The Women in European Economics Monitoring Tool:

Technical Description

Guido Friebel
Goethe University, Frankfurt, and
the Women in Economics Committee of the European Economic Association

Sascha Wilhelm
Goethe University, Frankfurt

Frankfurt, 20th May 2019

1. Introduction

There is evidence that women find it particularly hard to make careers in economics, compared to many other disciplines, and there is evidence that this is not just a taste issue, but may have a structural dimension. Auriol et al. (2019) summarize the state of the literature in the companion paper.

We here describe a tool that we have developed to monitor the situation in all European institutions that we could identify internet addresses of in real time. Our web-scraper collects information on researchers in European Institutions on a daily basis.

The obtained statistics may be interesting for the following groups of people:

1. For university presidents, deans and chairpersons in order to monitor how the own situation compares to the situation in other institutions in the same country, of a similar research standing or with similar challenges (think research departments in central banks or supra-national institutions). We use the standard hierarchical categories as, e.g., Assistant Professors, Associate Professors, and Full Professors.
2. For the broad public, because there is an increasing interest in gendered career outcomes in many realms of society. This research is of particular interests because it also has particularly important long-term effects on the orientation of the research world and the efficient use of human capital.
3. For job candidates because they would like to use information about the track record of an institution before deciding to accept job offers, and because they may have a genuine interest in a gender-balanced environment.
4. For researchers who would like to have access to a database reporting on who works where and on what level. This information can be matched with publication and citation records, information that allows to look at movements across and between institutions. This information can also be enhanced by survey evidence on the family situation of researchers such as number of kids, and profession of the spouse.

In what follows, we describe the methods we used, provide some first descriptive statistics, and discuss the limitations of the methodology.

2. Data collection algorithm

We designed an algorithm to monitor on a daily basis all known URL of European institutions that contribute to research in economics. The algorithm identifies the individuals listed on these websites and, where available, records the position titles these individuals hold. Gender is identified through first names, and a gender identification software analyzing pictures of the individuals. For the top 300 European research institutions (in terms of research output), these algorithms are complemented by our additional research classifying the obtained position titles into a generally accepted hierarchy of positions. Finally, we contacted the people responsible for managing the institutions and websites to verify the results of our work and provide us with feedback. Detailed information follows below.

Identifying research institutions

RePEc[1], a bibliographic database, provided us with a dataset of 4414 institutions contributing to the economic literature until December 2017. We manually identified all the institutions' websites containing a summary of affiliated researchers.[2] Importantly, we rely on RePEc's definition of institutions contributing to the field of economics. Therefore, in the data set, we do not only have institutions that primarily contribute to economics but also to neighboring research areas like finance, management, marketing or psychology.

A considerable number of institutions have multiple entries at RePEc. Spelling errors may result in multiple entries of institutions, and so does the fact that oftentimes sub departments from institutions have separate URLs. Therefore, to the extent possible, we remove duplicated and outdated institutions and exclude institutions that do not provide information of their researchers on their website. After this cleaning, 1801 different institutions remain.

Table 1 provides an overview over the countries in which the institutions reside. Column 1 summarizes the number of different institutions by country, as provided by RePEc. Column 2 counts the number of included institutions after data cleaning, and column 3 states the total number of different websites that we observe, given that some institutions have multiple websites.

[1] Accessible by http://repec.org/.

[2] Some institutions do not provide a comprehensive overview of researchers and cannot be monitored. In the majority of cases, these institutions are inactive or not related to academic research.

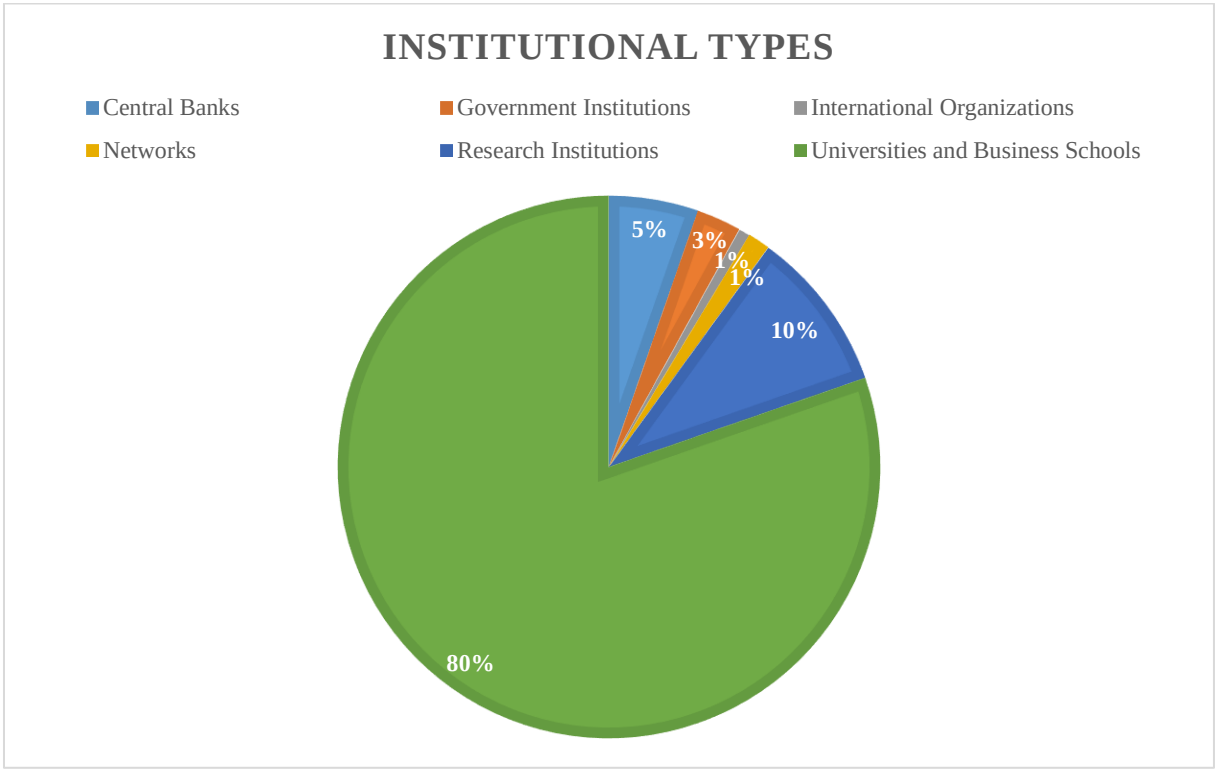
Table 1: Breakdown of institutions by country

	(1)	(2)	(3)
Country	**RePEc institutions**	**Included Institutions**	**Monitored Websites**
Aland	1	1	1
Albania	12	5	5
Armenia	15	1	1
Austria	89	45	62
Azerbaijan	13	3	3
Belarus	22	5	6
Belgium	126	39	89
Bosnia and Herzegovina	14	5	5
Bulgaria	31	11	12
Croatia	14	10	12
Cyprus	9	4	7
Czech Republic	37	22	48
Denmark	81	32	43
Estonia	11	9	9
European Union	0	2	2
Finland	44	18	19
France	384	159	212
Georgia	13	4	5
Germany	563	271	347
Greece	71	44	53
Hungary	79	28	32
Iceland	5	5	5
International Organization	0	16	17
Ireland	33	19	28
Italy	366	171	189
Kosovo	5	2	2
Latvia	18	11	12
Liechtenstein	4	2	2
Lithuania	17	10	14
Luxembourg	13	4	5
Macedonia	0	5	5
Malta	4	1	1
Moldova	11	1	1
Monaco	2	1	1
Montenegro	5	3	3
Netherlands	138	68	102
Norway	56	32	46
Poland	141	76	103
Portugal	102	48	57
Romania	115	32	57
Russia	357	19	23
San Marino	2	1	1

Serbia	13	6	6
Slovakia	33	21	22
Slovenia	12	9	9
Spain	353	152	232
Sweden	100	52	67
Switzerland	133	64	94
Turkey	221	76	80
Ukraine	0	1	1
United Kingdom	499	241	277
Vatican City	1	1	1

Figure 1 provides a breakdown of different types of institutions that we monitored. While the bulk of institutions are university organizations and business schools, we also find central banks, government institutions, international organizations and research institutions. A special case are research networks (CEPR, IZA, NBER, CESifo).

Figure 1: Breakdown by types of institutions



Web-Scraper

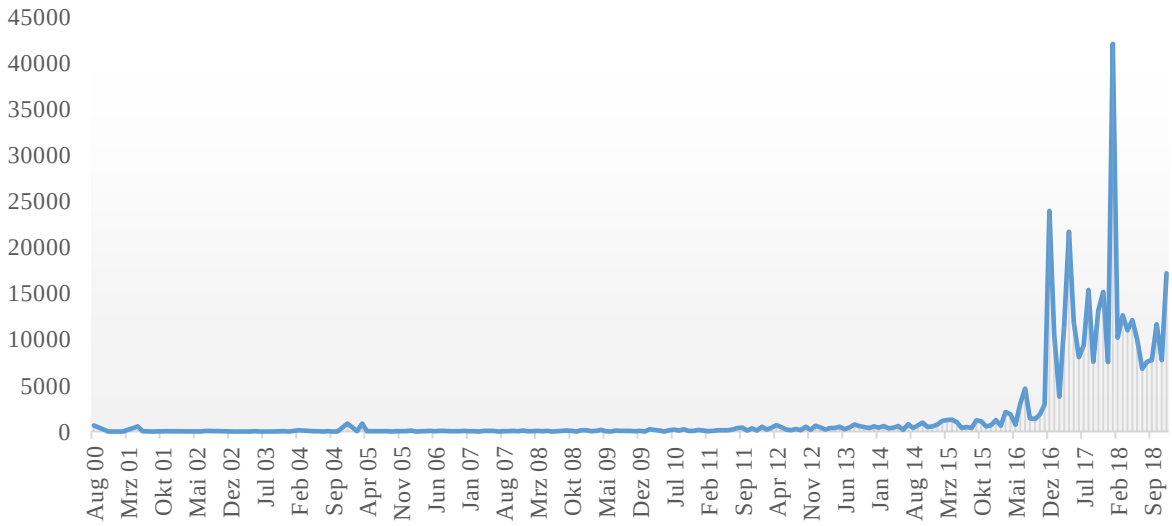
We program an individual web-scraper for every institution from the RePEc database. The web-scraper accesses the website and scrapes available information like name, photo, chair and position details. We use a Python[3] script with the BeautifulSoup[4] extension to parse websites. Importantly, our web scraping relies on institutions to provide a website with comprehensive lists of their researchers containing all relevant information. It is impossible to identify and web-scrape researchers' individual websites as this exceeds our resources for the number of parsing instructions we are programming for each website.

The web-scraper accesses all websites every 24 hours. We record changes like new positions, updated details, and record new information together with the time of access as an additional observation in our dataset. We permanently monitor failures of the web-scraper that emerge when institutions change their content or the address of their websites. In such cases, we try to update and adjust the parsing instructions within a month.

For historical information, we make use of the "Web Archive"[5] for the available institutions. Therefore, for a limited number of institutions, we obtain information about entries and exits from 2000 onwards. Every time we recognize a new position, we record the respective date and time. A position is "new" when the name, chair or position is not identified already. Hence, we also identify switches within an institution if it is mentioned on the institution's website. Figure 2 shows the frequency of information available about the institutions. It is noteworthy that the web archive does not provide encompassing information about the websites of research institutions in Europe. Arguably, because there is not much demand for this information.

[3] Python Software Foundation. Python Language Reference, version 2.7. Available at http://www.python.org.
[4] Leonard Richardson. Version 4.4.0. Available at https://www.crummy.com/software/BeautifulSoup/bs4/doc/.
[5] https://archive.org/web/

Figure 2: New positions by moment of time entered



Position characteristics

Based on the scraped information, we automatically process the obtained data to identify characteristics of the identified position.

We classify fields and columns from websites to position characteristics. If necessary, we also use regular expressions[6] to filter relevant information. Each position is described by the following characteristics:

- Name
- Picture
- Personal website
- Chair
- Position
- Others (like degree, room, etc.)

We rely on the name-parsing tool "nameparser"[7] to parse the name into its individual components. These components are title, first name, middle name, last name, suffix and nickname.

[6] Johnson, Walter L.; Porter, James H.; Ackley, Stephanie I.; Ross, Douglas T. (1968). "Automatic generation of efficient lexical processors using finite state techniques". *Communications of the ACM*. **11** (12): 805–813. doi:10.1145/364175.364185

[7] Derek Gulbranson. Version 1.0.2. Available at https://github.com/derek73/python-nameparser.

Concerning the position holder's first name, we determine their gender with the service by genderize.io[8]. The service utilizes big datasets from social networks to provide a probability for the gender of any given first name. Using the picture, we determine the gender of the position's holder with a convolutional neural network that heavily relies on the work of Levi and Hassner (2015)[9].

For few positions that cannot be classified well with our method, we did a manual online search and added the gender manually. Based on the determined genders from name, face recognition and manual online research, we are using the following procedure to infer the gender:

1. If available, the manual search suggestion is used. Otherwise, continue with 2.
2. If gender name probability of the name is 100%, the determined gender by the name is used. Otherwise, continue with 3.
3. If both gender and name coincide and their joint probability of certainty is above 95%, their suggested gender is used. Otherwise, continue with 4.
4. If only human name or face is recognized, and its squared probability exceeds 95%, accept the suggested gender. Otherwise, no gender is recognized until a manual search is performed and the algorithm starts again with 1.

Hierarchical levels

To draw conclusions about the status of researchers, we categorize positions into hierarchical levels. Inspired by the generally used definition of research hierarchies (e.g. in Wikipedia[10]), we define six types of hierarchical levels in descending order: (Full) Professors, Associate Professors, Assistant Professors, Lecturers, Research Fellows and Research Assistants. Non-academic positions do not belong to any of these levels and we exclude them in the analysis.

To map scraped position information to the six hierarchical levels, we developed an algorithm to classify levels based on position descriptions and name titles. We face the challenge that almost any country has its own non-standardized terms for its levels, oftentimes with multiple terms for the same hierarchical level. We use a text mining method to extract the hierarchical level from position descriptions and name titles.

[8] Available at https://generize.io.
[9] Gil Levi and Tal Hassner, Age and Gender Classification Using Convolutional Neural Networks, *IEEE Workshop on Analysis and Modeling of Faces and Gestures (AMFG), at the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Boston, June 2015
[10] Academic ranks in the United States: Most common hierarchy. Available at https://en.wikipedia.org/wiki/Academic_ranks_in_the_United_States#Most_common_hierarchy. February 5th, 2019.

We create a mapping between keywords and a representative level for each country. For all European countries, we determine around 50 keywords and map them to hierarchical levels. Every keyword that exists somewhere in the text is recorded as a potential level. Then, from all potential levels, we use the level which has the lowest order in Table 2.

For example, consider a typical position description of the equivalent of an assistant professor in a German university: "Juniorprofessorin in Industrieökonomie". Based on our mapping, we find two potential levels: (i) Assistant Professor, based on the existing position-keyword *Juniorprof*, and (ii) Professor, based on the keyword *Prof*. We look up the order from table 2, which yields an order of 2 for the Assistant Professor and 6 for Professor. Hence, we conclude that the position belongs to an Assistant Professor as its order has the lowest value.

Table 2: Hierarchical levels and their order in the position and title description algorithm

CLASS	HIERARCHICAL LEVEL	ORDER: POSITION	ORDER: TITLE
1	Professor	6	4
2	Associate Professor	1	2
3	Assistant Professor	2	3
4	Lecturer	3	5
5	Research Fellow	4	6
6	Research Associate	5	1

Concluding the hierarchical level by name titles works in a similar way. If a name title is for example "Prof. Dr.", we find two potential levels such as Professor, based on the existing title-keyword *Prof*, and Research Fellow, based on the title-keyword *Dr.* Professor has a lower order (4 vs. 6) for the title than Research Fellow and is, hence, concluded to represent the level based on the name title.

As information about positions is more precise and detailed, we determine our final classification by the classification of the position if information on positions and titles is available. If only the title is available, its classified hierarchical level is accepted. In case we do not identify any hierarchical level because neither title nor position is provided or informative, we do not conclude anything about the researcher's position.

The algorithm laid out succeeds in identifying hierarchical levels for a large majority of academic institutions. There are two exceptions. A number of UK institutions do not provide information about hierarchical levels and positions but only mention the title "Dr". For these people and in order not to jeopardize the validity of the algorithm, we manually identify the positions of those academics who are in the top 300 RePEc institutions. In France, "Mâitre de conferences" are tenured assistant professors; however, in some cases position holders also use the title associate professor in which case we called them associate professors.

Table 3 provides an overview of the positions we have identified and the result of the algorithms described above.

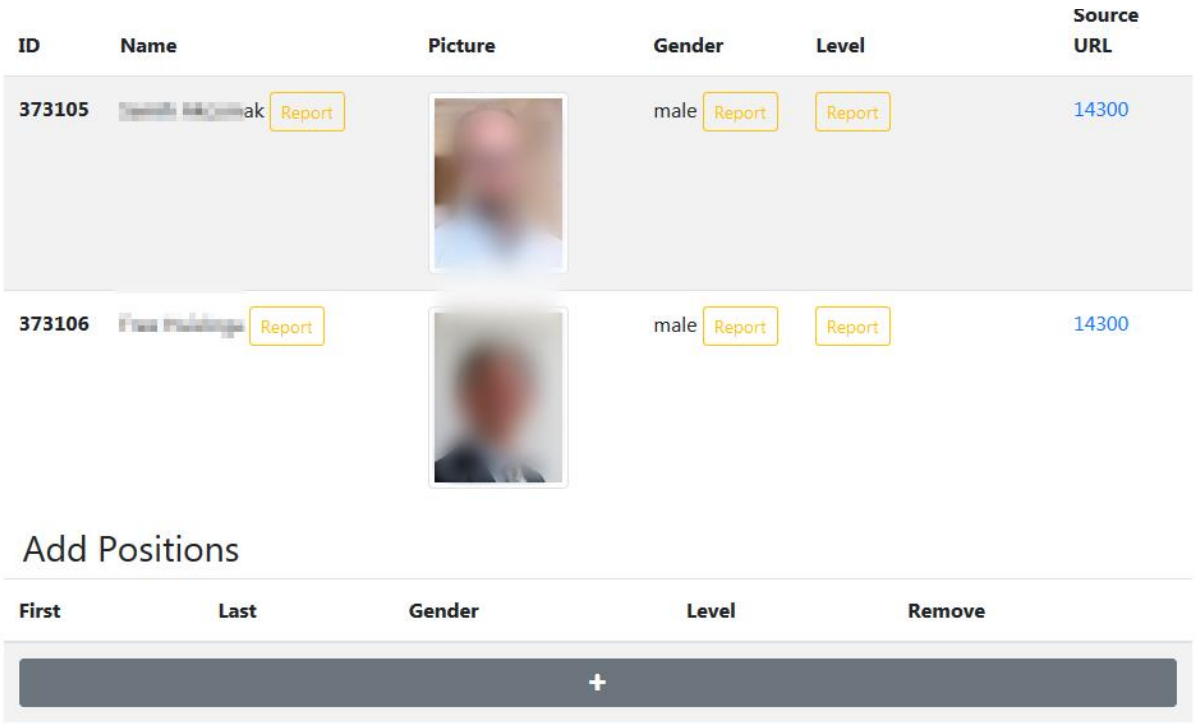
Table 3: Summary of identified positions

Active Positions	115,932	100%
Recognized Faces	28,467	24.6%
Recognized Names	101,918	87.9%
Recognized Genders	88,644	76.5%
Academic Level identified	50,725	43.8%

Additional verification

We perform crosschecks and verify manually RePEc's top 300 institutions with (i) a low number of identified positions, (ii) a low percentage of identified positions, or (iii) a low number of identified genders manually. Research assistants determine genders and hierarchical levels by online inquiries of these institutions. Furthermore, we contacted 294 contact persons of institutions, for instance, deans or head of departments, and asked them to check our results. We sent each contact an individual list of all researchers we found including our results concerning the identified gender and hierarchical status. Figure 3 provides an example of a data entry form sent to an institution.

Figure 3: Screenshot of verification form sent to an institution



If the contact persons identify errors, they can inform us of misclassifications or add missing positions. Table 4 provides a summary of the actions taken by the persons (or their staff) contacted. Besides requests to remove people who had left the institutions, and correction about the hierarchical positions, there are a small number of requests concerning the gender we identified.

Table 4: Summary of verification statistics

Reviews requested	290
Visits on review website	188
Position removal requests	423
Gender correction requests	142
Hierarchy correction requests	1579

Some institutions also sent us excel sheets with their requests. We took care to correct all issues that were reported but cannot summarize contents because the requests refer to quite different demands. We finally, in the top 300, exclude five institutions from the verification process that do not provide information about their researchers in table 5.

Table 5: Institutions excluded from the data set, and reasons

Institution	Country	Reason
Département Sciences Sociales, Agriculture et Alimentation, Espace et Environnement (SAE2), Institut National de la Recherche Agronomique (INRA)	France	Does not provide information about researchers' identity
Nationale Bank van België/Banque national de Belgique (BNB)	Belgium	Does not provide information about researchers' identity
Directorate-General Economic and Financial Affairs, European Commission	Belgium	Does not provide information about researchers' identity
Arbetsmarknadsdepartementet	Sweden	Does not provide information about researchers' identity
Russian Presidential Academy of National Economy and Public Administration (RANEPA)	Russia	Does not provide information about researchers' identity