Research Evaluation, 20(1), March 2011, pages 47–60
DOI: 10.3152/095820211X12941371876300; http://www.ingentaconnect.com/content/beech/rev

New uses of the institutional databases of universities: indicators of research activity

Emanuela Reale, Daniela Alejandra De Filippo, Isabel Gómez, Benedetto Lepori, Bianca Potì, Emilia Primeri, Carole Probst and Elias Sanz Casado

This paper explores the characteristics of the institutional databases of six universities in three European countries (Italy, Spain and Switzerland). Its aim is to test the extent to which databases can be considered valuable sources for building positioning indicators to describe different profiles of university research activities, rather than their competitive position along the single dimensions of scientific production and academic reputation. Project results support the evidence that institutional databases are social constructs, able to show a representation of the research performance of the universities, which is strongly affected by the interests of the different communities, influencing their development and evolution. Databases can also be valuable sources, when used in combination with international ones and with other information sources, to put together a broad picture of academic institutions and their scientific efforts.

PUBLICATIONS CONSTITUTE one of the major means for transmitting research results of higher education institutions, alongside formal teaching, direct (largely tacit) transmission of knowledge, mobility of people, and technological outputs. Since the 1960s there has been a considerable growth in bibliometric studies using publications to assess research performance of universities, to evaluate their productivity and to produce international bibliometric rankings (Van Raan, 2004, 2008; Bonaccorsi *et al*, 2007; Moed *et al*, 2004). These studies have been mainly based on databases collecting articles published in journals whose reputation and diffusion are generally internationally recognized, like the Web of Science and, more recently, the Elsevier database Scopus. Alongside internationally recognized databases, other sources for documented knowledge and knowledge dissemination can be identified, like CV databases, patents databases (used to measure university–industry collaborations), web-based publications databases, library catalogues like the Online Public Access Catalogue (OPAC), and other open archive systems or subject-specific databases (Van Raan, 2004; Torres-Salinas and Moed, 2009).

This paper explores the possibility of using institutional databases (i.e. databases that are created and maintained within higher education institutions), to build simple indicators on research activities of academic institutions, consistently with the positioning indicators rationale (Lepori *et al*, 2008). For this purpose, we use the results of an exploratory project, funded by the Network of Excellence PRIME within the EU FP6, which developed a survey on university

Emanuela Reale (contact author), Bianca Potì and Emilia Primeri are at Ceris CNR, Via dei Taurini, 19, 00185 Rome, Italy; Email: e.reale@ceris.cnr.it; b.poti@ceris.cnr.it; e.primeri@ceris.cnr.it; Tel: +39 6 49937853. Daniela Alejandra De Filippo and Isabel Gomez are at IEDCyT-CSIC (Instituto de Estudios Documentales sobre Ciencia y Tecnología- Consejo Superior de Investigaciones Científicas), Albasanz, 26–28, 28037 Madrid, Spain; Email: daniela.defilippo@cchs.csic.es; isobel.gomez@cchs.csic.es. Benedetto Lepori and Carole Probst are at the Centre for Organisational Research, Università della Svizzera italiana, via Lambertenghi 10a, CH-6904 Lugano, Switzerland; Email: benedetto.lepori@usi.ch; carole.probst@usi.ch. Elias Sanz Casado is at the Department of Library Science and Documentation, University Carlos III of Madrid, Calle Madrid 126-128, 28902 Getafe, Spain; Email: elias@bib.uc3m.es.

0958-2029/11/01047-14 US$12.00 © Beech Tree Publishing 2011

publications databases in Italy, Spain and Switzerland and analyzed the databases of six universities from these countries (Reale *et al*, 2009). Project results show strengths and weaknesses of the institutional databases. They support the evidence that institutional databases are social constructs, influenced by the interests of the different communities contributing to their construction; databases are also strongly shaped by the aim they want to pursue.

Section 1 of the paper introduces the theoretical framework. In Section 2, the characteristics of the publications datasets in the countries considered are discussed by looking at the results of a specific survey. Section 3 addresses the main methodological issues linked to the use of institutional databases to build research indicators. The content and aim of the databases are outlined by means of descriptors and indicators developed using institutional databases in six case studies. Finally, discussion, concluding remarks and suggestions are presented.

1. Theoretical framework

International publications databases allow for a rather fine-grained analysis of the production and productivity of countries, institutions and research groups and, thanks to very accurate citations indications, they are powerful instruments for mapping science, its evolution and dynamics (Van den Besselaar *et al*, 2007). Nevertheless, some methodological and practical limitations are discussed by the literature.

The first issue is that of limited coverage of non-journal publications, such as books, reports, proceedings, and publications published in national languages other than English. These outputs are likely to be extremely important at the national and regional level, for example in the interaction between academics and public authorities or local/regional small and medium-sized enterprises (Gómez *et al*, 2007, 2009; Hicks, 2004; Nederhof, 2006), and within less English-oriented scientific communities. Consequently, there is limited coverage in the Web of Science or SCOPUS of publications in sociology, political science, and generally in the humanities, but also in more science-related fields like engineering (Torres-Salinas and Moed, 2007; Hicks, 2004; Nederhof, 2006; Norris and Oppenheim, 2007; Iribarren-Maestro *et al*, 2009), and the reliability and validity of the bibliometric methodology is also affected.

The same holds true for strong interdisciplinary and applied research products, for which Web of Science data have to be integrated with more policy-oriented documents and reports (Merkx and Van den Besselaar, 2008). It should also be noted that citation behavior varies not only for what concerns the type of literature to be cited, but also in its meaning: in some fields, a substantive part of citations regards work that is criticized in the text, while other fields just omit the works they do not like. Lastly, we should also consider the difficulties in accessing and handling the data collected by international databases.

So far, international databases have provided a way to evaluate university research performance and academic reputation, mainly based on input/output indicators, but they have not been able to fully describe some patterns of academic knowledge, such as the more or less pronounced regional/national or international scientific research orientation of some domains, or the different types of public the scientific production is addressed to (Van Raan, 2001).

Other sources — such as researchers' CV databases, library catalogues like OPAC, and other public databases, like Google Scholar, open archive systems or subject-specific databases, like the ArXiv repository — are used experimentally to characterize academics' publication outputs (Lepori *et al*, 2008; Torres-Salinas and Moed, 2009).

CV-based analyses and surveys might be time-consuming and yield not easily comparable results (Bonaccorsi *et al*, 2007), yet they have proven to be useful techniques. As an example, they have been used for interesting exploratory studies aimed at outlining and tracking researchers' mobility and professional career trajectories, in a comparative perspective, providing further insights into S&T human resources and research systems (Cañibano *et al*, 2008; Lepori and Probst, 2009; De Filippo *et al*, 2009).

Web based databases (like Google Scholar, launched in 2004) often provide wider coverage of multidisciplinary content, publications and citations; however, the usability of Google Scholar for bibliometric analyses is often contested, as the procedures to include publications and citations are not public and the quality of data seems rather poor (Walters, 2007; Jacsó, 2008).

More recently, databases internal to higher education institutions have been analyzed as important means for representing the huge variety of university research outputs, since they include all the components of the research process and outputs (i.e. articles, books, reports, conferences papers, PhD theses), and accurately describe the diversity of research communities within academic institutions, their behaviors and practices (White, 2007; Van der Graaf and Van Eijndhoven, 2008). Some authors also highlight their value in supporting research assessment by providing access to the different outcomes produced by the academic institutions (Day, 2004; Harnard, 2001). Institutional databases can cover results that go beyond traditional publications, providing empirical evidence of the university's accountability toward society, thus its accomplishment of the so-called third mission. Databases and software, for example, which are midway between publications and technological outputs, or articles in newspaper, exhibits, working papers, notes, letters, manuals, reports, all show the commitment of the scientific academic community toward society.

Therefore, although it is argued that developing these publications databases could cause an overload of mediocre material and 'gray literature', institutional databases can also provide new means to access documents and improve the openness of research results, disclosing some of the aforementioned hidden patterns of academic knowledge (Van Raan, 2001).

In this paper we want to use the rationale of positioning indicators (Lepori *et al*, 2008), which describe the position of an academic institution within its complex, fragmented and multidimensional institutional space, its competitive behavior, and its collaborations and links with other actors, considered as relevant performance elements. These issues are particularly important when analyzing the publication activities of regionally oriented universities, of institutions in the non-university sector, or when assessing the publication activities of generalist universities (Bonaccorsi *et al*, 2007; Gómez *et al*, 2007) which cover all scientific domains.

Following the positioning indicators rationale, we aim at assessing the extent to which institutional databases can be used to produce simple research indicators at the level of whole higher education institutions and of different scientific fields within the institutions. We intend to explore the ability of institutional databases to become useful tools in representing the research activities of universities, by proving a broader view than that supplied by other sources of information, thus complementing existing publications databases. Publications show the diversities within the performance of universities, thus reflecting important aspects of their positioning and research specialization. A more comprehensive overview of the activities developed by the scholarly community within the institutions would be available to policy-makers and university managers by looking at the whole range of publication outputs.

At the same time, despite the differences across individual CVs and publications stored, we expect to find that databases are social constructs, revealing the interests of the different communities inside and outside the university (scholars, managers, funding agencies, developers, stakeholders). The coverage within each database is supposed to be strongly affected by the aim for which it was set up and by the way in which that aim was perceived and interpreted by the aforementioned communities within the academia.

A social construct is a concept or a practice created by a particular group. It is related to the ways in which individuals and groups participate in the creation of their perceived social status, which reflects the state of things, as they actually exist. A social construct involves looking at how social phenomena are built, institutionalized, and become part of the people's traditions. A socially constructed reality is seen as an ongoing, dynamic process that is reproduced by people acting to interpret and understand it, providing meaning and knowledge of it. In S&T studies, researches on social construction (Knorr Cetina, 1997; Latour and Woolgar, 1979; Searle, 1995; Mazzotti, 2008) have tried to demonstrate that what science has typically described as objective fact is instead related to the processes of social construction; thus, all the facts that we assume to be objective are largely shaped by human perspectives or opinions, particular feelings, beliefs, and desires, in much the same way as facts shape our ideas and beliefs.

When we refer to university publications databases as social constructs, we intend to analyze them by means of two propositions (Hacking, 1999):

- Databases are not objects taken for granted, nor are they inevitable;
- The structure of databases need not be as it is; rather it can be radically different or it can be transformed according to the process of social construction applied.

Coherently with the quoted literature, if university publications databases are social constructs, we can expect impacts on:

- The extent of their coverage, which will be limited to those publications seen as relevant by the different epistemic communities;
- The use of categories, which can be affected by the divergent views of scholars and librarians;
- The practices and incentives for updating, and the quality of the information stored;
- The organization of the databases (management, access, users, use, accessibility), which can vary on the basis of the purpose intended by the university managers.

Moreover, the databases do not allow us to compare the results from the various universities involved, but they make it possible to look at the different values and beliefs that underlie the specific social construction. Thus, the social construction of databases within the academic institutions shapes their basic characteristics; the evolution of the process can influence the way in which the different actors perceive the purpose and the added value of the databases.

2. Survey on diffusion, aims and uses of institutional databases

A survey was developed to answer the following questions:

- What is the situation of university publications databases in the selected countries?
- Where does their creation originate from and what aims do they address?
- What about completeness and use of these tools?
- Databases are constructed ad hoc, in a local

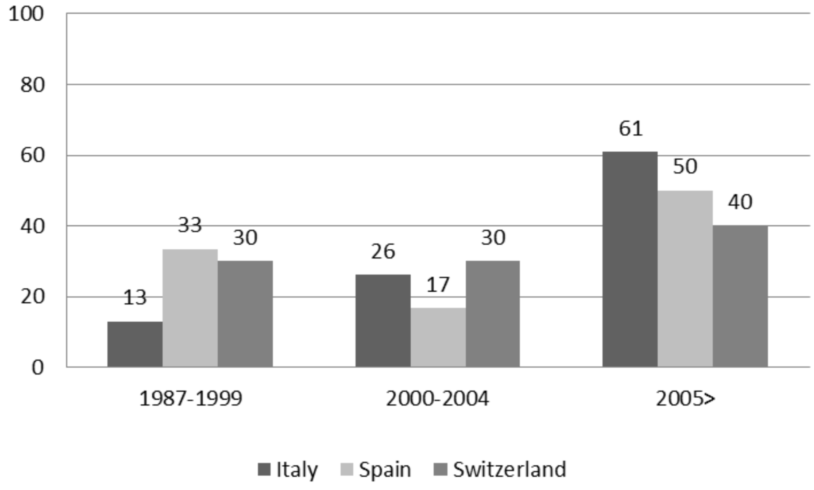

Figure 1. Year of construction of universities' institutional databases (percentage values per country/year)

community, based on the needs and purposes of the actors involved. Thus, which actors are involved?

- Which purposes are followed?
- How is this visible in the databases (e.g. through the selection of fields to be included)?

Our survey draws an interesting picture of institutional databases, addressing key issues such as their diffusion, publication years covered, initiative and purpose of their construction, main users and possible uses of database information, their accessibility and main shortcomings.

The sample of surveyed universities was selected according to the following criteria. They had to be public universities, polytechnics or technical universities, both generalist and specialized universities, not open universities, and representative, in size and location, of the national university system. In total, 57 universities answered the questionnaire: 12 universities from Switzerland, 26 from Italy, and 19 from Spain.[1]

The sample is thus fully representative for Switzerland (all Swiss universities are included), quite representative for Italy (some generalist public universities and a polytechnic, of different sizes and geographical location and with different types of institutional databases are included), but less representative for Spain where an important region, Andalucía, is not represented. Nevertheless, the collected data cover generalist, specialized and polytechnic universities from different regions.

The survey highlights that the setting-up of an institutional database is a very common and rather recent practice (Figure 1). Eighteen out of the 19 Spanish universities that answered the questionnaire (94.7%) have a database of scientific publications, as do most Italian (84.6%) and Swiss (83.3%) universities, and the databases were mainly implemented after 2005. The universities which do not have an institutional database, mostly because of administrative problems, are planning to build one. Also the coverage timeframe seems to be related to the year of creation: coverage is highest for recent years, from 2004 to 2008.

The initiative for institutional databases' construction is largely internal, often based on decisions taken within each university, and never related to government or regional initiatives. Although arising from internal initiatives, it is clear that the creation of databases is mainly due to the growing demand for information on output production by the government and funding agencies, in order to have the data needed for *ex-ante* and *ex-post* assessments. In fact, different aims prompt the construction of the databases; among them, the questionnaire indicated the following: evaluation, open access, visibility, monitoring, management and other aims to be specified. In Switzerland, the main objectives are open access and the chance to improve the visibility of academics products. Spanish universities indicate monitoring and management as relevant, while the Italian respondents consider evaluation and visibility as the two main aims (Figure 2).

At least three different types of uses can be identified for university managers, professors, funding agencies, and the government: *ex-post* evaluation, management and visibility. All answers point out that university managers use databases for *ex-post* evaluation and management purposes, professors

The initiative for institutional databases' construction is largely internal, often based on decisions taken within each university, and never related to government or regional initiatives

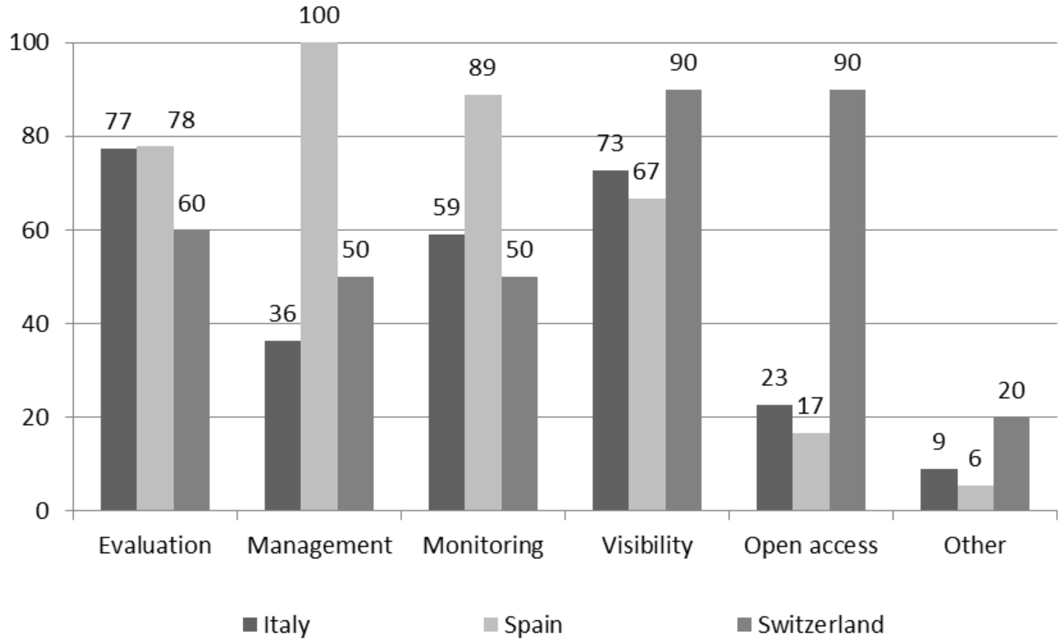

Figure 2. Aims of the dataset

and researchers for CV elaboration and visibility. Funding agencies mainly use institutional datasets for visibility in all three countries. Government use is generally related to visibility in Switzerland, to management and visibility in Spain and to *ex-post* evaluation and visibility in Italy.

The answers collected seem to support the evidence that when visibility and communication are the main aims, access to repositories is mostly free and accessible to different users: scientific commissions of disciplinary areas, registered external users, and search engines. Access is partial, that is, limited to (some) bibliographic information, when monitoring and management are the main aims (Figure 3).

Some shortcomings, which can limit the use and implementation of the datasets, are also indicated. Most respondents see the lack of data-cleaning and quality control as relevant shortcomings, and only a few do not consider them very important. In Italian universities there are also problems related to technical issues and to copyright, the latter pointed out by Swiss universities too. The limited propensity of the scientific staff towards archiving practices is another relevant shortcoming, while datasets' coverage is generally considered good and does not pose much of a problem. The expected evolution of the datasets is strongly related either to changes in the internal organization of research activities and academic structures or to different uses the datasets should be able to support (e.g. evaluation activities instead of dissemination purposes).

Comments in many questionnaires from the surveyed countries indicate the presence of datasets construction processes that are more complex than

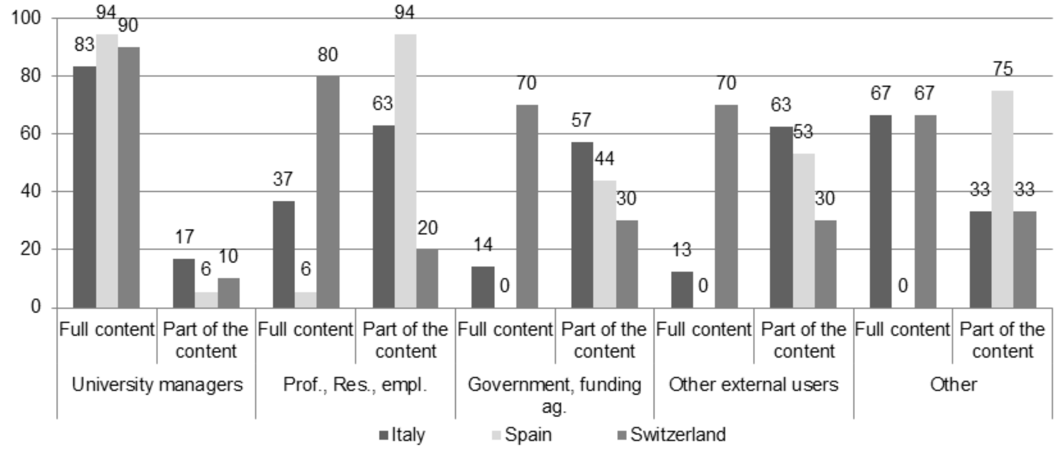

Figure 3. Access to datasets' full contents and part of the contents

they might appear at first glance, including negotiations among different actors and significantly embedded in their social context. Moreover, there are considerable differences even within the same country, as well as across different users. Interestingly enough, the respondents do not consider the construction of the database as a completed process, but rather as still ongoing. They acknowledge the need to overcome the drawbacks of the databases and to improve them by bridging the gaps among the different visions endorsed by the actors involved.

3. Methodological issues for the case studies

The UNIPUB project focuses on three countries: Italy, Switzerland, and Spain. It developed six case studies on university publications databases, two for each partner country (see Appendix 1).

The databases are analyzed according to a common framework made up of several dimensions: institutional information (university, department, researchers), bibliographic information (author information, scientific domain, publication year, condition and type of publication, language, aim of the publication, and type of audience), existence of procedures and guidelines for the construction, quality control and periodical updating of the databases. Table 1 shows the years and number of documents considered for each database.

This enables us to identify:

1. The data available for building indicators and the timeframe for their test on publications (types of publications, publication year, and bibliographic information);
2. The possible levels of aggregation (institutional information); and
3. Some common limitations to be dealt with when using institutional databases (differences in databases' organization, management, and content).

Table 1. Years considered in the test for each university

University	Years considered in the test	Number of documents considered for each dataset
Milano Bicocca – Italy	2006–2007	1,780
Napoli Federico II – Italy	2007	4,968
Università della Svizzera Italiana – Switzerland	2006–2007	2,645
University of Zurich – Switzerland	2008	8,609
Universidad Carlos III de Madrid – Spain	2005–2006	5,946
University of Barcelona – Spain	2005–2006	7,743

As for the first point, different typologies of publications are collected by institutional databases. Our analysis includes all the items that can be labelled as 'publications' under copyright rules.[2] Thus, the UNIPUB project does not include technological outputs, and cannot be considered fully descriptive of the universities analyzed, although the outputs included largely represent the results of their research performance.

Three main levels of aggregation are considered in our analysis — university, department and scientific field — and the data are normalized on the number of professors and researchers (head counts). We consider the research units below the faculty level (departments in the Italian and Spanish cases, and institutes and laboratories in the Swiss case), since they are the meso-organizational levels in charge of research activities within the universities.

The analysis of publications at the scientific field level includes various disciplinary areas, as the universities in the sample are mostly generalist universities. Publications are classified by disciplinary areas on the basis of the departments in which they originated. The departments are aggregated in wide disciplinary areas, following the definition of the first level of disaggregation provided by the *Observatoire des Sciences et des Techniques* (OST) classification, in order to have homogeneous and internationally comparable scientific fields. This scientific classification is acknowledged as the most suitable for the analysis, since it has already been used by the research teams, can be matched to the ISI subject fields, and is compatible with the OECD and UNESCO classifications (OECD, 2002; UNESCO, 1997). This choice implies that our analysis cannot capture the actual discipline of the publications. A better solution would be an individual classification of each document, but this is not feasible during this phase of the study. For collaborative publications, with authors affiliated to more departments and sectors, double counting or fractional counting criteria are applied, depending on the relevance of collaborative publications in the datasets.

Various limitations in the use of institutional databases reduce the chance to draw comparisons other than methodological ones. First, we should consider differences in the way the content of the databases is organized, which seem to be mainly due to the lack of a standardized model. In a number of countries, such as Denmark, Belgium, Norway and Australia, some efforts toward standardization have been made; see the analyses by Hicks and Wang (2009) and by Sivertsen (2010). Another important initiative is the EUROCRIS work, based on the Common European Research Information Format (CERIF) standard (EUROCRIS, 2009). Nevertheless, these initiatives have not been fully implemented by the universities in our sample.

Differences in archiving practices and in the guidelines for quality control as well as for data-uploading, -updating, and -cleaning affect the completeness of bibliographic information across universities and countries. Random checks on the correspondence between researchers' CVs and information stored in the datasets confirm that publications are regularly excluded and that scholars display different habits in data-storing, mainly related to the discipline and to the purpose of the dataset.[3] Finally, the lack of information on external authors (number, nationality, affiliation) prevents us from fully exploring collaboration patterns and the academic institutions' international standing. In brief, comparisons are not possible because the databases are built following a social construction mechanism.

4. Results of the case studies

According to our assumptions, entries in institutional databases should allow us to answer questions about the characteristics of universities' research efforts, which encompass knowledge production, users, diffusion at national or international level, and the collaborative or individual nature of publications, as well as publishing specificities. This analysis does not 'measure' total research outputs, but it indicates how the internal communities of key performers involved in the social construction of the databases represent and communicate said outputs.

With this purpose in mind, and considering the available data, we created four indicators providing descriptive information (number of publications per researcher, differentiation of scientific outputs, national/international orientation, articles in journals with and without impact factor [IF]) and three indicators focusing on assumptions about university research (reference community, scientific characterization of research outputs [i.e. basic or applied research], and publications with or without peer review). In principle, by rating academic departments and fields of science and by counting publications according to different purposes summarized by the indicators, normalizing them on the number of researchers, we can find out if there are differences in production, orientation, collaborations, external relations, and researchers' publishing choices. However, this analysis describes the characteristics of scientific productivity not as they actually are, but as the result of the social representation of the universities and their disciplinary areas provided by the databases. Given the mentioned limitations of the databases and the short timeframe covered, the results presented in the following sub-sections have the purpose of illustrating the suitability and utility of the descriptors and indicators.

4.1 Scientific differentiation

The academic institutions analyzed through the publication outputs of institutional databases show some interesting common features. First, a differentiation in scientific outputs can be observed, although journal articles and conference papers are still the most relevant types of documents in the so-called hard sciences, while books and book chapters are the most relevant products in humanities and social sciences. Yet, in some cases, it is evident that certain outputs have not been considered as relevant as others in representing the scientific value of research performance (e.g. there are universities whose guidelines establish the type of documents to consider, which types of books or newspapers to include or exclude, as part of the social construction of the databases).

In the examples provided by the University of Zurich (Figure 4) and the University of Barcelona (Figure 5), we can observe differences in the types of scientific production. However, the data might be strongly influenced by the varying propensity of scholars from different fields towards including some types of publications, such as conference papers, in institutional datasets.

4.2 National or international orientation

This indicator aims at pointing out whether the university's scientific production is more addressed to a national or international audience. Different criteria can be used to determine the orientation of different scientific products. For example, the language criterion is adequate for articles, but not representative

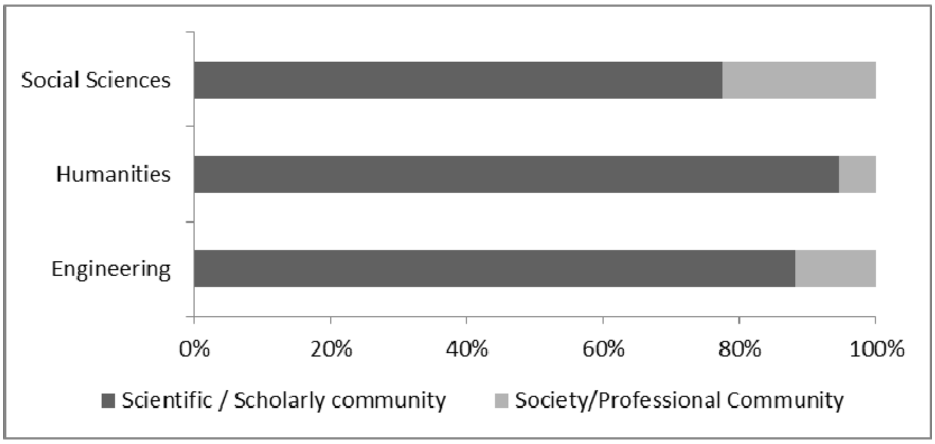

Figure 4. Scientific differentiation – University of Zurich

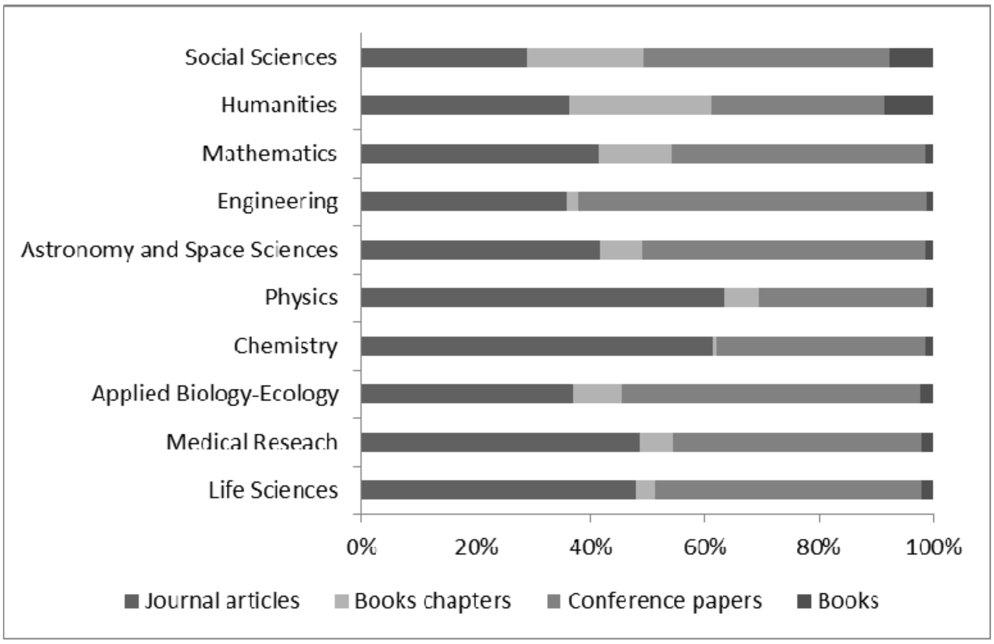

Figure 5. Scientific differentiation – University of Barcelona

enough for conference papers, since the location of the congress is also relevant; conversely, for books and chapters of books the country of the publisher is an important piece of information. In order to reduce limitations caused by the use of a single criterion, the national/international orientation descriptor is analyzed through two criteria: language of publication and, when available, relevance self-attributed by the researchers themselves.

The adopted solution still has some limitations regarding:

- The analysis of publication outputs in multi-language countries, like Switzerland: publications in the national languages of Switzerland can be considered as both national and international, since each linguistic region has a corresponding neighboring country;
- The analysis of publication outputs in English-speaking countries: the language criterion does not fit here (although no English-speaking country was included in this study). The same is true for Spain, as Spanish is a widely spoken language with a huge community;
- The need for further specifications when products like conference papers or books are analyzed.

Nonetheless, the indicator provides interesting information. Natural and medical sciences products display a strong international orientation, compared to the main national orientation which characterizes the social sciences and humanities. Moreover, we can observe consistency or discrepancies in the use of the two criteria (language and orientation self-attributed by researchers) in the analyzed scientific fields.[4]

The self-attribution and language criteria coincide to some extent in the hard sciences, where publications with international relevance are mainly in English, while in the humanities the percentage of publications seen as international is higher than that of publications in English. In fact, many publications in national languages are indexed as if they had international relevance (see Figures 6 and 7). This might mirror the different meanings attributed to international relevance in some scientific areas such as the humanities, where the national language is not an obstacle to the diffusion of the output beyond the national boundaries. However, differences can be detected in the hard sciences too (e.g. chemistry), showing that ‘international relevance’ can take on different meanings according, for instance, to the existing sub-areas, and to their prevailing values, ideas, and rules. Summing up, databases are interpreted differently across disciplinary areas, institutions and countries, and the choice of language is often subject-specific.

4.3 Characteristics of the research output

For articles in journals, publishing in refereed vs. non-refereed journals can be seen as a proxy of research quality, since the publication outputs undergo an explicit reviewing process. This indicator

Natural and medical sciences products display a strong international orientation, compared to the main national orientation which characterizes the social sciences and humanities

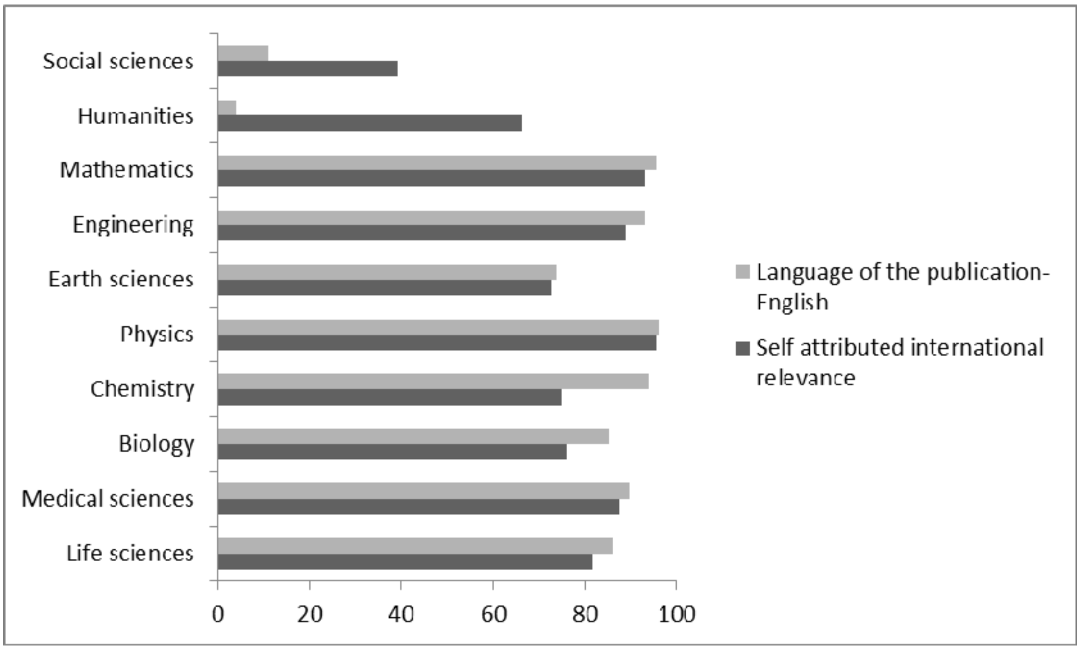

Figure 6. International relevance self-attributed and according to the language of the publication (%) – Napoli Federico II

presents some problems: it cannot be tested for Spanish universities because no information is available in the databases. As for the UZH ZORA database, the faculties have different rules regarding which articles can or cannot be considered as refereed (depending on the practices of the different scientific areas). In this case too, major constraints are represented by the uses and purposes of the databases, which shape decisions about the collection of scientific outputs, and by the attitude of researchers when adding their publications, which is influenced, for example, by the image of their output they wish to give.

Indications on refereed/non-refereed journals are associated with the IF/NON-IF descriptor, whose aim is to indicate whether university articles are mainly published in journals included in the JRC database. We do not intend to enter the discussion about the usefulness of the IF; we simply wish to analyze the possibility of looking at it through institutional databases, in order to describe the importance of IF journals within the overall university production.

Concerning the IF, the tests on publications databases support the evidence that international databases only partially cover the articles published in social sciences and humanities.[5]

Combining the aforementioned information with the analysis of publications that undergo a peer review process, we find confirmation that in the social sciences and humanities there is a lower propensity towards refereed articles in comparison to other

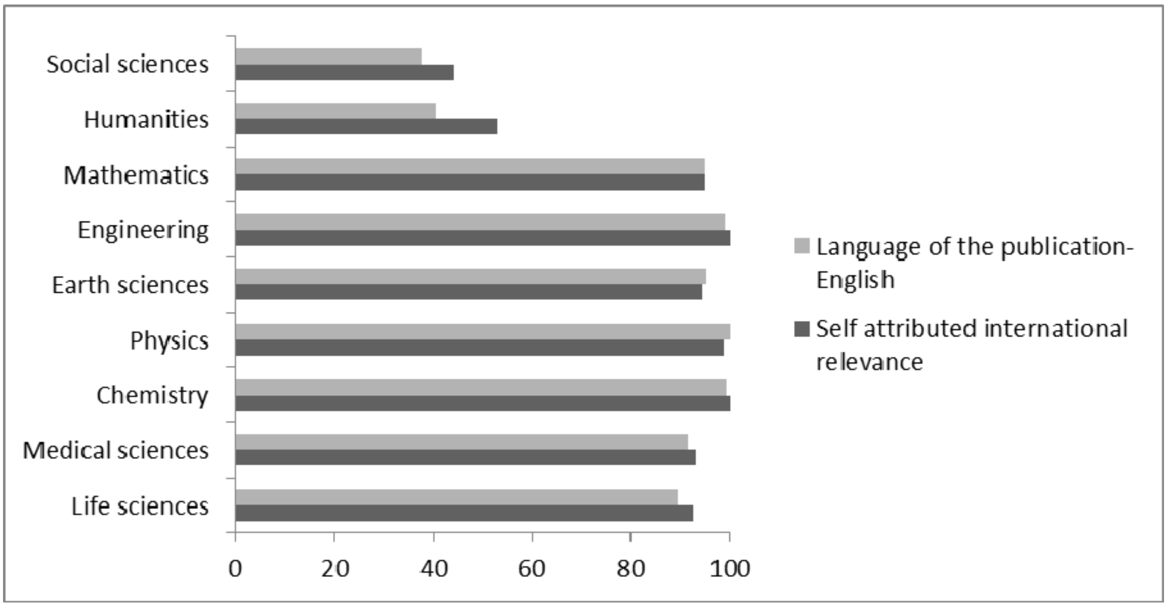

Figure 7. International relevance self-attributed and according to the language of the publication (%) – Milano Bicocca

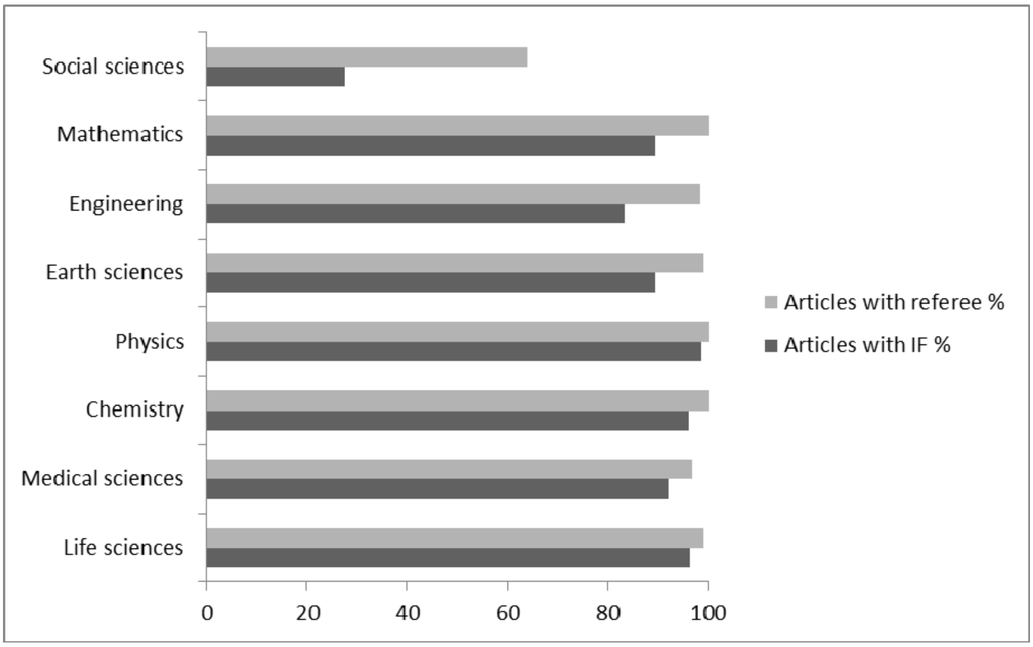

Figure 8. Articles with impact factor and with referee – Milano Bicocca

disciplinary areas, as proven also by the data on the Milano Bicocca University (Figure 8).

4.4 Scope of the publications

This indicator focuses on the main uses and types of users of university scientific outputs. This is certainly an interesting issue, especially in the discussion about the universities' third mission and their contacts with the non-academic environment, but the databases provide no direct information on the subject. We therefore analyze the audiences, by assigning the outputs to two different categories of audience as follows:

- Scientific/scholarly community: articles, books, chapters of books, PhD theses;
- Society/professional community: conference papers, proceedings, other publications.

The results prove that publications are mainly dedicated to a scientific audience, with limited differences across departments and scientific fields. Some examples of results generated by this indicator are shown for two of the universities analyzed (Figures 9 and 10).

Besides the legitimate debate about the basic assumption behind this indicator, it outlines an important bias, that is, the relevance of representativeness of outputs within the databases in relation to the whole scientific production of the universities. If the coverage is good, although not complete, the data on audience, measured through outputs, will be reliable. If the coverage is not too extensive, since only some

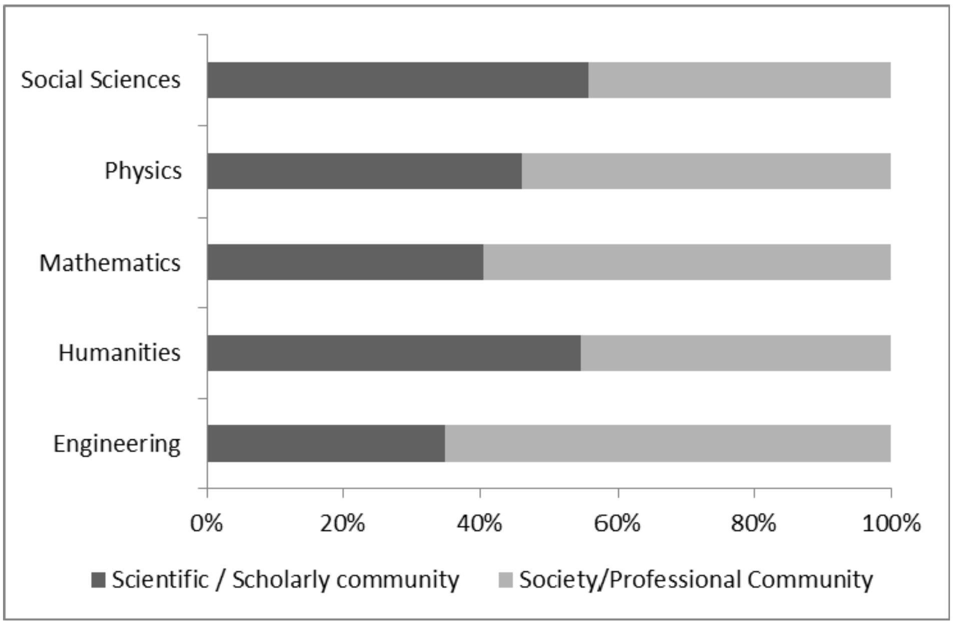

Figure 9. Reference community – University Carlos III of Madrid

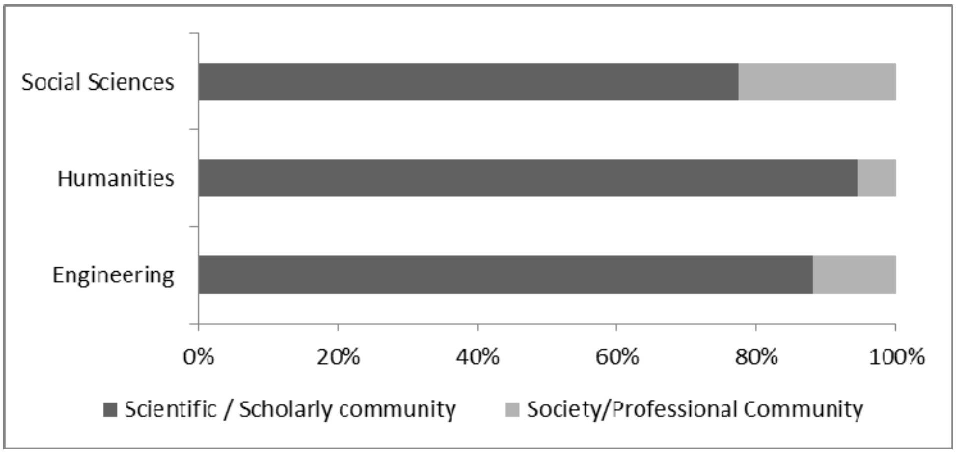

Figure 10. Reference community – Università della Svizzera italiana

products are included, the results will confirm a bias towards the scientific community, and the social role of the university might be misinterpreted. Again, we find evidence that databases are social constructs, showing a picture of the university that is driven by the way in which the different communities want to represent, communicate, and document their research activity.

5. Discussion and conclusions

This paper investigates the extent to which institutional databases of university publications can be considered a useful tool to analyze academic research results and to support the construction of research indicators. Institutional databases might help universities in attaining a rather complete picture of their scientific activities and in better understanding the role they play for different audiences, namely the scientific community and the social and economic context.

We consider institutional databases as social constructs, in order to highlight the extent to which they can represent university research performance, and how they can complement other available national and international sources of information on university publications outputs. This will allow university managers, the research community, the government and funding agencies to communicate the outcome of their research efforts; at the same time, it will supply information on the emerging research tendencies in different scientific fields.

Using a survey covering the three countries as well as case studies on a sample of universities, we have performed an empirical analysis, which has yielded some interesting results.

For each database, the survey considers where it comes from, its aims and users, access rules, quality control system, guidelines for updating, and the role of the different individuals involved in its creation and management. The survey highlights the low level of harmonization among institutional databases in each country as well as among different countries. It also supports the view that shared models for institutional databases and for management systems of research activities and products should be promoted (Sivertsen, 2009). This would overcome the current co-existence of different sets of data and the low interoperability of the software used, which are strong limitations to the employment of these tools for evaluation and management purposes and for comparative analyses. Moreover, it would avoid fragmentation and duplication of databases, since several databases often coexist at different levels within the same university (university, departments, laboratories).

The case studies point out the main opportunities and limitations in the use of institutional databases and their ability to provide a representation, rather than a comprehensive picture, of each university's profile of activities. Their most important limitations are that they do not provide information on the authors' institutional and disciplinary affiliation outside the university (nationality, scientific field they belong to) and they do not distinguish between internal and external authors.

Combining the results of the survey — which tell us how the process of construction was developed — and evidence from the case studies, we can understand how the structure of the databases and the organization of their content influence their completeness and the negotiation process behind their construction.[6] The evidence gathered confirms some results from other studies on the design and implementation of institutional repositories (King *et*

Combining the results of the survey and evidence from the case studies, we can understand how the structure of the databases and the organization of their content influence their completeness and the negotiation process behind their construction

al, 2006; Ferreira *et al*, 2008; Rieger, 2008). The more the process of negotiation evolves, integrating meanings and values to be represented through the databases, the more their content is modified. Hence, institutional databases can allow for an in-depth analysis of research performance and determinants. For example, they can help identify the various types of communities universities want to address. The same holds true for the analysis of co-authorships, which might include very interesting features, such as the characteristics of collaboration networks (through the authors' affiliation and nationality), interdisciplinary collaborations (looking at the scientific field of authors, besides that of the publication), and the role of researchers in the publications (first author, coauthor, last author). It might also allow for a mapping of authors' scientific production from a long-term perspective.

To conclude, databases are not objects taken for granted. Instead, they can evolve and overcome their current limitations, above all the lack of common criteria and clear guidelines addressed to authors (professors and researchers). The differentiation of products in institutional databases can help us analyze the audience and features of an institution's scientific production as well as its relations at the local level with different stakeholders. Institutional databases focus on the scientific production of different scientific areas and their characteristics are deeply influenced by the dimension of the departments, especially if we look at small ones, in which the productivity of a few individuals matters greatly. Furthermore, they are organized along many or few different disciplinary areas.

An interesting line of research could be the matching of information from institutional databases with that from international bibliographic databases. For instance, institutional databases might be very useful in overcoming the problem of author affiliation in international databases, thus allowing for the complete identification of all the articles produced by universities, although this would require a substantial amount of manual work. Moreover, problems related to homonymy and double counting might also be addressed. Lastly, institutional databases can be used in combination with other sources, such as Google Scholar, to analyze citations for non-article publications, and to provide information on the journals most used by scholarly communities in the humanities and social sciences, where the coverage of international databases is problematic.

All the aforementioned features might be valuable for different aims, such as research evaluation and university management, stimulating the academic debate about university research efforts and providing support in the decision-making processes.

Our results lead to the conclusion that, given the mentioned limitations, institutional databases of university publications are still a 'work in progress' and they need to be further developed and standardized, so that reliable research indicators can be produced for evaluation purposes. Nevertheless, considering their potential added value, their social construction and implementation process at the national and institutional level should be improved through good negotiations and incentives, for their coverage to be as large as possible and to overcome methodological problems. An issue for future research is the extent to which open access databases can reinforce the visibility and accountability of universities, contributing to their communication processes through free access to information on knowledge products, increasing the number of database users (academics, students, evaluators, university managers, etc.) and serving different purposes (e.g. grant applications, CVs), as well as making information and data easily transferable by professors and researchers when changing institution.

Appendix 1.

Our sample includes the Università Milano Bicocca and Università Federico II (Naples) for Italy, the Universidad de Barcelona (UB) and Universidad Carlos III de Madrid (UC3M) for Spain, and, for Switzerland, the Università della Svizzera Italiana (USI) and the University of Zurich (UZH). The universities were selected according to the following criteria:

- Size and location;
- Presence of a good database, with good differentiation of scientific outputs and bibliographic data, and adequate time and disciplinary publications coverage (verified through a random comparison between dataset outputs and professors' CVs). Each university had to provide access and support in using the data. Disciplinary and time coverage were also important factors, to allow for comparisons within areas and among different scientific areas and to have a consistent and somehow comparable set of data on a yearly basis, at least;
- Presence of a sample of datasets as representative as possible of the different institutional repositories existing at the national level.

As for the databases, the BOA (Bicocca Open Archive, the open access publications archive of Milano Bicocca) was created in June 2008 and is based on the SURplus system, promoted by CILEA, the Lombardy Inter-university Consortium for Automatic computation. The U-Gov Catalogue of the University of Naples Federico II is based on the U-Gov system, created by CINECA, the Inter-university Consortium for Automatic Computation, and managed by the three scientific schools of the university (to which the departments are affiliated for research activities). Both systems, of national relevance, aim at improving the management of research activities and products and they include, among other modules, institutional repositories.

(*continued*)

Appendix 1 (*continued*)

UNIVERSITAS XXI, the system adopted by the UC3M and by several other Spanish universities, was developed by the Office of University Cooperation, a company belonging to six Spanish universities and one bank, to provide an effective management system for universities. Instead, GREC is an informatics research management tool developed by the UB in 1987 and now used by several other universities and research centers; it includes different databases (curriculum vitae, projects, and publications).

The USI publications database was created in 2003/2004. It consists of an Oracle database with a PHP interface and, besides information on publications, it also contains individual information about researchers and professors (contact information, office hours, biography, CV, list of publications, and links to individual websites), and information on projects and courses.

The Zurich Open Repository and Archive (ZORA) has been used at UZH since the reporting year 2008. Its contents are directly transferred to the database used for academic reporting. ZORA works with the widely diffused open source software for repositories, EPrints. It does not contain information other than that concerning publications.

Notes

1. The questionnaire was addressed to the person responsible for the database or its administrator.
2. Articles, books, book chapters, proceedings, papers to conferences, working papers, PhD theses, software and datasets.
3. In the case of Italy, for instance, a coverage check for the years 2006 and 2007 was carried out through a random comparison between publications in the BOA and U-Gov datasets (the latter only for 2007) and the following sources: professors' CVs; yearly reports on research activities produced by the departments; and departments' repositories (where available).
4. For Napoli Federico II the two criteria – language of the publication and self-attributed relevance – are compared only for a limited sample of 2,451 publications.
5. It must be noted that AHCI has no calculated IF, no JCR. Therefore, only those humanities journals classified also in SSCI categories can have IF.
6. For instance, too many open fields make it more difficult for the researchers to archive their publications. If not mandatory, these fields are often left blank, or might even contain wrong information.

References

Bonaccorsi, A, C Daraio, B Lepori and S Slipersaeter 2007. Indicators for the analysis of higher education systems: some methodological reflections. *Research Evaluation*, **16**(2), June, 66–78.

Cañibano, C, J Otamendi and I Andújar 2008. Measuring and assessing researcher mobility from CV analysis: the case of the Ramón y Cajal programme in Spain. *Research Evaluation*, **17**(1), March, 17–31.

Day, M 2004. Institutional repositories and research assessment: a supporting study for the ePrints UK Project. Available at: <http://eprints-uk.rdn.ac.uk/project/docs/studies/rae/rae-study.pdf>, last accessed 12 December 2010.

De Filippo, D, E Sanz Casado and I Gómez 2009. Quantitative and qualitative approaches to the study of mobility and scientific performance: case study of a Spanish university. *Research Evaluation*, **18**(3), September, 191–200.

EUROCRIS 2009 CERIF 2008 1.0 Full Data Model (FDM) Introduction and Specification, April, available at: <http://eurocris.org>, last accessed 25 April 2009.

Ferreira, M, A Baptista, E Rodrigues and R Saraiva 2008. Carrots and sticks: some ideas on how to create a successful institutional repository. *D-Lib Magazine*, **12**(1/2).

Gomez, I *et al* 2007. Structure and research performance of Spanish universities. In *Proceedings of ISSI 2007*, eds D Torres-Salinas and H F Moed, pp. 334–345. Madrid: ISSI.

Gomez, I *et al* 2009. Structure and research performance of Spanish universities. *Scientometrics*, **79**(1), 131–146.

Hacking, I 1999. *The Social Construction of What?* Harvard University Press.

Harnad, S 2001. The self-archiving initiative. *Nature*, **410**, 1024–1025.

Hicks, D 2004. *The Four Literatures of Social Science*. In *Handbook of Quantitative Science and Technology Research*, eds H Moed, W Glänzel, and U Schmoch, pp. 473–496. Dordrecht: Kluwer Academic Publishers.

Hicks, D and J Wang 2009. Toward a bibliometric database for the social sciences and humanities. School of Public Policy, Georgia Institute of Technology, April. Available at <http://work.bepress.com/diana_Hicks/18/18>, last accessed 12 December 2010.

Iribarren-Maestro, I, M L Lascurain-Sánchez and E Sanz-Casado 2009. The use of bibliometric techniques in evaluating social sciences and humanities. In *Celebrating Scholarly Communication Studies: A Festschrift for Olle Persson at his 60th Birthday*, eds F Åström, R Danell, B Larsen and J W Schneider, pp. 25–37. International Society for Scientometrics and Informetrics.

Jacsó, P 2008. Google Scholar revisited. *Online Information Review*, **32**(1), 102–114.

King, C J, D Harley, S Earl-Novell, J Arter, S Lawrence and I Perciali 2006. *Scholarly Communication: Academic Values and Sustainable Models*. Berkeley, CA: Center for Studies in Higher Education, University of California–Berkeley.

Knorr Cetina, K 1997. Sociality with objects: social relations in post-social knowledge societies. *Theory, Culture and Society*, **14**(4), 1–30.

Latour, B and S Woolgar 1979. *Laboratory Life: the Social Construction of Scientific Facts.* Los Angeles, USA: Sage.

Lepori, B and C Probst 2009. Using curricula vitae for mapping scientific fields: a small-scale experience for Swiss communication sciences. *Research Evaluation*, **18**(2), June, 125–134.

Lepori, B, R Barré and G Filliatreau 2008. New Perspectives and challenges for the design of S&T indicators. *Research Evaluation*, **17**, March, 33–44.

Mazzotti, M 2008. *Knowledge as Social Order. Rethinking the Sociology of Barry Barnes*. Aldershot: Ashgate.

Merkx, F. and P van den Besselaar 2008. Positioning indicators for cross-disciplinary challenges: the Dutch coastal defense research case. *Research Evaluation*, **17**(1), 4–16.

Moed, H, W Glänzel and U Schmoch eds 2004. *Handbook of Quantitative Science and Technology Research*. Dordrecht: Kluwer Academic Publishers.

Nederhof, A J 2006. Bibliometric monitoring of research performance in the social sciences and the humanities: a review. *Scientometrics*, **66**(1), 81–100.

Norris, M and C Oppenheim 2007. Comparing alternatives to the Web of Science for coverage of the social sciences' literature. *Journal of Informetrics*, **1**(2), 161–169.

OECD 2002. The measurement of scientific and technological activities: proposed standard practice for surveys on research and experimental development. *Frascati Manual*, Paris.

Reale, E, D De Filippo, I Gómez, B Lepori, C Probst, B Potì , E Primeri and E Sanz Casado 2009. *Methodologies for the characterization of the publication output of higher education institutions using institutional databases, Final Report*. PRIME NoE. Available at <http://www.prime-noe.org>, last accessed 1 December 2010.

Rieger, O Y 2008. Opening up institutional repositories: social construction of innovation in scholarly communication. *Journal of Electronic Publishing*, **11**, 3.

Searle, J 1995. *The Construction of Social Reality*. New York: Free Press.

Sivertsen, G 2009. A bibliometric funding model based on a National Research Information System, ISSI Conference 14–17 July 2009, Rio de Janeiro. Available at <http://www.issi2009.org/agendas/issiprogram/public/documents/ISSI%202009%20Sivertsen%20Vista-100730.pdf>, last accessed 1 September 2009.

Sivertsen, G 2010. A performance indicator based on complete data on scientific publication output at research institutions. *ISSI Newsletter*, **6**(1), 22–28.

Torres-Salinas, D and H Moed eds 2007. *Proceedings of the Eleventh International Conference on Scientometrics and Informetrics*, Vol. 1, pp. 112–123. CSIC, Madrid, Spain: CSIC.

Torres-Salinas, D I and H F Moed 2009. Library catalog. analysis as a tool in studies of social sciences and humanities: an exploratory study of published book titles in economics. *Journal of Informetrics*, **3**(1), 9–26.

UNESCO 1997. *ISCED International Standard Classification of Education*. Paris: UNESCO.

Van den Besselaar, P and L Leydesdorff 1996. Mapping change in scientific specialties a scientometric reconstruction of the development of artificial intelligence. *Scientometrics*, **47**, 415–436.

Van den Besselaar, P, J Edler, G Heimeriks, L Henriques, P Larédo, T Luukkonen, M Nedeva, A Schoen and D Thomas 2007. Toward ERA configurations: an experiment on chemistry. Workshop, 'Beyond the dichotomy of national vs. European science systems: configurations of knowledge, institutions and policy in European research', Bonn, 30 May.

Van der Graaf, M and K van Eijndhoven 2008. *The European Repository Landscape Inventory Study into the Present Type and Level of OAI-Compliant Digital Repository Activities in the EU*. Amsterdam University Press

Van Raan, A 2001. Competition amongst scientists for publication status: toward a model of scientific publication and citation distributions. *Scientometrics*, **51**(1), 347–357.

Van Raan, A 2004. Measuring science. In *Handbook of Quantitative Science and Technology Research*, eds H Moed, W Glänzel and U Schmoch, pp. 19–50. Dordrecht: Kluwer Academic Publishers.

Van Raan, A 2008. Bibliometric statistical properties of the 100 largest European research universities: prevalent scaling rules in the science system. *Journal of the American Society for Information Science and Technology*, **59**(3), 461–475.

Walters, W H 2007. Google Scholar coverage of a multidisciplinary field. *Information Processing and Management*, **43**, 1121–1132.

White, W 2007. Opening access and closing risk: delivering the mandate for e-theses deposit. In ETD 2007 added values to e-theses, 10th International Symposium on Electronic Theses and Dissertations, Uppsala, Sweden, 13–16 June 2007, p. 6.