Research Policy 44 (2015) 1645–1658

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

SiSOB data extraction and codification: A tool to analyze scientific careers

Aldo Geuna[a,b,*], Rodrigo Kataishi[a,b], Manuel Toselli[a,b], Eduardo Guzmán[c], Cornelia Lawson[a,b,d], Ana Fernandez-Zubieta[e], Beatriz Barros[c]

[a] Department of Economics and Statistics Cognetti De Martiis, University of Turin, Italy
[b] BRICK, Collegio Carlo Alberto, Moncalieri, Italy
[c] Department of Languages and Computer Science, University of Malaga, Spain
[d] School of Sociology and Social Policy, University of Nottingham, UK
[e] Institute for Advanced Social Studies – Spanish Council for Scientific Research, Spain

ARTICLE INFO

Article history:
Received 26 January 2015
Accepted 28 January 2015
Available online 25 February 2015

JEL classification:
C81
C88
I23
O31

Keywords:
Information retrieval
Extraction and data integration
Academic careers
Research productivity
Mobility of research scientists

ABSTRACT

This paper describes the methodology and software tool used to build a database on the careers and productivity of academics, using public information available on the Internet, and provides a first analysis of the data collected for a sample of 360 US scientists funded by the National Institute of Health (NIH) and 291 UK scientists funded by the Biotechnology and Biological Sciences Research Council (BBSRC). The tool's structured outputs can be used for either econometric research or data representation for policy analysis. The methodology and software tool is validated for a sample of US and UK biomedical scientists, but can be applied to any countries where scientists' CVs are available in English. We provide an overview of the motivations for constructing the database, and the data crawling and data mining techniques used to transform webpage-based information and CV information into a relational database. We describe the database and the effectiveness of our algorithms and provide suggestions for further improvements. The software developed is released under free software GNU General Public License; the aim is for it to be available to the community of social scientists and economists interested in analyzing scientific production and scientific careers, who it is hoped will develop this tool further.

1. Introduction

Scientific and technological advances are acknowledged to be among the main drivers of social and economic development, and policy makers across the world are searching for strategies to encourage scientific production and the exchange of knowledge. Social scientists and economists have been trying to elicit the functioning of the process of scientific production, and to understand the contribution of science to innovation and economic growth (Antonelli et al., 2011).

The scientific research process is characterized by multiple research inputs and outputs, of which most studies collect only a small proportion. Publication and patent numbers as output measures are becoming more easily accessible and well recorded. However, academics also produce outputs such as teaching and consulting, and increasingly are required to operate in different work roles (e.g., teaching, services and administration) (Yuker, 1984; Enders, 2005), engage in collaborations with researchers from other countries (Glanzel et al., 2008) and sectors, (Ehrenberg, 2003) and cope with diverse job position changes in the course of their careers. Therefore, it is crucial to have access to more detailed databases that provide longitudinal information at the individual level, to achieve a better understanding of the relationship between the different factors that affect research production.

Against this background, the goal of this paper is to provide a tool (an example of big data management tool) for collecting and structuring information on researchers available on public websites and from academics' CVs. The SiSOB data extraction and codification tool will help scholars in economics, sociology, and related social science disciplines to gather data from different online sources and to assemble them into a system of structured databases to enable further statistical and econometric analysis.

* Corresponding author at: Department of Economics and Statistics Cognetti De Martiis, University of Turin, Lungo Dora Siena 100 A, 10153 Turin, Italy. Tel.: +39 116703891; fax: +39 11 6703895.
E-mail address: aldo.geuna@unito.it (A. Geuna).

http://dx.doi.org/10.1016/j.respol.2015.01.017

This paper describes the methodology and techniques used to develop the current version of the SiSOB environment; the software was released under GNU General Public License v3 in the GitHub[1] repository with the specific aim to stimulate and gather contributions from developers and users, to further develop the software and improve its performance. The SiSOB tool is validated using a sample of US and UK biomedical scientists but is applicable to any country where scientists' CVs are available in English. Our example case of the output database is an investigation of the main characteristics of a sample US scientists funded by National Institute of Health (NIH) and UK scientists funded by BBSRC (Biotechnology and Biological Sciences Research Council). We devote particular attention to the analysis of mobility and career patterns.

2. The analysis of researchers' mobility

2.1. Researchers' mobility

Researchers increasingly have to adapt to new institutions, sectors and work roles, while universities need to manage mobile researchers and their careers (e.g., OECD, 2008; European Commission, 2010a,b). The globalization of the research community which involves increasing levels of international mobility (OECD, 2003; Franzoni et al., 2012; Auriol et al., 2013) and collaboration (Glanzel et al., 2008), is making the geographical movements of researchers especially relevant for flows of knowledge across locations. The goal of improving the knowledge transfer process and encouraging relationships between research actors – university, industry and government (Powell et al., 1996; Leydesdorff and Etzkowitz, 1996; Bozeman and Ponomariov, 2009; Howells et al., 2012) – is making the movement of researchers between the public and private sectors particularly more germane. In addition, the increasing number of foreign PhD degree holders (Ehrenberg, 2003), the numbers of doctoral degree holders taking up post-doc positions (Gaughan and Robin, 2004; Zubieta, 2009) and joining firms (Mangematin, 2000), and the diversification of academic work roles (Yuker, 1984; Enders, 2005) also demand a better understanding of the labor markets for researchers, and the career consequences of mobility (Mangematin, 2000; Enders and Weert, 2004; Enders, 2005). Finally, the high levels of researcher mobility require a greater awareness of the different dimensions of researcher mobility in order to properly address its consequences.

Mobile researchers facilitate the knowledge and technology transfer process and also get access to knowledge, equipment, and networks (Martin-Rovet, 2003; Franzoni et al., 2012; Fernandez-Zubieta et al., 2013) that likely improve their research performance and career opportunities (Ackers, 2005). Therefore, individual researchers as well as the research system can benefit from increased levels of mobility. However, mobility might also be a reflection of a lack of job opportunities for researchers in their home countries (Ehrenberg, 2003; Gaughan and Robin, 2004; Stephan, 2012), and greater employment insecurity in the academic labor market (Smith-Doerr, 2006). Mobility might be a requirement for the pursuit of research careers in certain fields, and job experience abroad is sometimes a requirement for return to the home country (Ackers and Oliver, 2007). Mobility can also be associated with certain costs that might have a negative impact on the academic performance (Fernandez-Zubieta et al., 2013) and career development of researchers (Gaughan and Robin, 2004). Moreover, since patterns of mobility appear to vary considerably across types of mobility (e.g., postdoctoral mobility and tenure-track job mobility) (Zubieta, 2009), its effects might also vary.

In our case study, which provides an example of the information gathered using the SiSOB tool, and show that it is possible to distinguish between: non-tenured (forced) and tenured (voluntary) mobility, postdoctoral mobility, and job to job mobility. It further enables us to measure three mobility dimensions related to inter-institutional (job to job) labor mobility:

- International mobility: job transition to/from a foreign academic system,
- Sector mobility: job transition from academia to industry or vice versa (inter-sector mobility),
- Career mobility: job transition to a higher/lower position.

2.2. Measuring mobility using CVs

Several studies have exploited information contained in CVs to study various aspects related to the mobility of researchers (Bonzi, 1992; Dietz et al., 2000; Gaughan and Bozeman, 2002; Lee and Bozeman, 2005; Dietz and Bozeman, 2005; Cañibano and Bozeman, 2009; Fernandez-Zubieta et al., 2013). CVs and publicly available information on personal webpages constitute a rich source of longitudinal factual data on the major events in a researcher's career and their research contacts. While some dimensions of mobility can be inferred from bibliometric data, most of a researcher's activities are unobservable using traditional data sources. CVs have been found to be particularly useful for the analysis of academic careers since they provide reliable information on education, job transitions, and publications. Using data collected from CVs as well as pure bibliographic measures improves data accuracy since mismatches arising from name similarities and changes in researchers' institutional affiliations can be avoided.

The main problems related to using CVs have been identified as: availability, heterogeneity, truncation, missing information, and data coding (Dietz et al., 2000; Corley et al., 2003; Cañibano and Bozeman, 2009). Previous analyses based on CV information have either required the studied researchers to submit CVs (e.g., Dietz et al., 2000), or used electronic CV databases (e.g., Cañibano et al., 2008). Unavailability appears to be a problem if CVs are requested (Gaughan and Ponomariov, 2008), while access and standardization are problematic in the case of electronic databases (Cañibano and Bozeman, 2009). Heterogeneity refers to the different formats in which CVs are presented, the varying length and ordering of information (Dietz et al., 2000; Corley et al., 2003), and the inconsistency of information resulting from researchers being forced to use standard formats (Cañibano et al., 2008). CVs are often truncated (Dietz et al., 2000; Corley et al., 2003), including information only for the most recent years or the most relevant achievements. Certain information is excluded (e.g., grants and teaching), and many CVs need to be complemented with information from other sources (e.g., publications and patents). The coding of CV information and the cleaning of electronic CV databases for subsequent analysis by diverse coders applying similar criteria, have proven time consuming and error prone (Dietz et al., 2000; Corley et al., 2003; Cañibano and Bozeman, 2009). Thus, the main problems related to using CV information are the availability of CVs and related problems arising from a lack of standardization of the information and its processing.

The SiSOB data extraction and codification tool presented in the next section gathers information from publicly available sources on the web, and automatically extracts units of information and creates structured profiles following a semantic schema. In this research, it has been configured specifically to create a database of researchers' CVs, with the aim of overcoming some of the shortcomings described above.

[1] http://github.com/eduardoguzman/sisob-data-extractor.