Data identification, selection, gathering, and organization

Francesca Marzi, Diana Di Marco, Giovanni Stilo

In this document, we describe the process we used to identify data of the academic staff (researchers and professors) of the Italian public universities that we believe can be subjected to algorithmic bias. We show the data identification process in section 1. In section 2, we discuss how data were retrieved from publicly available databases and the encountered collecting difficulties. Finally, in section 3, we describe the organization of the database we built from the collected data within its organization.

1 Data Identification

To understand both the institutional and technical limitations useful to identify the data that can be employed for experimentation and susceptible to algorithmic bias, we started by studying the University of L'Aquila (UAQ) case. Then we extended the findings to other Italian public universities. During the initial phase, we iteratively interacted and interviewed two main institutional operative units:

- the Working Group that has defined the UAQ Gender Budgeting;
- the IT services of UAQ.

The interviews followed the following approach:

- we established a set of topics to discuss during the interviews;
- we conducted informal and conversational interviews.

After a critical analysis of the interviews, we decided to choose the data by evaluating those used in the literature review (our literature analysis is reported in *[FAIR-EDU] Gender Bias in Classic Academic Systems*) and those highlighted in the Gender Budgeting of UAQ for the year 2021 [5], which in turn follows the CRUI Guidelines [1].

The CRUI Guidelines were created to support the drafting of Gender Budgeting by identifying a set of data, indicators, indices, and their representations, allowing historical, national, and international comparability. The Guidelines suggest data for students, researchers, professors, technicians, administrative, and for institutional and government positions. Table 1 shows the proposed data and sources for researchers and professors grouped by composition, career, research, and teaching.

Composition	
Distribution by gender and role	http://dati.ustat.miur.it/dataset
Time series of academic staff by gender and role	http://dati.ustat.miur.it/dataset/dati-per-bilanciodi-genere
Distribution by gender, role, and age group	http://dati.ustat.miur.it/dataset/dati-per-bilanciodi-genere
Average age by role and gender	https://www.contoannuale.mef.gov.it/strutturapersonale/eta
Percentage of women by area and role: comparison with corresponding national data	http://cercauniversita.cineca.it/php5/docenti/cerca.php, http://dati.ustat.miur.it/dataset
Distribution of full professors among Fields of Research and Development in She Figures.	http://dati.ustat.miur.it/dataset/dati-perbilancio-di-genere
Femininity Report	http://dati.ustat.miur.it/dataset
Career	
Percentage of full professors to total teaching and research staff by gender	http://dati.ustat.miur.it/dataset
Ranges of University and Academic Careers	http://dati.ustat.miur.it/dataset/seriestorica-sul-personaleuniversitario
Glass Ceiling Index (GCI)	http://dati.ustat.miur.it/dataset
Promotions information by gender and CUN area	Internal Source
Percentage applications to NSQ by gender, percentage NSQ by gender	http://abilitazione.miur.it/public/candidati_2016.php?sersel=105
Full/definite time distribution by gender	Internal Source
Sabbatical year fruition	Internal Source
Gender composition of competition committees	Internal Source
Research	
PI in projects PRIN/SIR/ERC/Other Progects by gender and funding disbursed	PRIN http://prin.miur.it/index.php?pag=2017 SIR http://sir.miur.it/index.php/finanziati/index ERC https://erc.europa.eu/projects-figures/erc-funded-projects/results?search_api_views_fulltext
Funding projects PRIN/SIR/ERC/ Other Projects by scientific sector ERC and gender of the PI	PRIN http://prin.miur.it/index.php?pag=2017 SIR http://sir.miur.it/index.php/finanziati/index ERC https://erc.europa.eu/projects-figures/erc-funded-projects/results?search_api_views_fulltext
Average per capita of internal and external research funds	Internal Source
Teaching	
Percentage of degree thesis supervisors by gender	Internal Source

Table 1: Data suggested by CRUI Guidelines

Table 2: Data from public databases

Personal data	**Scientific Productivity**	**Academic Career**
Gender	National Scientific Qualification	Position
	Bibliografic information of publications	University and Faculty
	Bibliografic Citations	Current and Historical Affiliation (Departments)
	H-Index	Disciplinary Scientific Sector
		Macro Disciplinary Area
		Area of Expertise
		Carrer Advancements
		Academic Seniority

1.1 Data Selection

We identified a set of possible data that we considered that are susceptible to algorithmic bias. We grouped these data into three categories:

Personal Data: personal information about the professors belongs to this category, such as age, gender, and general leave (maternity, parental, sick).

Scientific Productivity: this category includes information concerning professors' scientific productivity, such as the list of publications and if they got the National Scientific Qualification. In addition, for each author in bibliometric sectors, we also considered some bibliometric information (the total number of papers, total citations, the h-index, publication range, papers per year, citations per year, and publication types, journal metrics).

Academic Career: the category includes information about professors' careers, such as the university and department of affiliation, career advancements, academic seniority, macro disciplinary area, scientific sub-sector they belong to, area of expertise, current academic appointment, academics managerial appointments, teaching activities, founded projects, committees, salaries, and sabbatical period.

We encountered several difficulties in retrieving some of this data, even in the case of UAQ. Many data are very sensitive and obtainable only through internal university databases after the informed consent of the interested persons. Currently, we are trying to understand the procedures for accessing these databases and if it is possible to obtain such personal information. Therefore, to ensure the project is completed on time, in our preliminary tests, we decided to use only data available from public databases and not consider such sensitive data. Table 2 shows the resulting set of data that we were able to obtain from public databases.

2 Data Gathering

We collected data identified in section 1 for the academic staff of UAQ and for all Italian public universities. We downloaded information relating to the academic career from the MIUR website [2], National Scientific Qualification from [3], and data on Scientific Productivity through the Scopus API [4].

2.1 MIUR

The MIUR website contains data on the whole population of the academic staff in Italian public universities. Data are available from 2001 onward and can be downloaded as Excel files (one file per year). For each year and for each university, we retrieved the list of all assistant, associate, and full professors with the information on name, surname, position, gender, macro disciplinary area, scientific sub-sector they belong to, and the university and department of affiliation.

2.2 National Scientific Qualification

The results of National Scientific Qualification are available for 2012 and onward. For each scientific sub-sector, we retrieved the information about who got the associated and full professorship qualifications through web scraping. Unfortunately, for each scientific sub-sector, only the lists of qualified individuals are available since those who still need to obtain the qualification have been removed from the website for privacy reasons 120 days after publication.

2.3 Scientific Productivity

To download data on scientific productivity, we used the Python library *pybliometrics* [6], a valuable wrapper for the Scopus RESTful API that allows access to the Scopus database via user-friendly interfaces. For each individual, we retrieved information about the total number of papers, total citations, the h-index, publication range, papers per year, citations per year, and publication types (Book Series, Conference Proceeding, Journal). Furthermore, for each academic journal of each author, we considered the CiteScore, SJR, and SNIP metrics, referring to the last year. Algorithm 1 illustrates the high level procedure for retrieving the above information given a professor's name, surname, and affiliation, starting from the list of professors obteined from the MIUR.

Algorithm 1 Scopus Search

```
Input: Name, Surname, AffiliationName.
result ← AuthorSearch(Name, Surname)
if size(result) $> 0$ then
    for each author in result do
        author_info ← AuthorRetrieval(author.id)
        affiliation_history ← author_info.affiliation_history
        if affiliation_history is not empty then
            if AffiliationName is in affiliation_history then
                total_papers ← author_info.document_count
                total_citations ← author_info.citation_count
                h_index ← author_info.h_index
                publication_range ← author_info.publication_range
                docs ← author_info.documents
                papers_per_year ← docs.groupby(year).count()
                citations_per_year ← docs.groupby(year, citedby_count).sum()
                paper_types ← docs.groupby(aggregationType).count()
                list_score ← empty list
                for each journal in docs do
                    source ← SerialTitle(journal.issn or journal.eIssn)
                    CitScore ← source.citescore
                    Sjr ← source.Sjr
                    Snip ← source.Snip
                    list_score.append(CitScore, Sjr, Snip)
                end for
            end if
        end if
    end for
end if
Output: total_papers, total_citations, h_index, publication_range,
Output: papers_per_year, citations_per_year, paper_types, list_score
```

The functions `AuthorSearch`, `AuthorRetrieval`, `SerialTitle` are pybliometrics classes that perform queries for search for authors, retrieve the complete author record and return basic information on registered serials (also called sources), like publisher and identifiers, but also metrics.

The result, done by individual's name and surname, could lead various sources of errors or missing values:

- `AuthorSearch` returns no value. The reason could be twofold: a) the author is not indexed on Scopus; b) Scopus uses a different alias for the name or surname (i.e., double name or surname, etc.), so the query fails in returning the author profile.

- `AuthorSearch` returns more than one value, i.e., there is more than one Scopus profile for the author's name. In this case, we select the correct profile based on the parameter *AffiliationName*. Nevertheless, there could be more than one Scopus profile for an author with the same affiliation. In this case, all the profiles are considered. However, some results may have no affiliation, or the Scopus affiliation history may not contain the AffiliationName parameter; in this case, the profile is discarded.

- `SerialTitle` could fail because of the absence of issn (or eIsnn) in the document list of authors. Even if the issn is available, sometimes it may happen that it doesn't match the journal issn due to some mistake.

3 Data Organization

Data gathered in section 2 were merged and organized into an SQL database. The merge between the various sources is done through first and last names; there are no other identifiers to allow precise cross-referencing. The relational model in Figure 1 show the logical organization of the database. It consists of a total of 7 entities and 6 relationships.

3.1 Relational Model Description

The following tables were created for the database structure:

Entity Tables:

- **Professor:** for each professor, the first name and surname are provided. This attributes concern personal information of each professor.
- **Year_By_Year:** the following table was merged with the *Publication* relation that participated with cardinality 1 to 1. The attributes cover all information concerning the number of professor's publications and citations for each year, from 2012 to 2022.
- **Career_Info:** the following table was merged with the *Has_Info* relation that participated with cardinality 1 to 1. The attributes relate to all the remaining information of the professor concerning the gender and the disciplinary scientific field but also the university career. The latter include the total number of published papers, the number of total citations received, the h-index, book series, book, conference

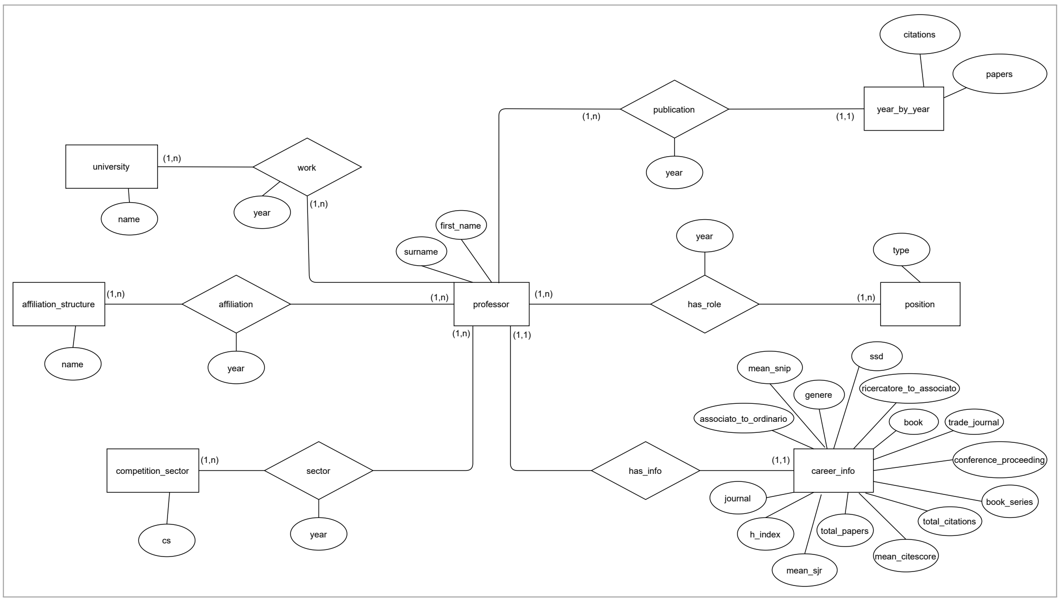

Figure 1: Relational model

proceeding, trade journal and the year in which a career jump from researcher to associate or from associate to full professor, if any, took place. Finally is also included the mean of the snip, the cite-score and the sjr on academic journals.

- **Position:** this table specifies the 3 possible positions that can be covered by the professors. These are represented with 3 integers: 0 for researcher, 1 for associate professor and 2 for full professor.
- **Competition_Sector:** this table specifies the name of the all competition sectors.
- **Affiliation_Structure:** this entity has as an attribute the name of the affiliation structures.
- **University:** in this table are the names of universities in Italy.

Relationship Tables:

- **Has_Role:** this table specifies for each professor the position held each year, from 2012 to 2022.
- **Sector:** this report specifies for each professor his or her area of expertise year by year, from 2012 to 2022.

- **Affiliation:** this table specifies for each year the affiliation structure to which each professor belongs, from 2012 to 2022.
- **Work:** for each year, from 2012 to 2022, the university at which a professor has served is indicated, with the aim of maintaining a history for each of them.

All other relations with entities that participated with cardinality 1 to 1 were appropriately merged with the entity tables.

Foreign keys: The following constraints were created for the foreign keys of the tables:

- **Professor** - has no foreign keys;
- **Year_By_Year** - has the *Professor ID* that must refer to an existing professor;
- **Career_Info** - has the *Professor ID* that must refer to an existing professor;
- **Position** - has no foreign keys;
- **Competition_Sector** - has no foreign keys;
- **Affiliation_Structure** - has no foreign keys;
- **University** - has no foreign keys;
- **Has_Role** - has the *Professor ID* and the *Position ID* that must refer to an existing professor and position respectively;
- **Sector** - has the *Professor ID* and the *Competition_Sector ID* that must refer to an existing professor and competition sector respectively;
- **Affiliation** - has the *Professor ID* and the *Affiliation_Structure ID* that must refer to an existing professor and affiliation structure respectively;
- **Work** - has the *Professor ID* and the *University ID* that must refer to an existing professor and university respectively;

These foreign keys make the different IDs refer to valid values.

References

[1] CRUI. Linee guida per il Bilancio di Genere negli Atenei italiani. https://www2.crui.it/crui/Linee_Guida_Bilancio_di_Genere_negli_Atenei_italiani.pdf.

[2] Ministero dell'Istruzione dell'Università e della Ricerca. Cerca Università. http://cercauniversita.cineca.it/php5/docenti/cerca.php.

[3] MIUR e Cineca. Abilitazione Scientifica Nazionale. https://abilitazione.miur.it/public/index.php.

[4] Elsevier. Scopus. https://dev.elsevier.com/.

[5] University of L'Aquila. Bilancio di Genere 2021. https://www.univaq.it/include/utilities/blob.php?item=file&table=allegato&id=5457.

[6] Michael E Rose and John R Kitchin. pybliometrics: Scriptable bibliometrics using a python interface to scopus. *SoftwareX*, 10:100263, 2019.