables, non-negative
Students enrolled ISCED 7 long degree - ISCED-F 00	check for missing variables, non-negative
Students enrolled ISCED 7 long degree - ISCED-F 01	check for missing variables, non-negative
Students enrolled ISCED 7 long degree - ISCED-F 02	check for missing variables, non-negative
Students enrolled ISCED 7 long degree - ISCED-F 03	check for missing variables, non-negative
Students enrolled ISCED 7 long	check for missing variables, non-negative

degree - ISCED-F 04	
Students enrolled ISCED 7 long degree - ISCED-F 05	check for missing variables, non-negative
Students enrolled ISCED 7 long degree - ISCED-F 06	check for missing variables, non-negative
Students enrolled ISCED 7 long degree - ISCED-F 07	check for missing variables, non-negative
Students enrolled ISCED 7 long degree - ISCED-F 08	check for missing variables, non-negative
Students enrolled ISCED 7 long degree - ISCED-F 09	check for missing variables, non-negative
Students enrolled ISCED 7 long degree - ISCED-F 10	check for missing variables, non-negative
Students ISCED 7 long degree - ISCED-F unclassified	check for missing variables, non-negative
Total students enrolled ISCED 7 long degree	alarm if "Highest degree delivered" = 2 and the value in this cell or in sum of ISCED 7 long degree students = "a" OR if cell is empty OR if "Highest degree delivered" < 2 and the value in this cell is not "a" OR if sum of men, women and gender unclassified (respectively citizenship, mobility and fields of education) is not equal to total students enrolled in ISCED 7 OR if men and women are text and gender unclassified is not (respectively citizenship, mobility and fields of education) OR if cell of men and women is numeric and gender unclassified is not (respectively national, foreigner and citizenship unclassified) OR if value is negative OR if "Lowest degree delivered" = 2 and total in this cell or in ISCED 7 long degrees is "a" OR if "Lowest degree delivered" = 3 and total is not "a" OR if breakdowns are numbers and total is text
Total students enrolled ISCED 5-7	alarm if number is not equal to sum of students ISCED 5, 6, 7 and 7 long degree OR if cell is empty OR if number is negative
Distance education institution	only 0, 1 or "m" possible
Graduates at ISCED 5 - men	check for missing variables, non-negative
Graduates at ISCED 5 - women	check for missing variables, non-negative
Graduates at	check for missing variables, non-negative

ISCED 5 - gender unclassified	
Graduates at ISCED 5 - national	check for missing variables, non-negative
Graduates at ISCED 5 - foreigner	check for missing variables, non-negative
Graduates at ISCED 5 - citizenship unclassified	check for missing variables, non-negative
Graduates at ISCED 5 - resident	check for missing variables, non-negative
Graduates at ISCED 5 - mobile	check for missing variables, non-negative
Graduates at ISCED 5 - mobility unclassified	check for missing variables, non-negative
Graduates at ISCED 5 - ISCED-F 00	check for missing variables, non-negative
Graduates at ISCED 5 - ISCED-F 01	check for missing variables, non-negative
Graduates at ISCED 5 - ISCED-F 02	check for missing variables, non-negative
Graduates at ISCED 5 - ISCED-F 03	check for missing variables, non-negative
Graduates at ISCED 5 - ISCED-F 04	check for missing variables, non-negative
Graduates at ISCED 5 - ISCED-F 05	check for missing variables, non-negative
Graduates at ISCED 5 - ISCED-F 06	check for missing variables, non-negative
Graduates at ISCED 5 - ISCED-F 07	check for missing variables, non-negative
Graduates at ISCED 5 - ISCED-F 08	check for missing variables, non-negative
Graduates at ISCED 5 - ISCED-F 09	check for missing variables, non-negative
Graduates at ISCED 5 - ISCED-F 10	check for missing variables, non-negative
Graduates at ISCED 5 - ISCED-F unclassified	check for missing variables, non-negative

Total graduates at ISCED 5	alarm if "Highest degree delivered" = 0 and the value in this cell = "a" OR if cell is empty OR if sum of men, women and gender unclassified (respectively citizenship, mobility and fields of education) is not equal to total graduates ISCED 5 OR if men and women are text and gender unclassified is not (respectively citizenship, mobility and fields of education) OR if cell of men and women is numeric and gender unclassified is not (respectively national, foreigner and citizenship unclassified) OR if value is negative OR if "Lowest degree delivered" = 1,2 or 3 and total is not "a" OR if "Lowest degree delivered" = 0 and total = "a" OR if breakdowns are numbers and total is text
Graduates at ISCED 6 - men	check for missing variables, non-negative
Graduates at ISCED 6 - women	check for missing variables, non-negative
Graduates at ISCED 6 - gender unclassified	check for missing variables, non-negative
Graduates at ISCED 6 - national	check for missing variables, non-negative
Graduates at ISCED 6 - foreigner	check for missing variables, non-negative
Graduates at ISCED 6 - citizenship unclassified	check for missing variables, non-negative
Graduates at ISCED 6 - resident	check for missing variables, non-negative
Graduates at ISCED 6 - mobile	check for missing variables, non-negative
Graduates at ISCED 6 - mobility unclassified	check for missing variables, non-negative
Graduates at ISCED 6 - ISCED-F 00	check for missing variables, non-negative
Graduates at ISCED 6 - ISCED-F 01	check for missing variables, non-negative
Graduates at ISCED 6 - ISCED-F 02	check for missing variables, non-negative
Graduates at ISCED 6 - ISCED-F 03	check for missing variables, non-negative
Graduates at ISCED 6 - ISCED-F 04	check for missing variables, non-negative
Graduates at ISCED 6 - ISCED-F 05	check for missing variables, non-negative
Graduates at	check for missing variables, non-negative

ISCED 6 - ISCED-F 06	
Graduates at ISCED 6 - ISCED-F 07	check for missing variables, non-negative
Graduates at ISCED 6 - ISCED-F 08	check for missing variables, non-negative
Graduates at ISCED 6 - ISCED-F 09	check for missing variables, non-negative
Graduates at ISCED 6 - ISCED-F 10	check for missing variables, non-negative
Graduates at ISCED 6 - ISCED-F unclassified	check for missing variables, non-negative
Total graduates at ISCED 6	alarm if "Highest degree delivered" = 1 and the value in this cell = "a" OR if cell is empty OR if "Highest degree delivered" = 0 and the value in this cell is not "a" OR if sum of men, women and gender unclassified (respectively citizenship, mobility and fields of education) is not equal to total graduates in ISCED 6 OR if men and women are text and gender unclassified is not (respectively citizenship, mobility and fields of education) OR if cell of men and women is numeric and gender unclassified is not (respectively national, foreigner and citizenship unclassified) OR if value is negative OR if "Lowest degree delivered" = 1 and total = "a" OR if "Lowest degree delivered" = 2 or 3 and total is not "a" OR if breakdowns are numbers and total is text
Graduates at ISCED 7 - men	check for missing variables, non-negative
Graduates at ISCED 7 - women	check for missing variables, non-negative
Graduates at ISCED 7 - gender unclassified	check for missing variables, non-negative
Graduates at ISCED 7 - national	check for missing variables, non-negative
Graduates at ISCED 7 - foreigner	check for missing variables, non-negative
Graduates at ISCED 7 - citizenship unclassified	check for missing variables, non-negative
Graduates at ISCED 7 - resident	check for missing variables, non-negative
Graduates at ISCED 7 - mobile	check for missing variables, non-negative
Graduates at ISCED 7 - mobility unclassified	check for missing variables, non-negative
Graduates at	check for missing variables, non-negative

ISCED 7 - ISCED-F 00	
Graduates at ISCED 7 - ISCED-F 01	check for missing variables, non-negative
Graduates at ISCED 7 - ISCED-F 02	check for missing variables, non-negative
Graduates at ISCED 7 - ISCED-F 03	check for missing variables, non-negative
Graduates at ISCED 7 - ISCED-F 04	check for missing variables, non-negative
Graduates at ISCED 7 - ISCED-F 05	check for missing variables, non-negative
Graduates at ISCED 7 - ISCED-F 06	check for missing variables, non-negative
Graduates at ISCED 7 - ISCED-F 07	check for missing variables, non-negative
Graduates at ISCED 7 - ISCED-F 08	check for missing variables, non-negative
Graduates at ISCED 7 - ISCED-F 09	check for missing variables, non-negative
Graduates at ISCED 7 - ISCED-F 10	check for missing variables, non-negative
Graduates at ISCED 7 - ISCED-F unclassified	check for missing variables, non-negative
Total graduates at ISCED 7	alarm if "Highest degree delivered" = 2 and the value in this cell or in sum of ISCED 7 long degree graduates = "a" OR if cell is empty OR if "Highest degree delivered" < 2 and the value in this cell is not "a" OR if sum of men, women and gender unclassified (respectively citizenship, mobility and fields of education) is not equal to total students enrolled in ISCED 7 OR if men and women are text and gender unclassified is not (respectively citizenship, mobility and fields of education) OR if cell of men and women is numeric and gender unclassified is not (respectively national, foreigner and citizenship unclassified) OR if value is negative OR if "Lowest degree delivered" = 2 and total in this cell or in ISCED 7 long degrees is "a" OR if "Lowest degree delivered" = 3 and total is not "a" OR if breakdowns are numbers and total is text
Graduates at ISCED 7 long degree - men	check for missing variables, non-negative
Graduates at ISCED 7 long	check for missing variables, non-negative

degree - women	
Graduates ISCED 7 long degree-gender unclassified	check for missing variables, non-negative
Graduates at ISCED 7 long degree - national	check for missing variables, non-negative
Graduates at ISCED 7 long degree - foreigner	check for missing variables, non-negative
Graduates ISCED 7 long degree-nation. unclassified	check for missing variables, non-negative
Graduates at ISCED 7 long degree - resident	check for missing variables, non-negative
Graduates at ISCED 7 long degree - mobile	check for missing variables, non-negative
Graduates ISCED 7 long degree-mobility unclassified	check for missing variables, non-negative
Graduates at ISCED 7 long degree - ISCED-F 00	check for missing variables, non-negative
Graduates at ISCED 7 long degree - ISCED-F 01	check for missing variables, non-negative
Graduates at ISCED 7 long degree - ISCED-F 02	check for missing variables, non-negative
Graduates at ISCED 7 long degree - ISCED-F 03	check for missing variables, non-negative
Graduates at ISCED 7 long degree - ISCED-F 04	check for missing variables, non-negative
Graduates at ISCED 7 long degree - ISCED-F 05	check for missing variables, non-negative
Graduates at ISCED 7 long degree - ISCED-F 06	check for missing variables, non-negative
Graduates at	check for missing variables, non-negative

ISCED 7 long degree - ISCED-F 07	
Graduates at ISCED 7 long degree - ISCED-F 08	check for missing variables, non-negative
Graduates at ISCED 7 long degree - ISCED-F 09	check for missing variables, non-negative
Graduates at ISCED 7 long degree - ISCED-F 10	check for missing variables, non-negative
Graduates ISCED 7 long degree- ISCED-F unclassified	check for missing variables, non-negative
Total graduates at ISCED 7 long degree	alarm if "Highest degree delivered" = 2 and the value in this cell or in sum of ISCED 7 long degree students = "a" OR if cell is empty OR if "Highest degree delivered" < 2 and the value in this cell is not "a" OR if sum of men, women and gender unclassified (respectively citizenship, mobility and fields of education) is not equal to total students enrolled in ISCED 7 OR if men and women are text and gender unclassified is not (respectively citizenship, mobility and fields of education) OR if cell of men and women is numeric and gender unclassified is not (respectively national, foreigner and citizenship unclassified) OR if value is negative OR if "Lowest degree delivered" = 2 and total in this cell or in ISCED 7 long degrees is "a" OR if "Lowest degree delivered" = 3 and total is not "a" OR if breakdowns are numbers and total is text
Total graduates enrolled ISCED 5-7	alarm if number is not equal to sum of graduates ISCED 5, 6, 7 and 7 long degree OR if cell is empty OR if number is negative
Research active institution	yes, only 0, 1 or "m" possible
Students enrolled at ISCED 8 - men	check for missing variables, non-negative
Students enrolled at ISCED 8 - women	check for missing variables, non-negative
Students enrolled ISCED 8 - gender unclassified	check for missing variables, non-negative
Students enrolled at ISCED 8 - national	check for missing variables, non-negative
Students enrolled at ISCED 8 - foreigner	check for missing variables, non-negative
Students enrolled ISCED 8-	check for missing variables, non-negative

citizenship unclassified	
Students enrolled at ISCED 8 - resident	check for missing variables, non-negative
Students enrolled at ISCED 8 - mobile	check for missing variables, non-negative
Students enrolled ISCED 8-mobility unclassified	check for missing variables, non-negative
Students enrolled at ISCED 8 - ISCED-F 00	check for missing variables, non-negative
Students enrolled at ISCED 8 - ISCED-F 01	check for missing variables, non-negative
Students enrolled at ISCED 8 - ISCED-F 02	check for missing variables, non-negative
Students enrolled at ISCED 8 - ISCED-F 03	check for missing variables, non-negative
Students enrolled at ISCED 8 - ISCED-F 04	check for missing variables, non-negative
Students enrolled at ISCED 8 - ISCED-F 05	check for missing variables, non-negative
Students enrolled at ISCED 8 - ISCED-F 06	check for missing variables, non-negative
Students enrolled at ISCED 8 - ISCED-F 07	check for missing variables, non-negative
Students enrolled at ISCED 8 - ISCED-F 08	check for missing variables, non-negative
Students enrolled at ISCED 8 - ISCED-F 09	check for missing variables, non-negative
Students enrolled at ISCED 8 - ISCED-F 10	check for missing variables, non-negative
Students enrolled ISCED 8 - ISCED-F unclassified	check for missing variables, non-negative
Total students enrolled at ISCED 8	alarm if "Highest degree delivered" = 3 and the value in this cell = "a" OR if cell is empty OR if "Highest degree delivered" < 3 and the value in this cell is not "a" OR if sum of men, women and gender unclassified (respectively citizenship, mobility and fields of education) is not equal to total students enrolled in ISCED 8 OR if men and women are text and gender unclassified is not (respectively citizenship, mobility

	and fields of education) OR if cell of men and women is numeric and gender unclassified is not (respectively national, foreigner and citizenship unclassified) OR if value is negative OR if "Lowest degree delivered" = 3 and total = "a" OR if breakdowns are numbers and total is text
Graduates at ISCED 8 - men	check for missing variables, non-negative
Graduates at ISCED 8 - women	check for missing variables, non-negative
Graduates at ISCED 8 - gender unclassified	check for missing variables, non-negative
Graduates at ISCED 8 - national	check for missing variables, non-negative
Graduates at ISCED 8 - foreigner	check for missing variables, non-negative
Graduates at ISCED 8 - citizenship unclassified	check for missing variables, non-negative
Graduates at ISCED 8 - resident	check for missing variables, non-negative
Graduates at ISCED 8 - mobile	check for missing variables, non-negative
Graduates at ISCED 8 - mobility unclassified	check for missing variables, non-negative
Graduates at ISCED 8 - ISCED-F 00	check for missing variables, non-negative
Graduates at ISCED 8 - ISCED-F 01	check for missing variables, non-negative
Graduates at ISCED 8 - ISCED-F 02	check for missing variables, non-negative
Graduates at ISCED 8 - ISCED-F 03	check for missing variables, non-negative
Graduates at ISCED 8 - ISCED-F 04	check for missing variables, non-negative
Graduates at ISCED 8 - ISCED-F 05	check for missing variables, non-negative
Graduates at ISCED 8 - ISCED-F 06	check for missing variables, non-negative
Graduates at ISCED 8 - ISCED-F 07	check for missing variables, non-negative
Graduates at	check for missing variables, non-negative

ISCED 8 - ISCED-F 08	
Graduates at ISCED 8 - ISCED-F 09	check for missing variables, non-negative
Graduates at ISCED 8 - ISCED-F 10	check for missing variables, non-negative
Graduates at ISCED 8 - ISCED-F unclassified	check for missing variables, non-negative
Total graduates at ISCED 8	alarm if "Highest degree delivered" = 3 and the value in this cell = "a" OR if cell is empty OR if "Highest degree delivered" < 3 and the value in this cell is not "a" OR if sum of men, women and gender unclassified (respectively citizenship, mobility and fields of education) is not equal to total students enrolled in ISCED 8 OR if men and women are text and gender unclassified is not (respectively citizenship, mobility and fields of education) OR if cell of men and women is numeric and gender unclassified is not (respectively national, foreigner and citizenship unclassified) OR if value is negative OR if "Lowest degree delivered" = 3 and total = "a" OR if breakdowns are numbers and total is text
R&D Expenditure	alarm if "Research active institution" = 0 and the value in this cell is not "a" OR if "Research active institution" = 1 and the value in this cell is "a" OR if this cell is empty OR if value is negative OR if R&D Expenditure is higher than total expenditure
Indicators	alarm if cell is empty OR if value is larger than one OR if value is negative OR if unclassified category of used value is larger than 25% of the total and not flagged with "ic" (if yes, flag it and make a note) OR (only for financial indicators) if one of the values used for calculation is negative (if yes, make a note)

Table 26: List of ratios for data control with R-script

Ratio	Rule
Tot exp / Tot Revenues	alarm if value is < 0.9 or >1.2
Personnel Exp / Tot staff	alarm if value is < 50,000 or >150,000
Academic staff (HC)/ Academic staff (FTE)	alarm if value is < 0.4 or >1
Administrative staff (HC)/ Administrative staff (FTE)	alarm if value is < 0.6 or >1
R&D exp / Tot exp	alarm if value is negative or >=1

11 References

Bonaccorsi, Andrea and Cinzia Daraio. 2007. Universities and Strategic Knowledge Creation. Specialization and Performance in Europe. Cheltenham: Edwar Elgar.

Bonaccorsi, Andrea, Cinzia Daraio, Benedetto Lepori and Stig Slipersaeter 2007. Indicators on individual higher education institutions: addressing data problems and comparability issues. *Research Evaluation,* 16, 66-78.

C. Batini and M. Scannapieco. 2006. Data Quality: Concepts, Methodologies, and Techniques. Springer.

Daraio, Cinzia, Andrea Bonaccorsi, Aldo Geuna, B. Lepori and et. al. 2011. The European university landscape: a micro characterization based on evidence from the Aquameth project. *Research Policy,* 40(1), 148-164.

European Commission. 2012. She Figures 2012. Women and Science. Statistics and Indicators.

EUROSTAT. 2007. Handbook on Data Quality Assessment Methods and Tools. Luxembourg: Eurostat.

Huisman, Jeroen 2000. Higher education institutions: as different as chalk and cheese? *Higher Education Policy,* 13, 41-53.

Jongbloed, Ben 2008. Creating Public-Private Dynamics in Higher Education Funding. A Discussion of three Options. In *Public-Private Dynamics in Higher Education Funding,* ed. J. Enders and B Jongbloed, pp. 113-138. Bielefeld: Transcript.

Kyvik, Svein and B. Lepori. 2010. The Research Mission of Higher Education Institutions Outside the University Sector. Striving for Differentiation. Dordrecht: Springer.

Kyvik, Svein 2004. Structural Changes in Higher Education Systems in Western Europe. *Higher Education in Europe,* 29 (3), 393-409.

Lepori, B., Carole Probst and Lukas Baschung 2010. Patterns of subject mix of higher education institutions: a first empirical analysis from the AQUAMETH database. *Minerva,* 48(1), 73-99.

OECD. 2015. Frascati Manual 2015. Guidelines for Collecting and Reporting Data on Research and Experimental Development. Paris: OECD.

OECD. 2002. Frascati Manual. Proposed Standard Practice for Surveys on Research and Experimental Development. Paris: OECD.

Scannapieco, Monica and Tiziana Catarci 2002. Data quality under a computer science perspective. *Archivi & Computer,* 2, 1-15.

Van Vught, Frans, Jeroen Bartelse, David Bohmert, Jon File, Christiane Gaethgens, Saskia Hansen, Frans Kaiser, Rolf Peter, Sybille Reichert, Jim Taylor, Peter West and Marijk van de Wende. 2008. Mapping diversity. Developing a European classification of higher education institutions. Enschede: CHEPS Center for Higher Education Policy Studies.