Chapter 8
Data Collection

Julie Callaert, Elisabeth Epping, Gero Federkeil, Ben Jongbloed, Frans Kaiser, and Robert Tijssen

8.1 Introduction

In this chapter we will describe the data collection instruments used in the development of U-Multirank. The first section is an overview of existing databases – mainly on bibliometrics and patents. The second describes the questionnaires and survey tools used for collecting data from the institutions (the self-reported data) – at the institutional and department levels – and from students. The next chapter outlines the design of the pilot test through which the feasibility of a multidimensional global ranking was assessed and presents the major outcomes.

8.2 Databases

8.2.1 Existing Databases

One of the activities in the U-Multirank project was to review existing rankings and explore their underlying databases. If existing databases can be relied on for quantifying the U-Multirank indicators this would be helpful in reducing the overall

J. Callaert (✉)
Center for Research & Development Monitoring (ECOOM),
Catholic University of Leuven, Leuven, Belgium

E. Epping • B. Jongbloed • F. Kaiser
Center for Higher Education Policy Studies, University of Twente,
Enschede, The Netherlands

G. Federkeil
Centre for Higher Education (CHE), Gütersloh, Germany

R. Tijssen
Science and Innovation Studies, Leiden University, Leiden, The Netherlands

F.A. van Vught and F. Ziegele (eds.), *Multidimensional Ranking: The Design and Development of U-Multirank*, Higher Education Dynamics 37,
DOI 10.1007/978-94-007-3005-2_8,

burden for institutions in responding to U-Multirank data requests. However, from the overview of classifications and rankings presented in Chap. 3 it is clear that international databases holding information at institution level or at lower aggregation levels are currently available only for particular aspects of the dimensions Research and Knowledge Transfer. For other aspects and dimensions, U-Multirank needs to rely on self-reported data. Regarding research output and impact, there are worldwide databases on journal publications and citations. For knowledge transfer, the database of patents compiled by the European Patent Office is available. In the next two subsections, available bibliometric and patent databases will be discussed.

To further assess the availability of data covering individual higher education and research institutions, the results of the EUMIDA project – which seeks to develop the foundations of a coherent data infrastructure at the level of individual European higher education institutions – were also taken into account (see Sect. 8.2.4). In addition, a group of international experts were asked to give their assessment of data availability in some of the non-EU countries to be included in the pilot study.

8.2.2 Bibliometric Databases

There are a number of international databases which can serve as a source of information on the research output of a higher education and research institution (or one of its departments). An institution's quantity of research-based publications (per capita) reflects its research output and can also be seen as a measure of scientific merit or quality. In particular, if its publications are highly cited within the international scientific communities this may characterize an institution as high-impact and high-quality. The production of publications by a higher education and research institute not only reflects research activities in the sense of original scientific research, but usually also the presence of underlying capacity and capabilities for engaging in sustainable levels of scientific research.[1] The research profile of a higher education and research institution can be specified further by taking into account its engagement in various types of research collaboration. For this, one can look at joint research publications involving international, regional and private sector partners. The subset of jointly authored publications is a testimony of successful research cooperation.

Data on numbers and citations of research publications are covered relatively well in existing databases. Quantitative measurements and statistics based on information drawn from bibliographic records of publications are usually called 'bibliometric data'. These data concern the quantity of scientific publications by an author or organization and the number of citations (references) these publications

[1] This is why research publication volume is a part of the U-Map indicators that reflect the activity profile of an institution.

have received from other research publications. There is a wide range of research publications available for characterizing the research profile and research performance of an institution by means of bibliometric data: lab reports, journal articles, edited books, monographs, etc. The bibliometric methodologies applied in international comparative settings such as U-Multirank usually draw their information from publications that are released in scientific and technical journals. This part of the research literature is covered ('indexed') by a number of international databases. In most cases the journals indexed are internationally peer-reviewed, which means that they adhere to international quality standards. U-Multirank therefore makes use of international bibliometric databases to compile some of its research performance indicators and a number of research-related indicators belonging to the dimensions of Internationalization, Knowledge Transfer and Regional Engagement.

Two of the most well-known databases that are available for carrying out bibliometric analyses are the Web of Science and Scopus.[2] Both are commercial databases that provide global coverage of the research literature and both are easily accessible. The Web of Science database is maintained by ISI, the Institute for Scientific Information, which was taken over by Thomson Reuters a few years ago. The Web of Science currently covers about 1 million new research papers per year, published in over 10,000 international and regional journals and book series in the natural sciences, social sciences, and arts and humanities. According to the Web of Science website, 3,000 of these journals account for about 75% of published articles and over 90% of cited articles.[3] The Web of Science claims to cover the highest impact journals worldwide, including Open Access journals and over 110,000 conference proceedings.

The Scopus database was launched in 2004 by the publishing house Elsevier. It claims to be the largest abstract and citation database containing both peer-reviewed research literature and web sources. It contains bibliometric information covering some 17,500 peer-reviewed journals (including 1,800 Open Access journals) from more than 5,000 international publishers. Moreover it holds information from 400 trade publications and 300 book series, as well as data about conference papers from proceedings and journals.

To compile the publications-related indicators in the U-Multirank pilot study, bibliometric data was derived from the October 2010 edition of the Web of Science bibliographical database. An upgraded 'bibliometric version' of the database is housed and operated by the CWTS (one of the CHERPA Network partners) under a full license from Thomson Reuters. This dedicated version includes the 'standardized institutional names' of higher education and research institutes that have been checked ('cleaned') and harmonized in order to ensure

[2] Yet another database is Google Scholar. This is a service based on the automatic recording by Google's search engine of citations to any author's publications (of whatever type) included in other publications appearing on the worldwide web.

[3] See: http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/

that as many as possible of the Web of Science-indexed publications are assigned to the correct institution. This data processing of address information is done at the aggregate level of the entire 'main' organization (not for sub-units such as departments or faculties). All the selected institutions in the U-Multirank pilot study produced at least one Web of Science-indexed research publication during the years 1980–2010.

The Web of Science, being both an international and multidisciplinary database, has its pros and cons. The bulk of the research publications are issued in peer-reviewed international scientific and technical journals, which mainly refer to discovery-oriented 'basic' research of the kind that is conducted at universities and research institutes. There are relatively few conference proceedings in the Web of Science, and no books or monographs whatsoever; hence, publications referring to 'applied research' or 'strategic research' are underrepresented. It has a relatively poor coverage of non-English language publications. The coverage of publication output is quite good in the medical sciences, life sciences and natural sciences, but relatively poor in many of the applied sciences and social sciences and particularly within the humanities. The alternative source of bibliographical information, Elsevier's Scopus database, is likely to provide an extended coverage of the global research literature in those underrepresented fields of science.

For the following six indicators selected for inclusion in the U-Multirank pilot test data can be obtained from the CWTS/Thomson Reuters Web of Science database:

1. total publication output
2. university-industry joint publications
3. international joint publications
4. field-normalized citation rate
5. share of the world's most highly cited publications
6. regional joint publications

This indicator set includes four new performance indicators (#2, #3, #5, #6) that were specially constructed for U-Multirank and have not been used before in any international classification or ranking.

8.2.3 Patent Databases

As part of the indicators in the Knowledge Transfer dimension, we selected the number of *patent applications* for which a particular higher education and research institution acts as an applicant and (as part of that) the number of *co-patents* applied for by the institution together with a private organization.

Data for the co-patenting and patents indicators can be derived from patent databases. For U-Multirank, patent data were retrieved from the European Patent Office

(EPO). Its Worldwide Patent Statistical Database (version October 2009),[4] also known as PATSTAT, is designed and published on behalf of the OECD Taskforce on Patent Statistics. Other members of this taskforce include the World Intellectual Property Organization (WIPO), the Japanese Patent Office (JPO), the United States Patent and Trademark Office (USPTO), the US National Science Foundation (NSF), and the European Commission represented by Eurostat and by DG Research.

The PATSTAT patent database is especially designed to assist in advanced statistical analysis of patent data. It contains patent data from over 80 countries; adding up to 70 million records (63 million patent applications and 7 million granted patents). The patent data are sourced from offices worldwide, including of course the most important and largest ones such as the EPO, the USPTO, the JPO and the WIPO. Updates of PATSTAT are produced every 6 months, around April and October.

PATSTAT is a relational database: 20 related tables contain information on relevant dates (e.g. of patent filing, patent publication, granting of patent), on patent applicants and inventors, technological classifications of patents, citations from patents to other documents, family links,[5] etc. Updates of PATSTAT are produced twice a year.

8.2.4 Data Availability According to EUMIDA

Like the U-Multirank project, the EUMIDA project (see http://www.eumida.org) collects data on individual higher education and research institutions. The EUMIDA project is meant to test whether a data collection effort can be undertaken by EUROSTAT in the foreseeable future. EUMIDA covers 29 countries (the 27 EU member states plus Switzerland and Norway) and has demonstrated that a regular collection of institutional data by national statistical authorities is feasible across (almost) all EU-member states, albeit for a limited number of mostly input indicators.

The EUMIDA and U-Multirank project teams agreed to share information on issues such as definitions of data elements and data sources, given that the two projects share a great deal of data (indicators). The overlap lies mainly in the area of data related to the inputs (or activities) of higher education and research institutions. A great deal of this input-related information is used in the construction of the indicators in U-Map. The EUMIDA data elements therefore are much more similar to

[4] This version is held by the K.U. Leuven (Catholic University Leuven) and was licensed to its ECOOM unit (Expertise Centrum O&O Monitoring).

[5] A patent family is a set of patents taken in various countries to protect a single invention (when a first application in a country – the priority – is then extended to other offices). In other words, a patent family is the same invention disclosed by (a) common inventor(s) and patented in more than one country (see: US Patent and Trademark Office: www.uspto.gov).

Table 8.1 Data elements shared between EUMIDA and U-Multirank: their coverage in national databases

Dimension	EUMIDA and U-Multirank data element	European countries where data element is available in national databases
Teaching & learning	Relative rate of graduate unemployment	CZ, FI, NO, SK, ES
Research	Expenditure on research	AT*, BE, CY, CZ*, DK, EE, FI, GR*, HU, IT, LV*, LT*, LU, MT*, NO, PL*, RO*, SI*, ES, SE, CH, UK
	Research publication output	AT, BE-FL, CY, CZ, DK, FI, FR, DE, GR, HU, IE, IT, LV, LT, LU, NO, NL, PL, PT*, RO*, SK, SI, ES, SE*, CH, UK
Knowledge transfer	Number of spin-offs	BE-FL, FR*, GR, IT (p), PT (p), ES
	Third party funding	CY, CZ, DE, IT, NL, NO, PL, PT, ES, CH
	Patents	AT, BE-FL, CZ, EE*, FI, FR*, GR, HU, IE*, IT, LU, MT*, NO, NL (p), PL*, SI, ES, UK
International orientation	(No overlap between U-Multirank and EUMIDA)	
Regional engagement	(No overlap between U-Multirank and EUMIDA)	

Source: Based on EUMIDA Deliverable D2 – *Review of Relevant Studies* (dated 20 February 2010 and submitted to the Commission on 1 March 2010).
* indicates: There are confidentiality issues (e.g. national statistical offices may not be prepared to make data public without consulting individual HEIs).
(p) indicates: Data are only partially available (e.g. only for public HEIs or only for [some] research universities).
The list of EUMIDA countries with abbreviations: Austria (*AT*), Belgium (*BE*), [Belgium-Flanders community (*BE-FL*)], Bulgaria (*BG*), Cyprus (*CY*), Czech Republic (*CZ*), Denmark (*DK*), Estonia (*EE*), Finland (*FI*) France (*FR*), Germany (*DE*), Greece (*GR*), Hungary (*HU*), Ireland (*IE*), Italy (*IT*), Latvia (*LV*), Lithuania (*LV*), Luxembourg (*LU*), Malta (*MT*), Netherlands (*NL*), Norway (*NO*), Poland (*PL*), Portugal (*PT*), Romania (*RO*), Slovakia (*SK*), Slovenia (*SI*), Spain (*ES*), Sweden (*SE*), Switzerland (*CH*), United Kingdom (*UK*).

the U-Map indicators, since U-Map aims to build *activity profiles* for individual institutions whereas U-Multirank constructs *performance profiles*.

The findings of EUMIDA point to the fact that for the more research intensive higher education institutions, data for the dimensions of Education and Research are relatively well covered, although data on graduate careers and employability are sketchy. Some data on scientific publications is available for most countries. However, overall, performance-related data is less widely available compared to input-related data items. The role of national statistical institutes is quite limited here and the underlying methodology is not yet consistent enough to allow for international comparability of data.

Table 8.1 above shows the U-Multirank data elements that are covered in EUMIDA and whether information on these data elements may be found in national databases (statistical offices, ministries, rectors' associations, etc.). The table shows that

EUMIDA primarily focuses on the Teaching & Learning and Research dimensions, with some additional aspects relating to the Knowledge Transfer dimension. Since EUMIDA was never intended to cover all dimensions of an institution's activity (or its performance), it is only natural that dimensions such as International Orientation and Regional Engagement are less prominent in the project.

The table illustrates that information on only a few U-Multirank data elements is available from national databases and, moreover, what data exists is available only in a small minority of European countries. This implies, once again, that the majority of data elements will have to be collected directly from the institutions themselves.

8.3 Data Collection Instruments

Due to the lack of adequate data sets, the U-Multirank project had to rely largely on self-reported data (both at the institutional and field-based levels), collected directly from the higher education and research institutions. The main instruments to collect data from the institutions were four online questionnaires: three for the institutions and one for students.

The four surveys are:

- U-Map questionnaire
- institutional questionnaire
- field-based questionnaire
- student survey.

The U-Map questionnaire had already been tested and fully documented in its design phase. The remaining three surveys were designed, pre-tested, modified where necessary and a full set of supporting instruments (data-collection protocols, glossaries, FAQ, help desk) were developed for their use in the pilot study.

8.3.1 U-Map Questionnaire

As explained earlier, the U-Map questionnaire is an instrument for identifying similar subsets of higher education institutions within the U-Multirank sample. Data is collected in seven main categories:

- general information: name and contact person; public/private character and age of institution;
- students: numbers; modes of study and age; international students; students from region;
- graduates: by level of program; subjects; orientation of degrees; graduates working in region;
- staff data: fte and headcount; international staff;

- income: total income; income by type of activity; by source of income;
- expenditure: total expenditure; by cost centre; use of full cost accounting;
- research and knowledge exchange: publications; patents; concerts and exhibitions; start-ups.

8.3.2 Institutional Questionnaire

The institutional questionnaire collects data on the performance of the institution. The questionnaire is divided into the following categories:

- general information: name and contact; public/private character and age of institution; university hospital
- students: enrolment
- program information: bachelor/master's programs offered; CPD courses
- graduates: graduation rates; graduate employment
- staff: fte and headcount; international staff; technology transfer office staff
- income: total; income from teaching; income from research; income from other activities
- expenditure: total expenditure; by cost centre; coverage
- research and knowledge transfer: publications; patents; concerts and exhibitions; start-ups.

8.3.3 Field-Based Questionnaire

The field-based questionnaire includes information on individual faculties/departments and their programs in the pilot fields of business studies, mechanical engineering and electrical engineering.

The following categories are distinguished:

- overview: name and address of unit responsible for organizing the field; contact person
- staff & PhD: academic staff; number of professors; international visiting/guest professors; professors offering lectures abroad; professors with work experience abroad; number of PhDs; number of post-docs
- funding: external research funds; license agreements/income; joint R&D projects with local enterprises
- students: total number (female, international degree and exchange students); internships secured; degree theses in cooperation with local enterprises
- regional engagement: continuing education programs/professional development programs; summer schools/courses for secondary students
- description: accreditation of department; learning & teaching profile; research profile.

A second part of the questionnaire asks for details of the *individual study programs* to be included in the ranking. In particular the following information was collected:

- basic information about the program (e.g. degree, length); interdisciplinary characteristics; full-time/part-time;
- number of students enrolled in the program; number of study places and level of tuition fees; periods of work experience integrated in program; international orientation; joint study program;
- credits earned for achievements abroad; number of exchange students from abroad; courses held in foreign language; special features;
- number of graduates; information about labor market entry.

8.3.4 Student Survey

The main instrument for measuring student satisfaction is an online survey. The student questionnaire uses a combination of open questions and predefined answers. Its main focus is on the assessment of the teaching and learning experience and on the facilities of the institution (see Table 7.1 in the previous chapter for more detailed information).