Bibliometric and patent indicators by gender: Is it feasible?

Angela Hullmann
(European Commission, Research Directorate-General)

Paper presented on the
First Workshop on Gender, Science and Technology Indicators,
Montevideo, Uruguay, 18. October 2001

This paper bases on the feasibility studies "Bibliometric indicators by Gender" and "Patent Indicators by Gender", funded by the European Commission and carried out by **Fulvio Naldi** and **Ilaria Vannini Parenti**, Biosoft sas, Milano, Italy. The paper summarises the main results of the two final reports.[1]

Introduction

Both publication and patent indicators are used as proxy to measure the scientific and technological productivity of a country. An application for a patent indicates that there has been a production of new knowledge linked to an invention, and, more importantly, that this knowledge might have potential economic returns. The bibliometric indicator that measures the number of publications gives an indication of the research capacity of a country or of a specific research community within the country as well as the changes in scientific productivity over time.

Given the importance attached to patent indicators as an output indicator, it might be worthwhile to investigate the gender issue. Are women more or less productive in the area of science and technology, are they associated with patents of higher economic value, in which technology fields do they tend to be present? Similar important questions can be asked about the publication indicator: do women publish more than men, in which disciplines and technology fields are they most active, do they publish higher quality papers?

Despite the importance of these questions – also with a view on the utilisation of a country's human resources – little information is systematically recorded on the gender of inventors and authors. In order to support the development of a methodology that enables registering gender aspects, the European Commission decided to commission two feasibility studies. Both studies were carried out by Biosoft in Milano, Italy. This paper relies on extracts from the final reports of the two studies.

This paper will describe, first, the methodological design of the so-called First Name Data Base (FNDB) which was developed in these feasibility studies to determine the gender of the inventor or author. Secondly, this paper will present the results emerging from the use of the FNDB methodology to generate patent and bibliometric data by gender for six EU countries, nine disciplines and 31 technology fields. Finally, the paper will assess the feasibility of the FNDB methodology and finish with an outlook if and how this methodology could be extended to all fifteen EU Member States.

[1] The final reports of the two feasibility studies will be published soon.

1. The First Name Data Base (FNDB)

The First Name Data Base (FNDB) is a collection of first names that at the present covers six European languages: English, French, German, Italian, Spanish and Swedish. The goal has been to set up a high quality database with two objectives: (1) perform gender analysis on any list of person names and (2) allow expansion to other languages.

FNDB has been specifically developed to perform gender analyses based on the names of inventors and authors of scientific publications. Its structure allows easily to extend its implementation to all the EU countries, following the methodology and the experience of this study.

Identification of the Sources

Sources can be divided in the following categories:
Dictionaries. They generally contain either commonly used names that can be found in most other sources, or unusual names extracted from literature and mythology.
Humanistic studies. They are few and restricted to the traditional names in a given language. Usually they are very accurate with very few spelling and gender assignment errors. However, the need of being formally correct may introduce problem with the names that are prevalently used for a gender and only exceptionally for the other. This aspect is discussed later in this section.
Calendars and lists of Saints. They include obsolete names and provide only lists of names without gender assignment. The intervention of a mother-tongue person is usually required.
Books or Internet web sites. These are addressed to parents either to suggest names for new-borns or to explain the meaning of the names. These sources are extremely rich and contain thousands of names. The quality in terms of spelling errors and gender assignment varies from one to another and is difficult to evaluate. Examples of poor reliability are the sites that allow the final user to add a name if it is not already included in the database. Some web sites publish the lists of the most used names in a country/linguistic area for a given year or 10 years period. Sometimes these lists go back to the beginning of 1900. These data comes from official lists and are very useful for our purposes.
Files from Record Offices and phone books. While the former are difficult to obtain, the latter represent an extended source but need the intervention of a mother-tongue person for gender assignment.

Criteria for Selection of the Sources and for Data Extraction

Spelling and gender assignment errors may appear in the original source or may be added during the manual input of data while creating FNDB. Spelling errors generally cause a proliferation of the records in the database with two consequences: (1) the database size increases, (2) a name which is misspelled in a given language may coincide with a name in another language leading to error in gender assignment.

Errors in the sources require special attention since the sources differ greatly in quality. Sources are classified in good, medium and poor quality.

Source quality	*Examples of sources*
Good	Dictionaries, lists of names obtained by Consulates or Centres of Culture, Internet files published by Academies, Universities, Governmental Organisations
Medium	Lists extracted from calendars and phone books (with manual gender assignment by a mother-tongue person), Internet sites where the addition of new names is controlled by an internal structure
Poor	Internet sites where new names can be directly added by the end-user

Fig. 1.1 Quality of the sources

Names have been included into First Name Data Base ver.0 only if they satisfied at least one of the following criteria:

- They came from a good quality source
- They appeared at least in two different medium-quality sources.
- Names found only in poor quality sources have not been included in FNDB ver.0.

This approach allows to remove almost all the spelling errors and to avoid useless record proliferation in the database. To validate and better evaluate the consequences of this strategy, an extended version of the database was built that contained all the names found in all the selected sources. Two poor quality sources with high coverage in terms of number of names and languages were also included in the extended database.

The following table shows the main characteristics of the two databases:

	FNDB ver.0	*Extended*
n. of different sources of names	20	22
Total number of source items	23.871	60.899
n. of different names	6.441	32.710
n. of “**B**oth” cases	739	1.543

Fig. 1.2: Size of FNDB ver.0 versus its extended version

The total number of records in the extended version is about 2,5 times the number of records of FNDB ver.0, the number of different names is about 5 times and the "**B**oth" cases are more than doubled.

Quality Improvement

Data quality is influenced by several kinds of errors that may occur in any phase of the processing. While some of them can be prevented and corrected, others can be detected but remain unsolved. This is the case of diminutives, nicknames or short forms, which represents the major part of "**B**oth" cases. Moreover, names are subjected to fashion; therefore names that were popular in the forties and fifties are now scarcely used. Nevertheless they represent a considerable portion of the analysed data sets, because the majority of the scientists, object of this analysis, was born at that time. In addition, the use of foreign names as well as the extension to both genders of a name used in the past for only one gender has become recently common.

With reference to this aspect, it is worthwhile to mention that in a few cases of rarely used names two mother tongue people of the same language classified the same name in different ways; when the gender assigned differ we reported "**B**oth" in FNDB, when one of the mother tongue person classified the name as "Not used" we applied the gender classification suggested by the other person.

There are three ways to improve the overall quality of FNDB:

- eliminating wrong gender assignments,
- adding the names of inventors and scientists still not included in FNDB,
- reducing the number of names classified as "**B**oth".

The first two points have been solved with the help of mother tongue people who checked the whole database against spelling and gender errors and classified manually the names of inventors/authors still not present in data base.

However, the most important action was the reduction of the number of names classified as "**B**oth". The importance of processing "**B**oth" cases is evident when considering for example the case of the data published by the US Social Security Administration (SSA). This official source contains a remarkable number (198) of very frequent names, classified as "**B**oth" that are generally used only for a specific gender. The following table reports the names that are marked as "Both" by SSA and that appear more than 100 times in EPO 98. The second column indicates the gender assigned by the other sources (**X** indicates "**B**oth" or language dependent).

Name	*Gender*	*N. of occurr. in EPO 98*
MICHAEL	M	2056
THOMAS	M	1768
JEAN	X	1643
DAVID	M	1138
JOHN	M	884
ROBERT	M	876
CHRISTIAN	M	825
RICHARD	M	728
PAUL	M	715
FRANK	M	661
MICHEL	M	634
PHILIPPE	X	552
WALTER	M	541
ANDREW	M	464
DANIEL	M	381
ERIC	M	347
GERD	X	345
JAN	X	339
MARK	M	336
CLAUDE	X	259
JAMES	M	253

Name	*Gender*	*N. of occurr. in EPO 98*
WILLIAM	M	250
ALEXANDER	M	245
ANTHONY	M	218
DOMINIQUE	X	217
BRIAN	M	213
CHRISTOPHE	M	211
CHARLES	M	155
JONATHAN	M	143
MARIE	F	141
CARL	M	136
GEORGE	M	128
FRANCIS	M	126
KENNETH	M	125
RENE	M	124
MARIA	F	115
JOSEPH	M	111
BO	X	109
KEVIN	M	107
PATRICE	X	104
RONALD	M	102
KARSTEN	X	100

Fig. 1.3: Frequently used names classified as "Both" in SSA source

It must be noticed that, since most of the 198 names defined as "**B**oth" in SSA are defined as "**M**" by the other sources, the inattentive application of SSA and of analogous sources could introduce a significant bias in the statistical analysis.

The following table shows the number of inventors classified as "**B**oth" with a version of FNDB that uses the SSA gender classification for the 198 names and with a modified version of FNDB where the 198 SSA names were classified with the specific gender according to all the other sources

Country	Number of Inventors	Both names in FNDB ver.0	%	Both names in modified version	%
Germany	55195	7406	13,4	1128	2,0
Spain	1383	155	11,2	53	3,8
France	16973	5794	34,1	1720	10,1
UK	15979	7464	46,7	262	1,6
Sweden	6718	919	13,7	379	5,6
Total	96248	21738	22,6	3542	3,7

Fig. 1.4: Results of the modified classification of "Both" names in FNDB ver.0

The reduction of "**B**oth" cases was obtained in two ways:

- assigning the prevalent gender code to the names classified as "**B**oth". For this purpose the mother tongue correctors have been asked to change into "**F**emale" or "**M**ale" the names that are prevalently used for a gender and only exceptionally for the other.
- assigning a gender code ("**F**emale", "**M**ale", "**B**oth" or "**N**ot used") for each language to every name in FNDB. This action is critical to improve quality since a name may have different genders in different languages. This is the case, for example, of "Andrea", which is male in Italian and female in Spanish and German. If the gender is not assigned to each language for every name, "Andrea" could only by classified as "**B**oth". Moreover if "Andrea" is found only in Italian sources, Spanish and German inventors called Andrea would be classified always as "**M**ale".

Diacritics and Multiple Names

The original design of FNDB allowed to distinguish between different inflections of the same name[2]. However no case of gender inflection[3] dependent was found. Besides the EPO database and several scientific journals rarely use diacritics in the names of authors and inventors, even if the correct spelling would require them. Sometimes tonic accents are used instead of graves ones and vice versa and the capitalisation rules of multiple names are not uniform.

To avoid useless proliferation of records, and to facilitate checking and gender assignment procedures FNDB ver.1 has been simplified grouping all the different

[2] Two spellings of the same name are called inflections when they only differ in stressed letters or special characters like "ñ", "ç", "Ǿ"). For example "Frederic" and "Frédéric" are inflections of "FREDERIC", while "Frederik" is considered as a different name.

[3] Cases like Michéle (French - Female) and Michele (Italian - Male) are effectively managed as "Language Dependent"

inflections in just one name. The German names that require umlauts (“ä ë ï ö ü”) and may be transliterated (e.g. with ä converted into “ae”) have been duplicated (i.e. both spellings appear in FNDB).

It is very difficult to obtain a good coverage of double names. The chance to miss some of the possible combinations is high. On the other hand, including all the combinations would lead to an exponential growth of the database with cost increase and loss of performance. In most cases there is no way to distinguish between a compound name and a second name: in French the double names are usually separated by hyphens (“-“) but similar and commonly accepted rules do not exist in other languages. Therefore it is necessary to process both cases in the same way. FNDB always uses hyphen as separator (no blanks are allowed). Multiple names are classified accordingly to the classification of the first component classified as "F" or "M".

Content and size of FNDB

The FNDB ver.1 structure is arranged as a table, (see the example illustrated in fig. 1.5) and is distributed as an Excel file or as a tab-delimited text file.

It is organised as follows:

Column 1 contains the names translated in plain ASCII-uppercase,

Column 2 contains the *generic* gender classification. If the name belongs to the same gender in all the languages in which it is used, the *generic* gender classification in column 2 contains that gender code, otherwise it contains "**x**" to indicate that the name is *language dependent*. Column 2 is also used to classify foreign names not used in any of the 6 languages.

Columns 3 - *n* contain the specific gender classification (**F**emale, **M**ale, **B**oth, **N**ot used) for the given language (Italian, French, English, German, Sweden, Spanish). Specific gender codes are usually written in uppercase. Lowercase codes indicate that the name has not yet be checked by at least two different mother tongue persons.

name	**general**	**GB**	**FR**	**DE**	**IT**	**ES**	**SE**
ANDRE	M	M	M	M	N	N	N
ANDREA	x	B	B	F	M	F	N
ANDREAS	M	M	M	M	N	N	M
ANDREE	x	N	F	B	N	N	N
ANDREES	M	M	N	N	N	N	N
ANDREI	M	M	M	M	M	N	N
ANDREINA	F	F	F	N	N	F	N
ANDREJ	M	M	N	M	N	N	N

Fig. 1.5: Example of FNDB ver.1 structure

Presently FNDB ver.1 contains 8.291 records. The following figures show the distribution by gender and country. FNDB includes 717 (8%) classified as **N**ot used by all mother tongue people. Even if they do not belong to any of the six countries, these names have not been discarded since they come from acceptable quality sources. and demonstrated to be particularly useful in gender assignment of inventors and authors of not European nationality.

Gender	*n. of different names*
Female	3634
Male	4115
Both and Language dependent	542

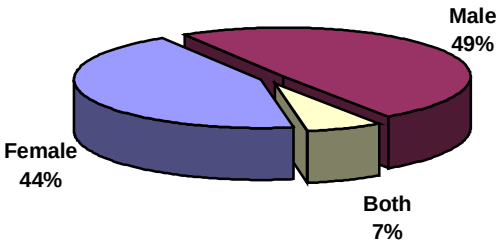
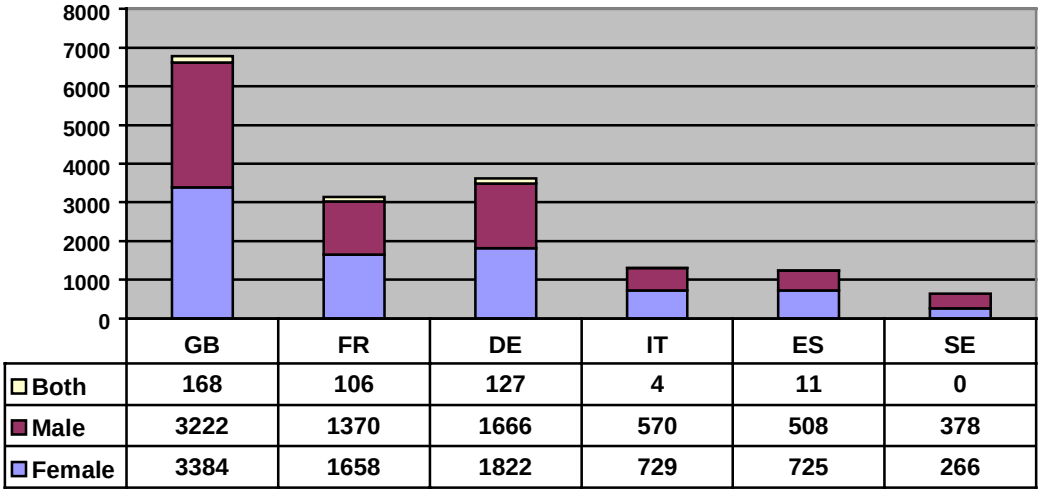


Fig. 1.6: Number of names by gender



	GB	FR	DE	IT	ES	SE
Both	168	106	127	4	11	0
Male	3222	1370	1666	570	508	378
Female	3384	1658	1822	729	725	266

Fig. 1.7: Number of names by gender and country

The sum of the names by country exceeds the total number of names in the database since many names belong to more than one country.

FNDB ver.1 was used to assign a gender to the authors/inventors names in the two feasibility studies for bibliometric and patent indicators.

2. Data Collection and Processing

2.1 Collection of Bibliometric Data

All the relevant international bibliographic databases contain only the initials of the names of the authors. Therefore, to perform gender analyses of scientists based on the first names, the authors' names have to be collected from the original publication in the library or on Internet.

The relevant number of processed items and the necessity of performing a feasibility study led to basing the sampling procedure on an "*a priori*" selection of the journals rather than on a random selection of the items.

Journals have been selected on the following basis:

- high availability of authors' first names
- high frequency of items written by authors of one of the six countries
- high scientific relevance
- balance of the geographical and disciplinary coverage

Some journals with a low relevance level and several journals of different fields of the same disciplinary sector (Medicine) have been added to the sample in order to identify whether the relevance or the subject matter may interfere significantly with the gender distribution.

The first selection was made from the journals with a high number of authors from Spain and Sweden, the less represented in SCI among the six countries, in order to reach a good coverage for all countries. Moreover, since it is impossible to predict in advance the amount of first names actually available in the chosen journals, the selection of the sample was built in a dynamic way, carrying out adjustments in real time during data collection.

The sampling technique adopted in this study could introduce a bias in the statistics when the variables are aggregated by country or discipline. For the indicators by country the bias introduced by the deterministic selection of the journals was significantly reduced by processing the journals thoroughly, thus obtaining a geographical distribution very close to SCI (fig. 2.1)[4]. The indicators by disciplines, on the other hand, need to be normalised since the selection was not balanced and even SCI is not universally accepted as representative of the disciplinary distribution of the papers published throughout the world in international journals.

Data collection of authors' names and nationalities was performed manually in the library. An online attempt to visit the web sites of the main publishers was initially made to verify the availability of the first names. The result was frustrating because 1995 editions are rarely available on the Internet, as the online offer only started in recent years. In any case most of the journals available on Internet publish only the contents of their issues and this is not sufficient for our purposes since first names and working addresses are usually printed only on the first page of the paper.

A precompiled form for data collection was produced for each journal selected. The forms include only the items written by authors from the six countries, sorted by issue and page number. For each item the following data are printed:
- the bibliographic information needed to identify the item,
- the surnames with initials of all the authors
- the country code of authors' working addresses. ISO codes were used for the 6 countries while all the other countries were coded as "nn". For the items written by authors of different countries, all the codes of the countries involved were listed since SCI do not provide links between authors and addresses.

[4] Even if the presence of US authors in the SCI database could be overestimated with respect to the European authors, the validity of the SCI geographical distribution inside and among the European countries is universally accepted.

The forms were filled in manually. The first names, when available, were written in beside the surname. The country code was checked only for the items with authors of different countries. In a few cases items were found containing the first name for some authors and only the initials for the others. These cases (called "Mixed items" in this report) were processed as well and were included in the data sample. If the first issues of a journal contained only items mixed or without first names, the journal was unselected.

As mentioned before, only the first issues of the journals with a high number of articles were processed to avoid an excess of items of the same discipline. Multidisciplinary journals were always processed in their entirety.

Certain difficulties encountered during data collection and processing deserve to be mentioned here because they make data collection more difficult and increase processing time. These are the following:

- The structure of the articles, the information available and the typographical formats change drastically from one publisher to another and may also vary from one journal to another.
- Within the same issue the authors' first names may be given, without any consistency, either as initials only or as full names. Sometimes both are used even in the same item. Surnames may follow first names or vice versa.
- The indication of the working addresses may or may not be linked to the authors' names. The way of associating addresses to names varies greatly item by item.

Journals usually report the name, address and country of the institutes where the authors work but they never report the authors' nationality. Therefore statistics on national productivity refer to the countries where the authors were working rather than to their actual nationality.

Sample information

Although the selection of the sample does not meet the strict criteria of statistical representation, the following figures show that the distributions of the sample by country, discipline and type of publication is similar to the SCI database. In Chapter 5.1, other possibilities will be mentioned to increase the representation.

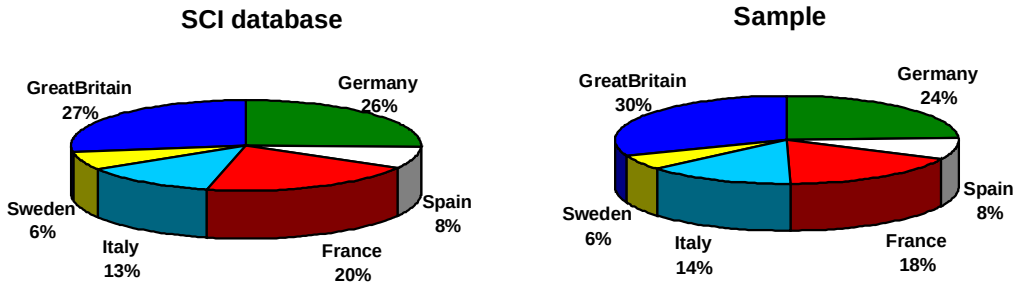


Fig. 2.1: Distribution by country in SCI and in the data sample

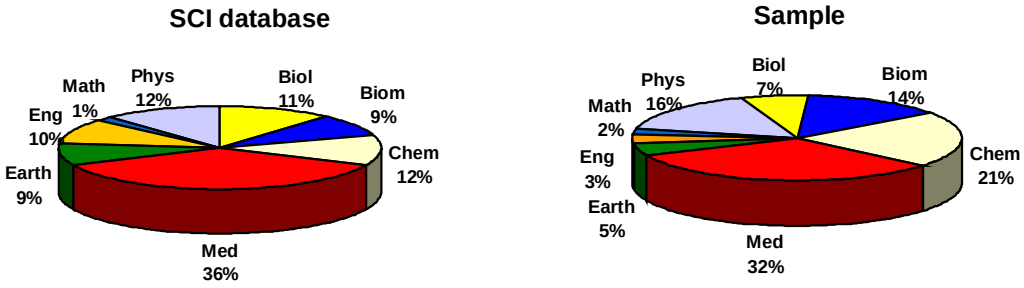


Fig. 2.2: Distribution by discipline in SCI and in the data sample

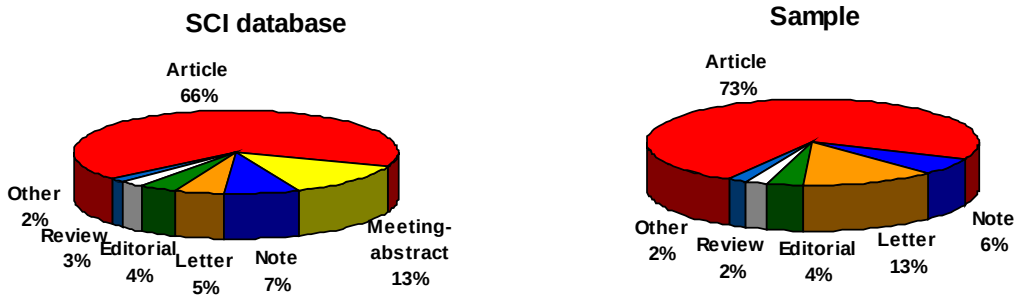


Fig. 2.3: Distribution by type of publication in SCI and in the data sample

2.2 Collection of Patent Data

Data were extracted from the *First '98* database, produced by a co-operation between the World Intellectual Property Organisation and the European Patent Office (EPO). This database contains the bibliographic data of all the European Patent Applications and the PCT International Applications published by EPO in 1998 and is distributed on 5 CD Rom.

The following fields from the EPO First 98 database were downloaded for each patent containing at least one inventor working in one of the 6 countries selected for this study:

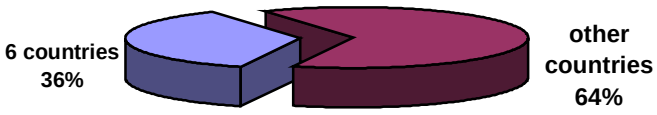
- Patent ID
- Filing and publication dates
- Full name of the inventor(s)
- Working address of the inventor(s), including country code
- Country code of the applicant(s)
- IPC classification code(s)

The preferred format for the inventors names (*family name, comma, first name(s), comma, optional titles*) may have several exceptions: initials may appear together/instead of first name; the second name, if present, may be separated by comma or space; double names may be separated by spaces or hyphenated; address and/or notes may appear after the name at new line.

While the working/residence address of the inventors is available for both European and PCT applications, the inventors' nationality is only included in the PCT applications.

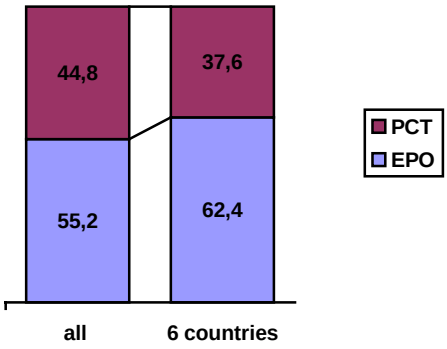
Nationality was downloaded and used to identify the gender when the inventor's name was classified in FNDB as "*country/language dependent*". Statistics are always based on the working/residence addresses of the inventors, therefore "*national production*" always refers to the country where the inventor was working and not to her/his real nationality.

Content of the database



The number of patents produced by the 6 countries and the number of inventors represents the 36% of the whole '98 EPO database. The mean number of inventor per patent is 2.4.

Fig. 2.4: Share of patents produced by the 6 countries



In the whole database PCT applications represent the 45% of the total number of patents. The share of PCT applications is smaller for the 6 countries with a percentage of 38%.

Fig. 2.5: Number of EPO and PCT applications

The following figure shows the distribution of the patents by country. It must be taken in account that the patents produced by inventors of different countries increase of one unit the total of each co-operating country.

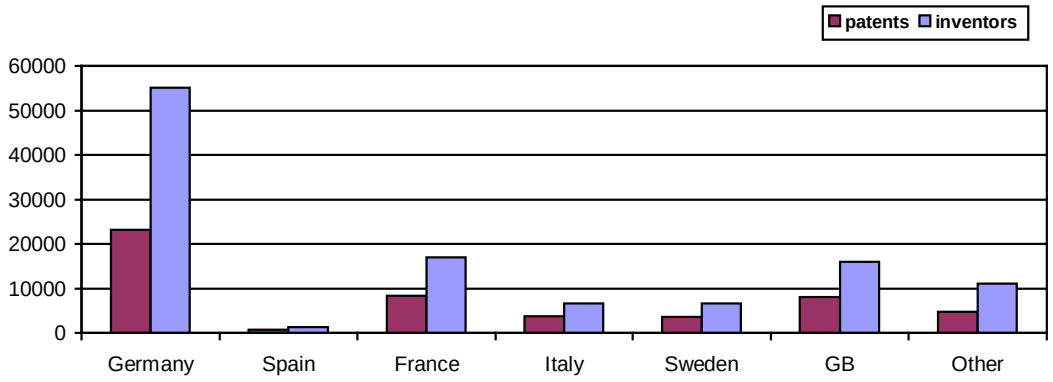


Fig. 2.6: Number of patents and inventors by country

German inventors are almost one half (48%) of the total and are involved in 44% of the patents. Germany is followed by France and Great Britain (both with 15% of inventors),

Italy and Sweden (6% of inventors). About 10% of the patents are produced in co-operation with inventors working outside the six countries.

3 Data Analysis: Methodology

Three indicators were introduced and used throughout this report in order to process publications produced by co-operation among authors/inventors of different countries and gender in a consistent manner.

Participation counts the number of publications/patents with at least one author/inventor of a given gender.

Contribution measures the involvement of each gender in the production of a publication/patent assuming that each author/inventor contributed the same amount.

Number of authors/inventors Total count of the authors/inventors of a given gender in each publication/patent.

The following chart exemplifies the calculation of the three indicators in the case of a publication produced by four authors/inventors.

Gender				Female Participation	Female Contribution	Female Total Count
F	M	M	M	1	1/4	1
F	F	M	M	1	2/4	2
F	F	F	M	1	3/4	3
F	F	F	F	1	4/4	4

Fig. 3.1: Calculation of Female Participation, Contribution and Total Count

It must be noted that the sum of the percentages of *participation* by gender or by country usually exceeds 100% since publications/patents are generally produced by authors/inventors of different gender.

The difference between the values of *participation* and *contribution* of the same subset of authors/inventors is an indicator of the level of co-operation between different gender while differences between *total count* and *participation* or *contribution* are measures of the level of co-operation within the same gender.

4 Summary of the Results

The bibliometric study is based on a data sample of 10,000 items published during the year 1995 in scientific journals of international relevance and written by about 35,000 authors from six European countries: Britain, France, Germany, Italy, Spain and Sweden. For the same countries, the patent study was done with patents published in the year 1998 by the European Patent Office (EPO). The EPO '98 database contains 132.845 patents produced by 313.463 inventors, whose working address is in one of the following six EU countries.

4.1 Publications by Gender

The gender classification of the authors was obtained using First Names Data Base (FNDB) that associates gender and language to more than 8000 first names used in the six countries subject of this analysis.

Results confirm the feasibility of the approach using the first name to identify the gender of the authors and the satisfactory coverage of FNDB ver.1: 91% of the authors whose first name was available (36,239) were classified as "Female" or "Male", 3% as "Both" (names used for both genders) and therefore unidentifiable, and about 6% were not classified because not found in FNDB.

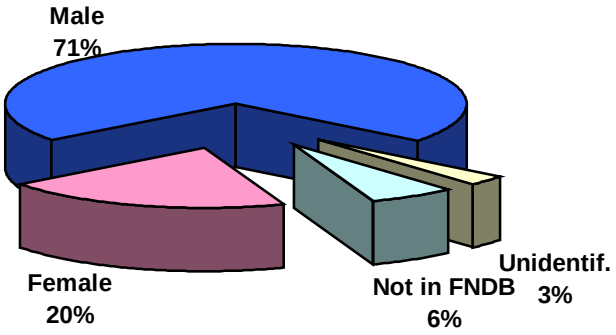


Fig. 4.1: Distribution of the gender classification of the authors in the data sample

In the following diagram *Participation* is represented as the percentage of items with at least one "Female" or "Male" relative to the total number of items in the sample. *Contribution* and *Number of authors* are shown as percentages of the total number of authors classified as "Male" or "Female" (excluding the cases of "Both", "Not in FNDB" and missing names).

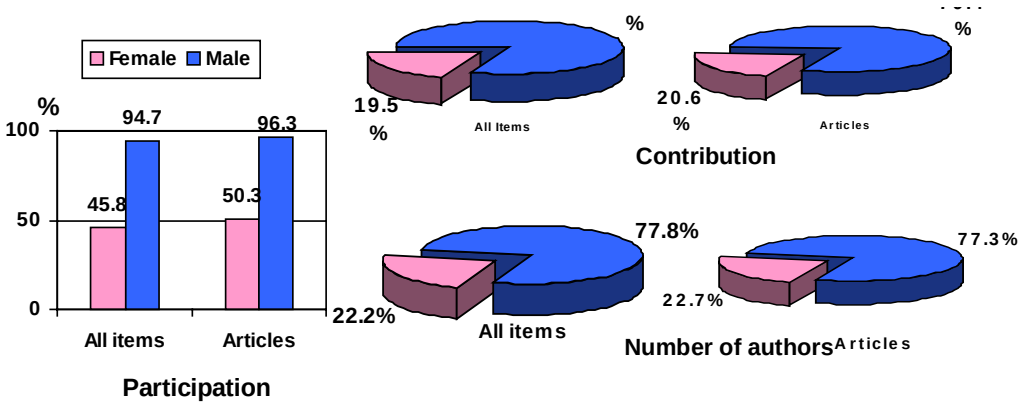


Fig. 4.2: Gender indicators for all countries and all disciplines

Publications by Gender and Discipline

In the following figures the gender indicators by discipline are shown for the whole sample. Only the authors classified as "Female" or "Male" and their respective items have been taken into consideration.

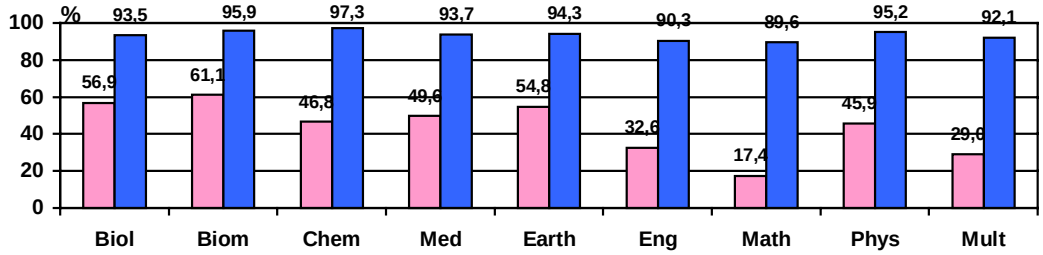


Fig. 4.3: Participation by gender and discipline

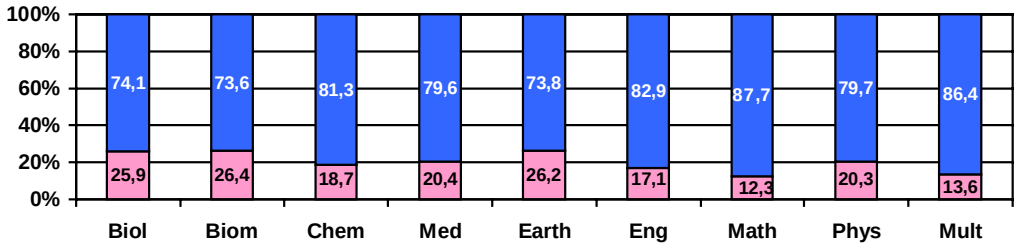


Fig. 4.4: - Contribution by gender and discipline

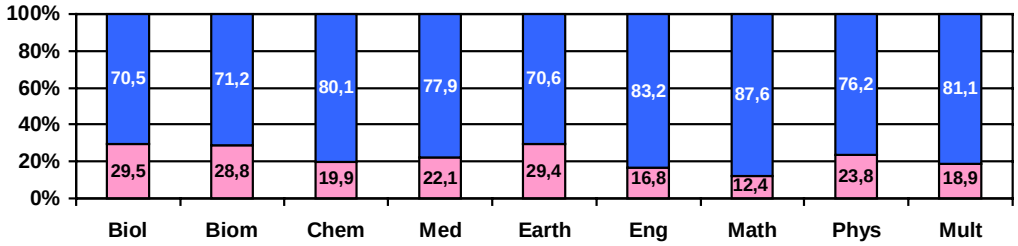


Fig. 4.5: - Number of authors by gender and discipline

Publications by Gender and Country

The following table shows the distribution of the authors by gender and country. Only the authors classified as "Female" or "Male" and their respective items have been taken into consideration.

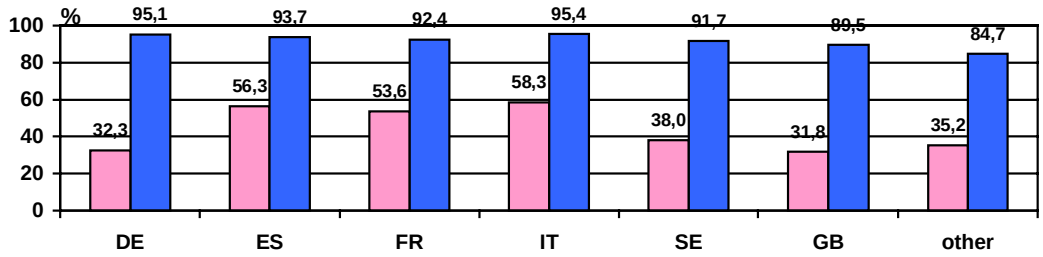


Fig. 4.6: Participation by gender and country

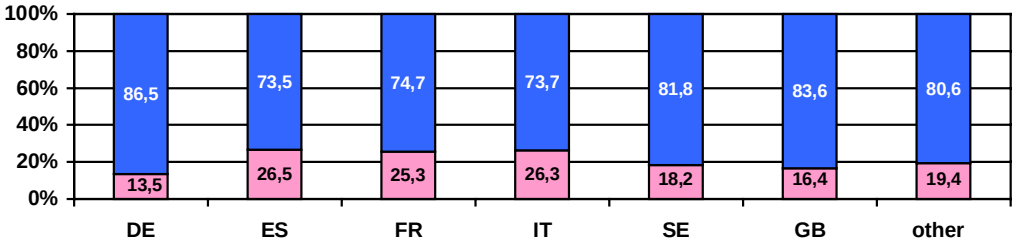


Fig. 4.7: Contribution by gender and country

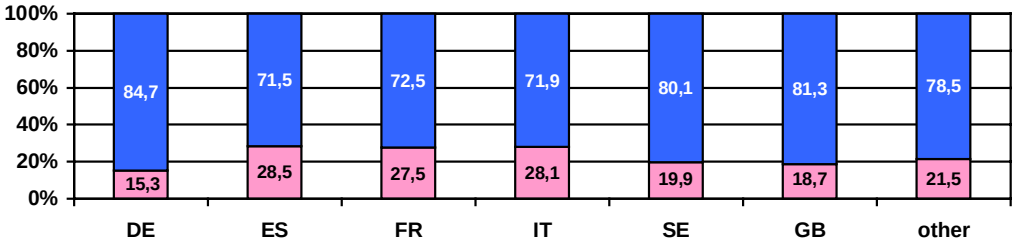


Fig. 4.8: Number of authors by gender and country

Gender Distribution by Journal

The gender distribution within the journals of the same discipline has been analysed in detail. This exercise has been purposely conducted to find out whether and how the selection of journals may influence gender indicators. While most of the disciplines are stable, some others show large differences between journals. Figure 4.9 shows the example of female and male participation in Clinical Medicine, which is the most heterogeneous field of all analysed disciplines. This example illustrates the specific problems in order to reach a representative selection of journals.

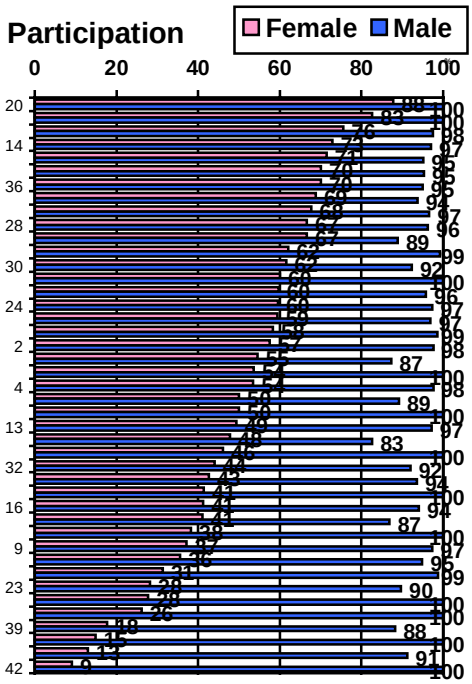


Fig. 4.9: Gender distribution by Journal – Example: Clinical Medicine

Gender distribution by impact factor

The Impact Factor (IF) Ranking List in each discipline has been used to measure the relevance of the journals. Journals and all their items are assigned to an IF level that ranges from 1 (the most relevant 10%) to 10 (the least relevant 10%). Since the journals with higher IF values were preferred in selecting the data sample, IF levels from 6 to 10 contain few items and have been grouped together in the following diagrams. The items published in journal classified in more than one disciplinary sector are counted twice.

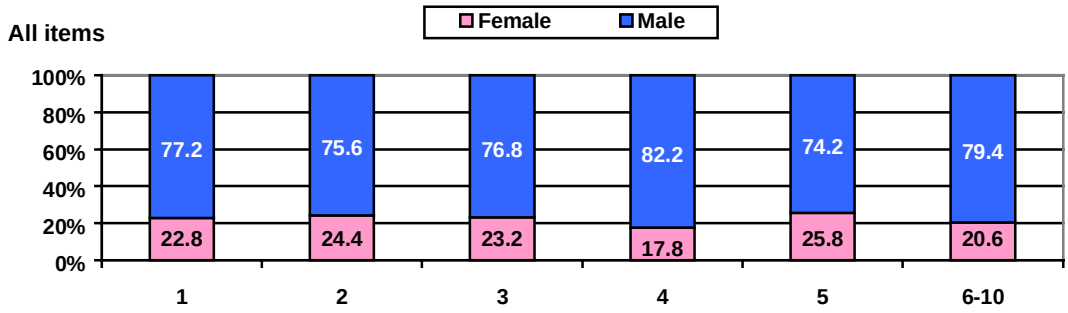


Fig. 4.10: Gender distribution by Impact Factor

The presence of women seems to be higher in the more relevant journals. Since the possible differences are marginal, the hypothesis should be verified analysing the data desegregated by country and discipline since the geographical and disciplinary distribution of the sample is different from that of the original population.

4.2 Patents by Gender

The main indicators extracted from the “Patent Indicators by Gender” study are reported in the following figures.

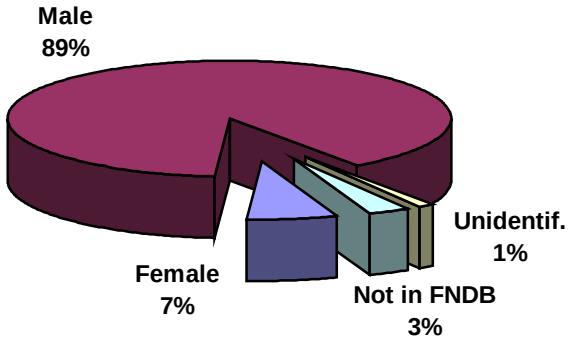


Fig. 4.11: Distribution of the gender classification of the inventors of the 6 countries

As already documented in section 1 a satisfactory coverage in gender assignment has been achieved: 96% of the 114.157 inventors have been classified as "**F**emale" or "**M**ale". 1% of the inventors has the first name classified as "**B**oth" and about 3% of the names have not been classified because not found in FNDB.

In figure 4.12 only the items classified as "**M**ale" or "**F**emale" are considered. 12,5% of the patents have at least one female inventor and 97,3% of the patents have at least one male inventor. As a consequence 87,5% of the patents have been produced only by men and 2.7% only by women. On the other hand female inventors are 7% of the total number and contribute to the overall production of patents with 5% of equivalent-patents.

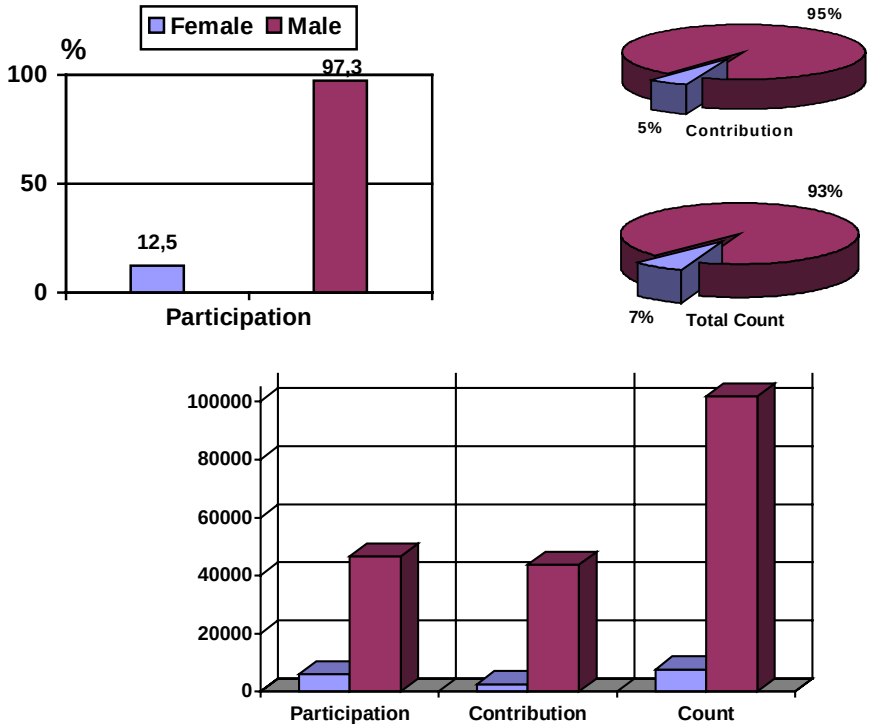


Fig. 4.12: Gender indicators for the six countries

Patents by Gender and Country

The following figures show the distribution of the inventors by gender and country. The country with the highest percentage of female inventors is Spain followed by France and Italy. Germany has the lowest percentage of female inventors (4,6% versus 15,8% of Spain). Since Germany produces about 50% of the patents, the low German percentage of female inventors influences significantly the global statistics.

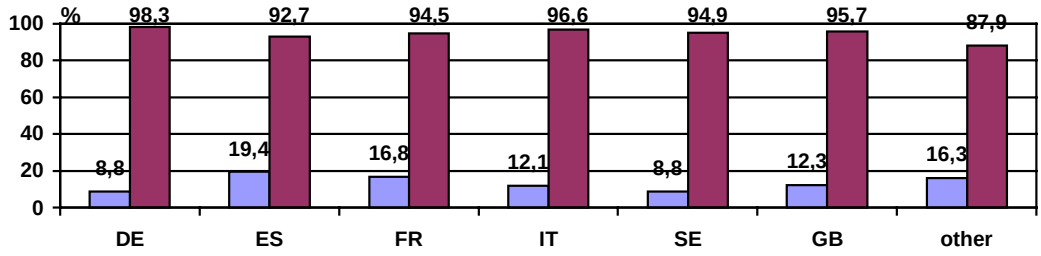


Fig. 4.13: Participation by gender and country

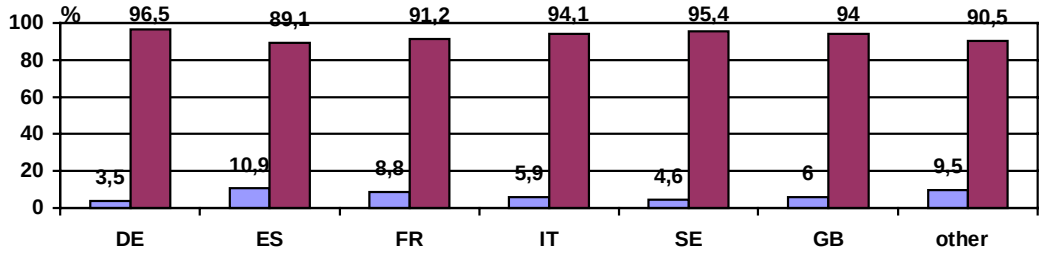


Fig. 4.14: Contribution by gender and country

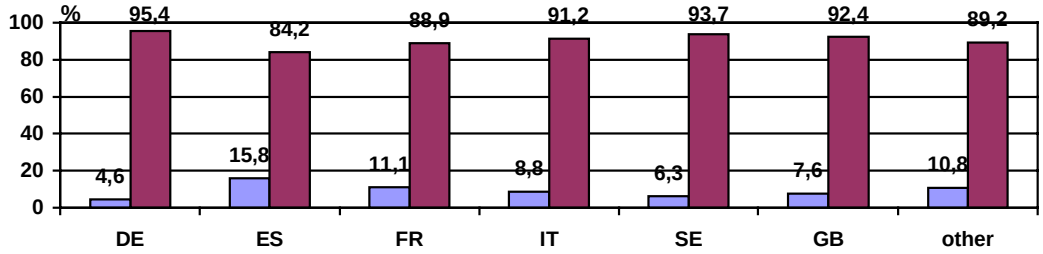


Fig. 4.15: Number of inventors by gender and country

5 Extension of the Studies to the 15 EU Member States and to Five Years

In this chapter, the extension of the two feasibility studies to the 15 EU Member States (additionally to the six analysed countries: Austria, Belgium, Denmark, Finland, Greece, Ireland, Luxembourg, Netherlands and Portugal), and to a time period of at least five years are discussed. The results of the studies show that there are big differences between the bibliometric and the patent studies. Therefore, they are treated again separately.

5.1 Bibliometric study

The most critical and expensive task of the bibliometric study is the collection of the authors' first names from the libraries. The main problems are the data entry procedure and the identification of an optimal sample, which provides the best geographical and disciplinary coverage with the minimum set of data. Especially important seems to be the disciplinary coverage within each disciplinary sector due to the presence of research fields, and perhaps journals, which are particularly "*gender dependent*".

The sampling strategies proposed are based on the following assumptions:

- Many journals never publish authors' first names and for those that do publish them the first names are actually only available in less than 60% of the items. On the basis of the results of this study the number of items with authors' first name is estimated to be about 1/3 of the total items published. The cost estimates take into account the fact that the selected items must be at least three times larger than the required data sample.
- According to the general statistics shown in Section 4.1, the average number of authors per items depends on the discipline as summarised in the following table

Discipline	Authors per item
Biology, Biomedical Research, Clinical Medicine	4.7
Chemistry, Earth and Space, Physics	4,3
Engineering	3,2
Mathematics	1,8

Fig. 5.1: Authors per item per discipline - Selection

- These values must be taken into consideration when designing the sample if, as in this study, the data collection is based on the items and the statistics are based on the authors.
- Section 4.1 shows that within each disciplinary sector there are disciplines -and maybe journals- particularly oriented to a specific gender. This is especially true for the less homogeneous sectors (Medicine, Engineering, Earth and Space Sciences) that contain disciplines where the percentage of female/male researchers is very different. As a consequence it is advisable to include in the sample items representative of all the disciplines of each sector. More precisely, the following procedure could be followed: *(a)* select a set of disciplinary groups (see next point) greater than the 9 sectors used in this study, *(b)* design the sample aiming to obtain a good coverage for each group, *(c)* produce the statistics for each group and - only at the end of processing - *(d)* aggregate the results into the 9 original sectors.
- The SCI classification schema (183 disciplines for the 1995 version) is probably too rich to be used for gender statistics desegregated into specific disciplines as suggested in the previous point. Many disciplines are very close to each other (for example there are 8 disciplinary codes relating to "Material Science" and 8 codes relating to "Psychology") and are often associated with the same journals. We estimate that to perform gender analyses SCI codes could easily be grouped into no more than 60-80 disciplinary groups. Grouping should be performed according to the similarity of the disciplines and the number of common journals. A correspondence table similar to the one used in this study could be used to link each group into the 9 disciplinary sectors.
- Gender distribution of an item may depend on the position of the item inside the actual publication. That is because many journals are organised into sections dealing with different topics (e.g. basic vs. applied research, field of interest, etc.) that may be gender-dependent. As a result, if no fully randomised sampling techniques are adopted, issues should be processed in their entirety.
- Gender distribution does not seem to depend on the issue (i.e. there is no reason why the January issue should have more female authors than the November issue or vice versa). As a result if a journal has a great number of items and is published frequently, it should be statistically correct to process only certain issues, provided that the selected issues are fully processed (see the previous point). For the same reason it is not strictly necessary to select the issues randomly. This rule does not apply to the multidisciplinary journals that generally publish thematic mono-disciplinary issues.

As already mentioned in the introduction, the collection of the authors' first names in the library is the most expensive task of the study.

The task can be split into 5 stages, each with a specific cost:

- Identification of the library / internet site where the full text of the journal is available
- Access to the journal and preliminary analysis to check the presence of first names. In our experience printed publications may be more complete than the internet version. For this reason it is advisable to check the printed copy before discarding a journal.
- Localisation of each selected item (number/year, issue and page)
- Collection of authors' first names and, where necessary, of their nationality.
- Input of names and nationalities into the sample database.

Correct planning and management of the job are crucial to contain costs. The best solution depends on the working environment and the local organisation. In any case and independently of the sampling technique adopted, it is important to produce the list of all selected items sorted by journal, year, issue and page, before starting the data collection. In this way all the items of the same journal can be collected at the same time without having to repeat the time consuming task of accessing the journals.

Data collection and inputting are critical since spelling errors can easily be introduced. In this study we:

- Kept data collection separate from data inputting, using paper forms filled in manually in the library. Alternatively inputting can be done directly in the library on portable PCs using electronic forms possibly equipped with manual scanners and OCRs.
- Developed a data entry program that presents the operator with the list of all the names extracted from FNDB with the right initial. The operator can then select a name from the list or type it in, if it is not in the list. A double check specifically made to evaluate the effectiveness of this procedure demonstrated that this method may induce the operator to select the closest name from the list when the actual name is not present in FNDB.

Three different sampling methodologies can be envisaged in order to extend the statistical analysis to all EU countries and to a period of several years:

A. **Fully random** selection of the **items** from all SCI records with at least one EU author

B. **A priori selection** of the **journals** to be included in the sample with the processing of all EU items/articles

C. **Semi-random** selection of **items** from all SCI records with at least one EU author, aiming to collect a similar number of items for each discipline and country (*quota sample*).

Task	*Cost* *min*	*Max*
Methodology A		
1,400 items/year	15,500	24,000
50,000 items/year	162,500	225,000
Methodology B		
150 journals fully processed	82,500	115,000
150 journals with threshold	46,500	67,000
250 journals with threshold	70,000	100,000
Methodology C		
20,000 items/year	85,000	120,000

Fig. 5.2: Cost estimate for data collection

5.2 Patent study

The extension of the analysis to 15 countries and to 5 years should not originate special problems as long as a first name database is available with the characteristics described in chapter 1. The study can be carried out at a total cost of 15,000-30,000 Euro and can be completed in 3-4 months. Additional costs are the updating of FNDB with the names of the new countries.

5.3 FNDB extension

Table 5.3 shows the coverage offered by FNDB ver.1 on data extracted from EPO '98 database for the 10 EU Member States not included in the present study.

	Inventors			Different names		
	total	not in FNDB	coverage	total	not in FNDB	coverage
Austria	2238	48	97,9	428	38	91,1
Belgium	3228	175	94,6	939	75	92,0
Denmark	2319	267	88,5	866	124	85,7
Finland	3431	1181	65,6	561	215	61,7
Greece	69	34	50,7	45	23	48,9
Ireland	456	24	94,7	254	17	93,3
Luxembourg	177	10	94,4	80	4	95,0
The Netherlands	7448	815	89,1	3623	429	88,2
Portugal	54	5	90,7	46	4	91,3
Present Coverage = percentage of inventors included in FNDB respect to the total number						

Fig. 5.3: Present coverage of FNDB ver. 1 for the other countries

Two strategies are feasible: **Strategy A** aims to reach a coverage greater than 95% for each language. The number of new names to be collected should be:

Language :	number of names :
Danish, Dutch, Finnish and Flemish	1500
Greek	1000
Portuguese	1000
Total	3500

Strategy B bases on the following steps:

1. Extraction of the names of the EPO inventors and/or of the authors of scientific papers
2. Preliminary assignment of gender using FNDB ver.1;
3. Manual gender assignment of the names not included in FNDB ver.1 with the assistance of two mother tongue persons for each language

The main advantage of Strategy B is the great cost reduction (3.500-6.500 Euro vs. 13.000-26.000 Euro). Nevertheless, Strategy A is more advantageous with other respects:

- While Strategy B requires to repeat the entire procedure to extend the analysis to other time periods or to other lists of names with additional costs, this is not required for Strategy A.
- Although gender analysis quality might be reasonably good using Strategy B, it is predicted to produce less reliable data than Strategy A.
- A good quality first name database, as the outcome of Strategy A, would be used without extra costs for several other purposes as, for example, gender identification of members of boards, committees, working groups, etc.
- A good coverage allows to extend the gender analysis without extra costs to any other country where the national language is already included in FNDB.
- The availability of a database where names and genders are identified by language would allow various types of cross-analyses with the goal, for example, of producing indicators on the international mobility of scientists.

6 Conclusions

As already mentioned in the beginning, this paper describes the results of a feasibility study, carried out for the European Commission in order to analyse the feasibility of measuring gender in science and technology. The focal point was put on the methodology and the assessment of costs in comparison to a representative coverage of bibliometric and patent data. Results according to the gender issue are not yet analysed and interpreted.

The creation of the First Name Data Base (FNDB) is initiated and can be applied for gender analyses in bibliometric data and patents, but also for all other gender analyses which rely on first names. At the moment, it is done for six languages, but its extension is only a question of money. The methodology seems to be sufficiently developed and proved.

The implications for the collection of gender indicators are twofold. The patent indicators by gender can easily be analysed because of the completeness of the first names in patent databases. The results presented in section 4.2 can therefore be considered as a good starting point for deeper analyses and interpretation of gender differences in technological performance. The extension to 15 EU Member States and other relevant countries can easily be done.

For the bibliometric analyses, the feasibility of the collection of gender indicators relying on first names can be questioned. As only one third of the journals provide initials or full first names, automatically two third of all journals cannot be included in the gender analyses. Distortions due to this cannot be examined. Therefore, the creation of a representative sample is the most important problem to be solved. Given that this can be done, the extension to more countries should be feasible, but requests a larger sample in order to get a good coverage for smaller and less publishing countries for all disciplines. This kind of analyses is time-consuming and expensive, the feasibility depends on the amount of money, which can be spent on this exercise.

The presented studies are an important step towards the better understanding of gender differences in scientific and technological output and therefore a new tool of analysing the gender question with sophisticated indicators more deeply. The European Commission will follow this subject in the future.