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

UNIVERSITÀ DEGLI STUDI DI TORINO

SiSOB Data Extraction and Codification:

A tool to analyse scientific careers

Aldo Geuna[(*ab)], Rodrigo Kataishi[(ab)], Manuel Toselli[(ab)], Eduardo Guzmán[(c)], Cornelia Lawson[(abd)], 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 (Turin), Italy
[c] Department of Languages and Computer Science, University of Malaga, Spain
[d] School of Sociology and Social Policy, University of Nottingham
[e] Institute for Advanced Social Studies - Spanish Council for Scientific Research

Acknowledgements

We would like to thank Riccardo Beltrame, Paolo Cecchelli, Emma Gabos, Raimondo Iemma and Daniel Lopez for their help with building the database, software and the dictionaries. Financial support from the European Commission (FP7) Project "An Observatorium for Science in Society based in Social Models – SISOB" Contract no.: FP7 266588 and the Collegio Carlo Alberto Project "Researcher Mobility and Scientific Performance" is gratefully acknowledged. Ana Fernandez-Zubieta acknowledges financial support from the JAE-Doc "Junta para la Ampliación de Estudios" Programme that is co-financed by the Social Structure Funds (SSF).

*: *Corresponding author* - Department of Economics and Statistics Cognetti De Martiis, University of Turin, Lungo Dora Siena 100 A - 10153 Turin, Italy, Tel: +39 0116703891, Fax: +39 011 6703895; email: aldo.geuna@unito.it

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 analysing scientific production and scientific careers, who it is hoped will develop this tool further.

Keywords: Information retrieval, Extraction and data integration, Academic careers, Research productivity, Mobility of Research Scientists.

JEL codes: C81; C88; I23; O31

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 characterised 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.

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

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; EC, 2010a, 2010b). The globalisation 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

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

(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 labour 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, 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) labour 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.

2.1. 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 standardisation 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 many CVs need to be complemented by other sources of information (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 and Cañibano and Bozeman, 2009). Thus, the main problems related to using CV information are the availability of CV information and related problems arising from a lack of standardisation 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 configure specifically to create a database of researchers' CVs, with the aim of overcoming some of the shortcomings described above.

3. The SiSOB data extraction and data structuring tool

The SiSOB tool is an open-source web application that provides social sciences and economics researchers with a reliable means for detecting, storing, and coding information on academics.

The tool could be used to collect data on industrial scientists, though there is much less public information available for this group of researchers. It enables consistent collection of data which, although it still needs some human input, drastically reduces the amount of time required to code. In this section we describe the structure and development of the tool. It is structured in five main modules namely: Crawler (M1), Emails Extractor (M2), CV Extractor (M3), CV Survey (M4) and Text Analyser (M5). The first four modules have been fully developed and tested. Module M5 has been structured and a first development of the algorithm codes has been proposed, it is hoped that the community of users and developers will contribute further developments to increase its performance. Figure 1 depicts the SiSOB tool environment; Figure 2 details the data crawling and processing layer flowchart.

Figure 1: SiSOB Data Extractor Environment

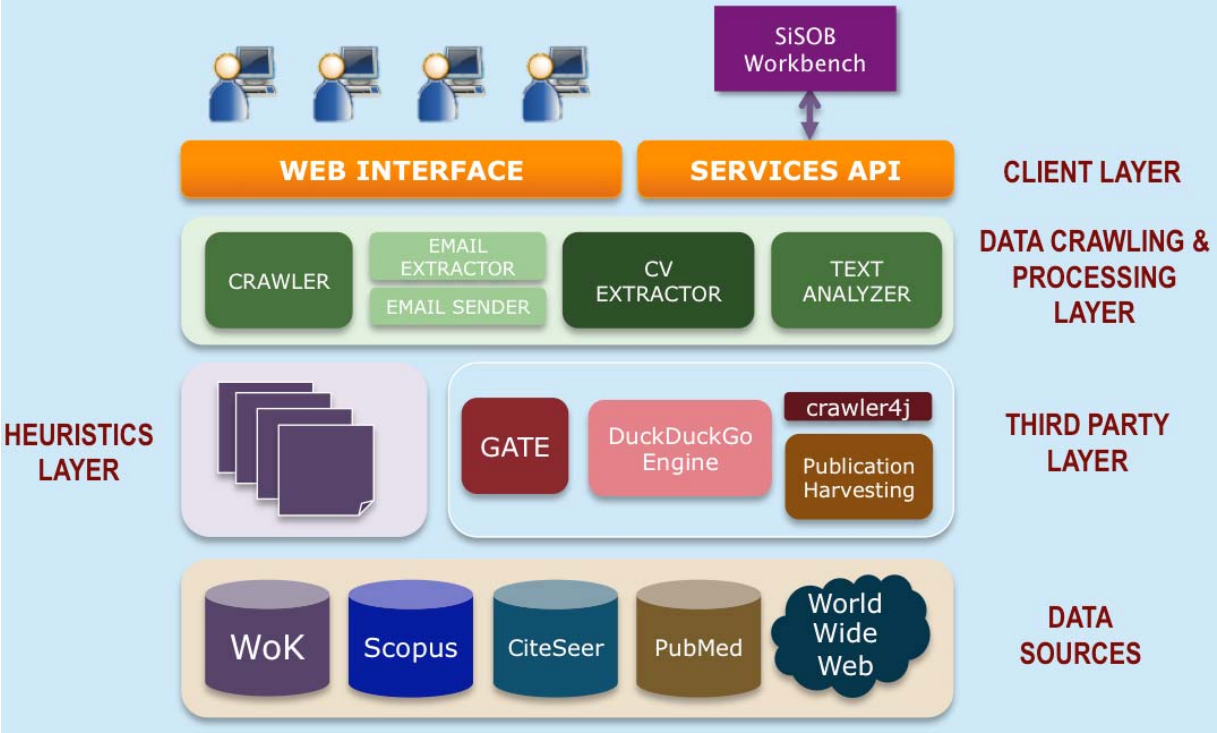
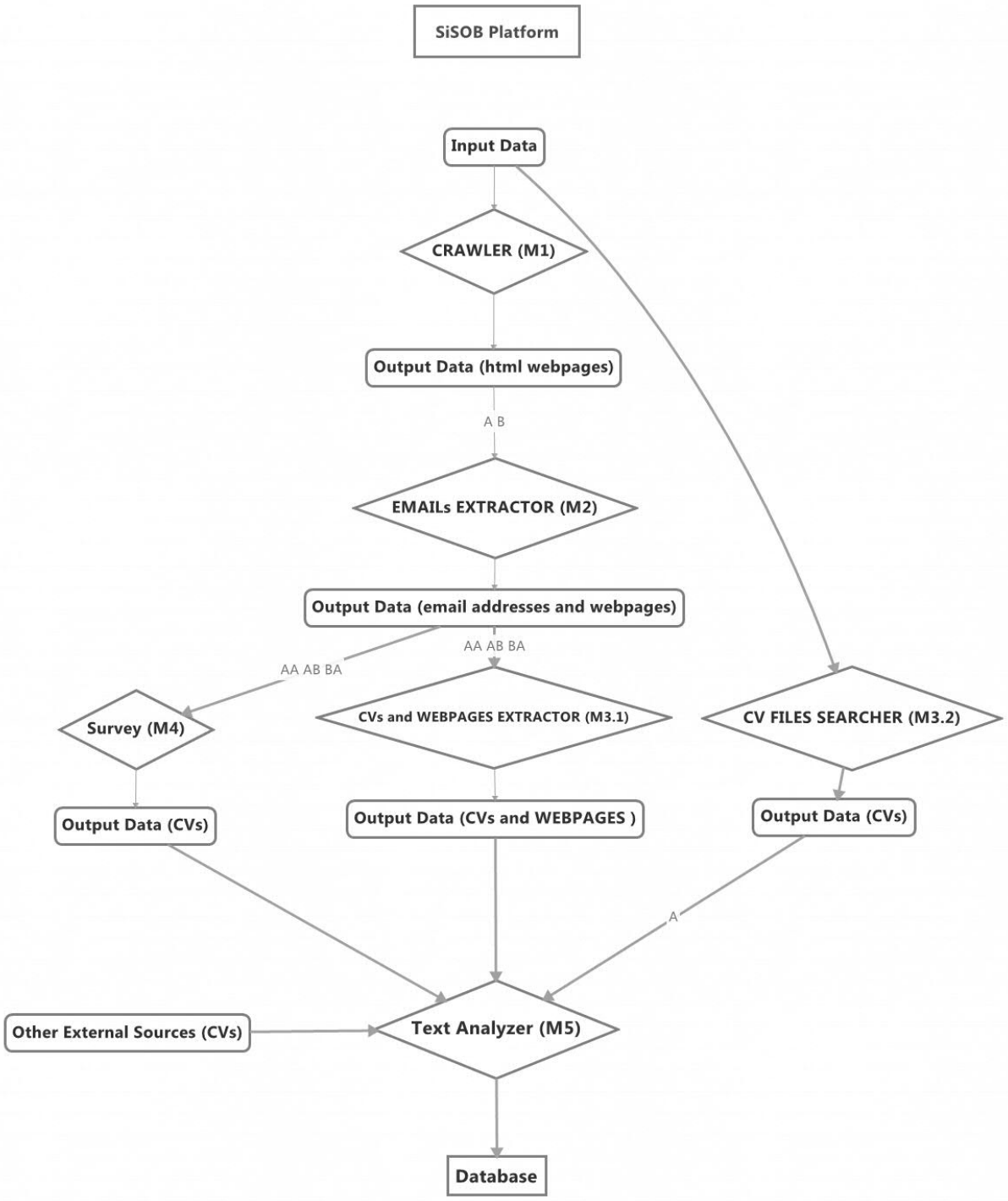


Figure 2: SiSOB Data Extractor Flowchart



3.1. Crawler (M1)

The first module, Crawler (M1), requires as input a small set of information about individual academic scientists. The input for this module is a CSV[2] file that includes a list of scientists, their names, surnames, initials (one or more), research field[3], and optionally, university affiliation. The objective of M1 is to locate researchers' personal webpages. Although several alternatives were studied along the project, in its final version the crawling procedure uses the open software search engine DuckDuckGo.[4] We rely on the DuckDuckGo algorithm to search webpage URLs and metadata for a combination of the search terms. The result of the search is a list of links but only the first (non-sponsored) link is taken up by the M1 process.

The M1 crawling process combines the five input variables in different patterns in order to allow the user to choose between different search strategies: from highly accurate but less broad, to very broad but less accurate. Although all combinations of the variables are possible, trial and error identified six patterns as the most useful:[5]

1. Name, Surname, Initials (target: the Internet)
    - *Example: John J Smith*
2. Name, Surname, Initials, Research Subject (target: the Internet)
    - *Example: John J Smith Chemistry*

[2] CSV files should be in UTF-8 format, with semicolon as word separator. For further information see examples provided on the web application.

[3] This field can take many forms, it can be a general definition of the scientific field or, when available, the departmental affiliation or any other specificities of the research field.

4 The DuckDuckGo (www.duckduckgo.com) platform provides its API for free which allowed us to integrate our open-source platform with theirs. It is a crowdsourced open-search engine, which focuses on privacy and provides open-source advantages. It uses more than 50 different sources including Bing, Yahoo, Wikipedia, Wolfram Alpha, and Yandex. For more information go to http://duckduckhack.com/

5 All the variables within the patterns are technically combined with the "AND" logic operator. Although DDG replaces this automatically in its algorithms, the M1 query explicitly adds the logic operator.

3. Name, Surname, Initials (target: Institution Domain)
    - *Example: John J Smith [only within *.sussex.ac.uk/]*
4. Name, Surname, Initials, University Affiliation (target: the Internet)
    - *Example: John J Smith Sussex*
5. Name, Surname, Initials, Research Subject, University Affiliation (target: the Internet)
    - *Example: John J Smith Chemistry Sussex*

Pattern 1 represents the broader type of search. The search set is the entire World Wide Web. A simple search based on name, surname and additional initials leads to a large number of possible false webpages. Pattern 2 adds the research field which leads to a narrow search but where homonymy can still be a problem. For example, one could perform a search for the personal webpage of a known researcher called John J Smith, doing research in chemistry at Sussex. John Smith is a relatively common Anglo Saxon name, thus the likelihood of finding a false positive outcome of a certain John J Smith doing chemistry at Cambridge or at Stanford (USA) instead of at Sussex is very high. Pattern 3 is very restrictive. It searches for name and surname plus initial only on a specific website (i.e. the institution of affiliation of the researcher) rather than the entire Internet. Pattern 4, in addition to name, surname and additional initials, searches for the university or institution of affiliation of the researcher. Pattern 5 provides the narrowest search on the open Internet and considers name, surname, additional initial, scientific field and affiliation. None of these methods provide 100% accuracy. For example, if the input data are out of date, as in the case of a list of grant recipients (our case study), and the researcher has moved to another university or other employment, it will be impossible to find the researcher using the domain restrictions.

Once the pattern of search is chosen, the tool performs the search. The final output consists of two CSV files: one reporting the cases where no information was found, and one containing the successful searches, that is, containing the search terms plus the identified URL. In this CSV file, each entry is assigned a score ("A" or "B") - "A" if the webpage contains the name, surname and initials of the academic, and "B" in all other cases. The following two modules of the tool further filter false positives included in the Crawler (M1) search, and extract information from the correctly identified webpages.

3.2. Email Extractor (M2)

The output file generated by M1 is the input for the Email Extractor (M2). The extraction of email addresses has two objectives. First, email is used to administer the survey in Module M4 to collect further information on family and children, which usually is not included in CVs, and to request a copy of the scientist's CV. Second, an email address on the webpage allows further validation of the accuracy of the academic webpage identified in M1.[6]

Email identification is based on a search of email-like structures within the websites obtained from M1. To select the correct email address a library of email address models has been developed. The patterns vary because each university can decide the structure of the email addresses allocated to its employees. Table 1 shows the different email address models ; the user can supply the standard academic extension used in the country analysed, such as "*.ac.uk", "*.es", "uni*.it" and "*.edu", in the cases of the UK, Spain, Italy and the US (regular expressions

[6] Though locating the email address on a webpage reduces the probability of a false positive, it does not insure that all the pages identified are correct. It is a necessary but not sufficient condition for having found the right personal web page; correct mail addresses can be displayed in different locations such as university news, abstracts, laboratory homepages, etc. Although in our manual testing (see following section) we achieved over 95% positive results.

can be used to limit the search space). Score “A” is assigned to email addresses having some combination of name and surname. Score “B” is assigned to email addresses having combinations of initials and numbers and truncated surnames. Score “Z” is assigned to general institutional email addresses that do not report a trace of name and/or surname.

Table 1: Email address models and scores

A	B	Z
john.smith@####	jis@####	webmaster@####
smithjohnc@####	sji@####	web@####
smith.j.john@####	js@####	mail@####
smith.john@####	sj@####	wmaster@####
johnsmith@####	jis7r@####	contact@####
smithj####	r7jis@####	webfeedback@####
smith@####	**jis**@####	info@####
msmith@####	*jis*@####	HelpDesk@####
john.s@####	jis*@####	help@####
johno@####	*jis@####	desk@####
jjohnsmi@####		*helpdesk*@####
jsmith@####		staff@####
smithjj@####		publicaffairs@####
smithj@####		public@####
		public@####
		neurobiology@####
		#department@####
		#university@####
		#field@####
		support@####
		support@####
		news@####

These patterns represent an initial standardisation to the email address extraction approach. This library can (and should) be complemented by additional patterns to enable exhaustive recognition of all the different possibilities.

We use the intermediate output of M2 to validate the webpages collected through M1. Using the scores from M1 and M2, the system is able to re-rank the webpages according to the probability of their being a correct match. Table 2 presents the website and email address scores that, combined, decide acceptance or rejection of the webpage.

Table 2: Example of webpage selection criteria

Website Score	Email Score	Final Evaluation
A	A	AA, AB, BA: Accepted
B	B	BB: to be manually checked
	Z	BZ, AZ: rejected

The “accepted” group (“AA”, “AB”, “BA”) is the final validated output of M2. The output of M2 is a CSV file containing the web addresses and email addresses identified, along with their scores, which form the input for M4 (survey) and M5 (text analyser).

3.3. CV and webpage information extractor (M3)

Module M3 consists of two components, *CV and Webpage Extractor* and *CV File Searcher*, which identify and download URLs and pdf (doc, docx, rtf) files containing curricular information.

CV and Webpage Extractor

The results of M2 form the input for this component, which has two subtasks. Firstly, for each URL in the CSV input file, the CV Extractor navigates the links searching for "cv", "curriculum", "vitae", "cvitae", or "curricula" and downloads the webpages identified, if any. It searches similarly for "pubs" and "publications", and downloads the identified publications pages. Finally, it downloads the home page URL. Secondly, for each URL in the CSV input file, the CV Extractor navigates the home page and all its sub-links searching for files (with "pdf", "doc", "docx", "rtf" extensions) with any of the previous words (cv, curriculum, vitae, vitae, cvitae, curricula, pubs, publications) in their titles, which if found are downloaded. The output of this subtask consists of one zip folder, containing the above mentioned files, and two CSV files, one reporting the URLs of homepages, CVs, and lists of publications, and the other reporting the downloaded filenames within the zip folder.

CV File Searcher

Since researchers' CVs may be found on other webpages such as old webpages, conference sites, pages of organisations where the researcher acted as an advisor, etc., we implemented an additional CV file search not limited to the web addresses found in M2. This component searches the Internet using a pattern very similar to that employed in M1. It searches for researchers using: <full name> + <research field> + "pdf/doc/docx/rtf" + "(cv OR curriculum OR vitae OR cvitae OR vita OR curricula)". The output is two files: a CSV with the URLs of the documents found, and a CSV with a list of those researchers whose CVs were not found. In the former CSV file, the list of potential CVs is qualified by two scores: "A" if the URL of the CV or its filename

contains a keyword referring to the name, surname or initials of the researcher plus a CV-denoting term (such as cv, curriculum, etc.), and "B" if it does not contain these keywords.

3.4. Email survey (M4)

The email addresses identified in M3 are used to implement an email survey. The survey is designed to collect CVs and to gather personal information such as marital status and the number and age of children, which are not normally included in CVs or on webpages. This further information is useful for econometric analyses and gender studies. The email survey consists of two components: the *Email sender* and the *Online-form with a CV repository*. In the current version of the tool we have not included the specific components described below as there are a number of different approaches to surveys that different users may prefer.

Email sender

We used a simple email client able to manage comma separated lists, and send emails in an automated way. This method, although standard, should be implemented with care. Email server restrictions are common in email services provided by research institutions (such as universities). Hence, implementation of a method to avoid outgoing email congestion is a significant issue. Also, the email containing the survey should be delivered to the researcher's email "inbox" and not into a "spam" folder. Several open source email programs are able to work under these constraints; most require an email-server to be set up on the sending machine. Alternatively, a client-side solution can be implemented. We used Mozilla Thunderbird with the "Mail-Merge" extension. This program has the capability to handle CSV lists and schedule outgoing mails. The input database, the output of M2, is automatically edited to contain the email address and name

of the researcher, using a specific architecture that allows them to be inserted as part of the body of the email text (which can be modified by the user as needed) in CSV format, with header columns in the form “FirstName”, “LastName” “Email”.

Online form with a CV repository

The second component includes the form which the respondent must access to respond to the survey. A web front-end has to be set up to present the questionnaire for response, and a hosting service is required to handle the information received and to save the uploaded data (CV files and information requested in the questionnaire). In the surveys we conducted, the implementation allowed respondents to send the information by replying to our original email and attaching their CV and responding to some short questions. Although potentially simpler for some respondents, this implementation implies an additional step to codify the information received. The temporary online database (that allows uploading of files to a server) is a faster, simpler, and easier process, and would avoid additional codification of email responses.

The final output of M4 is a CSV file with the scientist’s ID, CV information transformed into plain text, and personal information collected via the online form.

3.5. Text Analyser and Codifier (M5)

This module extracts information from curricular and other HTML sources identified in the previous steps. Its inputs are the CSV files from M3, from both the *CV extractor* and *CV*

searcher, as well as from M4.[7] It is also able to accept other external files containing career information that may be produced either manually or by automatic processes such as the new OECD software for the creation of career information from SCOPUS (OECD, 2013).

This is the most complex module of the SiSOB tool and comprises two main components. The first is fully developed, the second requires a crowd source approach to complete its implementation and increase its performance, and to reduce the human time requirement for the codification process. The first component splits the input file into four main content blocks: Personal Data, University Studies, Professional Activities, and Publications, generating four CSV output files with the information organised by line. These files can be used for manual codification of the information, and as inputs to the second component which performs automatic analysis, codification, and extraction of information. Splitting the CVs into four blocks reduces the chance of false recognition of specific curricular terms in the automatic codification process.

The first and second components are implemented using the open-source software, GATE.[8] GATE needs to be properly customised with a set of dictionaries to provide the extracting engine with the appropriate semantics. The dictionaries consist of a set of semantic expressions, which may either be provided if available or generated as a result of an iterative process in which the user evaluates the output of M5 and feeds back the errors so that GATE "learns" what to search for, and how. For structured databases, these dictionaries can easily be built in advance; however, unstructured or semi-structured data require a long process of testing and validation to produce a comprehensive dictionary for the data extractor. This process results in the creation of

[7] Current version of the tool requires manual preparation of the CSV input files in line with M5 minimum requirements, for example: "ID";"LASTNAME";"INITIALS";"RESEARCHER_PAGE_URL". For more details see information and examples on the web application.

[8] GATE is a tool for processing information involving human language developed by computer scientists at the University of Sheffield (Cunningham et al. 2011).

a set of dictionaries containing expressions and words corresponding to the different variables in the final output files such as job positions, organisations, scientific disciplines, funding agencies, grants, universities, countries, provinces, etc. During development of the SiSOB tool a large set of patterns for curricular expression detection were created and written in JAPE[9] (a component of GATE) and now are part of the Text Analyser and Codifier. Specifically, we built dictionaries for the extraction of data from the CVs of US and UK biomedical scientists (the case study discussed in this paper) and for UK engineers and natural scientists. All dictionaries will additionally be available to allow users to further contribute to their development and improve the performance of this module[10].

The current implementation of the Text Analyser based on the JAPE files splits the input file into the following categories:

- *Personal data*: gender, nationality, birth city, birth region, birth country, birth date, email, phone number, etc…
- *University studies*: field, qualifications, institution, city, region, country, etc.
- *Professional*: position name, start date, duration, institution, city, region, country, etc.
- *Publications*.

The Text Analyser then analyses each textual block and automatically codifies the text on the basis of the JAPE files, into a standardised data output. The current version of the second component of M5 is preliminary and a work in progress, and will require further development of

[9] JAPE allows the definition of templates that are used in the codification phase to identify the curricular items inside pieces of text from which the information is being extracted.

[10] Extensions of the project to other countries or scientific fields will require the development of new sets of dictionaries. For the full list of variables extracted and dictionaries see: https://github.com/eduardoguzman/sisob-data-extractor/tree/master/gate-data-extractor-service/GATE-6.0/plugins/annie/resources/gazetteer

both the dictionaries and the routines to properly codify the information. The current outputs are four CSV files, one for each main curricular category, to facilitate visualisation of the results and allow users to provide feedback to the system and fix errors in the detection of curricular items.[11]

The main obstacles to automatic codification are the semi-structured format of CVs and webpages, and the variance in the terms used in CVs. The more structured the CVs and the more comprehensive the dictionaries supplied, the higher the success of the codification process. Due to the variety of career positions and their descriptions in CVs (for instance, sabbatical leave, visiting period, secondment, etc.) automatic codification of professional activities is currently quite problematic, although we have obtained interesting results for the other curricular blocks and especially for the identification of publications (see further discussion of the validation on the tool in section 4).

3.6. Publication Harvesting

Information on published output can vary significantly across CVs and webpages, for example some researchers include only their most recent publications, others include their most cited works, while a minority include all publications. In order to collect complete information, the output of the SiSOB tool can be complemented with information on publications collected directly from publication databases. In the case of biomedical researchers we used Publication Harvester (henceforth PH), an open-source software that automates the process of gathering publication information for individual scientists (Azoulay et al., 2006). PH searches for

[11] In accordance with European General Data Protection Regulation, these output files need to be downloaded to the user's server and anonymized by deleting researchers' names, retaining only the ID. The user needs to delete these files from the SiSOB server as well the output files of previous tasks.

publications in Medline,[12] for example, based on the scientist's name, surname and research field. PH combines a range of algorithmic searches to build datasets stored in a CSV-formatted output file that can be fed to the Text Analyser for integration with web and CV information.

During the process of generating publication counts at the individual level, the researcher is faced with the problem of uniqueness and accuracy of scientist's names.[13] Various approaches are possible. First, the problem could be ignored since name frequency should be orthogonal to other determinants of scientific productivity (Azoulay et al., 2006). Alternatively, one could exploit common name frequencies to weight the obtained results. Finally, more advanced search algorithms could be used employing the information gathered with the SiSOB tool; for example publication time span (after MA and before death), or scientific headings from html webpages and/or CVs, could be used to implement a more restrictive search. Information collected from the publications extracted in M5 can also be used to improve accuracy.

4. Testing and validation of the SiSOB tool

The validation of each module was performed on a sample of 9,903 biomedical researchers in the US who received funding from the National Institute of Health (NIH) and 2,426 biomedical researchers in the UK who received funding from the Biotechnology and Biological Sciences Research Council (BBSRC) (see Section 5 for more details on the construction of the samples).

[12] http://www.pubmed.gov/

[13] Frequently used names make difficult to disambiguate authors, while for unusual names use can be inconsistent.

M1: Crawler

We run a validation procedure for Patterns 2 and 5, which use the Internet as the target of search,, comparing a broader search (Pattern 2) with a more restrictive one (Pattern 5). Table 3 presents the results of the Crawler extraction. The proportion of "A" scores (surname given at the URL address) differs only marginally between the two patterns. The Crawler, through Pattern 2, scored around 70% of our US sample and 62% of our UK sample as "A". On the other hand, using Pattern 5, M1 scored 67% of the US sample and 63% of the UK sample as "A". Summing scores "A" and "B" (no surname given in the webpage) the tool was able to retrieve websites for 99% of researchers[14].

Table 3: Validation Results for M1

Pattern	Data	Input	Score A	%	Score B	%	A+B	%
P2	US	9903	6967	70,4%	2912	29,4%	9879	99,8%
P5	US	9903	6668	67,3%	3166	32,0%	9834	99,3%
P2	UK	2426	1518	62,6%	893	36,8%	2411	99,4%
P5	UK	2426	1537	63,4%	885	36,5%	2422	99,8%

M2: Email Extractor

As a validation of the Email Extractor we consider the sum of scores for M1 and M2: "AA", "AB", and "BA". Validation of personal webpages reporting "BB" scores must be done manually.[15] Table 4 shows the results of the email extraction from Patterns 2 and 5. The SiSOB tool identifies a webpage and email address for about 40% of our sample. Interestingly, after subtracting "BB" scores, the final set of potentially correct personal webpages ("AA"+ "AB"+ "BA") is higher for the more restrictive Pattern 5 than for Pattern 2.

[14] At this stage of the process the authors discourage users to discard webpages with score B as the identified page may be correct even if the surname is not automatically detected. See limitations for further details.
[15] Score B is assigned to emails containing a combination of initials and numbers, as well as to truncated surnames, making it difficult to achieve a complete set of possible email address models.

In order to validate the automatic approach used to score pages, we did a manual check of 500 random personal webpages of UK researchers resulting from M2, based on the M1 Pattern 5 search. 95.5% of the webpages checked were correct personal home webpages.

Table 4: Validation Results for M2

Pattern	Data	AA	%	AB	%	BA	%	BB	%	AA+AB+BA+BB
P2	US	3591	36,3%	227	2,3%	215	2,2%	98	1,0%	36,3%
P5	US	3587	36,2%	215	2,2%	185	1,9%	106	1,1%	41.3%
P2	UK	781	32,2%	45	1,9%	88	3,6%	13	0,5%	38,2%
P5	UK	924	38,1%	48	2,0%	76	3,1%	11	0,5%	43,7%

M3: CV and Webpage Information Extractor

Module M3 has two components - *CV and Webpage Extractor* and *CV File Searcher*. The first component performed well in successfully downloading the whole set of home pages identified in M2. In terms of identified publications pages, for the US sample, the CV extractor was able to retrieve 2,183 lists through Pattern 2 and 2,183 through Pattern 5, and for the UK sample, 553 through Pattern 2 and 457 through Pattern 5.[16] The success rate for CV download differs widely between the two samples. For the US sample the tool retrieved 232 (Pattern 2) and 226 CVs (Pattern 5), for the UK sample it was able to detect and download only 4 CVs (Pattern 2) and 1 CV (Pattern 5).[17] In most cases in the UK, curricular information was included in the text of the home page rather than as a CV file or link, which may explain the very poor result.

The second component, *CV File Searcher*, was able to retrieve 5,025 CVs of US researchers (1,459 with score “A” and 3,566 with score “B") and 1,156 CVs of UK researchers (223 with

[16] The total list of publication do not match the total output form M2 because most of the hompages already include the list of the publication. Moreover these results take into account both publication pages with score A and B.
[17] These figures refer to any type of CV format including pdf, doc, docx and rtf with score AA and BA.

score “A” and 933 with score “B”). However, as expected, due to the very broad search pattern (the institution is omitted here) and because only one of the two modules provides an input (CVs are assigned with “A” or “B” according to the presence of the recipient surname within the file name), results are not as robust as in the previous cases where we were able to assess the quality through both M1 and M2 scores. For example, in the UK case, among the “A” scored CVs, 209 are false positives[18] and we were only able to retrieve 14 correct CVs.

M4: Email Survey

The email addresses collected by M2 (and through manual searches in the case of the BBSRC) were used to survey the NIH and BBSRC researchers. The email survey showed a delivery success rate of 95% in the case of the NIH sample and 92.9% for the BBSRC sample (i.e. a 7.1% email delivery failure rate). The high delivery success rate confirms the output quality of M2. The 7.1% failures may be linked to identification of an outdated webpage.

M5: Text Analyser

Due to the ongoing development of the third component in the Text Analyser it is difficult to provide robust validation of current performance. Our expectation, and most important reason for publishing this paper, is that the contributions of users and developers will allow us to develop a much better performing module M5. The version of M5 at publication time was able to correctly handle large sets of CVs and/or webpages and to correctly split the files into homogeneous sub blocks. Furthermore, the codification of personal data (if available) such as address, contact information, marriage status and date of birth was quite successful. Finally, after converting CVs

[18] Files with score “B” are not CVs or do not contain the surname in the filename. Files with score “A” have the surname in the filename but do not belong to the correct researcher.

from PDF to RTF format, the tool is able to correctly retrieve publication data from CVs and other web sources. Identified publications are reported by row in the output CSV files.

5. The NIH and BBSRC case studies[19]

Our case studies examine the mobility and career patterns of researchers funded by the NIH and the BBSRC. The NIH is the leading funding agency for academic research in biomedicine in the US. NIH grant award data cover grants awarded by the NIH since 1970 and include personal identifiers (ids) for principal investigators (PIs) since 1985. They also provide information on university and subject affiliation for all funded researchers. Research project grants (R01) are assigned to around 230,000 PIs which include university researchers as well as researchers from NIH institutes and industry. We limited our sample to researchers that received at least one R01 grant during the period 2001 to 2010 and were working for a university at the time of grant award. We further limited the sample to academics that worked at schools of medicine, arts and science, graduate colleges or schools of engineering and in departments of biology, chemistry, neurology, genetics or their sub-fields at the time of grant award. This left us with an initial sample of 10,221 PI identifiers. NIH information on names, institutions and subject areas was used as input for the SiSOB tool.

BBSRC data cover grants awarded by the BBSRC, the leading funding agency for academic research and training in non-medical bioscience in the UK, from 1994 to 2010, and include personal identifiers for 7,527 researchers. The database includes both PIs and Co-investigators (Co-I). We limit the sample to those researchers that received at least two grants during the period 1994 and 2010, resulting in a list of 3,615 researcher IDs, which include academics but

[19] The data used for these case studies were collected up to May 2014, while the most recent version of the tool was used to perform the validation in December 2014. This explains discrepancies in numbers of researchers and CVs between sections 4 and 5.

also researchers working in industry and public research laboratories. In order to gather more thorough and up-to-date information (the most recent grant received by some researchers was in the 1990s) and to identify academics, we cross-referenced these researchers with the 2008 Research Assessment Exercise (RAE).[20] RAE 2008 includes a comprehensive listing of all research-active staff in all UK universities, for 2007. Amongst the 3,615 researchers that received at least two BBSRC grants since 1994, we identified 2,426 submitted to RAE 2008 by their university departments. Thus, they could be identified as working at a UK university in 2007. The RAE database contains the researcher's name, university and discipline in 2007, information which was used as input for the SiSOB tool.

SiSOB Tool

To collect CVs for the BBSRC sample, we firstly collected email addresses manually to support the development of the SiSOB tool and gathered valid email address for all 2,426 researchers. For the NIH sample, we utilised the SiSOB tool. We crawled the personal web pages of researchers (M2) and identified 4,037 valid email addresses, representing 40% of the original sample. All researchers were surveyed (M4) to ask them for their CV and additional personal information (family situation, nationality). The BBSRC survey consisted of nine rounds, from September 2011 to January 2014, resulting in 296 (12.2%) complete CVs. The NIH survey was conducted in five rounds from October 2013 up to April 2014 and resulted in 169 valid CVs (a response rate of 4%).

[20] The RAE was the UK evaluation exercise conducted by the UK Higher Education Funding Councils, to measure the quality of the research activities undertaken at UK universities and determine funding allocations for the succeeding 5 to 7 years. The RAE was replaced by the Research Excellence Framework (REF) in 2014.

We then directly crawled the web for researchers' CVs (M3). For the BBSRC this process resulted in 13 correctly identified CVs. The final UK database then consists of 309 CVs, corresponding to a response rate of 12.7% from the initial set of 2,426 academics. For the NIH sample, this process resulted in 215 correctly identified CVs.[21] The final set of US CVs consists of 384 entries representing 3.8% of the initial population of 10,221 academics.

Response Bias Analysis

To test for non-response bias we used the institutional composition of the full population and the sample of respondents. For the NIH sample we rely on university affiliation as an indicator involving a number of dimensions such as geographic distribution, size and institutional quality of the two groups. The analysis of institutional distribution revealed a total of 309 universities in the full population and 135 in the respondents' sample (42%) which account for 80% of the most important institutions in the full population. In order to formally address the representativeness of the sample we used the Wilcoxon Rank Test. As a result, we found a significant match between the two distributions (population and respondents) with a 5% degree of tolerance, suggesting that the sample is not significantly different from the total population. To address additional concerns over sample bias at the individual level we compare the distribution of subject areas, number of years actively involved in NIH sponsored research and number of grants in the full population and the sample population. We perform Kolmogorov–Smirnov tests of the equality of distributions and find that there is no significant difference between the years of grant activity in the respondent sample and those that did not answer (15.93 vs. 16.01 years since first grant[22], $\rho = 0.539$). However, we find some difference in the number of grants (2.7 vs.

[21] The most recent version of SiSOB tool has a better performance in the identification of NIH CVs.
[22] Years active are truncated at 28 years (=2013-1985).

2.8 grants, $\rho = 0.003$) and in the field distribution ($\rho = 0.035$). As a robustness check, we test the hypothesis excluding the field of chemistry, which has the highest response rates, and no longer find significant field differences ($\rho = 0.177$). Among our respondents 76% are life scientists and 24% are chemists (compared to only 10% chemists in the original population).

The BBSRC-based sample allows us to conduct different kinds of tests for sample representativeness based on RAE and BBSRC information. Academics in the full sample population come from 81 universities and respondents from 52 (64%) which account for 80% of the top institutions. Again we find no difference in population based on universities represented (Wilcoxon Rank Test). We also compare the distribution of the amount of funding received, grant numbers, years actively involved in BBSRC sponsored research, and subject areas by the full and the sample populations (using Kolmogorov–Smirnov tests). We find no significant difference in grant value (£1.82 million vs. £1.78 million, $\rho = 0.352$) or in grant numbers (5.2 vs. 5.1, $\rho = 0.491$) between the respondent sample and those that did not answer. There is a small difference in years since first grant (12.7 vs. 13.3, $\rho = 0.040$) with the respondents sample being slightly younger than non-respondents. We find no differences in the subject area distribution ($\rho = 0.763$).

CV Codification

CVs were coded by hand supporting the development of the M5 Text Analyser particularly in relation to the development of dictionaries and validation of results.

Personal details, education history, and career paths up to 2012 were recorded. We excluded academics with incomplete career data and those who had retired within the five years prior to 2012. For the NIH sample this leaves us with 331 researchers with complete CV information and

29 researchers with partial information (e.g. missing birth or education information), and for the BBSRC with 277 researchers with complete and 14 researchers with partial information.

Journal publications were collected from the Medline database using PH (Azoulay et al. 2006). The Medline database includes bibliographical information for articles published in the life sciences and biology. We collected publications for all the academics in our sample. Those with common name-surname combinations and those with Asian last names were excluded and publications reliably collected for 297 NIH academics and 244 BBSRC academics.

5.1 A brief analysis of mobility and productivity

Demographic information

The descriptive statistics for all 360 NIH and 291 BBSRC researchers are displayed in Table A1 in Appendix A. For the US sample in 2012, researcher average age is 55; 21% of researchers are women; 76% of those that reported their nationality were US citizens; 79% studied for a BA degree in the US; and 88% were awarded a PhD degree by a US institution (average year of PhD award is 1986). The average NIH researcher was appointed to a first tenure-track position in 1989. In 2012, 71% of researchers had the rank of professor and 1% had left academia.

For the UK sample, the average age is 52; 22% are women, 76% of those that reported their nationality are UK citizens; 83% studied for a BA degree in the UK; and 84% were awarded a PhD degree by a UK institution (average year of PhD award is 1987). The average BBSRC researcher was appointed to first permanent position in 1992; in 2012, 68% had the rank of full professor and 1% had left academia. The two samples thus show some striking similarities.

Mobility

Based on CV information we can reconstruct the mobility paths of researchers from career start until 2012. On completion of their PhD studies, 76% of NIH researchers did a postdoc, 14% of them outside the US. Following appointment to the first tenure-track or tenured post, 52% were mobile - 57% of which moved just once, but in 14 cases there were four or more moves. Most mobility is between universities: 44% of researchers move between higher education institutions, and moves are accompanied by promotion in 38% of cases. However, 12 researchers moved from industry into academia or public research organizations (PROs), while 23 researchers moved from PROs into academia, or vice versa. 17% of researchers moved to the US from abroad after completing their postdoc (9%) or after having been in a full job position for a few years (8%). Most job mobility is between US institutions and 44% of researchers moved between US universities at least once.

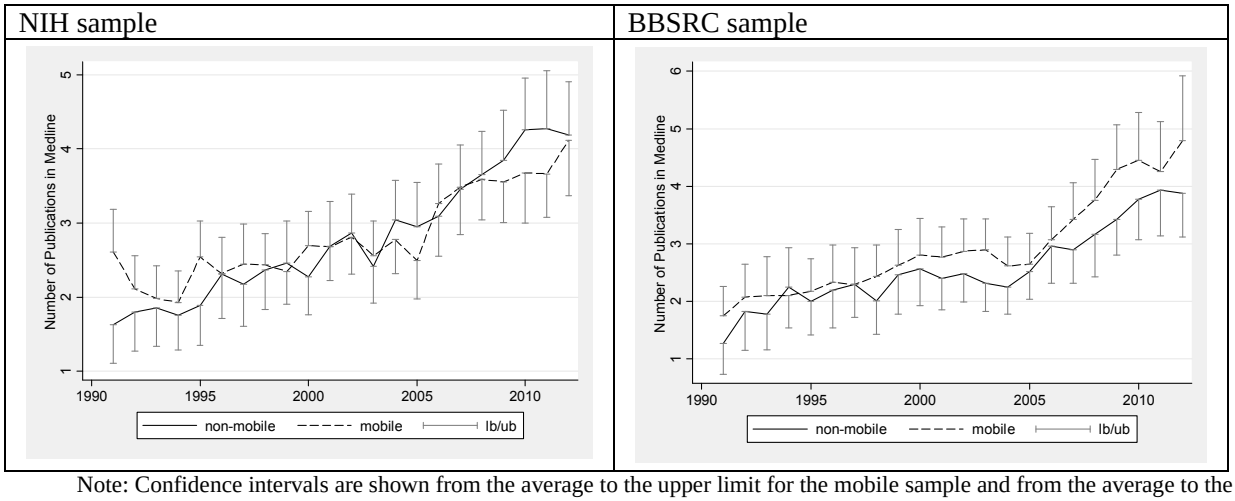
For the BBSRC sample we find that 84% undertook a postdoc following their PhD education, half of them outside the UK. After first permanent appointment, 52% were mobile, the same as in the NIH sample, with 56% moving just once. Most mobility is between universities (77%) but 16 academics move to academia or public research from industry and 29 researchers from PROs into academia. Mobility between universities is accompanied by promotion in 45% of cases. Just as mobility in the US sample was primarily between US universities, most mobility in the UK sample is between UK universities (30% of BBSRC academics move between UK universities at least once). International mobility is more widespread in the BBSRC sample compared to the NIH sample. 49% of researchers move to the UK either after their postdoc elsewhere (33%) or after attainment of their first permanent position (16%). The majority of these internationally

mobile academics (78%) did their PhD in the UK, but spent some years abroad, before returning to the UK.

Productivity and mobility

To exemplify the type of analysis enabled by the data collected using the SiSOB tool we look at the link between publications and mobility for the 297 NIH and 244 BBSRC academics for which publications could be reliably identified. On average, these researchers produce 2.9 publications per active working year. We also recorded co-author profiles and found that the average researcher has 4.2 co-authors.[23]

Figure 3: Performance of mobile and non-mobile researchers



Note: Confidence intervals are shown from the average to the upper limit for the mobile sample and from the average to the lower limit for the non-mobile sample. Other limits are now shown to improve the visualization of yearly averages.

We first compare the publication histories of mobile and non-mobile researchers, considering any type of mobility. We limit the analysis to papers published since 1991 since the number of researchers already active before 1991 is very small. Figure 3 shows that NIH researchers that move at least once during their careers do not perform better than those that never move;

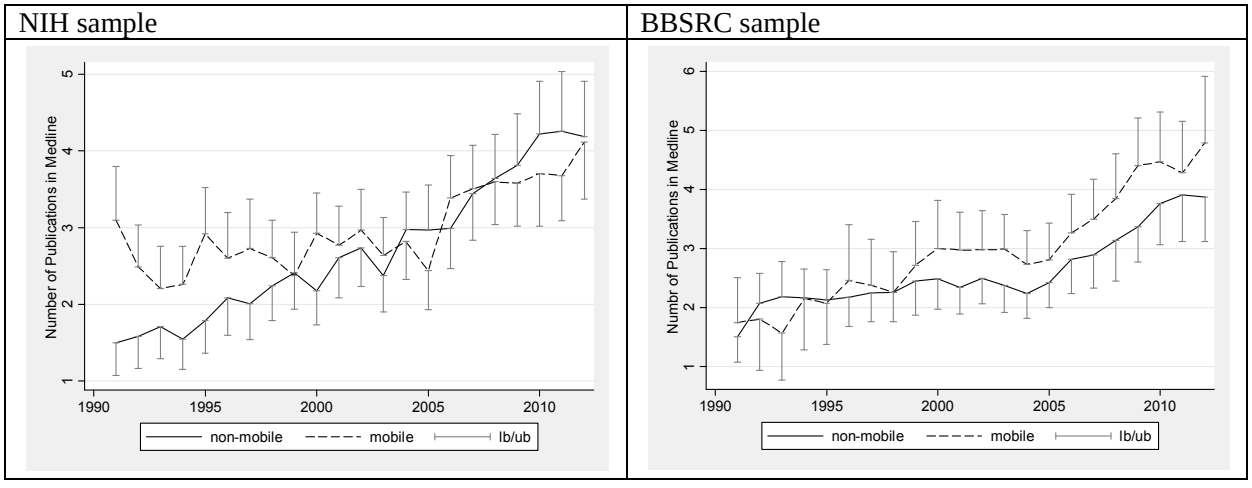
[23] Publication and co-author numbers are the same for both samples.

however, in the BBSRC sample we do find a significantly higher number of publications for mobile academics in recent years.

Researchers coded as non-mobile may move in future and thus become mobile researchers. Figure 4 therefore adds the pre-mobility observations of mobile academics to the group of non-mobile researcher. It shows that for the NIH sample the average number of publications published by mobile researchers is above the average for non-mobile researchers in the early 1990s. However, from 1998, we no longer find a significant difference. In the BBSRC sample the opposite is true. While we find no difference in performance during the early 1990s, from 1998 onwards, mobile researchers outperform their peers. This development follows the 1996 RAE and may be directly related to the assessment of researchers and their departments which has increased the incentives for both mobility and research output.

The number of publications by non-mobile and mobile researchers in both samples has increased significantly since 2005, but the increase has been strongest for mobile BBSRC researchers.

Figure 4: Performance of mobile and non-mobile researchers with pre-mobility years considered as non-mobile



The lack of a clear difference in performance may arise because we conflate different mobility types. As Ferndandez-Zubieta et al. (2015a) argue: different mobility types affect performance differently due to varying opportunity and mobility costs. In a highly competitive research labour market, such as the US and the UK, job mobility can be expected to be driven by research-related motivations. Once researchers are granted a permanent position, job mobility can be expected to happen voluntarily and institutional selection to be better informed. Information on research-related factors will be visible for researchers in permanent positions, and both individuals and institutions would be better able to take informed mobility and hiring decisions. We could therefore expect mobility to only increase performance in cases where academics move voluntarily between higher education institutions (Fernandez-Zubieta et al. 2015b). Voluntary mobility is defined as a move after an academic is granted a permanent (tenured) academic post. In the UK, assistant professorships are considered permanent positions subject to a three year probation period. In the US, assistant professorships are tenure-track positions and, thus, not permanent. If an academic moves before achieving associate professor status, this is considered forced mobility.[24] These differences in the academic markets result in significant differences in the number of forced and voluntary moves in the NIH and the BBSRC samples. Amongst the BBSRC sample, 35% of academics move voluntarily to a different HEI, while only 7% are forced to move.[25] In the US, 24% change jobs while holding a permanent position, but 26% move while holding a fixed term position.

[24] Of course we cannot rule out that non-tenured mobility also happens voluntarily or that mobility of senior staff is forced. For example, in the UK, academics could be forced out of their permanent position through restructuring.
[25] Some of the forced mobility happens when academics are working abroad (for example in the US). In the UK forced mobility is only measured if academics are on teaching fellow contracts that are non-permanent.

In Figures 5 and 6 we compare the publication performance of voluntary mobile academics to that of all other researchers in the sample.[26] The graphs for NIH researchers show a clear difference in performance between those that move voluntarily and the others. Voluntarily mobile academics publish significantly more for all years in the sample. The difference is even more pronounced if we count the pre-mobility observations as non-mobile. For the BBSRC sample the graphs look similar to those in Figures 3 and 4, with a better performance of voluntarily mobile academics only for years since 2007 (ρ<0.001). If we limit the sample to those that never leave academia and differentiate forced mobile from non-mobile (Figure A1 in Appendix A) then we find for the US sample, that forced mobile academics publish least and the difference between voluntary mobile and non-mobile is smaller, but still significant. For the BBSRC sample we still find a significantly higher performance for voluntary mobile compared to non-mobile academics for the years since 2007 (ρ<0.001), but there is no difference between forced and voluntary mobile academics, perhaps due to the very low number of forced mobile academics (17 researchers).

As we are examining two different types of academic system with very different definitions of voluntary mobility, it may be more appropriate to consider the performance of academics that are mobile at associate professor level or higher also for the UK sample. This addition does not change the results significantly.

Overall, the graphs presented here show that it is very important to take account of the mobility type when looking at performance differences.

[26] Graphs are similar if we exclude those that have been mobile outside academia.

Figure 5: Performance of voluntarily mobile researchers compared to others

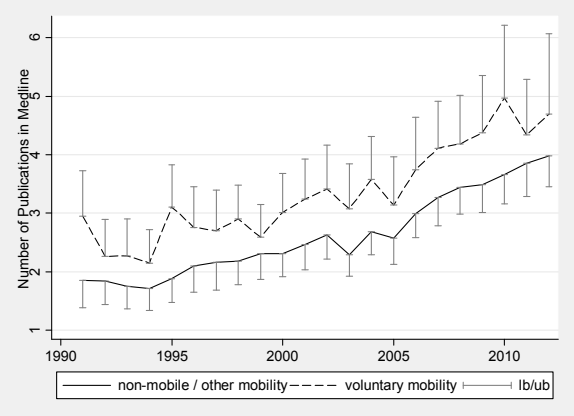
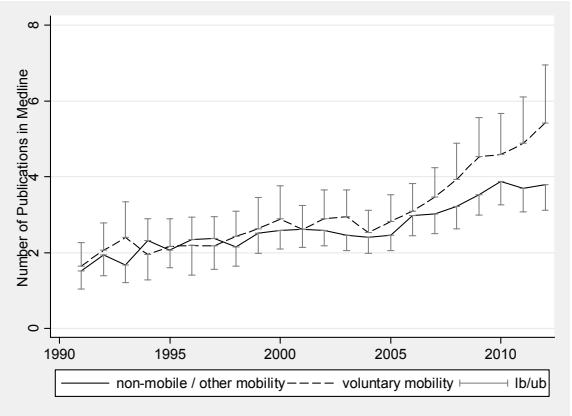
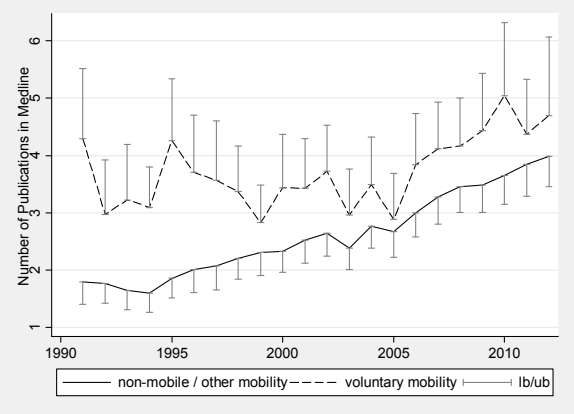
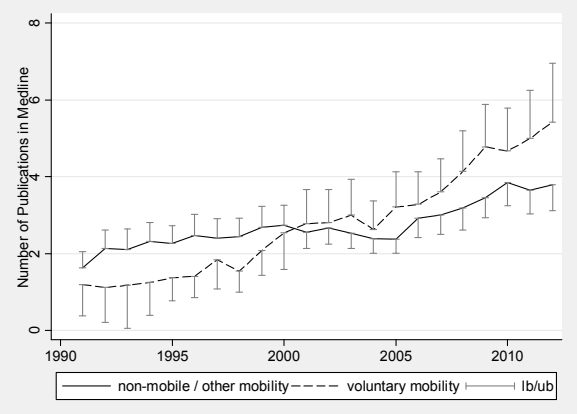
NIH sample	BBSRC sample
	

Figure 6: Performance of voluntarily mobile researchers compared to others with pre-mobility years considered as non-mobile

NIH sample	BBSRC sample
	

Finally, we look at the performance of researchers in the years surrounding a mobility event. Figures 7 and 8 show average publication numbers over the six years before and after the move for all types of mobility (including sector mobility) and for voluntary higher education mobility only, and the average productivity of the average non-mobile researcher in a similar 12 year window (accounting for average publication increase over the observation period). In the NIH sample we see that on average mobile researchers are less productive surrounding the years of

the move compared to non-mobile. However, if we consider voluntary mobility separately, we find that they have a performance surge the year before the move, followed by a decline in performance in the year of the move which is only recovered in year three after the move. By year four they outperform all non-mobile researchers.

Figure 7: Pre- and post-mobility publication numbers of NIH researchers

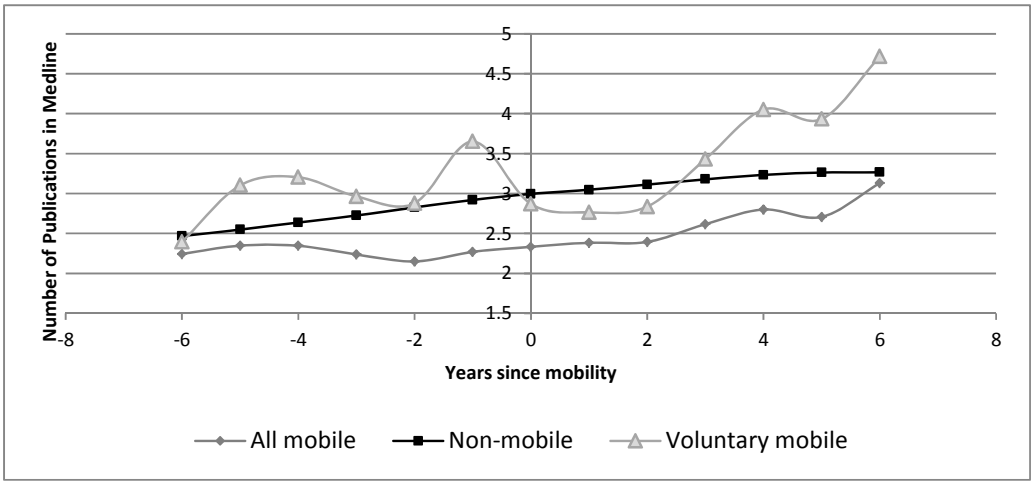
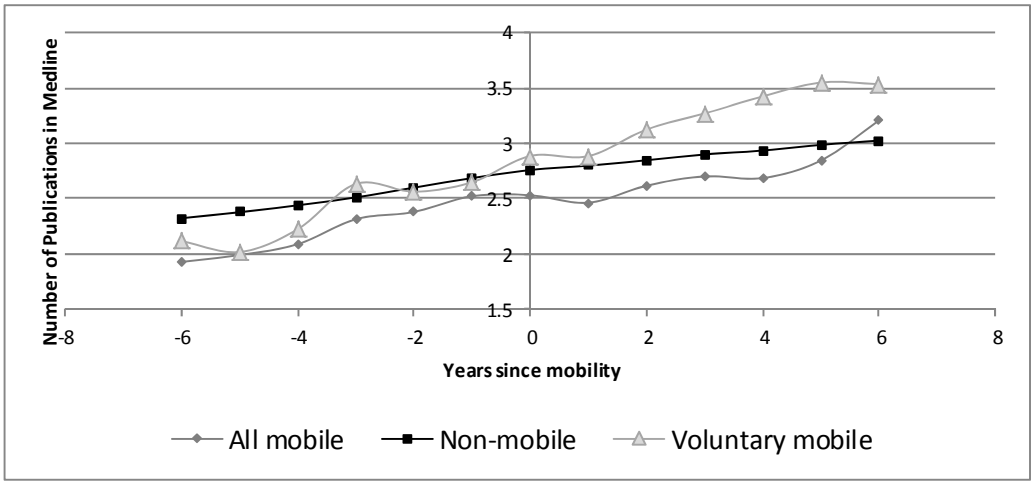


Figure 8: Pre- and post-mobility publication numbers of BBSRC researchers



In the BBSRC sample (Figure 8), the publication performance trend prior to the mobility event is very similar for the average mobile and non-mobile researchers. Mobile researchers then

experience a slight drop in productivity in the year directly following the move, after which publications increase. Voluntary mobile already outperform non-mobile researchers from year two after the move. Other mobile researchers experience a slower increase but catch up with the non-mobile by year six after the move.

Overall this indicates the importance of considering different types of mobility when investigating the relationship between mobility and performance (Fernandez-Zubieta et al., 2015a). It also shows how CV data can be used to identify the exact point of mobility, which has been shown to be related to performance. Other types of mobility qualifiers may hold very different results. For example, Fernandez-Zubieta et al. (2013) differ between upward and downward mobility and find that only those that move to a better institution increase their performance.

6. Conclusions

The SiSOB Data Extraction and Codification Tool aims to provide a system for collecting and structuring information on scientific researchers from publicly accessible websites, and complementing them with CV information. We tested our tool on a sample of biomedical researchers in the US and UK with very satisfactory results. Much work is still needed to accomplish a fully automatic procedure for data collection and codification but the current state of development of the tool makes an excellent starting point. We are specifically calling for help in developing the last component of module M5, the Text Analyser, which tries to develop heuristics to extract reliable information on scientists' careers, from CV and webpage information collected by the previous modules.

The SiSOB tool was released under GNU General Public License v3 in the GitHub repository making it available for future improvements and developments. In addition the current first release of the system and successive updates will be available to users at the SiSOB website server http://sistractor.lcc.uma.es/extractor and its mirror at University of Torino http://sisobcv.unito.it.[27] Users in economics, sociology, and other social sciences are encouraged to use the tool, report any problems, and provide feedback. The next release is to include some form of automatic feedback, including for the creation and archiving of country-specific dictionaries.

The paper provides a brief example of the kind of information that can be extracted from the structured final data files produced by the SiSOB tool, and offers a first brief case study of its application to a sample of biomedical researchers in the US and UK. Our example shows that the data allow analysis of the career trajectories of researchers and investigation of the interactions between mobility and publication and general career development. This research area has thus far been neglected due to the problems involved in identifying and collecting reliable relevant data. We show that CVs are a valuable source of data to identify the exact point and type of mobility, information which is related to academic performance. Indeed, in the analysis of the mobility-productivity relationship we show that it is important to differentiate between voluntary and forced mobility, the former being the mobility of researchers with a permanent/tenured position; only voluntary mobility is associated to higher research performance (especially in the US).

The algorithms used to locate, extract, and structure information from personal und university websites will be useful not only for measuring the link between mobility and publications but

[27] To request an account contact either eduardo.guzman@lcc.uma.es or aldo.geuna@unito.it.

also for investigating the innovation capacity of individuals as well as scientific research social impact. For example, the algorithm can be used to track researchers' footprints on the web and evaluate their general visibility. It can also be exploited to measure the importance of scientific fields or inventions, going beyond traditional publications and patent measures and employing context-based web analysis.

The paper shows that the SiSOB automated searching and codification tool (an example of big data management tool) is a very promising and powerful tool for building more comprehensive databases (with longitudinal and cross-sectional information at individual level). However, automated searching and codification algorithms have limitations. For example, disambiguation algorithms are problematic when dealing with very frequent names or "roots", which can affect a specific population. In our analysis we had to exclude very frequent roots and Asian names. The tool also exhibits some technical limitations. Firstly, as it employs a web search service (M1 relies on DuckDuckGo), changes in the way this search engine provides results would inevitably require updates of M1. Secondly, webpages output by DuckDuckGo must be post-processed, making access difficult in some cases when: a) the webpage obtained from the search no longer exists and thus cannot be accessed, or b) the webpage exists but the server on which it is located does not allow the scrapping process and therefore no results can be retrieved. Finally, at the time of publication, the current version of the tool does not allow a crawling process for more than 5,000 researchers simultaneously, and cannot handle multiple parallel tasks without problems due to hardware requirements. In the future a more powerful hardware infrastructure should allow us to overcome these problems

Many governments have endorsed open data principles. The OECD declaration of the 30th January of 2004 demands that publicly funded archive data should made publicly available.

These principles could be extended to other types of data provided when applying for public funds, including academic CVs. Public access is already granted to names, field and institutions of publicly funded scientists (e.g. in the UK, USA). In some countries, CV information is required to be provided through standardised web platforms (e.g. in Norway, Spain, and Portugal) when applying for grants or positions but these databases are usually not publicly available. An exception is Italy where the CVs of candidates for national professor certification are publicly available. As we have seen, CV information is very rich, unique and relevant for research and policy making purposes. Granting access to CVs or CV databases, with the previous consent of the scientists, could contribute both to open data principles and to improving the knowledge of the research system.

References

Ackers, H.L. (2005) Moving people and knowledge, the mobility of scientists within the European Union. *International Migration* 43, 99-129.

Ackers, H.L, Oliver E. (2007) From Flexicurity to Flexsecquality? The impact of the fixed-term contract provisions on employment in science research. *International Studies of Management and Organization* 37(1), 53-79.

Antonelli C., Franzoni, C., Geuna, A. (2011) The organization, economics, and policy of scientific research: What we do know and what we don't know—an agenda for research. *Industrial and Corporate Change* 20(1), 201-213.

Auriol, L., Misu, M., Freeman, R. (2013) Careers of doctorate holders: Analysis of labour market and mobility indicators. *OECD STI Working Papers* 2013/4.

Azoulay, P., Stellman, A., Zivin, J. (2006) PublicationHarvester: An Open-Source Software tool for science policy research. *Research Policy*, 35(7), 970-974.

Bonzi, S. (1992) Trends in research productivity among senior faculty. *Information Processing and Management* 28(1), 111–120.

Bozeman, B., Ponomariov, B. (2009) Sector switching from a business to a government job: Fast-track career or fast track to nowhere? *Public Administration Review* 69(1), 77-91.

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

Cañibano, C., Bozeman, B. (2009) Curriculum vitae method in science policy and research evaluation: The state-of-the-art. *Research Evaluation*, 18(2), 86–94.

Corley, E., Bozeman, B., Gaughan, M. (2003) Evaluating the impacts of grants on women scientists' careers: The curriculum vita as a tool for research assessment. In *Learning from Science and Technology Policy Evaluation: Experiences from the U.S. and Europe*, eds. P Shapira and S Kuhlmann, pp. 293– 315. Cheltenham, UK: Edward Elgar.

Cunningham, H. et al. (2011) Text processing with GATE (Version 6). University of Sheffield Department of Computer Science http://gate.ac.uk/.

Dietz, J.S., Chompalov, I., Bozeman, B., Lane, E.O., Park, J. (2000) Using the curriculum vita to study the career paths of scientists and engineers: An exploratory assessment. *Scientometrics* 49(3), 419–442.

Dietz, J.S., Bozeman, B. (2005) Academic careers, patents, and productivity: Industry experience as scientific and technical human capital. *Research Policy* 34, 349-367.

Ehrenberg, R.G. (2003) Studying ourselves: The academic labour market. *Journal of Labor Economics* 21(2), 267-287.

Enders, J. (2005) Border crossings: Research training, knowledge dissemination and the transformation of academic work. *Higher Education* 49, 119-133.

Enders, J., Weert, E. (2004) Science, training and career: Changing modes of knowledge production and labour markets. *Higher Education Policy* 17, 135-152.

European Commission (2010a) *Europe 2020. A European Strategy for Smart, Sustainable and Inclusive Growth*. Luxembourg: Office for Official Publications of the European Communities.

European Commission (2010b) *Europe 2020 Flagship Initiative Innovation Union. SEC(2010) 1161*. Luxembourg: Office for Official Publications of the European Communities.

Fernandez-Zubieta, A, Geuna, A., Lawson C. (2013) Researchers' mobility and its impact on scientific productivity. University of Turin Working Papers No. 13/2013

Fernandez-Zubieta, A. Geuna, A., Lawson, C. (2015a) What do We Know of the Mobility of Research Scientists and of its Impact on Scientific Production. In A. Geuna. Global mobility of research scientists: the economics of who goes where and why. Elsevier

Fernandez-Zubieta, A. Geuna, A., Lawson, C. (2015b) Mobility and Productivity of Research Scientists. In A. Geuna. Global mobility of research scientists: the economics of who goes where and why. Elsevier

Franzoni, Ch., Scellato, G., Stephan, P. (2012) Foreign-born scientists: Mobility patterns for 16 countries. *Nature Biotechnology* 30(12), 1250-1253.

Gaughan, M., Bozeman, B. (2002) Using curriculum vitae to compare some impacts of NSF research center grants with research center funding. *Research Evaluation* 11(1), 17–26.

Gaughan, M., Ponomariov, B. (2008) Faculty publication productivity, collaboration, and grants velocity: Using curricula vitae to compare center-affiliated scientists and unaffiliated scientists. *Research Evaluation*, 17(2), 103–110.

Gaughan, M., Robin, S. (2004) National science training policy and early scientific careers in France and the United States. *Research Policy,* 33(4), 109-122.

Glanzel, W., Debackere, K., Meyer M. (2008) 'Triad' or 'Tetrad'? On global changes in a dynamic world. *Scientometrics* 74(1), 71-80.

Howells, J., Ramlogan, R., Cheng, S. (2012) Innovation and university collaboration: Paradox and complexity within the knowledge economy. *Cambridge Journal of Economics* 36(3), 703-721.

Lee, S., Bozeman, B. (2005) The effects of scientific collaboration on productivity.*Social Studies of Science*, 35(5), 673-702.

Leydesdorff, L., Etzkowitz, H. (1996) Emergence of a Triple Helix of university-industry-government relations. *Science and Public Policy* 23, 279-286.

Mangematin, V. (2000) PhD job market: professional trajectories and incentives during the PhD, *Research Policy* 29(6), 741-756.

Martin-Rovet, D. (2003) Opportunities for Outstanding Young Scientists in Europe to Create an Independent Research Team. Strasbourg: European Science Foundation.

OECD (2003) The International Mobility of Researchers: Recent Trends and Policy Initiatives. Paris: OECD.

OECD (2008) The Global Competition for Talent: Mobility of the Highly Skilled Directorate for Science Technology and Industry. Paris: OECD.

OECD (2013) Researchers on the Move: The Impact of Brain Circulation. OECD Brief. http://www.scribd.com/doc/178378893/Researchers-on-the-Move-The-Impact-of-Brain-Circulation.

Powell, W.W. Koput, K.K., Smith-Doerr, L. (1996) Inter-organisational collaboration and the locus of innovation: network learning in biotechnology. *Administrative Science Quarterly* 41, 116-145.

Smith-Doerr, L. (2006) Stuck in the middle: Doctoral education ranking and career outcomes for life scientists. *Bulletin of Science, Technology & Society* 26(3), 243-255.

Stephan, P. (2012) *How Economics Shapes Science*. Cambridge, MA: Harvard University Press.

Yuker, H.E. (1984) Faculty workload: Research, theory and interpretation. ASHE-ERIC Higher Education Research Report No. 10, Association for the Study of Higher Education, Washington, DC.

Zubieta, A. (2009) Recognition and weak ties. Is there a positive effect of postdoctoral positions in academic performance and career development? *Research Evaluation* 18(2), 105-115.

Appendix A

Table A1: Descriptive Statistics

	NIH					BBSRC				
	mean	sd	min	max	N	mean	sd	min	max	N
Age in 2013	54.62	10.78	36	89	352	51.93	8.39	35	71	287
Female	0.21	0.41	0	1	360	0.22	0.41	0	1	291
Born in the US (Born in the UK)	0.76	0.43	0	1	209	0.76	0.43	0	1	186
Year of BA	1979.31	10.77	1943	1998	345	1981.83	8.64	1962	1999	280
Year of PhD	1985.78	11.24	1950	2004	343	1986.60	9.03	1967	2006	288
BA at US university (UK university)	0.79	0.41	0	1	349	0.83	0.38	0	1	277
PhD at US university (UK university)	0.88	0.33	0	1	347	0.84	0.37	0	1	288
Year of permanent/tenure-track position	1988.87	11.85	1956	2008	360	1991.55	9.86	1968	2011	291
Position in 2013										
Non-academic	0.01	0.07	0	1	360	0.01	0.10	0	1	291
Assistant professor	0.05	0.22	0	1	360	0.04	0.19	0	1	291
Associate professor	0.23	0.42	0	1	360	0.27	0.44	0	1	291
Full Professor	0.71	0.45	0	1	360	0.68	0.47	0	1	291
Mobility measures										
Postdoc	0.76	0.43	0	1	360	0.84	0.37	0	1	291
Postdoc outside US (outside UK)	0.11	0.31	0	1	360	0.41	0.49	0	1	291
Job mobile after postdoc	0.52	0.50	0	1	360	0.52	0.50	0	1	291
Job mobile between universities	0.44	0.50	0	1	360	0.40	0.49	0	1	291
Job mobile between US universities (UK universities)	0.41	0.49	0	1	360	0.30	0.46	0	1	291
Career job mobility (university mobility only)	0.17	0.38	0	1	360	0.18	0.39	0	1	291
Geographical job mobility	0.08	0.27	0	1	360	0.16	0.37	0	1	291
Sector job mobility (industry to university/PRO)	0.03	0.18	0	1	360	0.05	0.23	0	1	291
Sector job mobility (PRO to university)	0.06	0.24	0	1	360	0.10	0.30	0	1	291
Times job mobile	0.88	1.13	0	6	360	0.77	0.94	0	4	291
Voluntary mobility (university mobility only)	0.24	0.43	0	1	360	0.35	0.48	0	1	291
Forced mobility (university mobility only)	0.29	0.45	0	1	360	0.07	0.25	0	1	291

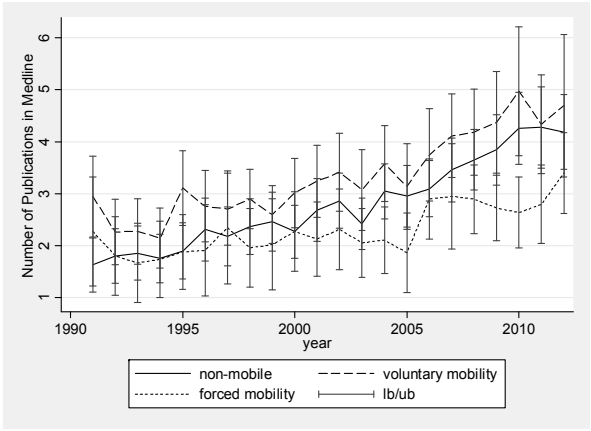
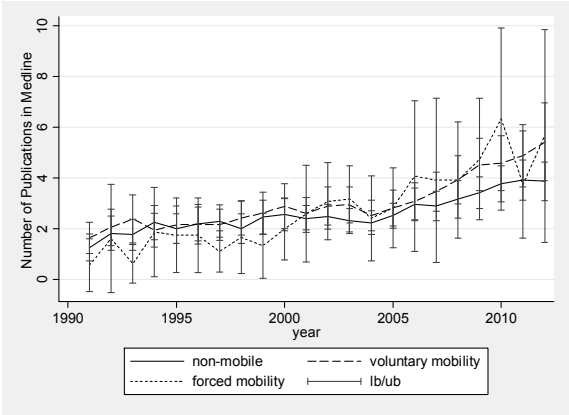
NIH – only those that never left academia	BBSRC – only those that never left academia
	

Figure A1: Performance of voluntarily mobile researchers compared to others that never left academia