Introduction

Andrea Bonaccorsi

In a paper published almost two decades ago in the *American Economic Review*, Zvi Griliches, the father of econometrics of research, innovation and productivity complained about the lack of statistical data on the most interesting aspects of the economy (Griliches, 1994). His famous opening remark as President of the American Economic Association in 1993 was: 'our understanding of what is happening in our economy (and in the world economy) is constrained by the extent and quality of the available data' (ibid., p. 2). After examining some unresolved empirical puzzles, he asked why statistical agencies and government offices do not collect relevant data. Among other factors, Griliches interestingly noted:

> We ourselves do not put enough emphasis on the value of data and data collection in our training of graduate students and in the reward structure of our profession. It is the preparation skill of the econometric chef that catches the professional eye, not the quality of the raw materials in the meal, or the effort that went into procuring them. (Ibid., p. 14)

And after a discussion of the limitations of official data he offered the following recommendation:

> We need also to make observation, data collection, and data analysis a more central component of our graduate teaching . . . We also need to teach them to go out and collect their own data on interesting aspects of the economy and to rely less on 'given' data from distant agencies. (Ibid., p. 15)

Griliches' recommendations have been followed by an army of talented researchers who have 'gone out' all over the world, and to a certain extent also by statistical offices and international organizations. We now have much better and comparable data on firm R&D activities and expenditures, innovation activities and expenditures, patent extension, quality and litigation, output in services, intangible investment and other areas relevant to the analysis of productivity and growth. In addition, statistical offices have started to make available to researchers the raw data collected at the level of individual entities, such as firms, in the form of anonymized

microdata. This has opened an exciting stream of research, in which, among others, the best graduate students of Griliches (and their students as well) have produced great advancements in recent years.

There is an area, however, where microdata are surprisingly missing. Despite the large emphasis on the knowledge economy and society, there are actors that produce and diffuse knowledge that are invisible in official statistics – namely, universities. Irrespective of their size or importance, universities still suffer from a severe data constraint.

Only very aggregated data on universities are available at national and international level (UNESCO, OECD and Eurostat), without any breakdown. In addition, data on education and students are collected under the framework of educational statistics, while data on research are the domain of R&D statistics. They follow different statistical standards and manuals, and are typically organized in separate departments at statistical offices. Data on education follow the notion that students can only be enrolled into one programme at a time, so that the collection of data is organized around programmes, and not the institutions delivering the programme. Data on research, on the other hand, follow the notion that researchers at higher education institutions allocate part of their time budget to research and part to teaching. Consequently, the aggregate expenditure can be split using an average formula, while the expenditure at the level of individual institutions cannot be split reliably by differentiating across types of institutions.

The result is that it is impossible to reconcile information on the two main missions of universities – education and research – allocating the corresponding activities to a recognizable statistical entity. Even basic data such as the number of academic staff and the number of students are not available at the micro-level. This limitation is particularly severe in Europe, due to the lack of a common framework and the institutional fragmentation of national higher education systems, which started to harmonize their educational structures only after the Bologna Declaration.

Moving from this consideration, in 2004 I submitted a proposal to a newly created European research scheme, called PRIME Network of Excellence, of which my university was a member. The idea, developed jointly with Cinzia Daraio, then a postgraduate student, was to test whether it could be possible to integrate administrative microdata from ministries or government agencies into a coherent and comparable framework. Several researchers from six European countries agreed to join the project, labelled AQUAMETH (Advanced Quantitative Methods for the Evaluation of the Performance of Public Sector Research). At a later stage, three other countries joined the project, then relabelled AQUAMETH 2. At that point, the idea was quite pioneering, somewhat similar to the recommendation

by Griliches that graduate students should 'go out and collect their own data'. Nevertheless, the idea of integrating several national administrative datasets seemed audacious. After the project was completed, I was told that the reports from external referees were extremely sceptical about the feasibility of the idea, given the fragmentation of administrative situations across European countries. With the benefit of hindsight, being awarded a grant for this project was only possible because of a certain propensity to risk, for which we must praise the PRIME board.

As a matter of fact, we succeeded in a collective effort to create the first integrated dataset of microdata on universities in nine European countries, which was the basis for a book in 2007 (Bonaccorsi and Daraio, 2007) and later on for a *Research Policy* article in 2011 (Daraio and Bonaccorsi et al., 2011). Incidentally, we believe the latter, with 26 co-authors, is one of the papers with the largest number of authors ever published in social sciences.

At this point we could simply stop and let the things go, but we then conceived a new idea. Given that the integration of administrative data proved feasible, why not try to address the statistical offices directly? There was clearly a need to move from a pioneering stage to an institutional stage, in which microdata could be made available officially. The notion of a census gradually took shape. The basic idea was that universities are publicly relevant institutions, for which a set of basic information should be available at microdata level, without violating statistical secrecy. We started to circulate this idea in conferences and meetings.

Again, some scepticism on the value of addressing statistical offices worked against us. There was (and still is) a tradition of studies in higher education in which the statistical apparatus was kept to the minimum level. Descriptive and historical analyses seemed to be entirely appropriate to the field. There was also a good tradition of comparative studies in Europe in which national experts built narratives on the evolution of higher education and of national policies, on top of which highly influential interpretive frameworks were built. In these communities comparative quantitative analyses were simply considered not credible.

Another difficulty was that roughly in the same period the European Commission had started the idea of a multidimensional mapping and ranking of universities, based on rich survey data collected directly at university level. The official data provided by statistical offices seemed relatively poor compared with survey data.

To these arguments we opposed two simple points. First, there was clearly a need for establishing a census of higher education institutions in Europe. A simple list with basic indicators was extremely valuable for this purpose. Without such a list any effort to collect questionnaire data was meaningless, because whatsoever statistical representativeness could not

be estimated. Thus, our idea to establish a census was to be considered complementary to the effort to build up multidimensional exercises based on survey data. Second, the advantage of building a census with basic indicators is that it becomes possible to integrate the data with other sources of data, at institutional or geographic level. Having the names of the institutions made it possible to integrate official data with data on scientific publications, citations, patents, webometrics, or participation in framework programmes. Having the names and location of institutions made it possible to integrate official data with NUTS 2 and NUTS 3 covariates of all kinds, from economic and social data, to entrepreneurship, patenting or innovation data at regional and, when available, local level. Thus, our idea was very simple: build up the statistical infrastructure and you will have a number of spillovers.

But, again, ideas eventually find their own way. In 2008, two Directorates of the European Commission (DG EAC and DG RTD, that is, education and research) joined their effort, together with Eurostat, to launch a feasibility study to explore the notion of a European register of higher education institutions. Apparently the idea of a census had gained legitimation. The perimeter of the data collection was much larger than universities, since it included a significant share of tertiary education institutions, or institutions delivering only diploma or bachelor degrees. In a few months, we were able to build up the conceptual framework, organize a core group of universities leading the project and collect a crew of experts in 29 countries (EU-27 plus Norway and Switzerland).

The fact that you are reading this book, and perhaps asking yourself where the saga will end, is demonstration that we were awarded the contract for the feasibility study. The study was labelled EUMIDA and was completed in 18 months, leading to a Final Report published in 2010 on the website of DG Research (EUMIDA, 2010). It included a core set of data, called Data Collection 1, referring to 2457 institutions. Of these, 1364 were identified, after an innovative statistical procedure, as research active and were the object of a second data collection (Data Collection 2) with a much larger set of indicators. Data were provided by a network of national correspondents established by the EUMIDA Consortium. The correspondents cooperated with national representatives in the National Statistical Offices and/or Ministries of Education. However, the data have not been formally approved by these bodies.

One year after the Final Report, the European Commission also published the data included in Data Collection 1 on its website. Data Collection 2 was not considered sufficiently mature to be published. We were permitted to use both Data Collections exclusively for research purposes and without using the names of the institutions, unless for illustra-

tive purposes. The preliminary results of these studies are presented in this book. Since the readers may be interested in specific topics and use chapters separately, we have left short descriptions of data in each of them, allowing for a certain duplication of content. Also, when using data from the extended dataset from Data Collection 2 the authors may have worked on samples of different size. In fact, while the basic set of data (DC 1) is complete across all countries, there are still missing data in the extended dataset (DC 2), which are commented upon in the chapters.

The EUMIDA project itself has been a major scientific and political experience. Issues of classification and definition have been prominent since the beginning. While we started with the assumption that these issues could be settled in a professional way, following purely academic standards, we were eventually forced to accept the notion that statistical definitions are politically sensitive and are deeply embedded in the institutional texture. To give an example, after months of discussion, the official authorities in France preferred to leave their section of data blank, as will become apparent in the following chapters, rather than being obliged to draw (highly sensitive) statistical boundaries between higher education institutions and public research organizations. This is in itself an indication of how apparently technical issues have a political dimension. We have tried to rationalize this experience in a reflexive paper, co-authored by Benedetto Lepori, in *Minerva* (Lepori and Bonaccorsi, 2013).

After the delivery of the Final Report, there was a long period before the European Commission presented and then implemented its plans for a follow-up. This was due partly to a lack of resources and the need to ensure Eurostat's participation in the process. The European Commission will, however, in cooperation with Eurostat, improve data on European higher education learning mobility and employment outcomes, and work towards a European Tertiary Education Register (EC, 2011, p. 11).

At the time of writing this Introduction (March 2013) a new tender had been issued by Eurostat and the European Commission, with the goal of publishing the census and extending it to accession countries and to a few non-EU countries. This is at least a positive development.

Meanwhile, we collected a number of studies that show the potential of integrating official EUMIDA data with other sources of data. The book is divided in three sections. In Part I, issues of structure and governance of European higher education institutions are addressed. In Chapter 1, Andreas Niederl and co-authors offer a broad picture of the European higher education landscape. They offer detailed empirical evidence for the existence of a large non-university sector, particularly in dual systems, which, however, is able to attract only a marginal share of students. It seems that a number of students, whose educational needs are likely to be satisfied

by institutions of vocational training, still prefer to undertake a seemingly prestigious university curriculum. At the same time they show the existence of an emerging non-university research sector, formed by institutions that, while not being allowed to deliver the doctoral degree, still claim for themselves a role in research production. This confirms the well-known prediction of the higher education literature on the existence of the so-called 'academic drift'. In terms of future evolution of the European higher education landscape, however, it can be asked whether this drift may help to give prestige and legitimation to vocational training institutions, making them more able to attract large number of students, or rather is a subtle way to satisfy the need of teaching staff to be recognized academically.

This dilemma is also addressed by Torben Schubert and co-authors in Chapter 2. They use advanced clustering techniques to identify structural models of higher education. There are two interesting findings here. First, there is a robust demarcation between the university and the college models across most European countries. There is no evidence of a reduction in institutional barriers among the two types (as happened in the UK), although there is also no evidence that countries with a unitary structure (Italy and Spain in particular) will converge towards the dual model. Non-university institutions are smaller, more specialized, less internationalized, and have a concentration of private initiatives. Second, the authors fail to identify a structural model of 'research university'. When looking for a cluster of universities that are not legally or institutionally different from other universities, but have indicators of research and teaching activity that are structurally different, they did not find evidence for it. This is a warning message, to be discussed widely.

Finally, the issue of private initiative is at the core of Chapter 3 by Pedro Teixeira, Vera Rocha, Ricardo Biscaia and Margarida Cardoso. They show evidence of the growth of a private higher education sector, particularly in countries that achieved the massification at a later stage, or entered into a political transition as in Eastern Europe. The private sector identified a need for mass education in professional fields, some of which were suddenly opened by the introduction of market institutions after the fall of the Berlin Wall. While this role can be beneficial, there are issues of accreditation and quality that must be seriously addressed. A major result of the analysis is that the private sector is much less diversified than the public sector in terms of the span of fields of education offered. It does not seem to fill many niches of educational opportunities. In addition, the private sector in Europe, somewhat differently from what happened in the past in the United States and is happening currently in Asia, does not seem to invest in research. Summing up, the first part of the book opens a number of critical issues for the future of European higher education.

Part II is dedicated to the missions of universities, research, education and the third mission. The interest here is on the output of universities, in particular their scientific publications, patenting, start-up companies, or the internationalization of human capital. In Chapter 4, Ulrich Schmoch carries out a bibliometric analysis of the published output of universities, pointing to the need to select carefully the sources of data. Since the two most used bibliometric databases (ISI Thomson Web of Science and Scopus) treat scientific disciplines with different degrees of coverage, the choice of a particular source may produce quite different representations of the strength of universities, depending on the subject mix. The creation of a census of higher education institutions will open the way for a number of studies that integrate bibliometric information, disaggregated by scientific field, with structural indicators.

Chapter 5 exploits another direction for the use of EUMIDA data, the breakdown of undergraduate and graduate students and of academic staff by nationality. This places a number of delicate definitional issues, carefully discussed by Marco Seeber and Benedetto Lepori. By calculating the share of international students and staff at university level and by controlling for a number of geographic and institutional factors at country level it becomes possible to address another challenging empirical issue, that is, the relation between internationalization and performance of universities. While the data do not allow for the analysis of the quality of education, for example in terms of learning outcomes or employability, the internationalization of both students and staff offers an interesting window on the educational activity of universities in a connected world.

In Chapter 6, Attila Varga and Márton Horváth examine another output of universities, traditionally labelled under the Third Mission, namely academic patents. A large literature has addressed the issue of the impact of university research on valorization activities, such as patenting, licensing, or creation of spinoff companies. Most studies use national samples, while the chapter exploits the cross-country nature of the dataset. The chapter is also an important example of the potential of EUMIDA data for addressing a number of hot issues in economic geography and regional growth. The authors have undertaken a valuable work of georeferentiation of university data. The geographic unit is not the traditional regional level (NUTS 2), but the NUTS 3 level, corresponding to small regions, provinces or municipalities. This is a welcome novelty in the literature, since it is largely recognized that the large regional scale does not capture the spillover effects of public research that take place at the level of cities and metropolitan areas.

A similar approach has been taken by Massimo Colombo, Massimiliano Guerini, Cristina Rossi Lamastra and myself in Chapter 7. The authors

integrate the EUMIDA dataset for Italy with a rich array of variables at province level in an effort to measure the impact of university research on the creation of new firms. The findings have interesting policy implications. The two chapters together, on patenting and on entrepreneurship, are a good example of the potential for integrating census data with spatial indicators. There is a clear need for the large literature on third mission for reaching a stage of scientific maturity, in which the test of hypotheses is carried out not on small samples (of unknown statistical representativeness) or case studies, but on a large set of comparable data across countries.

Part III is dedicated to issues of efficiency and productivity, or the relation between inputs and outputs. In Chapter 8, Tasso Brandt and Torben Schubert raise an interesting and general question – why do we see universities organized in similar ways despite large institutional differences across countries? In a nutshell, universities are loosely connected federations of research units, typically organized in departments. Research units enjoy large operational autonomy, while the university level has authority on administrative and financial decisions. The authors offer a compelling argument and test the existence of economies or diseconomies of scale at two different levels – the research unit and the university. In this case the EUMIDA data have been integrated with data at a lower level of resolution, collected in a German study on research units in four scientific fields.

Chapters 9 and 10 address the issue of efficiency of European universities from two diverse yet complementary methodological perspectives. In Chapter 9, Zara Daghbashyan, Enrico Deiaco and Maureen McKelvey use the notion of cost efficiency and adopt a stochastic frontier approach, while in Chapter 10 Cinzia Daraio, Léopold Simar and myself use the notion of technical efficiency and introduce a new non-parametric technique for the estimation of directional distance. While using different definitions and estimation methodologies, the chapters argue that European universities have somewhat low cost efficiency and produce educational outputs at a lower level than it would be possible keeping the research output constant. Coming from such different empirical strategies and from cross-country data, these findings call for attention in policy circles. This is another example of the potential of EUMIDA data: by using observations from several countries it becomes possible to reach a higher level of generality than achieved so far in the large but still somewhat inconclusive literature on efficiency in higher education.

As I end the Introduction, it is still unclear whether there will finally be a day in which microdata on European universities will be officially available. It is our hope. Good data are not only the necessary meal for good economic analysis, as Zvi Griliches stated. They are also a necessary ingredient for good decisions in democratic societies.

ACKNOWLEDGEMENTS

This book was possible due to the collaboration of many people. Francesco Molinari provided excellent support in the creation of the EUMIDA Consortium and in the management of the project. Peter Whitten at DG RTD and Lene Mejer at Eurostat have been challenging the Consortium for months on many substantive issues of definition and coverage of data. It is nice to work with people deeply interested in the results of a study.

For a year and half Tasso Brandt, Daniela De Filippo, Benedetto Lepori, Andreas Niederl, Ulrich Schmoch, Torben Schubert, Stig Slipersaeter, Francesco Molinari and myself have been working hard to meet the challenging deadlines of the project, as well as to deliver in some cases national data. When we delivered the Final Report we were told that no one had made a bet we could eventually meet the deadlines. The work could not have been done without a group of dedicated experts in all European Union countries, as well as Switzerland and Norway. They accepted our continuing pressure and managed the relations with national statistical authorities with high professionalism. Their names and countries are as follows: Michael Ploder (AT), Michele Cincera, Reinhilde Veugelers (BE), Alexey Pamporov (BG), Josef Benes, Helena Sebkova, Karel Sima (CZ), Achilleas Mitsos (CY, EL), Inna Haller, Thomas Stenkhen (DE), Peter Lotz (DK), Jaan Masso (EE), Adela Garcia Aracil (ES), Tarmo Juhani Räty, Eija Paakko (FI), Patrick Llerena, Laurent Bach, Mireille Matt (FR), Laszlo Conka, Annamaria Inzelt (HU), Cinzia Daraio, Emanuela Reale, Alessandro Daraio (IT), Linas Eriksonas, Juras Ulbikas (LT), Anda Adamsone-Fiskovica (LV), Jonathan C. Borg (MT), Ben Jongbloed (NL), Krzysztof Leja (PL), Pedro Teixeira (PT), Amza Catalin Gheorghe (RO), Enrico Deiaco, Maureen McKelvey (SE), Anton Lavrin, Lubomira Srhankova (SK), Franc Mali (SI), Aldo Geuna and Federica Rossi (UK, IE).

After the publication of Data Collection 1 we decided to invite a few colleagues external to the Consortium to join us in writing empirical papers based on the data. We thank them for accepting the invitation and contributing to the volume.

REFERENCES

Bonaccorsi, A. and C. Daraio (eds) (2007), *Universities and Strategic Knowledge Creation. Specialization and Performance in Europe*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publisher, Cheltenham, PRIME Series on Research and Innovation Policy in Europe.

Daraio, C. and A. Bonaccorsi et al. (2011), 'The European university landscape: a micro characterization based on evidence from the AQUAMETH project', *Research Policy*, **40**(1), 148–64.

EUMIDA (2010), *Feasibility Study for Creating a European University Data Collection, Final Study Report*, European Commission, EUMIDA Consortium, last accessed 22 September 2013 at http://ec.europa.eu/research/era/docs/en/eumida-final-report.pdf.

European Commission (2011), *Supporting Growth and Jobs – An Agenda for the Modernisation of Europe's Higher Education Systems*, COM(2011) 567 Final, Brussels: EC.

Griliches, Z. (1994), 'Productivity, R&D, and the data constraint', *American Economic Review*, **84**(1), 1–23.

Lepori, B. and A. Bonaccorsi (2013), 'The socio-political construction of a census of higher education institutions: design, methodological and comparability issues', *Minerva*, **51**(3), 271–93.