Mapping Excellence in Science and Technology across Europe

Life Sciences

Final Report

Centre for Science and Technology Studies
Leiden University, the Netherlands

ECM Noyons
RK Buter
AFJ van Raan

Fraunhofer Institute Systems and Innovation Research
(Fraunhofer ISI), Karlsruhe, Germany

Ulrich Schmoch
Thomas Heinze
Sybille Hinze
Rebecca Rangnow

October 2003
Report of project EC-PPLS CT-2002-0001 to the European Commission

Executive summary

This report describes the potential of bibliometric and patent studies to identify centers of excellence in Europe in science fields. In particular four Life Sciences fields of are taken as case studies. The contractors describe the methodological and practical problems of the used data and analyses as well as the potential and feasibility of such an exercise.

To summarize, the major questions of the present report are whether

- it is possible to handle the large data sets linked to the considered areas in a satisfying way,
- such an exercise can be executed with a reasonable amount of work and costs,
- it is possible to identify relevant institutions in the areas considered,
- the results can be presented in a way that is adequate to the needs of different user groups.

Based on the experiences of the present exercise, various measures to improve the methods of analysis and presentation are suggested. The entire project with the developed tool is available at www.cwts.nl/ec-coe.

We succeeded to create a tool that enables various types of users to identify on the level of 'main organization' (university, company) centers of excellence. Still, in the course of this project we encountered some problems that should be solved in order to carry out the identification in a cost-effective way and with reliable results on a larger scale that just one field.

Expert input is crucial for collecting the proper publication and patent data to be used as the basis for the analyses and particularly the delineation of fields. At present it is not possible to collect these data without experts who are able to compile an effective search strategy to retrieve the relevant data. These experts should not only 'know' the field but also have knowledge about search strategies and their use. These experts can be supported by a search interface to see the effect of specific search strings.

In view of the difficulties we encountered with the input of experts in order to delineate the fields, we wonder whether this is the best approach to meet the objective. In practice, there are two types of 'fields'. One concerns a 'known' field, i.e., a field established for years and years (e.g., genetics, immunology, and neuroscience). For these fields, delineation is much easier because experts know the journals that cover the core of the field and they have ample experience using search terms. The second type is the 'developing' field (e.g., bioinformatics). In this field there appears to be much less consensus among experts about what should and should not be covered, there are hardly or no journals specifically for that field, and the experts have a rather limited experience as to what search terms to use for collecting relevant data. Furthermore, the results for the 'known' fields are much easier to validate than the results for the 'developing' fields. As the results should in some way refer to what the field experts expect, validation appears to be considerably easier for the 'known' fields.

It should also be noted that field delineation on the basis of patents differs from publications, because patent databases contain a well-developed classification system (the International

Patent Classification, IPC). This classification creates an additional and powerful facility to collect the proper data. The publication databases essentially lack such an overall generic scheme.

The above observations lead to the conclusion that implementation of our approach as described in this report on a larger scale (i.e., applying it to hundreds of 'fields') is not feasible, simply because we expect that it is impossible to get experts involved on such a large scale in a reasonable way, without losing control over the results. Moreover, we know that the science landscape is changing, and that particularly new and developing fields will attract interest to identify centers of excellence. But precisely in these developing fields delineation of the field on the basis of expert input is problematic, as discussed above. However, there are good prospects to deal with this delineation issue, but this has to be investigated in more detail. In principle it should be possible to start with a limited set of publications and to enlarge this set on the basis of co-citation relations, similar keyword patterns and other bibliometric characteristics.

With respect to the use of address data in publication and patent data, we conclude that they may be used at the level of 'main organization' (university, company, research institute) in most member states of the EU and associated states. At that level, cleaning of data by national experts is certainly feasible. It seems however, that problems with cleaning are not the same in every country. Apart from the size of the country, it is well known that the science system in countries like France (particularly, the 'interwoveness' of the CNRS) differs considerably from the system in the Netherlands. The complexity of the system in France makes it almost impossible, also for national experts, to clean the data, even on the level of organization. Cleaning of these address data would be easier in a 'bottom-up approach'. This means that beforehand a limited list of organizations has to be compiled within each country. Then the address data could be cleaned using this basic list of organizations.

With respect to linking patent and publication indicators, we have made in this project a huge step forward as we were able to identify inventors as authors in the same field. Hence, we were able to identify the 'research address' of inventors and thus to build indicators for institutions having both patent and publication data. This enables us to find 'bridges' between scientific and technological performance within an R&D field.

In this project, we created a tool for different users to enter the fields chosen for this study. The design of this tool had to be flexible enough to be used by different types of users. Because of the variety of users (from scientific experts to policy makers), we are not yet completely able to determine whether the requirements of all users are satisfied. Still in view of the purposes of this study we are convinced that we indeed have. The tool enables users to determine their own criteria and thresholds to identify research entities of a certain productivity or impact. In particular, the possibility to combine different indicators enhances the utility of the tool for the different user groups considerably. On a large scale we were able to combine patent and publication indicators, which can be considered as a major step forward to explore the multiple aspects of excellence.

With respect to the size of research entities, we were within the scope of this project not able to go a step below the level of 'main organization', e.g., from university to department. It appeared that the quality of address data in publications on the level of departments and (if available) faculty, was so low, that we do not provide results on department level systematically. Moreover, the cleaning efforts for experts in the different national science systems would be huge. Especially in larger countries like Germany, France and the UK, we could not ask to clean the address data at any lower level than the main organization.

It should be noted that the activity and performance of these organizations are only measured within the field. The name of the organization as mentioned in the tables and rankings do not refer to the entire organization but only for the part active in a particular field.

Apart from these data problems, we mention the debate on the validity of performance indicators on the level of departments. For some purposes and within particular contexts, the entity to focus on should be even below the departments. In these cases the ‘group’ seems more appropriate. In this study we were not able explore this, but we have ideas as to how to deal with this. We suggest that combination of author names and organization name could be used effectively to define groups. A combination of groups may be used to define a department or even a faculty.

Still, as mentioned above, we were able to provide information on research in a specific field in an efficient interactive tool, enabling the user to use his/her own criteria and thresholds to identify research entities at the level of organization, with a particular performance. Moreover, we provide the tool at different levels of aggregation (world, EU, and national level).

The geographical interface can be used to localize the identified organizations. This enables a specific user to search for entities together with the information of its geographical position.

Table of contents

1 Introduction

In January 2000, the European Commission adopted a communication preparing the creation of a European Research Area (ERA).[1] This project essentially aims at creating favorable conditions to increase the impact of research efforts by strengthening the coherence of research activities and policies conducted in Europe. In particular, it recognizes that world-class excellence exists in practically all areas and disciplines in Europe. These competencies, however, are not always sufficiently well-known across national borders, for example by companies. At the Lisbon European Council on March 2000, the Heads of State or Government endorsed this project and set a series of objectives and an implementation timetable. In particular, they requested to map research and development excellence in all Member States, in order to foster the dissemination of excellence.

In order to ensure close co-operation with Member States on the subject of excellence, Commissioner Busquin convened a group of nominated representatives from Member States, the High-Level Group (HLG). On November 2000, a specialized workshop was organized, where experience in various countries was presented regarding methods for assessing excellence. Based on these experiences, the commission published a paper on the methodology how to map excellence in research and technological development in Europe (CEC 2001).

In 2000, the commission additionally assigned an expert group in order to discuss appropriate indicators for a Europe-wide mapping of excellence. Various indicators were suggested, but the experts agreed on publication indicators as starting point. The main argument for this approach was that all other indicators have to cope with the problem of comparability across national borders in a substantial way. In addition, the experts recommended to complement the analysis by patent indicators for including the technological perspective. They emphasized that patent indicators are not only useful to describe excellence of industry-based research, but also to depict the orientation of scientific research on application. So the combination of both indicators should show up different dimensions of excellence. Later on, these lead indicators should be complemented by other indicators.

Against this background, the Commission decided in 2001 to analyze the three areas of economics, life sciences and Nanotechnology in more detail. The major aim of this exercise was to examine the feasibility of such analyses in various dimensions. For example, it should be tested whether it is feasible at all to provide data necessary for an appropriate mapping, but also aspects of an appropriate presentation of a large amount of data with regard to user needs or the costs were of concern.

In November 2001, the Commission awarded two short preparatory studies as to life sciences and Nanotechnology for exploring first methodological issues. In March 2002, it commissioned two broader studies for analyzing patents and publications in these two areas. The present report describes the outcome of these studies wherein the Centre for Science and Technology Studies (CWTS) of the University of Leiden was responsible for publications, the Fraunhofer Institute for Systems and Innovation Research (Fraunhofer ISI) for patents. The results of this study are made available by a special interface at the WWW (http://www.cwts.nl/ec-coe) where patents and publications data at institutional level can be

[1] The following description is based on CEC (2001: 3).

browsed, and geographical maps of centers of excellence can be generated. This report should be read as a twin sister of the nanotechnology report of the project EC-PPN-CT-2002-0001.

1.1 Objective of this study

This report primarily describes the methodological and practical problems of the analyses and addresses the various aspects of feasibility in more detail. In addition to the institutional analysis, it discusses some macro-statistical issues of the exercise. The work of the two main contractors was supported by recommendations of the High Level Group and in particular by a stakeholders' panel with relevant experts in the areas considered. The Commission is presently going to evaluate the outcome of the studies and to supplement the publication and patent indicators by other data. In the study, the area of life sciences was broken down into the four sub-areas Genetics/heredity, Neurosciences, Immunology and Bioinformatics. As to the geographical coverage, the study focuses on institutional structures in the member countries of the EU and in the so-called associated countries; all in all, on 32 countries.

To summaries, the major questions of the present report are whether

- it is possible to handle the large data sets linked to the considered areas in a satisfying way,
- such an exercise can be executed with a reasonable amount of work and costs,
- it is possible to identify relevant institutions in the areas considered,
- the results can be presented in a way that is adequate to the needs of different user groups.

Based on the experiences of the present exercise, various measures to improve the methods of analysis and presentation are suggested. With regard to institutional details, the reader is kindly asked to refer to the website mentioned above.

1.2 General starting points

Mapping of excellence in science and technology (S&T) is one of the central goals in the *European Research Area* strategy of the European Commission. Excellent scientific work is the origin of breakthroughs, and therefore it is crucial to identify centers of excellence and to promote them. Moderate scientific work will not lead to important breakthroughs. Only excellent work really counts. This is particularly important as scientific breakthroughs are, be it in an unpredictable way (Airaghi et al 1999), the driving forces of socio-economic change and development.

S&T scientific excellence is not always sufficiently well known across and even within national borders. Mapping of excellence and particularly of evolving patterns of excellence in all EU Member States (and EU Associated States) will strongly enable the EC strategy of fostering the dissemination of S&T excellence.

Identification of S&T excellence is a matter of evaluation. Most of the R&D evaluation processes worldwide rely heavily or almost completely on expert panels and other forms of peer review. Undoubtedly, opinions of experts are of crucial importance. Nevertheless there may be severe problems in peer review (Horrobin 1990; Moxham and Anderson 1992). New developments in the field of quantitative studies of science and technology offer methods to support peer review in order to keep it objective and transparent.

This study aims at the application of these advanced quantitative methods. Its main objective is 'S&T excellence mapping', i.e., the application of a sufficiently powerful analytical instrument based on bibliometric analysis as well as on patent analysis to identify excellence in science and technology.

The methodology described in this study is based on the following crucial aspects of mapping S&T excellence.

1. *Definition of the selected fields* on the basis of concepts. A group of experts commissioned by the EC provided keywords for the selected life science fields. In addition, we used field definitions based on journal sets and, where possible, field definitions based on classification systems of relevant data sources;
2. *Collection of relevant documents*: on the basis of the above field-defining elements, we focus on scientific documents published in international journals covered by citation indexes[2], documents covered by Medline, Current Contents, and other relevant data sources. Using similar field-defining elements (keywords, classification codes of patent databases) we retrieved patent data from the European Patent Office (EPO) and the World Intellectual Property Organization (WIPO);
3. Construction of maps on the basis of the data in the collected documents, to represent the cognitive, scientific structure ('bibliometric map') of the selected fields as a 'basic landscape' in order to visualize the mutual relations between sub-fields and (often problem-oriented) themes, as well as the interdisciplinary relations with other fields;
4. Identification of the major actors in the selected fields, both in scientific (based on publication and citation analysis) as well as in technological (based on patent analysis) terms, and positioning of these centers of S&T excellence on the map of the field;
5. Representation of the centers of S&T excellence on a geographical display in order to create a clear overview of European competencies in science and technology;
6. Validation of the maps and inclusion of further surveys in order to enhance the findings of this study, made possible by making the maps accessible in a user-friendly design.

The study will cover all the EU Member States and the associated countries: Bulgaria, Cyprus, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia, Slovenia, Malta, Iceland, Liechtenstein, Norway, Israel and Switzerland.

1.3 Basic Principles of Bibliometric Mapping

Each year about a million scientific articles are published. How to keep track of all these developments? Are there specific patterns 'hidden' in this mass of published knowledge, at a 'meta-level', and if so, how can these patterns be interpreted (Van Raan and Noyons 2002)?

We discussed above that a research field can be defined by various approaches: on the basis of selected concepts (keywords) and/or classification codes in a specific database, selected sets of journals, a database of field-specific publications, or any combination of these approaches. Along these lines, we collected for each selected field titles and abstracts of all

[2] CD-Rom versions of the Science Citation Index, SCI; if necessary the Social Science Citation Index, SSCI; and Arts & Humanities Citation Index, AHCI; and all 'specialty' indexes such as Neurosciences, Biochemistry and Biotechnology, etc., published by the Institute of Scientific Information's (ISI) in Philadelphia.

relevant publications, for a series of successive years, thus operating on many ten thousands of publications per field. With a specific computer-linguistic algorithm we parsed the titles and abstracts *of all these publications*. This automated grammatical procedure yields all nouns and noun-phrases (standardized) that are present in the entire set of collected publications (Noyons 1999).

An additional algorithm creates a frequency-list of these many thousands of parsed nouns and noun-phrases while filtering out general, trivial words. We consider the most frequent nouns/noun phrases as the most characteristic concepts of the field (this can be 100 to 1,000 concepts, say ***N*** concepts). The next step is to *encode* each of the publications with these concepts. In fact this code is a binary string (yes/no) indicating which of the ***N*** concepts is present in title or abstract. This encoding is as it were the 'genetic code' of a publication. Like in genetic algorithms, we now compare the encoding of each publication with that of any other publication by calculating the 'genetic code similarity' (here: *concept-similarity*) of all publications in a specific field pair-wise. The more concepts two publications have in common, the more these publications are related on the basis of concept-similarity and thus can be regarded as belonging to the same sub-field, research theme or research specialty. In a biological metaphor: the more specific DNA-elements two living beings have in common, the more they are related. Above a certain similarity threshold, they will belong to a particular species.

The above procedure allows clustering of *information carriers* -the publications- on the basis of similarity in *information elements* - the concepts ('co-publication' analysis). Alternatively, the more specific concepts are mentioned together in different publications, the more these concepts are related. Thus, information elements are clustered ('co-concept' analysis). Both approaches, the co-publication and the co-concept analysis are related by matrix algebra rules. In practice, the co-concept approach (Noyons and Van Raan 1998) is most suited for science mapping, i.e., the 'organization of science according to concepts'.

Intermezzo: For a super market 'client similarity' on the basis of shopping lists can be translated into a clustering of either the clients (information carriers, where the information elements are the products on their shopping lists) or of the products. Both approaches are important: the first gives insight into groups of clients (young, old, male, female, different ethnic groups, etc.), and the second is important for the spatial division of the super market into product groups.

In main lines the clustering procedure is as follows. We first construct for each field a matrix composed by co-occurrences of the N concepts in the set of publications for a specific period of time. We normalize this 'raw co-occurrence' matrix in such a way that the similarity of concepts is no longer based on the pair-wise co-occurrences, but on the co-occurrence 'profiles' of the two concepts in relation to all other concepts. This similarity matrix is input for a cluster analysis. In most cases, we use a standard hierarchical cluster algorithm including statistical criteria to find an optimal number of clusters. The identified clusters of concepts represent in most cases recognizable 'sub-fields' or research themes. Each sub-field represents a sub-set of publications on the basis of the discussed concept-similarity profiles. If any of the concepts is in a publication, this publication will be attached to the relevant sub-field. Thus, publications may be attached to more than one sub-field. This overlap between sub-fields in terms of joint publications is used to calculate a further co-occurrence matrix, now based on sub-field publication similarity.

To construct a map of the field, the sub-fields (clusters) are positioned by multidimensional scaling. Thus, sub-fields with a high similarity are positioned in each other's vicinity, and sub-fields with low similarity are distant from each other. The size of a sub-field (represented by

the surface of a circle) indicates the share of publications in relation to the field as a whole. Particular strong relations between two individual sub-fields are indicated by a connecting line.

A next step (Noyons et al 1999) is the integration of mapping *and* performance assessment. It enables us to position actors (such as universities, institutes, R&D divisions of companies, research groups) on the worldwide map of their field, and to measure their influence in relation to the impact-level of the different sub-fields and themes. Thus a strategic map is created: who is where in science, and how strong? We discuss the methods applied for performance assessment in Section 3. Furthermore, indicators on patent data are added which provides us with information on technological strength.

A series of maps of successive time periods reveals trends and changes in structure, and even may allow 'prediction' of near-future developments by extrapolation. Such changes in maps over time (field structure, position of actors) may indicate the impact of R&D programs, particularly in research themes around social and economic problems. In this way, our mapping methodology is also applicable in the study of the socio-economic impact of R&D (Airaghi *et al* 1999).

Bibliometric maps provide an instrument that can be used optimally in an electronic environment. Moreover, there is a large amount of detailed information 'behind the maps'. For this study it is of crucial importance that this underlying information, particularly on research performance above the 'excellence threshold' (Section 3) and technological performance can be retrieved in an efficient way, to provide the user a possibility to explore the fields and to judge the usefulness of maps against the user's own expertise. We provide in this study an advanced internet-based user-interface (Noyons 1999; Noyons *et al* 2000) to enable this further exploration of the maps and the data 'behind the maps'. Thus, the bibliometric maps and their internet-based user-facilities will enable users to compare the scientific performance of groups/institutes with other (EU and foreign) 'benchmark' institutes. Likewise, the maps can be used for the selection of benchmark institutes, for instance institutes chosen by the experts.

1.4 Measurement of Scientific Performance

Why bibliometric analysis of research performance? Peer review undoubtedly is and has to remain the principal procedure of quality judgment. But peer review and related expert-based judgments have serious shortcomings and disadvantages (Moxham and Anderson 1992). Subjectivity, i.e., dependence of the outcomes on the choice of individual committee members, is one of the major problems. This dependence may result in conflicts of interests, unawareness of quality, or a negative bias against younger people or newcomers to the field.

We absolutely do not plead for a replacement of peer review by bibliometric analysis. Subjective aspects are not merely negative. In any judgment there must be room for the intuitive insights of experts. We claim however that for a substantial improvement of decision-making our bibliometric method has to be used in parallel to a peer-based evaluation procedure.

The most crucial parameter in the assessment of research performance is *international scientific influence*. We consider international influence as an important, measurable aspect of scientific quality and therefore we developed standardized, bibliometric procedures to assess research performance within the framework of international influence or impact.

Undoubtedly, the bibliometric approach is not an ideal instrument, working perfectly in all fields under all circumstances. But our approach works very well in the large majority of the natural, the medical, the applied and the behavioral sciences. These fields of science are the most cost-intensive and the ones with the strongest socio-economic impact. The most central question we want to answer, is whether the performance is *high* or *low*, and speaking about scientific excellence, *very high*.

The rationale of our bibliometric approach is as follows. Scientific progress can be defined as the substantial increase of our knowledge about 'everything'. In main lines we discern basic knowledge ('understanding') and applicable knowledge ('use'). This knowledge can be tacit ('craftsmanship') or codified ('archived & publicly accessible'). Scientists communicate (and codify) their findings in a relatively orderly, well-defined way since the 17th C. Particularly the phenomenon of serial literature is crucial: publications in international journals. Thus, communication, i.e., exchange of research results, is a crucial aspect of the scientific endeavor. Publications are not the only, but certainly very important elements in this knowledge exchange process. Although not perfect, we adopt a publication as a 'building block' of science as a structure -which is the basis of our mapping methodology- and as a source of data.

Thus, bibliometric assessment of research performance is based on one central assumption: scientists who have to say something important do publish their findings vigorously in the open, international journal ('serial') literature. This choice introduces unavoidably a 'bibliometrically limited view on a complex reality'.

Journal articles are not in all fields the main carrier of scientific knowledge; journal articles are not 'equivalent' elements in the scientific process, they differ widely in importance. Even in the fields where this is the case, journal articles are challenged as 'gold standard' by new types of publication behavior (electronic publishing).

However, the 'daily practice' of scientific research shows that inspired scientists in most cases go for publication in the better and -if possible- the best journals. This observation is confirmed by many years of experience in research evaluation procedures with peer review as core. Each year about 1,000,000 publications are added to the scientific archive of this planet. Certainly this number but also numbers for sub-sets of science (fields, institutes) are in many

cases sufficiently high to allow quantitative analyses yielding statistically significant findings. Publications offer usable elements to 'measure' important aspects of science: author names, institutional addresses, journal (which indicates not only field of research but also: status!), references (citations), concepts (keywords, keyword-combinations or 'noun-phrases').

Work of high quality provokes reactions of colleague-scientists. They are the international forum, the 'invisible college', by which research results are discussed. Often, these colleague-scientists play their role as a member of the invisible college by referring in their own work to earlier work of other scientists.

This process of citation is a complex one, and it certainly not provides an 'ideal' monitor on scientific performance. This is particularly the case on a statistically low aggregation level, e.g., the individual researcher. But the application of citation-analysis to the work, the 'oeuvre' of *a group as a whole over a longer period of time*, does yield in many situations a strong indicator of scientific performance, and in particular of scientific quality. An important, absolutely necessary condition is that the applied citation-analysis is part of an advanced, technically highly developed bibliometric method.

In the Appendix A of this report we discuss in more detail the basic elements of our advanced bibliometric methodology. We here claim that we are able to construct one specific, powerful 'crown' indicator, which normalizes the measured impact of a research group or institute to a worldwide, field- (or sub-field-) specific reference value. It is the *internationally standardized impact indicator* ***CPP/FCSm***. The normalization is based on specifically fitted publication- and citation time-windows in which article type, e.g., review papers, normal papers, or letters, are taken into account. Sub-fields are defined according to the structures within a specific field as found in the mapping structure. This 'crown' indicator enables us to observe immediately whether the performance of a research group or institute is significantly far below (indicator value < 0.5), below (indicator value 0.5 - 0.8), around (0.8 - 1.2), above (1.2 - 2.0), or far above (>2.0) the international (western world dominated) impact standard of the field or sub-field (which by definition equals 1).

In order to take geopolitical aspects in a clear and objective way into account, our approach enables to calculate indicators not only normalized to international reference values, but also to *EU-reference values*, or *EU-region reference values*.

The search for scientific excellence should start systematically at the 'meso'-level of larger institutions, such as universities or major parts of universities, like faculties or large institutes. Bibliometric analyses performed at the macro-level (e.g., a whole country) yield at best general assessments of fields as a whole, for instance, how good a country's performance is in physics, chemistry, psychology or immunology, *without* a reliable breakdown to the individual research groups or programs.

The reason for the choice of the meso-level, i.e. the institution, to start the assessment procedure, is that only at the *input*-side all necessary information, particularly data on personnel and on the composition of groups and programs, is available to a sufficiently accurate extent. Such institutional infrastructure data are not available in general publication databases and must be collected separately in relation the institutions concerned. After an assessment of these larger institutions as a whole, the performance analysis can be narrowed down to research groups and programs within these institutions.

We stress that in the measurement of scientific impact one has to take into account the *aggregation level of the entity* under study. The higher the aggregation level, the larger the volume in publications and the more difficult it is to have an impact significantly above the international level. Based on our long-standing experiences, we can say the following. At the

'meso-level' (e.g., a university, faculty, or large institute, about 500 or more publications per year), a ***CPP/FCSm*** value above 1.2 means that the institute's impact as a whole is significantly above (western-) world average.

Particularly with a ***CPP/FCSm*** value above 1.5, the institution can be considered as a scientifically strong organization, with a high probability to find very good to excellent groups. As discussed above, the next step in our *search for excellence* is the breakdown of the institution into smaller units, i.e., research groups and/or programs. Therefore the bibliometric analysis has to be applied on the basis of institutional input data on personnel and composition of groups.

The bibliometric algorithms can now be repeated efficiently on the lowest but most important aggregation level, that of the research group or research program. In most cases the volume of publications at this level is between 10 and 20 per year. At the group level a ***CPP/FCSm*** value above 2 indicates a very strong group, and above 3 the groups can be, generally, considered as excellent and comparable to top-groups at the best US universities. If the threshold value for the ***CPP/FCSm*** indicator is set at 3.0, we filter out the excellent groups with high probability (van Raan 2000a). It is important to focus on a *recent period of time*, thereby 'allowing' scientists to recognize important work and to let this important work 'take roots'. This means that such a recent period should not be too 'short'. We take 1996-2001 as a suitable period.

In order to carry out the above discussed sub-field normalization procedure of our crown indicator, we also performed an impact-analysis of all sub-fields and themes (as a whole) found in the bibliometric mapping procedure of each field. Such a sub-field impact analysis also has an important value of its own, as it indicates the 'status' of specific sub-fields in terms of international topicality, relevance, influence ('hot' or 'cold' sub-fields). This approach enables users to compare international developments *within* the four life science fields in terms of national (EU member states) strengths and weaknesses, as well as to observe whether specific groups or institutes are located in 'hot' or 'cold' sub-fields.

We developed a second indicator of scientific excellence in the following way. For each of the selected fields we calculated for the impact distribution function of all publications and determine the field-specific top-10% of this distribution (Van Raan and Van Leeuwen 2001). Next we identified those R&D entities with a number of publications above a specific threshold in the top-10% of the impact-distribution function. As we deal with specific fields - and not with a very broad scientific domain such as 'the life sciences'- we expect that most of the identified R&D entities will be at the aggregation level of an institute or department within -mostly- a university or a large public/private institution or company, possibly a group within an institute or department.

Finally, we position these centers of research excellence on the 'landscape of the field' as provided by the bibliometric mapping procedure described in Section 2.

The results of the bibliometric analyses need to be verified and validated against the results of *evaluation schemes and prior analysis at the national level* and against the knowledge of experts. *Verification* primarily means an investigation of evidently 'missing' centers of excellence. This is important, as it is always possible that scientific excellence is not directly measured by bibliometric analyses. Moreover, verification is important to remove errors and incompleteness of addresses of research organizations, departments, groups. In addition, we can identify networks of collaborating groups. *Validation* is primarily a careful comparison of the expert's opinion about scientific quality and the assessment results by bibliometric performance analysis as discussed in the next section.

1.5 Discussion of major recurring issues concerning bibliometric analysis

In this section we focus on five 'frequently asked questions' concerning bibliometric analysis, in particular issues related to (1) the reliability of the initial data, (2) search for quality, (3) the timeliness of the analysis, (4) the comparability of the different research systems, and (5) the (supposed) US bias of the citation index.

1.5.1 Reliability of the initial data

A crucial element of any quantitative study is the reliability of the initial data. This problem is often expressed in two questions: *'Do we have the 'right' field?'* and if so, *'Do we have enough documents (publications and patents) to represent the field in a statistically significant way?'* The answer to the first question depends on the quality of the field delineation. If there is a high degree of consensus within the group of experts providing the concepts to delineate the field in term of keywords and keyword-combinations, the first question can be answered positively. We state however, that in specific cases, particularly in relatively new fields -also in this study, for instance bio-informatics- there is not such a high consensus among experts. However, our mapping methodology presented in this study offers the possibility to 'confront' experts with the effect of their choices of concepts on the field-delineation. Thus, maps of fields can be improved in an iterative way.

Within an agreed delineation, all relevant publications -covered by the citation indexes and/or other databases if agreed- and patents -covered by EPO- can be identified. This means that as soon as there is an agreement on the 'right' field in terms of concepts used for delineation, it will be no problem to represent the field with a statistically significant number of publications and patents

1.5.2 Search for quality

A core problem of this study is formulated by the question: 'Do we focus on 'quality'? often followed by 'instead of output volume'. Scientific performance relates to achieved quality in the contribution to the increase of our knowledge ('scientific progress') as perceived by others: 'peers'. That is, 'knowledgeable others' have to judge the work. This is daily practice in science: colleague-scientists act as reviewers in accepting publications for a scientific journals; awarding PhD's and other degrees; appointments to scientific positions; granting research proposals; prizes, awards, invitations. But also by the statistically most frequent action: citing the published work of colleague-scientists, which highly correlates with all above aspects.

There are about 15,000,000 references about per year, thus providing a wealth of 'communication-based' linkages between 'science today' and earlier work.

As discussed in the foregoing sections, the above phenomenon of referencing is the basis of *citation analysis*. Undoubtedly, scientific quality is a multi-facet characteristic. Citation-analysis based bibliometric analysis provides indicators of international impact, influence. This can be regarded as -at least- one crucial aspect of scientific quality, and thus a 'proxy' of quality as follows from a longstanding experience in bibliometric analysis.

A first and good indication whether bibliometric analysis is applicable to a specific field is provided by the publication characteristics of the field, in particular the role of *international*, refereed journals. If international journals are a dominating or at least a major means of communication in a field, then in most cases bibliometric analysis is applicable.

1.5.3 Timeliness of the analysis

A frequently posed question concerns the 'delay problem': Does bibliometric analysis suffer from a substantial 'delay' in the measurement of research performance? An answer to this question first needs a further refinement: delay as compared to what? To the average 'processing time' of a publication? To the average 'running time' of a project? Or to peer review 'time cycles'?

The entire process starting with scientific activities and leading to successively 'publishable' results, writing of an article, submission of the article, publication of the article, citations to the article, varies considerably for the different fields of science and often within a field or even research theme. Depending on type of activities and type of results it may take years. But during that time the work is improved, not the whole process time can be regarded is a 'delay' or a 'waste of time'. Furthermore, the average duration of a major research project is about 4 years, and the same is the case for most peer review time cycles. Also, during the publication process the awareness of scientific community (and peers!) evolves (e.g., average time between field-specific conferences etc.).

The above means that 'bibliometric awareness' does not necessarily take more time than 'peer awareness' as illustrated in Appendix A2 with a recent example of a publication in physics.

Moreover, the bibliometric system itself proves empirically the robustness of the method simply by showing that citation-analysis based indicators are remarkably stable, which means that recent past performance is a reliable predictor for near-future performance.

1.5.4 Comparability of the different research systems

It is often quite problematic to understand the structure of a research organization in terms of 'realistic' units such as departments or research groups. There are major differences in research systems between countries. The University of London is not a university anymore in the usual sense. It is an 'umbrella organization' covering several different relatively autonomous universities. In Paris and other French cities not such an umbrella structure exists, there we deal with completely autonomous universities that were part of originally one 'mother-university'. It is often cumbersome to distinguish between departments of these different universities within a city. The two 'Free Universities' of Brussels (Vrije Universiteit Brussel, VUB, and the Université Libre de Bruxelles, ULB) are a notorious example in this sense. Another well-known problem is the 'interwoveness' of the French CNRS and French universities.

This problem is in fact a 'fine structure' problem: matching bibliometric data ('external') with the 'real fine-structure' ('internal') of a main organization (e.g., a university). In order to do this, we need accurate 'fine-structure' data per organization. Moreover, this internal structure is 'dynamic': new departments, schools and certainly new research groups are created all the time.

Bibliometric analysis provides a two-fold first-but-good approximation:

Narrowing down of fields: the smaller the bibliometric 'refining' of fields (e.g., from neuroscience as a whole to brain infarct research as a specific research theme within neuroscience), the more we approach 'real' units such as research groups within the internal structure of a main organization: "convergence principle". In this study the bibliometric mapping methodology (finding the structure of a specific field) is particularly suited for this approach.

Networks of co-operating scientists: the analysis of collaborating researchers provides the *internal structure* of that specific (sub-)field in terms of co-authors. Thus the real, '*working floor' groups* are identified. This identification is completely *independent* of the quality of information on main organization addresses. It is, as it were, based on a 'bibliometrically-driven' *self-organization* of science.

1.5.5 Supposed US bias of citation index

If a bibliometric analysis would be demonstrably or at least plausibly affected by a US bias we can apply in the calculation of our indicators EU averages, EU distribution functions and/or benchmarking with other EU-institutes.

2 Patent analyses

2.1 Preliminary remarks

The aim of the present report is to identify centers of excellence in specific fields of science and technology by publication and patent indicators. Centers of excellence comprise universities and public research establishments as well as industrial companies. As companies primarily strive for the private appropriation of knowledge, their number of publications is modest, so that publications are not appropriate indicators for assessing their research output. Against this background, the patent analyses complement the publication analyses with regard to the structures of research in industry. In addition, patents are increasingly used in public (non-profit) organizations, i.e., in universities and in public non-university research organizations, so that the public entities can be analyzed by the two indicator sets of publications and patents. In a simplified way, publications represent the orientation of public research organizations towards basic research and the expansion of knowledge, patents their orientation towards application and technological realization. Therefore, patents bring in an additional dimension for the assessment of excellence also for public research organizations.

Patents are generally seen as output indicators of applied research and technological development (Schmoch 1999a). They represent intellectual property rights and are thus legal documents. The owner of a patent is given – for a limited period – the exclusive right to exploit a technical invention commercially. In this perspective, competitors are excluded from producing and selling on the basis of the invention for which a patent is granted. However, through publication of the patent document, the competitors get access to the knowledge linked to the invention and are stimulated to create own solutions for the problem in question. Patent documents are published in paper form and are registered in public databases that can be used for statistical analysis.

A patent application has to fulfill various criteria for being granted. First, the described invention has to be new on a world-wide level. It is not sufficient that an invention is new for the company, or new in a specific country. Secondly, the new product or process must be distinctly different compared to the state of the art, i.e., it must clearly imply an inventive step. So for someone experienced in the state of the art, the solution suggested by the invention must not be obvious. Thirdly, the invention has to be exploitable in commercial terms. Scientific discoveries without a practical purpose are not patentable.

The third criterion implies that most patent applicants are industrial companies. This is reinforced by the fact that patent applications are costly so that their issuance is only reasonable, if a commercial exploitation is aimed at. Against this background, in most areas of technology the relative number of patent applications individual inventors and research institutes hold are quite small. However, in knowledge-based, research-intensive areas, the participation of public research institutes and universities is often significant.

Patents are normally registered immediately after the invention is made. Clearly, if an invention becomes public without a registered patent application, it is no longer possible to achieve patent protection, as the invention is not considered to be new any more. Further, the date of first registration, the *date of priority*, is important, in case a similar invention is

registered by a competitor.[3] In any case, the registration of a patent is closely linked to the results of research and development having good prospects for commercial exploitation.

The priority or first application of a patent is generally made at the domestic patent office, as this is the cheapest way of registration. Within a period of one year after submission the applicant can decide to go abroad. This is necessary, if he is interested in patent protection in foreign markets, because so far his patent has a national coverage only. The first choice for European applicants is an application at the European Patent Office (EPO). At the EPO, a central examination procedure is carried out that is valid for all member countries of the European Patent Convention (EPC). If the patent is granted, the applicant has to transfer it to all destination countries for achieving national protection.

The analysis proposed for this study is based on applications at the EPO. As the expenses for an EPO application and examination are high, EPO applications represent inventions of high technological and commercial value. Furthermore, the same legal rules apply for all applicants so that a statistical distortion by specific national rules is avoided. The major bias at the EPO can be seen in its different strategic relevance of the European market for European countries and for overseas countries (USA, Japan, etc.). However, in this study, only member countries of the European Union and associated states/candidate member countries are examined in more detail.

In recent years, foreign applications are increasingly filed as an international application at the World Intellectual Property Organization (WIPO) or as a PCT application (PCT = Patent Co-operation Treaty) (Schmoch 1999b). This is a central application procedure, often including a preliminary examination. However, the relevant examination with respect to legal status is made by the 'destination offices'. The EPO can be a destination of a PCT application; the respective applications are then called Euro-PCT applications. The proposed study will cover direct applications at the EPO as well as Euro-PCT applications. In both cases, the applications are published strictly 18 months after the priority date. This is a decisive advantage compared to an analysis of patents at the United States Patent and Trademark Office (USPTO), because only granted patents are published there, and the delay between application and grant is uncertain and can be several years.[4]

The above implies that in June 2002 a relevant number of EPO and PCT applications of the invention/priority year 2000 is already available, whereas at the USPTO only the priority year 1996 is more or less complete (Schmoch 1999b, Barré et al. 2001). As the proposed project aims at the assessment of the *present performance* of institutions, the USPTO data have to be considered in fact too old. The consequence of this difference is illustrated in Figure 2.1-1 by the example of a search for the term "nano" in early March 2002. Whereas the EPO and PCT applications show a considerable growth of patent applications in Nanotechnology since 1996, this decisive increase does not shine up in the US data. This means for analytical purposes that for recent years, the samples which can be identified are not representative at all.

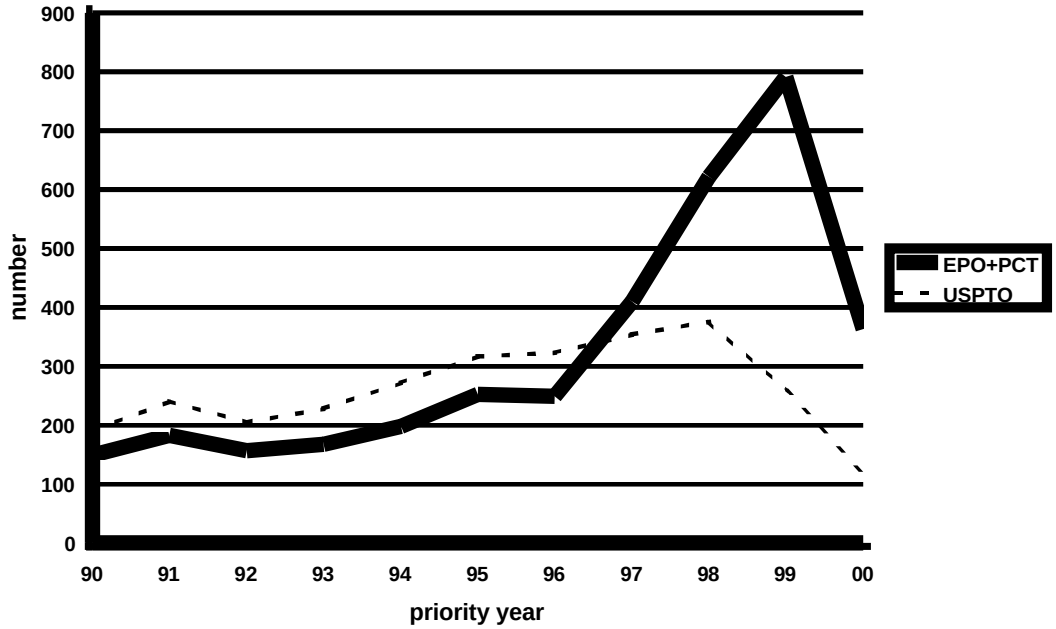
Since 2002, USPTO applications can also be published 18 months after the first application, equivalent to the European ruling, but they must not be published at this time. This leads to a mixture of early and late publications, and according to first statistical analyses, US inventors still prefer late publications. Therefore, this study is exclusively based on EPO and PCT applications. It can be assumed that important patent applications of US origin are also registered at the EPO or the WIPO, so that they are included in the analysis.

---

[3] These regulations apply to the European context; in the United States, the legislation is slightly different.

[4] The publication rules at the USPTO have changed recently, but this has not had a visible effect on the statistical data yet.

Figure 2.1-1: Patent applications identified by the search term “nano” with open right-hand truncation in the database World Patent Index, Search on March 10th, 2002
Source: WPINDEX (STN), computation by Fraunhofer ISI



2.2 Definition of data sets

2.2.1 Basic rules of the sample definition for statistical analyses

The analyses of the present study refer to five areas of science and technology. In order to achieve appropriate results for the mapping of excellence, an adequate definition of samples referring to these areas is crucial. For this purpose, some basic rules of sample definition need to be followed.

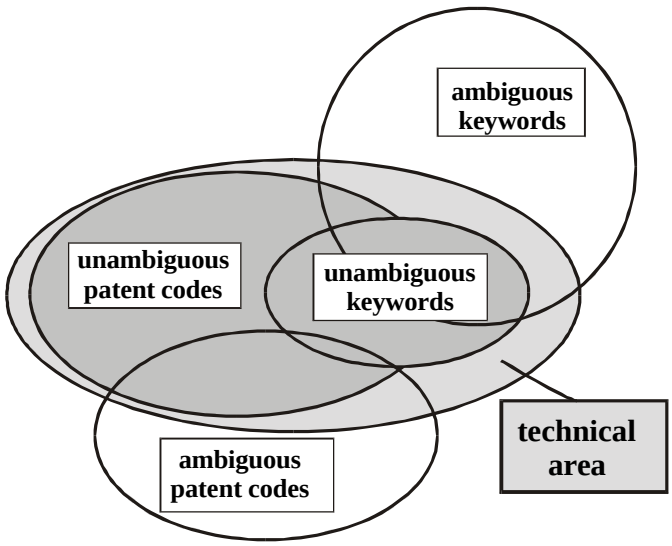
In database searches, it is nearly impossible to cover all patents[5], as patents can appear in unforeseen contexts. Therefore, the aim of a patent search is the definition of samples which are statistically representative for the area considered. Therefore, the samples should be as large as possible and include as few unsuitable documents as possible. This orientation implies a major difference between statistical searches and novelty searches for legal purposes, for instance, in the examination process of a patent application. The aim of legal searches is to identify all relevant documents, even if they appear in unforeseen, marginal contexts. Against this background, legal searches often include unsuitable documents, which are excluded by a process of intellectual examination of each document covered by the original sample. This can be realized in legal searches, as the search areas are more precisely defined and much smaller than in statistical searches. In statistical searches, the number of documents is much larger, so that they have to rely primarily on automatic searches without detailed intellectual checks.

The principles of statistical searches are illustrated in Figure 2.2-1 wherein the light grey area represents the complete document set of the technical area considered. The standard approach

[5] The term "patent" includes "patent applications" and is used for simplifying the wording.

in patent searches is to look at appropriate codes of the International Patent Classification (IPC), by which the examiners of the patent office classify each patent application, so that a high quality of classification can be assumed. Each patent document has at least one IPC code, the primary code and often also a secondary code. In the case of the existence of suitable IPC codes, it is strongly recommended to use primary and secondary codes in order to achieve an optimal yield. The IPC is a hierarchical classification with 8 sections on the highest level of aggregation covering all areas of technology (WIPO 2000). The next levels of disaggregation are 22 sub-sections, 120 classes and 630 sub-classes. The finest classification is made by groups and sub-groups with about 67,000 different items. All in all, patent documents are generally classified in a very fine and accurate way. However, with regard to new areas, the classification is less useful, because the IPC is revised, after lengthy processes of international agreement, every five years only. For instance, the area of Nanotechnology was introduced in the year 2000 with the 7th edition of the IPC, so that only documents published after 1999 are included.

Figure 2.2-1: Illustration of data sets identified in statistical patent searches



In general, the share of the complete technical area covered by IPC codes is very high, though there are still relevant documents not encompassed. For finding additional documents, keyword searches can be performed. It is important that keywords used in such additional searches are unambiguous with reference to the area considered, so that exclusively relevant documents are selected. In many cases experts in a specific field cite keywords which are often used in their field, but also in other contexts. These keywords would lead to a too high number of unsuitable documents, so that the sample would not be representative for the field considered. This situation is illustrated in Figure 2.2-2 by the area of ambiguous keywords. The same applies to ambiguous classes covering to a certain extent relevant documents, but also documents of other areas.

In some cases, however, it can be useful to include ambiguous keywords or IPC codes, if the share of unsuitable documents in limited and does not call the representativity of the total sample into question. Furthermore, the precision of ambiguous keywords and IPC codes can be improved by combining them with other keywords or IPC codes. This can be achieved by the inclusion of positive associations (by intersection) or the exclusion of negative

associations. For instance, documents identified by a certain keyword are included only, if they are classified in specific technical areas defined by IPC classes, or excluded if they appear in other unsuitable classes. In the illustration of Figure 2.2-1, the final patent sample is represented by documents with unambiguous patent codes and keywords[6] highlighted in dark grey color. This area covers the largest part of the total set of relevant documents, shown in light grey.

Keyword searches in official patent documents published by national or regional patent offices are generally less productive, as the legal requirements of disclosure with regard to titles and abstracts are not very strict. If a search has to be based on keywords to a considerable extent, it should be executed by the so-called World Patents Index (WPI) which is produced by the company Derwent. The staff of Derwent prepares improved titles and abstracts describing the technological content of each patent application. The effect of the enhanced quality of the Derwent titles and abstracts may be illustrated by a search for the term "nano" with right-hand truncation within EPO and PCT patent applications in the priority year 1999. By a search in the official databases, about 200 documents are identified, by a search in WPI about 800.

To sum up, patent searches are generally based on IPC codes complemented by keywords. Statistical searches aim at the definition of representative samples and cannot cover all relevant documents in most cases. It is decisive to avoid ambiguous IPC codes and keywords, if the share of unsuitable, misleading documents is too high. Finally, it is important to look at the use of appropriate databases in the case of keyword searches.

2.2.2 Genetics / heredity

The definitions of the samples of patents and publications cannot not based on identical search strategies due to substantial differences between both approaches. First, the abstracts of patents and publications are formulated with different orientations implying a different use of keywords. Second, the patent classification is an important search instrument in the case of patents, whereas the classification of scientific areas in the Science Citation Index is less elaborated and can be only used as supporting tool. Third, the areas of the present exercise are defined in a scientific, not a technological perspective. For instance, Genetics/heredity is a sub-discipline of the scientific discipline biology and not a technology. With regards to the patent search, it has to be asked what the potential output of the scientific research in Genetics/heredity may be in terms of technology. For genetics, it is obvious that the major technological output is genetic engineering, classified by the IPC group C12N015. In addition, other IPC codes may be included such as A61K048 referring to medical preparations containing genetic material. The search terms "gene", "genetic" etc. generally appear in documents which are already classified in C12N015. However, in some cases, they are registered outside this group, so that a keyword search can identify a few additional documents. Furthermore, other IPC codes are used to complement the basic strategy as documented in Table A-1. It would be too complex to explain all steps of the search strategy in detail, as it would be necessary to clarify the meaning of all IPC codes used. This would require the consultation of the complete manual of IPC codes with nine volumes (WIPO 2000). As substitute, the most relevant classes and sub-classes which are used in the search strategies of the five areas are listed in Table A-2.

By the complete strategy, a total number of about 21.500 documents was identified for the priority period 1996 to 2000 with reference to patent applications at the EPO or the WIPO (direct EPO applications plus Euro-PCT applications). Thereof, 90 percent were already

[6] The case of keyword and patent code comibnations is not shown for reducing the complexity of the pictures.

found by the IPC group C12N015, so that the various refinements of the search strategy had a visible, but minor effect.

For the preparation of the present study, the European Commission issued a short study in order to identify potential methodological problems (Dybkaer/Bauin 2002). As to patents in Genetics/heredity, it solely describes possible areas of technological application (ibid.: 33) and mentions that the searches were performed on the basis of IPC codes in the databases EPODOC and EUREG of the European Patent Office. The authors analyzed application years, not priority years and found for the period 1996 to 2000 only 7,560 patents, thus much less than the 21,500 documents of the present study. It can be assumed that the preparatory study also used the IPC code C12N015, so that the reasons for this enormous difference are unclear. Perhaps, the authors only looked for grated patents instead of applications, but this is less probable. Rather, they may have exclusively considered primary IPC codes, instead of including secondary codes. However, for achieving high yields, all IPC codes should be taken into account.

2.2.3 Neurosciences

The patent search strategy for Neurosciences is primarily based on keywords as documented in Table A-3. The most important yield is achieved by the first search command with terms such as "nerve", "brain" or "meningitis" and the terms "alzheimer" and "multiple sclerosis". These three search commands cover about 90 percent of the total data set. In addition, typical neurological diseases were identified by manuals (Hoffmann-La Roche 1998, WHO 1993) and included in the search strategy by keywords. In addition the IPC code A61P025 refers to the therapeutic activity of chemical compounds with regard to neurological diseases. However, the sub-class A61P was introduced but in 2000 with the 7th edition of the IPC classification and does not cover patent applications of former years. The results of all search commands are combined in command 20 by the search string "L1-L19". In the last search step, this set is limited to specific technical areas such as A61B (medical diagnosis) or A61N (electrotherapy). In the case of A61K (preparations for medical purpose), the search was limited on IPC groups with direct reference to biological methods and processes such as A61K039 (preparations containing antigens or antibodies). In the first phase of the project, the stakeholder panel decided to introduce this limitation as the analysis aims at life sciences and not chemical sciences. This decision proved to be appropriate with regard to the patent activities of companies in the perspective to identify of relevant research centers in life sciences. In the case of public research entities, however, the inclusion of chemical drugs may have led to useful results. For the scientific research in Neurosciences can lead to the definition of targets referring to neurological diseases and support the development of chemical drugs. By the limitation of the data set of command 20 to specific technical fields, the number of included documents is reduced by about 50 percent.

The quality of the search results primarily depends on the precision of search terms such as "neural". As the string "neural network" is used for specific types of computers, the classification G06 (computing) was not included in the last search step defining technical areas of validity. A definite exclusion would have been misleading, because some methods of diagnosis of neurological diseases work with computers. However, some cases were found in G01N (analyzing materials) were neural networks were used, leading to an inclusion of unsuitable documents. So it would have been more precise to definitely exclude the term "neural network", but the quantitative effect would be marginal.

All in all, about 10.500 patent applications were identified for Neurosciences (period 1996 to 2000). In the preparatory study, about 4.200 documents were found, a number which is again much lower than in the present study. The main reason is the decision of the authors to limit the search on IPC codes[7] and not to include keywords, as the risk to include too many unsuitable documents was assessed high. In this study, we coped with this problem by the described limitation of keywords on specific technical areas. Although the authors of the preparatory study exclusively used IPC codes, they found a high number of "irrelevant patents" (Dybkaer/Bauin 2002: 70). In many cases, they could not appropriately assess the content of a document due to the poor description in the "official" titles and abstracts. In other cases, the used IPC codes seem to be have been too general. In the present search strategy, the IPC codes are explicitly linked to neurological purposes; multiple-use codes were not included.

2.2.4 Immunology

Comparable to Neurosciences, the patent search strategy for Immunology is primarily based on keywords as documented in Table A-4. The term with the largest number of hits is "immun" (with open right-hand truncation). Together with "autoimmun" and "HIV", about 75 percent of all documents are determined. The additional effect of other search terms related to specific immune diseases is higher than in the case of Neurosciences. Again, the documents identified by keywords are limited to specific technical areas. In Immunology, the number of applications found by IPC symbols is relevant, for instance, by A61K038-19 (cytokines, lymphokines, interferones) or C12N015-24 (interleukins produced by biotechnical processes). However, most of these documents were already identified by keywords, so that the additional effect is limited.

With 22.700 applications in the period of 1996 to 2000, Immunology is the largest area of this study. In the preparatory study, the authors decided to adopt a pure IPC-based strategy and identified about 4.350 patents. In the present strategy, already the IPC-based parts were more effective than in the preparatory study, but the relevant yield was achieved by keyword searches in the WPI database.

2.2.5 Bioinformatics

In the first stage of the project, the definition of Bioinformatics could not be easily delineated. Three versions were suggested:

- hardware systems with biological elements
- informatics for biological and medical purposes
- software and hardware based on biological analogies (e.g., neural networks, genetic algorithms etc.)

Finally, an expert of the stakeholders' panel suggested to adopt the definition of the European Science Foundation (ESF).

"The use of computational techniques to handle, analyze, and add value to the flood of data doming out of modern genomics and proteomics research."

[7] Again, the search strategy is not documented.

Based on this definition, Bioinformatics is focused on software development. In recent years, it is possible to patent software as long as it is linked to the improvement of hardware. Furthermore, it is increasingly possible to get patent protection for "pure" software inventions (Blind et al. 2003: 112ff), but the area of patentability is still limited. Against this background, the authors of the preparatory study concluded.

"No patent analysis has been undertaken for the field of Bioinformatics, as it is not considered relevant. Innovations from this field are generally speaking, not patentable " (Dybkaer/Bauin 2002: 39).

Nevertheless, in the present study, the attempt was undertaken to identify patent applications in Bioinformatics. The search terms "Bioinformatics" and "biocomputing" brought some hits, but the majority of documents was identified by search strings on "gene analysis", "gene structures", "gene production", or "metabolic pathway" in combination with the technical area of computing (G06) (see Table A-5). The majority of documents identified primarily refers to software and not to hardware. For the search procedure more sophisticated proximity operators were used such as "W" or "A" in order to achieve a higher yield of appropriate documents. For instance, the simple combination of "gene" and "analysis" by the standard Boolean operator "and" may identify documents with a high distance between the keywords in the abstract, so that the term "analysis" refers to other things, but not genes. By proximity operators, the sequence and distance of keywords can be defined more precisely.

All in all, nearly 500 patent applications were identified for the period of 1996 to 2000. This number is modest compared to the other fields considered, nevertheless, it is relevant. The low number may be due to the fact that Bioinformatics is an emerging field in an early phase of development and that its focus is still primarily on science and less on technological exploitation. In addition, patent applications in the software sector are still not a common standard, and many firms have not realized this new possibility. Against this background, the patents identified by the database search represent a limited sample with regard to the institutions which are active in the technological use of Bioinformatics. The firms and research entities in the sample are probably a selection of progressive actors with high awareness of new instruments of competition. In any case, the results of the patent analysis in Bioinformatics have to be interpreted with caution.

2.2.6 Lessons of the sample definition process

The experiences of this study show that the appropriate definition of samples is crucial for all other downstream work steps. In new dynamic areas, it is not sufficient to exclusively build the patent search strategies on IPC codes; rather keyword searches play a considerable role. Therefore, it is important to use databases with good facilities for keyword searches, as official database are less effective in this regard.

The remarks on the outcome of the preparatory studies should not be seen as criticism to their authors, because they achieved substantial results within a short time. Rather, the less encouraging outcome of their efforts shows that patent search strategies cannot be built as side products of search strategies for publications, but require substantial knowledge of the patent system, for instance, as to legal rules and classifications, and additional efforts to achieve appropriate specific definitions.

The strategies must be developed in close interaction of experts in the fields and experts for patent searches. The patent experts cannot be neutral with regard to the field content, but must acquire sufficient knowledge about the field so that they are able to assess the relevance of suggestions of the field experts and to identify appropriate IPC symbols. The setting of the present project with short stakeholders' panel meetings and distributed field experts proved to be less effective. The suggestion for keywords often came in too late or were contradictory. In the project, we got substantial input from the external experts for Bioinformatics, but only for some aspects.

For the 4 Life Sciences (LS) fields, experts of the biotechnology department of Fraunhofer ISI were consulted in order to achieve suitable decisions within at short notice. This approach was necessary to cope with the limitations of time and funds in this project. An optimal approach would be the nomination of small field expert groups of about 3-5 members each. They should meet at the institute of the patent experts for discussing the field definitions. This first agreement is necessary, as in many cases the definition of the field is fuzzy. In a second step, the experts can suggest keywords. The patent experts can explain the techniques of patent searches and comment on the precision of the keywords suggested. As next step, the first suggestions for a search strategy can be transferred into search strings and directly executed in databases, so that the output is immediately shown to the patent and field experts. The accuracy of search terms can be checked by titles and abstracts of sub-samples of documents. Then, by an iterative process, the search strategy can be gradually improved. It is realistic to come to a final definition within an expert meeting of about two days. By such an approach, it is possible to define search strategies with high yield and accuracy with a limited amount of time and costs.

The in-house database at Fraunhofer ISI was constructed on a Microsoft ACCESS platform. The database consists of 57.835 records whereof the distribution on the five areas is shown in Table 2.3-1. There is a certain overlap between the fields, as illustrated by the example of Figures 2.3-2 and 2.3-3. These records with multiple field references are included several times in the database in order to facilitate the analysis process. All in all, the cooperation with the EPO proved to be very helpful in solving the specific problems of the present project.

Table 2.3-1: Total number of patent documents in the in-house database

Field	***Number of patent documents***
Genetics	21507
Neurosciences	10488
Immunology	22724
Bioinformatics	493

2.3 Data matching and data cleaning

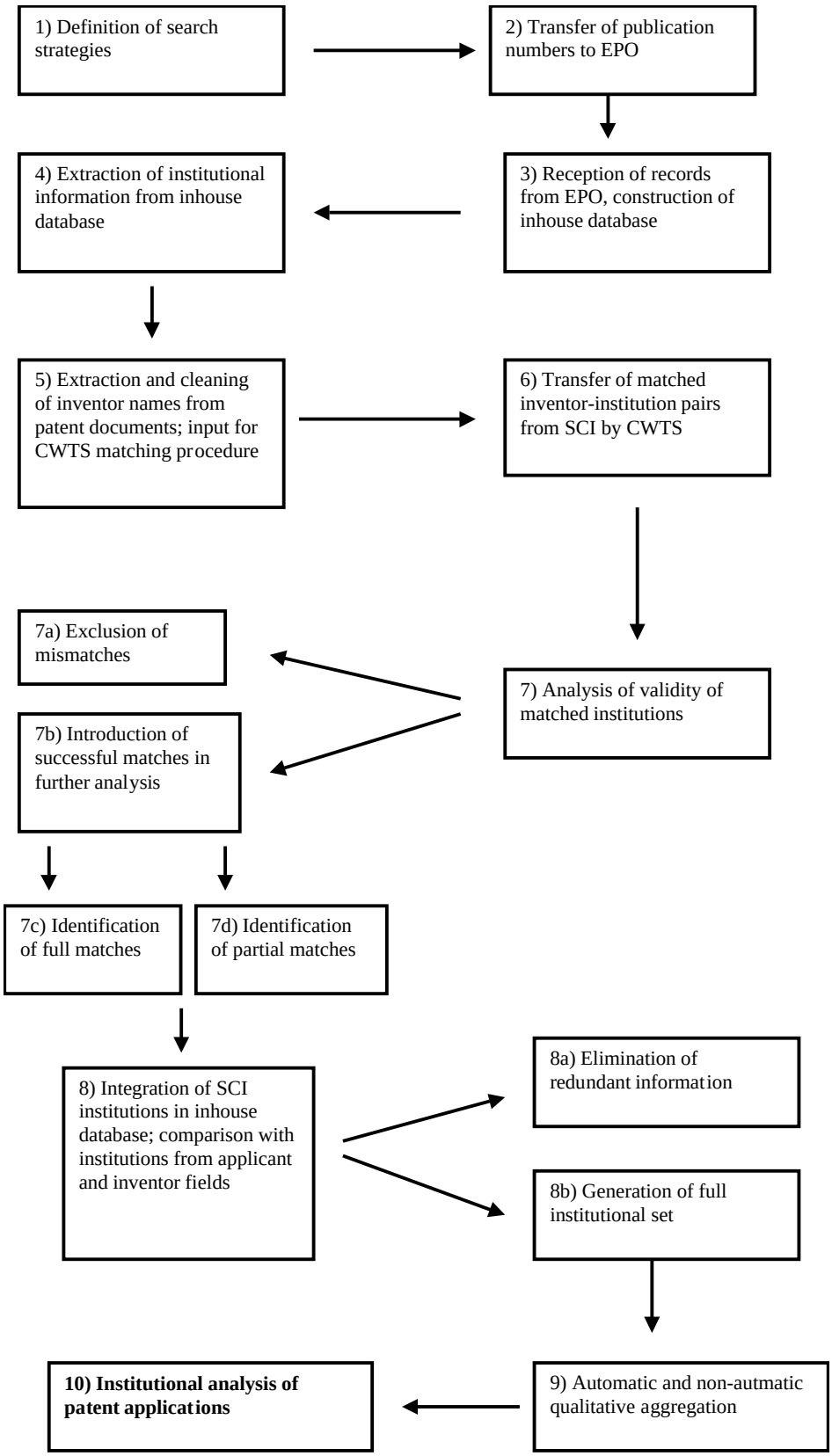
2.3.1 Flowchart data generation/matching/cleaning

The major aim of the patent analysis is the identification of excellent institutions. Therefore it is necessary to link the patent applications to specific laboratories/institutes. First, the analysis serves for identifying industrial enterprises, as patents are more appropriate than publications for this type of institution. However, in research-intensive fields, patents are also taken out by public research establishments, so that specific efforts have been undertaken to look at them in more detail. In the standard approaches of institutional analysis, an analysis of the patent applicants is performed, i.e., of the applying firms. But scientific institutions will be involved in a more complex way. As to non-university research institutions, a patent is generally applied by the mother organization, not by the institute or laboratory. In the case of universities,

- sometimes the university – not the institute or department – appears as applicant,
- sometimes the inventor himself, mostly the professor, is also the applicant,
- sometimes the property rights are transmitted to a firm which appears as applicant.

In the latter, very frequent case, the linkage to the university is only documented by the name of the inventor. This means that applications of a company may originate in scientific institutions, a fact which is not visible in the applicant field. Consequently, it was necessary to identify the specific institutional affiliation of the inventors by additional methods. Against this background, the patent analysis consisted of following working steps illustrated in Figure 2.3-1.

Figure 2.3-1: Flowchart of working steps of the patent analysis



The flowchart presents ten major working steps:

1) The definition of search strategies has already been discussed in section 2.2.
2) The publication numbers identified by the search strategies were transferred to the EPO which identified the referring patent documents in its database.
3) The EPO transferred files with the records of the patent documents to Fraunhofer ISI. The record information was used to construct an in-house database with regard to the five areas considered in this project.
4) The institutional information available in the records were extracted, in particular from the applicant and also from the inventor field.
5) The inventor names were extracted and cleaned in order to transfer them to CWTS for matching them with authors in the Science Citation Index (SCI).
6) CWTS linked the identified authors to their institutions and send this information back to Fraunhofer ISI.
7) The validity of the matched institutions was assessed, mismatched institutions were excluded, the successfully matched institutions were separated into full and partial matches.
8) The institutions identified by the SCI matches were integrated into the in-house database. The new institutional information was compared with the already existing information from the inventor and applicant fields. Redundant information was eliminated, the final institutional set was generated.
9) The cleaned list of institutional entries had to be aggregated by an automatic and non-automatic qualitative process, so that the number of patents per institution/ organization could be generated.
10) On this basis, the complete institutional analysis of the patent applications could be performed as final step.

The different steps will be explained in more detail in the following sections. If the analysis of institutions by patents had been limited to enterprises, it would have been finished after the first step (definition of search strategies) and complemented by a statistical online analysis of the applicants. The additional steps were necessary to cope with the problem of identifying scientific institutions in an appropriate way.

2.3.2 Interaction with European Patent Office

All search strategies described in section 2.2 were developed on the basis of the WPI database, as in all of them, keyword searches are relevant. Furthermore, it is easier to check the validity of search strings by the help of sub-samples, as the WPI records provide an improved text disclosure compared to official documents. Figure 2.3-2 illustrates the content of a WPI record whereof the title is already much more informative than the official title (cf. Figure 2.3-3). The document is part of the samples for Genetics/heredity as well as Bioinformatics. The reference to Genetics/heredity is based on the IPC code C12N015 (IC field), the reference to Bioinformatics on the additional classification in the IPC sub-class G06F. The WPI abstract shows that the inclusion of the document in the Bioinformatics sample is justified. The WPI record additionally comprises legal information, e.g. the first priority application at the USPTO (PRAI field) or the parallel applications at the WIPO, the EPO, and the Japanese and Australian offices (PI field). The USPTO patent is not published yet, due to the delay of the grant process (cf. section 2.1). In addition, the record contains data on IPC codes (field IC), and inventors (IN) and the patent applicants (PA), but not the nationality of the inventors or patent applicants/assignees. However, the information on the national origin is important for linking the records to specific countries. Some researchers use

the priority country for this purpose, but the example illustrates that this approach can be misleading. The priority country of the example in Figure 2.3-2 are the United States, but the inventors and the applicant are from France, as shown in Figure 2.3-3.

Figure 2.3-2: Example of a record in the WPI database
Source: WPINDEX (STN)

```
L1   WPINDEX  (C) 2003 THOMSON DERWENT
AN   2000-038446 [03]   WPINDEX
TI   Novel secreted protein 5' expressed sequence tag sequences used in
     diagnostic, forensic, gene therapy, and chromosome mapping procedures.
DC   B04 D16 T01
IN   DUCLERT, A; DUMAS MILNE EDWARDS, J; GIORDANO, J
PA   (GEST) GENSET
PI   WO 9953051    A2 19991021 (200003)* EN          C12N015-11
       RW: AT BE CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE
        W: AU CA JP US
     AU 9930501    A  19991101 (200013)
     EP 1068312    A2 20010117 (200105)  EN          C12N015-11
        R: AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
     JP 2002511259 W  20020416 (200242)      924p    C12N015-09
ADT  WO 9953051 A2 WO 1999-IB712 19990409; AU 9930501 A AU 1999-30501 19990409;
     EP 1068312 A2 EP 1999-912007 19990409, WO 1999-IB712 19990409; JP
     2002511259 W WO 1999-IB712 19990409, JP 2000-543599 19990409
FDT  AU 9930501 A Based on WO 9953051; EP 1068312 A2 Based on WO 9953051; JP
     2002511259 W Based on WO 9953051
PRAI US 1998-69047    19980428; US 1998-57719    19980409
IC   ICM  C12N015-09; C12N015-11
     ICS  C07K014-47; C07K016-18; C12M001-00; C12N001-15; C12N001-19;
          C12N001-21; C12N005-10; C12N015-10; C12P021-00; C12P021-02;
          C12Q001-68; G01N033-53; G01N033-566; G06F017-50
ICA  G06F017-30
AB   WO   9953051 A UPAB: 20000118
     NOVELTY - Novel 5' expressed sequence tag (EST) sequences, corresponding
     to human secreted proteins, are disclosed.
          DETAILED DESCRIPTION - A purified nucleic acid (I), comprising one of
     the 811 polynucleotide sequences given in the specification (and sequences
     complementary to these sequences), is new. INDEPENDENT CLAIMS are also
     included for the following: (1) A purified nucleic acid comprising at
     least 15 consecutive nucleotides of (I). (2) A purified or isolated
```

polypeptide (II) comprising one of the 788 polypeptide sequences given in the specification. (3) A method of making a cDNA: (a) contacting a collection of mRNA molecules from human cells with a primer comprising at least 15 consecutive nucleotides of (I); (b) hybridizing the primer to a mRNA in the collection that encodes the protein; (c) reverse transcribing the hybridized primer to make a first cDNA strand from the mRNA; (d) making a second cDNA strand complementary to the first cDNA strand; and (e) isolating the resulting double-stranded. (4) A method of making a cDNA: (a) obtaining a cDNA comprising (I); (b) contacting the cDNA with a detectable probe comprising at least 15 consecutive nucleotides of (I) under hybridizing conditions; (c) identifying a cDNA which hybridizes to the detectable probe; and (d) isolating the cDNA which hybridizes to the probe. (5) A method of making a cDNA comprising: (a) contacting a collection of mRNA molecules from human cells with a first primer capable of hybridizing to the polyA tail of the mRNA; (b) hybridizing the first primer to the polyA tail; (c) reverse transcribing the mRNA to make a first cDNA strand; (d) making a second cDNA strand complementary to the first cDNA strand using at least one primer comprising at least 15 consecutive nucleotides (I): And (e) isolating the resulting double-stranded cDNA. (6) A method of making a polypeptide comprising: (a) obtaining a cDNA which encodes a polypeptide encoded by (I) or a cDNA which encodes a polypeptide comprising at least 10 consecutive amino acids of a polypeptide encoded by (I); (b) inserting the cDNA in an expression vector such that the cDNA is operably linked to a promoter; (c) introducing the expression vector into a host cell whereby the host cell produces the protein encoded by the cDNA; and (d) isolating the protein. (7) In an array of discrete ESTs or fragments thereof of at least 15 nucleotides in length, the improvement comprising inclusion in the array of at least one sequence (preferably 5) selected (I), its complement or fragments thereof comprising at least 15 consecutive nucleotides. (8) An enriched population of recombinant nucleic acids, the recombinant nucleic acids comprising an insert and a backbone, wherein at least 5% of the insert nucleic acids in the population comprise a sequence selected (I), its complement or a fragment thereof. (9) An antibody composition capable of selectively binding to an epitope-containing fragment of a polypeptide comprising a contiguous span of at least 8 amino acids of any (II). (10) A computer readable medium having stored thereon (I) and (II). (11) A computer system comprising a processor and a data storage device wherein the data storage device has stored thereon a sequence selected from (I) and (II). (12) A method for comparing a first sequence to a reference sequence wherein the first sequence is selected from (I) or (II), comprising: (a) reading the first sequence and the reference sequence through use of a computer program which compares sequences; and (b) determining differences between the first sequence and the reference sequence with the computer program. (13) A method for identifying a feature in (I) and (II), comprising: (a) reading the sequence through the use of a computer program which identifies features in sequences; and (b) identifying features in the sequence with the computer program. (14) A vector comprising (I). (15) A host cell containing the vector of (14).

USE - The 5' expressed sequence tags (ESTs) and can be used for producing secreted human gene products. They can be used to identify and isolate 5' untranslated regions (UTRs) and upstream regulatory regions which control the location, development stage, rat, and quantity of protein synthesis, as well as stability of mRNA. The ESTs are also useful as probes for chromosome mapping, and to obtain full length cDNA clones. The ESTs can also be used in forensic procedures to identify individuals,

or in diagnostic procedures to identify individuals having genetic diseases resulting from abnormal gene expression. The products may also be used in gene therapy protocols. The nucleic acids encoding the signal peptide can be used for directing extracellular secretion of a polypeptide or the insertion of a polypeptide into a membrane, or importing a polypeptide into a cell. The proteins encoded by the EST sequences may be useful in treating a variety of human conditions.

ADVANTAGE - Secreted proteins have therapeutic value, and the identification of new secreted proteins is valuable. Prior art methods to identify and characterise the 5' portions of such genes is limited, due to lack of specificity and comprehensiveness. The present invention provides 5' ESTs which meet this nee, and can be used to easily identify and isolate 5' potions of genes.

Dwg.0/10

TECH WO 9953051 A2 UPTX: 20000118

TECHNOLOGY FOCUS - BIOTECHNOLOGY - Preferred computer medium: The computer system of (11) further comprises a sequence comparer and a data storage device having reference sequences stored thereon. The sequence comparer especially comprises a computer program which indicates polymorphisms. The computer system of (11) even further comprises an identifier which identifies features in the sequence. Preferred method: In the method of (12), the step of determining differences between the first sequence and the reference sequence comprises identifying polymorphisms.

Due to the complexity of the institutional affiliation, it was not possible to simply perform the institutional analysis by the help of online databases. Rather, it was necessary to construct a special in-house database where additional institutional information could be integrated and analyzed. It would have been not useful to construct the in-house database by WPI records, as they do not provide information on the nationality of inventors and applicants. In addition, the download of WPI records is very expensive and would not have been covered by the available project budget. Therefore, the European Patent Office was asked to provide the necessary data. The EPO disposes of a broad patent file with records of various sources, including EPO and international PCT applications.[8] These files called EPODOC and EUREG record the official data, in particular the nationality of inventors and applicants. However, the content quality of the titles and abstracts is poor, as they are prepared by the applicants. This characteristic is not relevant for the present project, as the EPO data were not used for keyword searches. Rather, Fraunhofer ISI executed the searches in the WPI database and transferred the identified publication numbers to the EPO. The EPO matched the publication numbers with their data set and transferred the records back to Fraunhofer ISI. The contents of the EPO databases EPODOC and EUREG are similar with the major difference that EUREG includes additional information on the legal status about the patent examination process. In the project, the EUREG version was selected because the structure of the inventor field allowed for an easier integration in an in-house database. Figure 2.3-3 documents the EUREG version of the WPI record shown in Figure 2.3-2.

[8] PCT = Patent Co-operation Treaty.

Figure 2.3-3: Example of EPO's Document Structure

```
328/493 - (C) EUREG / EPO
 PN  - ---EP1068312--- A2 19991021 [2001/03]
 PPN - ---WO9953051---  19991021 [2001/03]
 PR  - ---US19980057719--- ---19980409--- ; ---US19980069047---
     ---19980428--- [2001/03]
 AP  - EP19990912007 19990409
 PAP - WO1999IB00712
 XIC - C12N-015/11 ; C12N-015/10 ; C07K-014/47 ; C12P-021/00 ; C12Q-001/68 ;
     C07K-016/18 ; G06F-017/30 ; G06F-017/50 [2001/03]
 ET  - 5' ESTS AND ENCODED HUMAN PROTEINS [2001/03]
 INA - 01 / DUMAS MILNE EDWARDS, Jean-Baptiste / 8, rue Gr goire-de-Tours /
     F-75006 Paris / FR
     02 / DUCLERT, Aymeric / 6 ter, rue Victorine / F-94100 Saint-Maur / FR
     03 / GIORDANO, Jean-Yves / 12, rue Duhesme / F-75018 Paris / FR
     [2001/03]
 PAA - FOR ALL DESIGNATED STATES
     GENSET
     24, rue Royale
     75008 Paris/FR [2001/03]
```

2.3.3 Extraction of institutional information from EPO data

The raw data files from EPO were prepared for further analysis at both the document and the institutional level. Both aspects are discussed in the following sections.

2.3.3.1 Applicant/inventor information at the document level

First of all, the data files were cleaned for patent applicants from the USA, Japan and other countries leaving a total number of documents with applicants and inventors respectively from EU and associated countries as depicted in Table 2.3-2.

Table 2.3-2: Total number of patent documents with applicants or inventors from EU and associated countries

	Applicant-determined documents	Inventor-determined documents	Total number
Genetics/ heredity	6288	6622	7111
Neurosciences	3120	3258	3526
Immunology	7027	7406	7909
Bioinformatics	133	139	151

The total number of patent documents used for further analysis was on the union of documents of the applicant and the inventor sets which gives the number of patent documents where at least both one patent applicant and/or one inventor come from EU and/or associated countries. Consequently, this number is smaller than the sum of the inventor and the applicant sets. We will generally refer to this union of documents.

2.3.3.2 Applicant/inventor information at the institutional level

Not only patent documents can be treated as units of analysis but also institutional entries in these documents. Therefore, the number of applicants and inventors in patent documents were counted. As presented in Table 2.3-3, there is abundant information with regard to these two sources of information. Genetics/heredity and Immunology prove to be the largest fields of analysis, both of them making up more than three quarters of the inventor data. In Table 2.3-3, the applicants and inventors are counted several times, if they appear in more than one document. So, the number of applicants in Table 2.3.3 his higher than the number of documents in Table 2.3-2, as in some patents have multiple applicants. Data entries from the applicant field give full information about an applicant. In contrast, the inventor field gives, in principle, only the full name of the inventor without systematic provision of institutional information.

Table 2.3-3: Number of applicant and inventor entries in patent documents

	Applicants	Inventors
Genetics	7215	18930
Neurosciences	3549	8541
Immunology	8101	20324
Bioinformatics	137	374

2.3.4 Extraction of inventor/author names

As already explained in section 2.3-3, public research institutions often do not appear as patent applications, even if they are the actual origin of an invention. Rather, patents are applied by private companies or the individual researchers. In order to show these hidden linkages to public research institutions, the inventor names were matched with author names

in the Science Citation Index (SCI) which records their institutional affiliation. Therefore, the next analytical step was to extract inventor names, to clean them if necessary, to store them in a separate file and to send them to CWTS as input for a matching procedure with the Science Citation Index.

2.3.4.1 Coping with spelling errors

A closer look at the data from the EUREG file of the EPO revealed a high number of spelling errors in the inventor field (INA, see Figure 2.3-4). For correcting these errors, the inventor data were uploaded into a particular Table with separate columns for surname, primary first name, institutional affiliation[9], town, postal code and country. As shown in Figure 2.3-4, in some inventor fields, parts of the scientist's full name or address are spelled incorrectly or are even incomplete. Many languages use vowel and consonant mutations (such as the German umlauts ö, ü, ä; the French accents é, è, ç; or the Czech c, g) that have adequate and comprehensive spelling codes neither in patent databases nor in common character sets used for data storage (e.g. ASCII, OEM, ISO, ANSI). Hence, the data from EPO was first standardized by using a special text editor software (EditPad Classic 3.5.3). Second, remaining misspellings were identified and rectified by using an in-house text editor tool and a manual case by case procedure. Data was cleaned as far as possible, but due to time and budget constraints perfectly cleaned data cannot be expected to emerge from this step of analysis. The main effort was invested into identifying and rectifying common error types, less attention was paid to very particular, low number types of errors. It can be assumed that at least 90 percent of the spelling errors in the inventor names were corrected. As to potential follow up activities, this effort for cleaning inventor names has to be taken into account.

[9] Somtimes, the institutional affiliation is documented in the inventor field (see further below).

Figure 2.3-4: Selected examples of spelling errors in EPO documents (highlighting by Fraunhofer ISI)

647/21507 - (C) EUREG / EPO
INA - 01 / NEUNER-JEHLE, Martin / 3, rue de la Gruerie / F-91100 Gif Sur Yvette / FR
02 / VAN DEN BERGHE, Loic / 6bis, rue de Viroflay / F-75015 Paris / FR
03 / BONNEL, **S,bastien** / 1, rue Maublanc / F-75015 Paris / FR
04 / UTEZA, Yves / 32, rue Henri Barbusse / F-94800 Villejuif / FR
05 / MENASCHE, Maurice / 7, ruelle des Oulches **d'H,rivaux** / F-95400 Villiers Le Bel / FR
06 / DUFIER, Jean-Louis / 149, **rue de SUEvres** / F-75015 Paris / FR
07 / ABITBOL, Marc / 36bis, rue Balard / F-75015 Paris / FR

685/21507 - (C) EUREG / EPO
INA - 01 / BILL, Roslyn / **Danska v"gen** 15 / S-463 71 **L"d"se** / SE
02 / BOLES, Eckhard / Roentgenweg 5 / 40591 **Dusseldorf** / DE
03 / GUSTAFSSON, Lena / Vitmossegatan 12 / S-431 69 M"lndal / SE
04 / HOHMANN, Stefan / Norra **H"cks"b"cksv"gen** 46 / S-443 32 Lerum / SE
05 / LARSSON, Christer / **Kik sgr"nden** 23 / S-431 64 **M"lndal** / SE
06 / OTTERSTEDT, Karin / Sterlingsgatan 2A / S-414 81 **G"teborg** / SE

3508/21507 - (C) EUREG / EPO
INA - 01 / FUCHSBERGER, Norbert / Institute of Virology, Dobravsk 9 / 84246 Bratislava / SK
02 / HAJNICK , **Val,ria; / Sl dkovičova** 2 / 811 06 Bratislava / SK
03 / KOC KOV , Paula / Institute of Virology, Dobravsk 9 / 84246 Bratislava / SK
04 / SLOV K, Mirko / Institute of Zoology, Dobravsk 9 / 84206 Bratislava / SK
05 / **GAŠPER K, Juraj / Odbor rske N** m. 2 / 81108 Bratislava / SK

2.3.4.2 Transfer format

After successful data cleaning, field-specific lists of inventors were filed to CWTS. As illustrated in Figure 2.3-5, every inventor name was linked to an ID code, to his or her country of origin and further address information. The ID code was necessary for being able to re-link

information sent back by CWTS to the original patent file. All in all, more than 51.000 names were extracted, cleaned and sent to CWTS.

Figure 2.3-5: List of inventors (extract)

16283-9	AALKJAER, Christian	DK	DK-8220 Brabrab
14713-2	AALTONEN, Johanna	FI	FIN-00400 Helsinki
16724-2	AANDAHL, Einar, Martin	NO	N-2600 Lillehammer
5429-2	AAPOLA, Ulla	FI	FIN-33720 Tampere
8747-2	AARNOUDSE, Corlien, A.	NL	NL-2224 XM Katwijk
2722-1	AARSKOG, Nina, K.	NO	N-5700 Voss
8491-2	AASE, Karin	SE	S-171 77 Stockholm
8800-2	AASE, Karin	SE	S-17177 Stockholm
2667-4	AASE, Karin	SE	171 77 Stockholm

2.3.5 Linkage of inventor names and institutions in SCI

As to the match of inventor names to author names in the SCI, different approaches were tested. The following selection criteria proved to be the most appropriate:

- Match with publications of the period 1996 to 2002
- Match with publications in the area considered.
- Match of surname and initial of primary first name
- Match of author and inventor country
- Match of first authors in case of publications with multiple institutions
- Match of all authors in case of publication with one institution

The aim of the matching process was to achieve as many matches as possible, on the one hand, and as precise matches as possible, on the other hand. The limitation on the publication period is necessary to refer to scientific activities at the time as the patent applications issued; otherwise the danger of inaccurate matches is too high. A search in the total SCI leads to a high number of inappropriate matches due to many identical names in other research fields. With the restriction to the specific area, for instance, Genetics/heredity, the quality of match improves considerably. It was not possible to look for the full first names of the authors, as the SCI only documents the initials of the first names. In the case of multiple first names, they are registered in a systematic way neither in patent applications nor in publications. Therefore, a restriction of the match on the initial of the primary first name was unavoidable. It proved to be less useful to match the city of inventors and authors, as patents generally document the private address and publications the institutional one. It cannot be assumed that all authors live in the city of their institution, so that the match of the country was the only reasonable solution. In case of multiple authors and institutions, the SCI only allows for a clear reference between the first author and the first institution, However, if only one institution is indicated, all authors can be referred to it. All in all, the combination of the different matching criteria led to satisfying results.

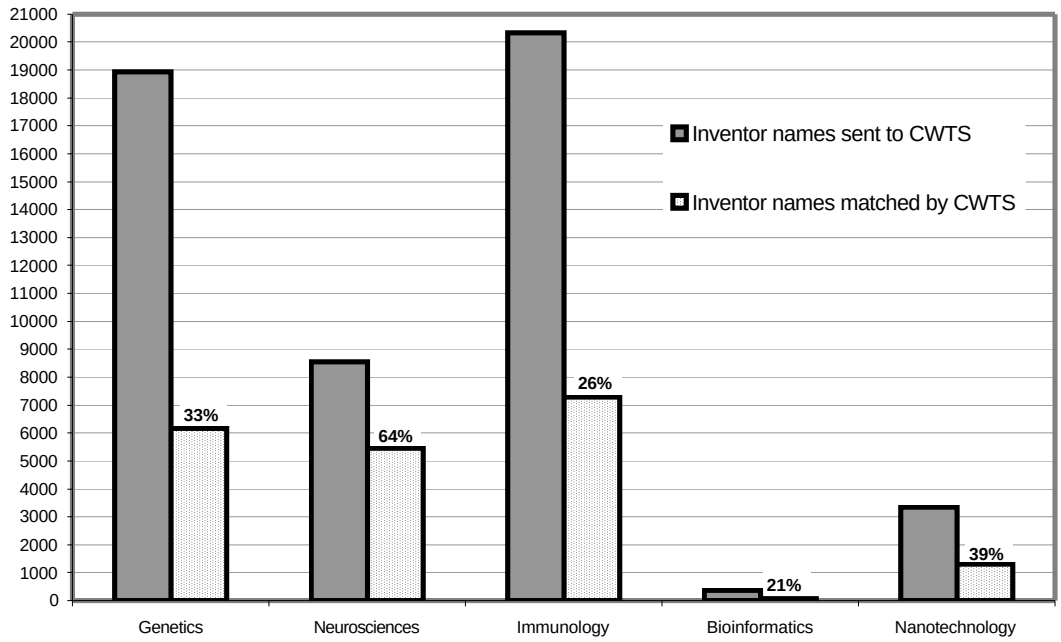
Fraunhofer ISI received the matched data from CWTS in a format illustrated in Figure 2.3-6. The first column displays all first names, the second column shows the surname and the truncated initial of the first names used by CWTS as input for the matching process in SCI ("%" stands for open truncation), the third and fourth column gives the author name and the institutional affiliation respectively, as retrieved in the SCI.

Figure 2.3-6: List of matches for inventors (extract)

Johanna	AALTONEN J%	AALTONEN J	FINLAND/HELSINKI/NATL PUBL HLTH INST
Ulla	AAPOLA U%	AAPOLA U	FINLAND/TAMPERE/UNIV TAMPERE
Nina K	AARSKOG N%	AARSKOG NK	NORWAY/BERGEN/UNIV BERGEN
Pierre	ABAD P%	ABAD P	FRANCE/ANTIBES/INRA
Nacer Eddine	ABBAS N%	ABBAS N	FRANCE/EVRY/AVENTIS
Sonia	ABDELHAK S%	ABDELHAK S	FRANCE/PARIS/INST PASTEUR
Marc	ABITBOL M%	ABITBOL M	FRANCE/PARIS/UNIV PARIS 05

The data in Figure 2.3-7 give an impression how many inventor names could be matched in SCI. It is a startling result to see matching coefficients vary between 21 and 64 percent. First of all, the low overall match for Bioinformatics seems to be related to a problematic definition of the field in SCI. Intermediate evaluation results obviously indicate that the delineation of this particular field was less successful than in the others. In contrast, Neurosciences reaches a remarkable 64 percent match. This means that about two out of three inventor names were matched successfully in the SCI, obviously due to specific field structures explained further below. For Genetics/heredity and Immunology, about one third of the inventors was matched in SCI.

Figure 2.3-7: Matched inventor names in SCI
(all fields, total numbers and percent)



2.3.6 Qualitative analysis of matched CWTS inventor-institution pairs

Before implementing CWTS matching results into the Fraunhofer ISI in-house database, an extensive qualitative analysis had to be carried out in order to ascertain whether the matched pairs could be used for further analysis. For this purpose a threefold classification scheme was applied which allowed inventor-institution pairs to classify into "non-match", "partial match" and "full match". This analysis was carried out by a qualitative analysis of all cases.

2.3.6.1 Matches and non-matches

First, it had to be examined whether CWTS matches were either non-matches and had to be excluded from further analysis, or whether the matched inventor-institution pairs qualify for further steps of analysis. Table 2.3-4 shows two instructive examples of non-match and a match cases. The first example (Andersen) displays a case in which the initials of all first names do not correspond to SCI initials. A correct match would have been "Andersen LN". Such cases were given the code "non-match" and were not considered anymore. The second example (Aaltonen) displays a case where the SCI initial corresponds perfectly to the capital letter of the first name. Such cases were given the code "match" and were implemented in the in-house database.

Table 2.3-4: Non-match and match examples of inventor/author names

First names in patent	Truncated name in SCI	Matched name in SCI
Lene Nonboe	ANDERSEN L%	ANDERSEN LW
Johanna	AALTONEN J%	AALTONEN J

2.3.6.2 Partial and full matches

Within the group of matches, two different types were differentiated. First, there were "full matches" such as the Aaltonen case in Table 2.3-4, where the initials correspond fully to the first names. Second, there are "partial matches" with small deviations between initials and first names. Consider Table 2.3-5 for three common examples of "partial matches".

The first example in Table 2.3-5 (Andersson) shows that the matched initial "M" is incomplete, as the initials of the first names are "MK". These cases, however, cannot be classified as "non-match" since it seems very plausible that "Maria Kristina Andersson" has mentioned only her primary first name in the publication document. In the light of other clues "M" seems favorable though: a second match gives "ML" as initials and is certainly a "non-match". Therefore, Karolinska Institute at Stockholm appears not to be the correct institutional affiliation of the person examined, but University Gothenburg has a high probability to be a match and is used for implementation into the in-house database, classified as "partial match".

Table 2.3-5: Partial matches from SCI (extracts)

First names in patent	Truncated name in SCI	Matched name in SCI	Matched institution in SCI
Maria Kristina	ANDERSSON M%	ANDERSSON M	SWEDEN/GOTHENBURG/UNIV GOTHENBURG
Maria Kristina	ANDERSSON M%	ANDERSSON ML	SWEDEN/STOCKHOLM/KAROLINSKA INST
ChristopheJR	ANDRE C%	ANDRE C	FRANCE/RENNES/UNIV RENNES 1
Catherine	ANDRE C%	ANDRE C	FRANCE/RENNES/UNIV RENNES 1
Joseph	COHEN J%	COHEN JL	FRANCE/PARIS/UNIV PARIS 06
Jose	COHEN J%	COHEN JL	FRANCE/PARIS/UNIV PARIS 06
Jean	COHEN J%	COHEN JL	FRANCE/PARIS/UNIV PARIS 06

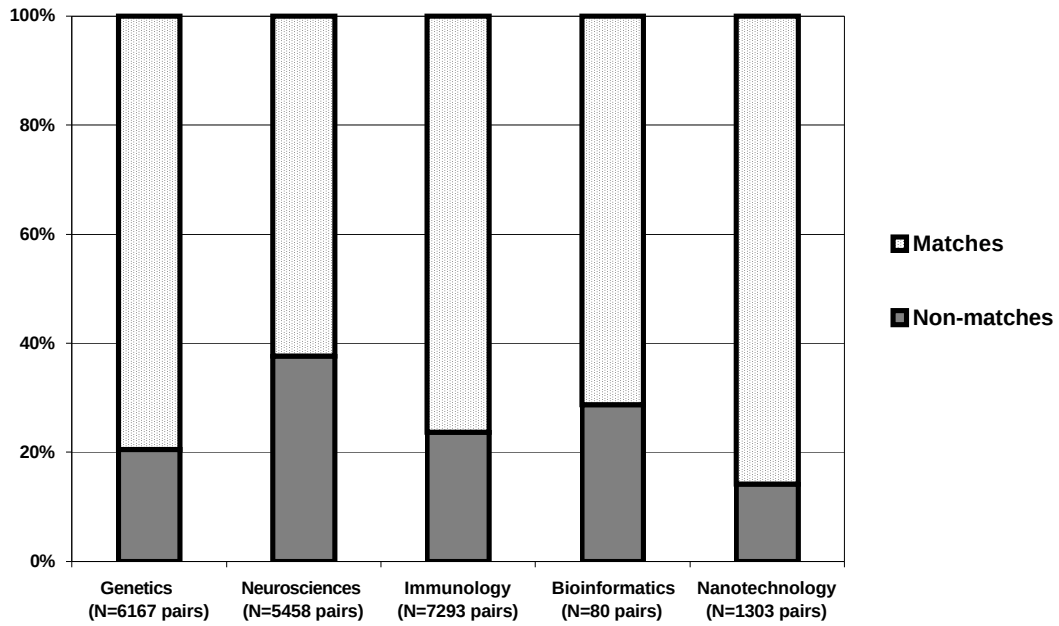
The second example in Table 2.3-5 (Andrec) displays a case where an institution has a good chance to be a "full match" for two persons. As can be seen, the initial "C" might refer to "ChristopheJR" or "Catherine". Though the latter person appears as a better match, both persons live in French-speaking areas (Belgium and France), and both might work for University Rennes 1. Because one cannot determine the "correct/true" person, it was decided to use both, but to treat them as "partial matches".

The third example in Table 2.3-5 (Cohen) combines the second and third matching problem. Hence, both the matched initial does not correspond perfectly to the initial of all first names, and there is more than one person whom the institutional affiliation might refer to. Compared to the second example, however, matched initials are "overcomplete" insofar as they indicate the existence of a second first name that cannot be found in the patent document. Again, as in case three, the home town of the inventor gives no additional clue as to the "true" inventor. Consequently, it was decided, too, to attach the label "partial matches" to these three persons and to implement them into the Fraunhofer ISI in-house database.

2.3.6.3 Results for all types of inventor-institution matches

The count of non-matched and matched inventor-institution pairs shows that the CWTS matching procedure can be deemed as an overall success. As shown in Figure 2.3-8, for Genetics/heredity and Immunology about eighty percent of the inventors could be properly assigned to an institution in SCI.

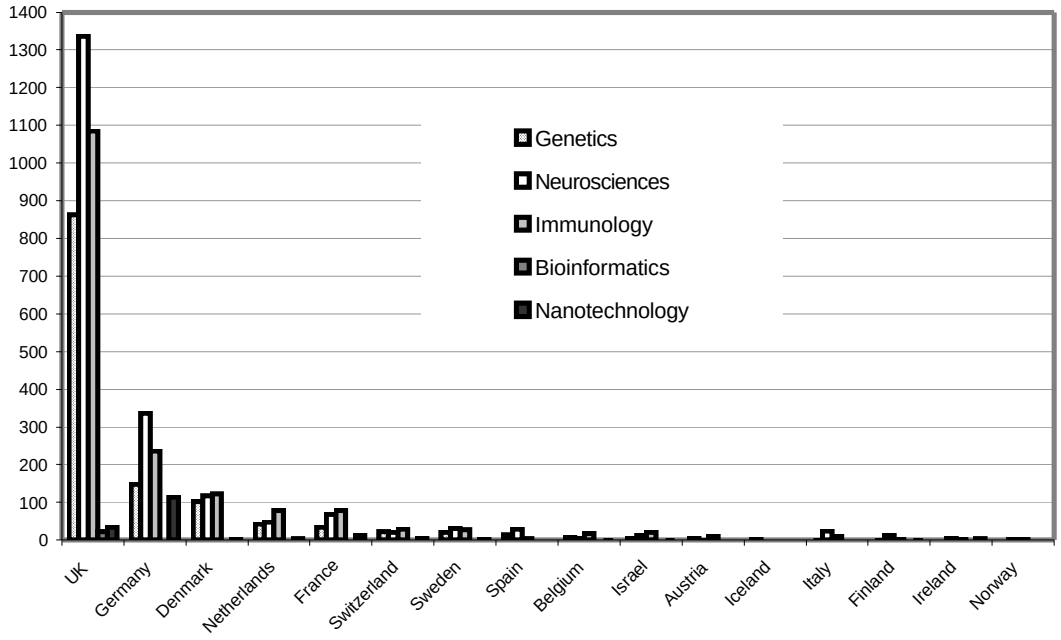
Figure 2.3-8: Non-matches and matches (inventor-institution pairs) across all fields (percent)



Question marks are left for Bioinformatics and particularly Neurosciences where non-matches are substantially higher than in the aforementioned fields. As reminder, the share of original matches in Neurosciences was quite high (Figure 2.3-7), obviously linked to a broader share of non-matches. At the country level (Figure 2.3-9), the UK has most non-matches (64% of

all non-matches) irrespective of the field, followed by Germany, Denmark, the Netherlands and France (together 29% of all non-matches). A detailed Table with all relevant numbers can be found in the Annex of the report.

Figure 2.3-9: Non-matches (inventor-institution pairs) across fields and countries (total number)



The authors of this study could not examine systematically the reasons underlying the field- and country-specific distributions of matches and non-matches. However, the high numbers of non-matches for the United Kingdom call for further explanation. We suggest two answers: First, an inspection of the UK data reveals that British scientists list their second and third first names (and their initials respectively) more often in patent documents than their continental colleagues. Consequently, as there are more combinations of both initials and full names, the probability to have a non-match increases for British researchers. Second, this finding is reinforced by the above average coverage of English-language journals in the SCI (see Grupp et al. 2001) which enlarges the stock of potential British authors and thus, leads to a further increase of the probability for non-matches in the British case.

To give an instructive demonstration of this phenomenon, consult Table 2.3-6 where for the common family name "Brown" three first name variations are considered. For the truncated initial "A" there are five possible matches in SCI, namely "A", "AJ","AF","AK" and "AJP". These five entries occur ten times because there are two persons, "Andrew James" and "Anthony M". As "AJ" is a full match for the former, and "A" a partial match for the latter (each refers to one patent document, indicated by the ID code), eight non-matches remain to be deleted. Likewise, there are six non-matches for "Daniel" because two out of three combinations of initials (all are partial matches) must be removed from further analysis. For other countries, such as Germany, France or Sweden, these types of non-matches occur to a much smaller extent. To sum up, the high number of non-matches for the UK can be attributed to the extensive use of initials, second and third first names as a cultural peculiarity and is reinforced by the high coverage of the English-language journals in SCI.

Table 2.3-6: Non-matches (inventor-institution pairs) across fields and countries

ID code	First names in Patent	Truncated name in SCI	Matched name in SCI
15256-1	Andrew James	BROWN A%	BROWN AJ
15256-1	Andrew James	BROWN A%	BROWN AJP
15256-1	Andrew James	BROWN A%	BROWN AF
15256-1	Andrew James	BROWN A%	BROWN A
15256-1	Andrew James	BROWN A%	BROWN AK
17871-1	Anthony M	BROWN A%	BROWN AJ
17871-1	Anthony M	BROWN A%	BROWN AJP
17871-1	Anthony M	BROWN A%	BROWN AF
17871-1	Anthony M	BROWN A%	BROWN A
17871-1	Anthony M	BROWN A%	BROWN AK
11128-3	Daniel	BROWN D%	BROWN DM
21013-1	Daniel	BROWN D%	BROWN DM
21014-1	Daniel	BROWN D%	BROWN DM
11128-3	Daniel	BROWN D%	BROWN DT
21013-1	Daniel	BROWN D%	BROWN DT
21014-1	Daniel	BROWN D%	BROWN DT
11128-3	Daniel	BROWN D%	BROWN DJ
21013-1	Daniel	BROWN D%	BROWN DJ
21014-1	Daniel	BROWN D%	BROWN DJ

With regard to full and partial matches, the results of the matching procedure seem, too, an overall success (Figure 2.3-10). Apart from Bioinformatics with none, partial matches are evenly distributed around 18-19 percent across fields. This means that at least more than eighty percent of all matched inventor-institution pairs are unambiguous ones. Among the countries with most partial matches, again the UK, Germany are the most important ones and also, at some distance, France the Netherlands and Denmark (Figure 2.3-11). The overall distribution of partial matches seems not to diverge distinctly from the distribution of non-matches. Apart for the aforementioned reasons as to the UK, no evidence for a systematic pattern of the distribution of partial matches across fields and countries could be identified. A detailed Table with all relevant numbers can be found in the Annex of the report.

To conclude, non-matches are not equally distributed by field and country and the share of non-matches is relevant, so that this additional non-automatic check must be carried out to achieve reliable results.

Figure 2.3-10: Full and partial matches (inventor-institution pairs) across all fields (percent)

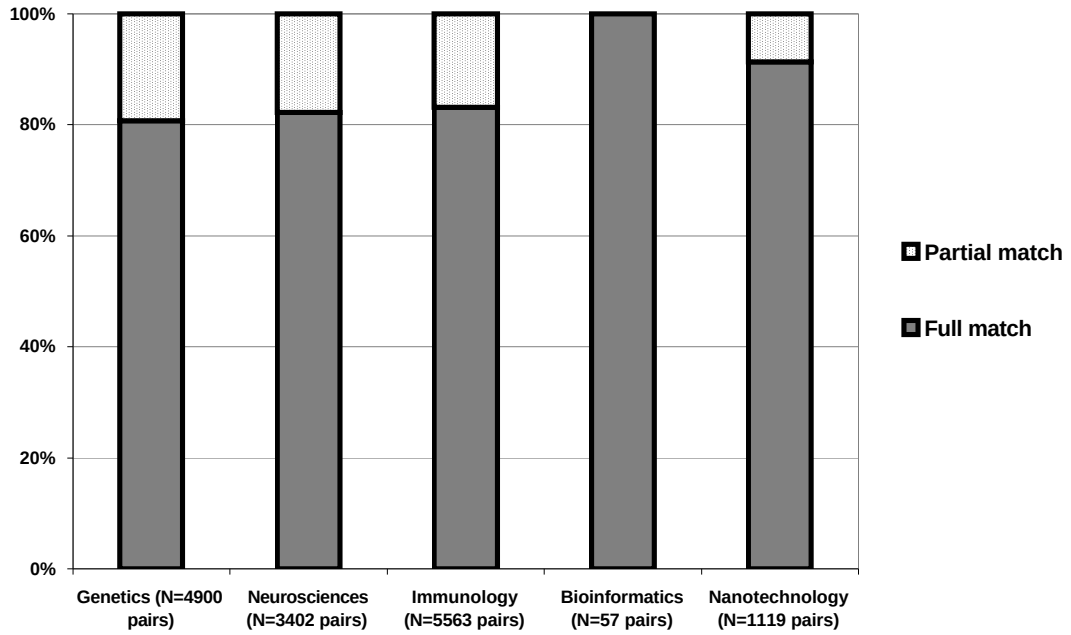
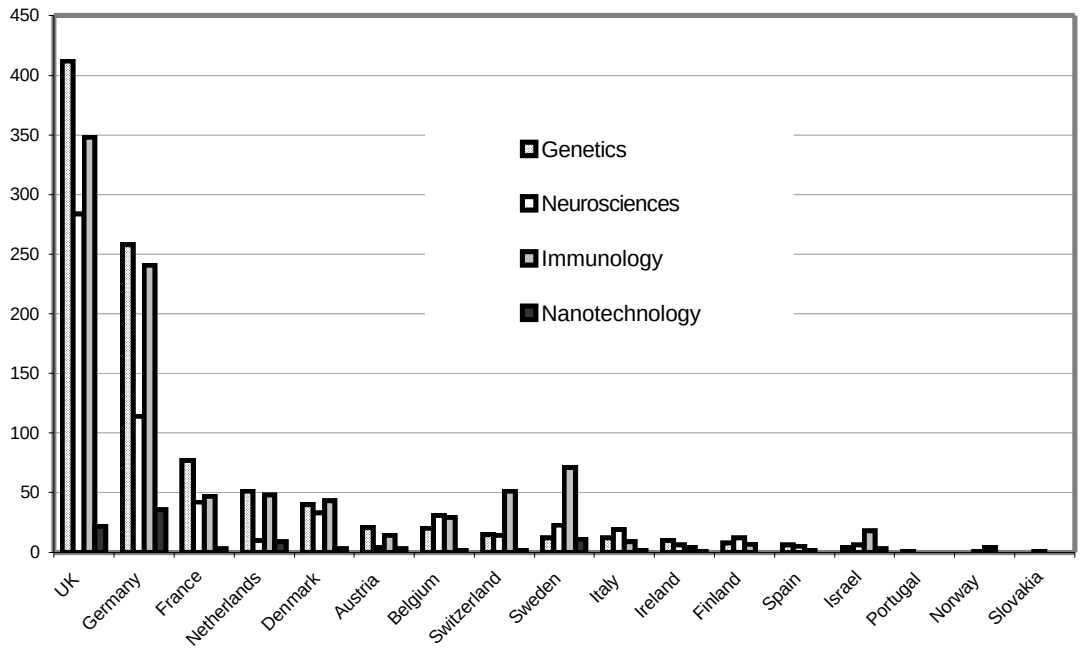


Figure 2.3-11: Partial inventor-institution matches across countries and fields (total number)



2.3.7 Implementation of matched inventor-institution pairs into the in-house patent database

Fraunhofer ISI implemented all full and partial CWTS inventor-institution matches into its in-house database. In addition to CWTS matches, institutional information from the aforementioned inventor field was taken systematically into account, for this field was found

to contain a relevant number of institutional data entries. As an example consult Figure 2.3-12 depicting that "Lorenz Poellinger" works at "Karolinska Institutet" (extract from original EPO document), i.e., the inventor does not indicate his or her private address, but the address of his institution in this specific patent document. This extraction of institutional information from the inventor field had again to be done by a non-automatic process, as only a limited share of inventors record their institutional address, and if available, the institutional address has not a standard format. Therefore, this extraction was quite labor-intensive.

Figure 2.3-12: Example of institutional information in INA field (extract from original EPO document)

1451/22724 - (C) EUREG / EPO

PPN - ---WO0212326--- 20020214 [2002/15]

PR - US20000223480P 20000807

AP - EP20010970061 20010807

PAP - WO2001IB01775

XIC - C07K-014/47

ET - MECHANISM OF CONDITIONAL REGULATION OF THE HYPOXIA-INDUCIBLE FACTOR-1

BY THE VON HIPPEL-LINDAU TUMOR SUPPRESSOR PROTEIN

INA - 01 / POELLINGER, Lorenz / **Karolinska Institutet** / S-171 77 Stockholm /

SE

02 / PEREIRA, Teresa / **Karolinska Institutet** / S-171 77 Stockholm / SE

03 / RUAS, Jorge / **Karolinska Institutet** / S-171 77 Stockholm / SE

PAA - FOR ALL DESIGNATED STATES

Angiogenetics Sweden AB

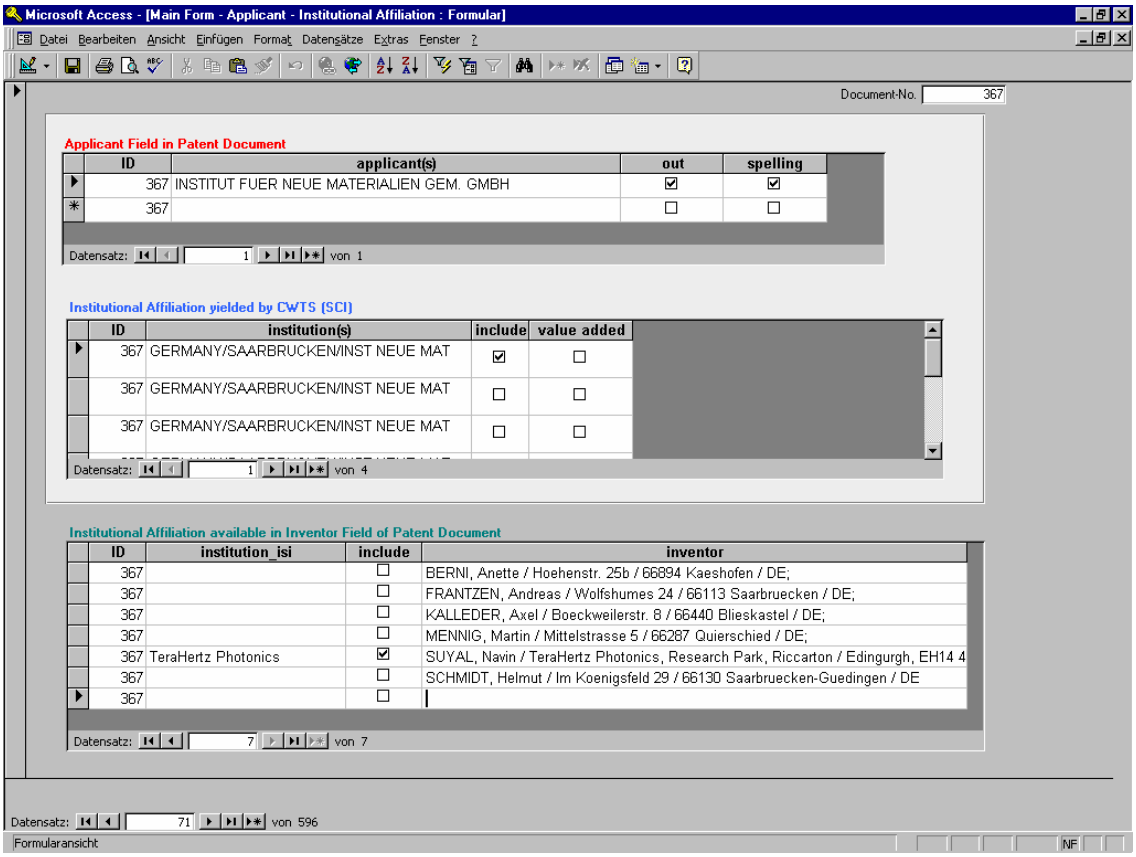
P.O. Box 440, Medicinaregatan 7B

405 30 Gothenburg/SE

2.3.7.1 Proceeding of implementation

Institutional data both available in the inventor field and extracted from SCI make up a pool of institutional information that was examined systematically. The screenshot of the MS Access interface (Figure 2.3-13) depicts from top to down the applicant institution entry, or source 1 ("Applicant Field..."), institutional information derived from SCI, or source 2 ("Institutional Affiliation...(SCI)") and finally, institutional information as available in the inventor field of the patent document, or source 3 ("Institutional Affiliation...Document").

Figure 2.3-13: MS Access interface screenshot for determining institutional affiliations of a patent document



First, it was checked whether there is any additional information at all. For source 2 and source 3 in Figure 2.3-13 there are two entries that can be considered for further analysis. The inventor-institution couples yielded by SCI generates the same information as was available in the applicant field, namely "Institut für neue Materialien GmbH Saarbrücken, Germany". In contrast, there is additional institutional information in the inventor field, or source 3: "TeraHertz Photonics" which could be implemented ("include"-box is ticked).

Secondly, for implementing new institutional information from the second and third source, some rules were applied regarding the **level of institutional aggregation**:

- **Universities** are taken at the organization level (e.g., University of Brussels), not the faculty or department level.
- For **non-university research organizations** the institute's level is taken instead of the umbrella organization (e.g., for Germany "Fraunhofer-Institute for Biomedical Engineering" is used instead of "Fraunhofer-Gesellschaft zur Förderung der angewandten Wissenschaften", and for France "CNRS Institute de Biologie de Lille" is taken instead of "CNRS").
- **Private corporations** are taken at the organization level.

Further, rules were applied with respect to the data used for the analysis **in case of identical institutional information**:

- In case of identity of source 1 (applicant field) and source 2 (inventor field), the former is chosen.
- In case of identity of source 1 (applicant field) and source 3 (SCI match), the former is chosen.
- In case of identity of source 2 (inventor field) and source 3 (SCI match), the latter is chosen.

These procedures were carried out for every patent document, every field of science and every country. Because the systematic check of informational overlap, by definition, could not be operated by a computer program, a qualitative analysis had to be carried out. The whole procedure comprised many ten thousand manual checks!

2.3.7.2 Data redundancy

Data redundancy occurs less often within the applicant field, but for source 2 and 3, either because the inventors work at the same institution yielding one institutional information out of two or more data entries (case 1). In other records, institutional affiliations identified in the inventor field or the SCI reproduced an already existing institution from the application field (case 2). Examples for both, case 1 and 2 are presented in Figure 2.3-14 and 2.3-15.

Figure 2.3-14: Data redundancy, case 1: Two inventors work at the same institution

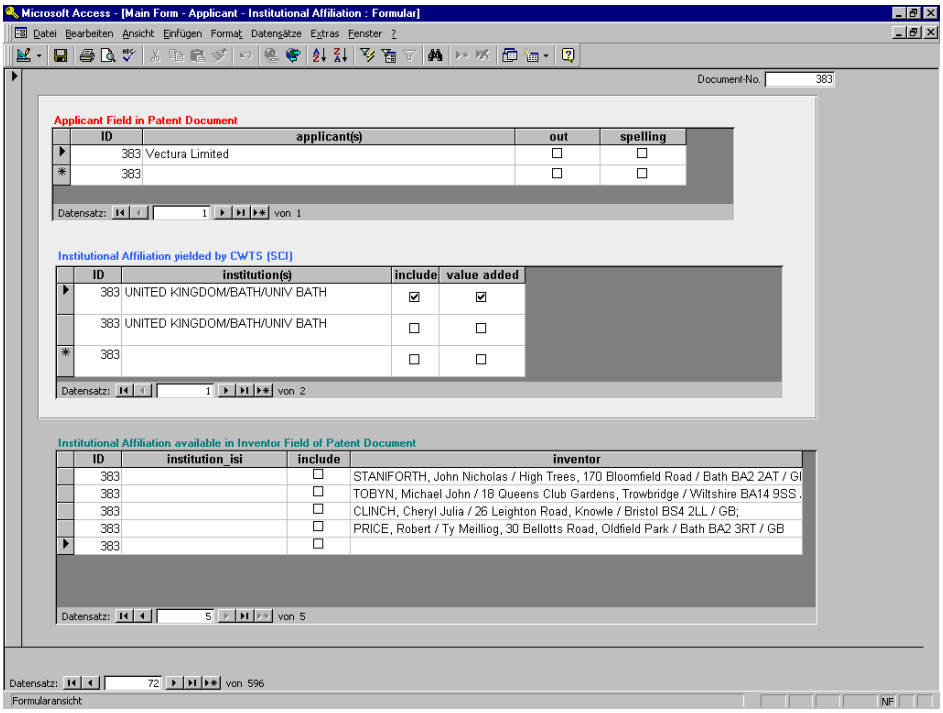
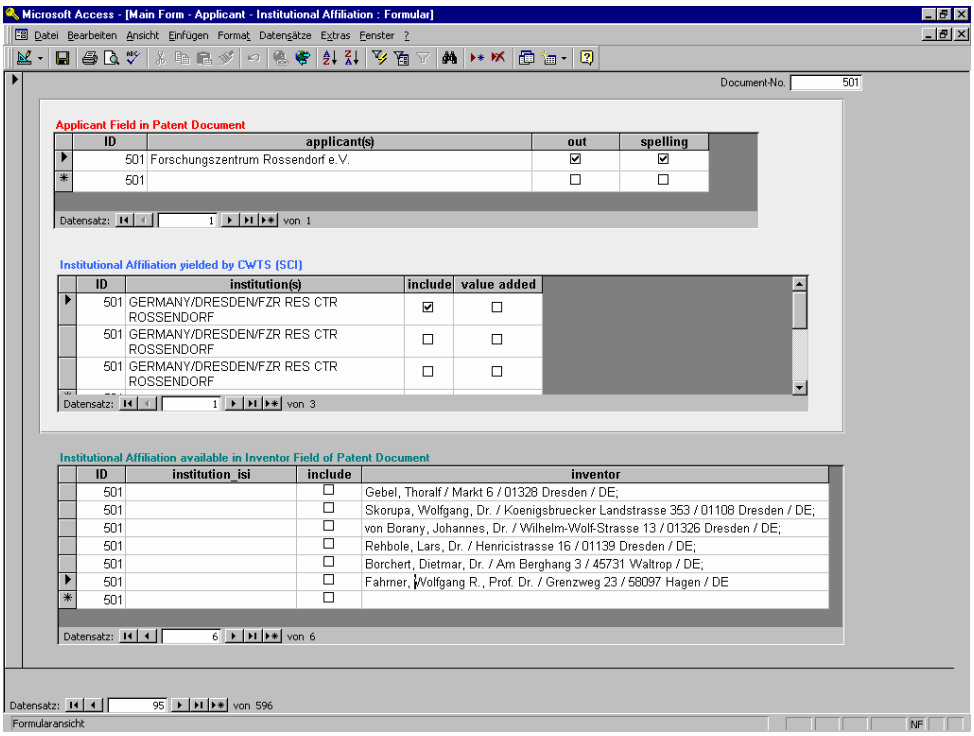


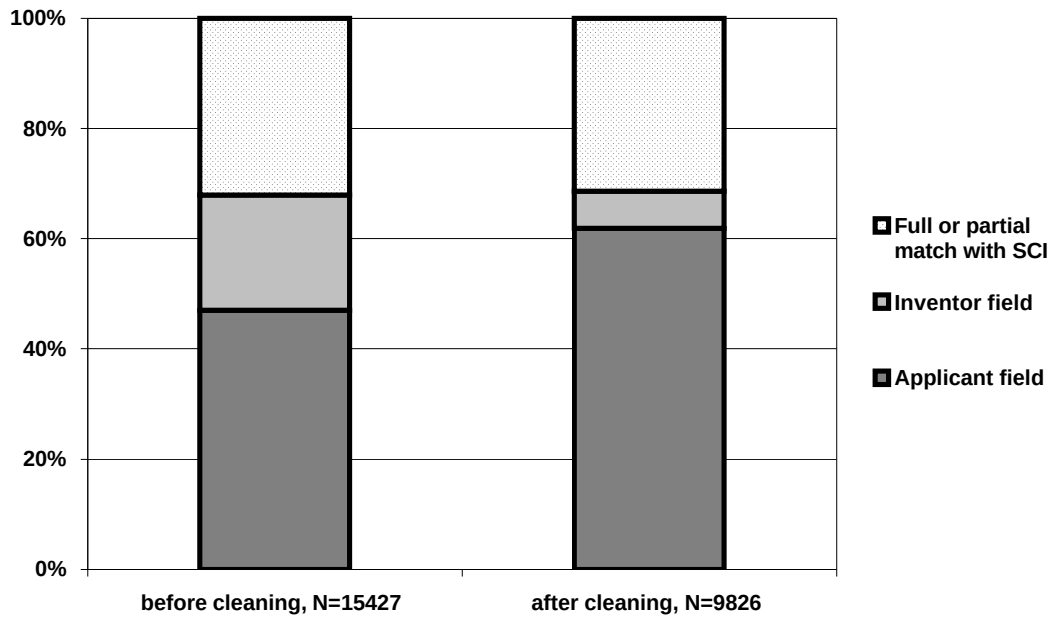
Figure 2.3-15: Data redundancy, case 2: SCI-extracted institution reproduces information from applicant field


Microsoft Access - [Main Form - Applicant - Institutional Affiliation : Formular]
Datei Bearbeiten Ansicht Einfügen Format Datensätze Extras Fenster ?
Document-No. 501
Applicant Field in Patent Document
ID applicant(s) out spelling
501 Forschungszentrum Rossendorf e.V.
501
Datensatz: 1 von 1
Institutional Affiliation yielded by CWTS (SCI)
ID institution(s) include value added
501 GERMANY/DRESDEN/FZR RES CTR ROSSENDORF
501 GERMANY/DRESDEN/FZR RES CTR ROSSENDORF
501 GERMANY/DRESDEN/FZR RES CTR ROSSENDORF
Datensatz: 1 von 3
Institutional Affiliation available in Inventor Field of Patent Document
ID institution_isi include inventor
501 Gebel, Thoralf / Markt 6 / 01328 Dresden / DE;
501 Skorupa, Wolfgang, Dr. / Koenigsbruecker Landstrasse 353 / 01108 Dresden / DE;
501 von Borany, Johannes, Dr. / Wilhelm-Wolf-Strasse 13 / 01326 Dresden / DE;
501 Rehbole, Lars, Dr. / Henricistrasse 16 / 01139 Dresden / DE;
501 Borchert, Dietmar, Dr. / Am Berghang 3 / 45731 Waltrop / DE;
501 Fahrner, Wolfgang R., Prof. Dr. / Grenzweg 23 / 58097 Hagen / DE
501
Datensatz: 6 von 6
Datensatz: 95 von 596
Formularansicht
NF

2.3.7.3 Results of the implementation process

Figures 2.3-16 and 2.3-17 display the share of institutional information at the outset of the exercise and the final distribution respectively. Information concerning the applicant field is provided in the dark grey area, while the bright grey shows the relative amount of institutional entries available in the inventor field. The amount of inventors, for which no further institutional affiliation could be found in patent documents, but which were full or partial matches with SCI are displayed in the white dotted area (see legend).

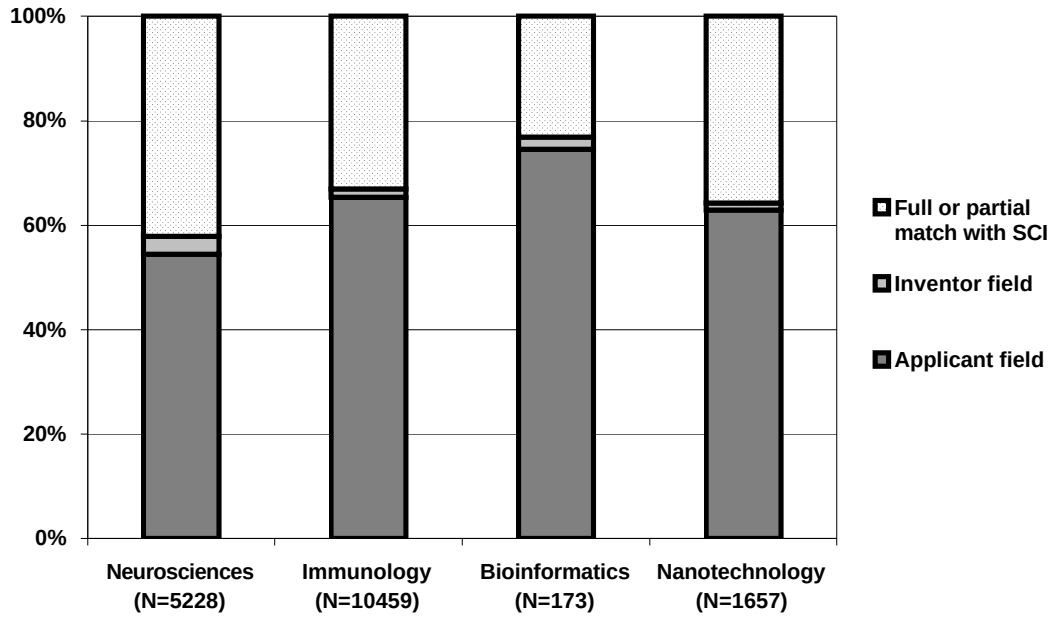
Figure 2.3-16: Results of implementation procedure as to the source of institutional information (Genetics/heredity)



The cleaning process reduced the number of data entries considerable, mainly due to redundant information. In sum, data redundancy leads to a reduction of data entries between 36 and 13 percent (Genetics/heredity: 36%, Neurosciences: 25%, Immunology: 23%, Bioinformatics: 13%).

Results from the starting set suggest that data both from the inventor field in patent documents and matching with SCI yield considerable additional information. Except for Bioinformatics, between a third and a half of all institutional information stem from sources other than the applicant field. This holds also after deleting redundant information. Contribution from the SCI matching procedure varies between 31 and 42 percent (Bioinformatics: 23%). To conclude, the systematic use of institutional data from the SCI makes available new information to a considerable, but varying degree.

Figure 2.3-17: Results of the implementation procedure as to the source of institutional information



It is important to note that institutional information from the inventor field is highly redundant to inventor-institution matches from the SCI and also to the applicant field. Therefore, the share of additional inventor field information in the field of Genetics/heredity reduced from 21 to 7 percent after cleaning. In the other four fields even less information was yielded from the inventor field, namely 3 percent for Neurosciences, and 2 percent for Immunology.

The overlap was particularly substantial between SCI matches and institutional information in the inventor field. For Genetics/heredity a manual check produced an overlap of over 60 percent between these two sources. First, this points to a considerable validity of the SCI match, for information from the inventor field is treated as redundant only for identical inventor-institution pairs in SCI.

Secondly, with regard to future mapping projects, a systematic exploitation of the inventor field seems not to be recommendable. As the additional institutional information varies between 1 and 7 percent, extensive efforts to retrieve such information manually from the inventor field (as done for Genetics/heredity) appear not to be justified. Future studies should rely solely on institutions from the applicant field and from the SCI. Also in this case, a systematic comparison of both sources as to redundancy can be regarded as useful, but the necessary amount of work will be considerably reduced compared to the integration of three sources.

2.3.8 Aggregation and categorization of institutional entries

For mapping excellent research entities, it is necessary to show the total number of patent applications for each institution. However, after the cleaning and matching on the level of institutional entries, it is still not possible to show this specific information. For this purpose, an additional working step has to be executed, the aggregation of the institutional entries. To a large extent, the institutional entries are identical and can be easily aggregated. But in various cases, there are small spelling differences or the institutions appear in quite different versions, as they were introduced from different sources. For instance, the applicants in the patent

databases refer to the national languages, whereas in the Science Citation Index (SCI), all institutional names are transferred into an English-language version. For a unified presentation, all versions linked to the same institution have to be brought together and the referring patent applications have to be added up. This process has to cope with several problems.

- The institutional versions in different languages have to be fit. As in the project, 32 countries were covered, this is sometimes quite problematic.
- In various cases, different units of an institution/organization appear in the database of institutional entries. For instance, many university clinics have specific names and the decision is needed whether they are independent units or are sub-units of the universities; or the names of various firms do not directly show that they are affiliations of larger companies.
- In some cases, patents originating from universities are taken out by license or transfer units which appear as applicants.

For solving this problems of institutional aggregation, specific expertise is needed, in particular specific knowledge on the national innovation systems. Therefore, we asked the members of the High Level Group (HLG), to support us in the unification of the institutional entries. The representatives of many countries agreed to clean the institutional lists referring to the area Neurosciences (countries). We sent an institutional, country-specific list to each national expert with a detailed explanation how the data should be unified.[10] The data for all missing countries and the missing three areas were unified by experts of Fraunhofer ISI, partly based on own experience, partly on manuals and information from the Internet. In addition, the Fraunhofer experts checked the output of the national experts with the result that in most cases, only minor oversights were detected due to the very high number of institutions. The specific knowledge of the experts on the national institutional structures lead to a high quality of the results. In some cases, however, a complete revision was necessary. As to the institutional aggregation, the same rules were applied as in Section 2.3.7.

The decision for a institutional aggregation at the university level was taken, as in many cases in the inventor fields or the SCI, the university is given and not the sub-unit. Furthermore, many departments or centers of universities appear with different names so that the exact association is problematic. In the case of non-university organizations, the linkage to institutes/centers is much clearer and only difficult in few cases. In particular, the disaggregation of the French CNRS, the Italian CNDR and the Spanish CSIC proved to be delicate, but in general, this lower level was useful to illustrate the regional distribution of research activities. The amount of work in this step is shown in Table 2.3-7.

The meaning of the table may be explained by the example of Genetics/heredity. The cleaning process started with 7215 applicants, as documented in Table 2.3-7. By a first cleaning, independent inventors which could not be linked to an institution were eliminated from the data set, leading to a reduction of the applicant numbers. But the inclusion of additional institutions form the inventor field and the SCI brought an increase of the institutional entries, so that 9826 institutions had to be analyzed after the matching process. Their automatic aggregation substantially reduced the number of institutions to 2719. The non-automatic qualitative aggregation, in this case by Fraunhofer ISI, implied a further reduction to 1079 institutions/ organizations, i.e., to 39 percent of the list after the automatic cleaning. Thus, the improvement of the institutional information by the non-automatic qualitative aggregation is substantial and cannot be replaced by a purely automatic cleaning. The amount of work in this

[10] This explanation is documented in the Annex.

step is considerable, as long lists of institutions have to be checked with care. In the case of Germany, the original list given to the national expert comprised 552 different entries, in the case of France 369, the United Kingdom 788, Switzerland 114, or Denmark still 80. For all fields, 10499 automatically pre-aggregated institutional entries had to be checked and thus far more than the expected some hundreds. The final number of 1079 institutions in genetics refers to 6363 patent applications,[11] i.e., one institutions applied for 5.9 patents in 1996 to 2000 on average.

As to further activities, we see no realistic alternative to this step of non-automatic qualitative institutional aggregation, as substantial knowledge on the institutional structure of specific national innovation systems is needed.

Table 2.3-7 Data volume processed in different working steps

Level of aggregation	**Genetics/ heredity**	**Neuro-sciences**	**Immu-nology**	**Bioinfor-matics**	**Total**
Number of applicants[12]	7215	3549	8101	137	19002
Number of institutions after matching process	9826	5228	10459	173	25686
Number of institutions after automatic aggregation	2719	1854	4972	106	9651
Number of institutions after non-automatic qualitative aggregation	1079	1366	1832	89	4366
Number of patent applications[13]	6363	2984	6763	151	16261

[11] The number of 6363 patent applications is lower than the 7111 applications according to Table 2.3-2, as patents of independent inventors were excluded.
[12] Including independent inventors.
[13] Without independent inventors.

2.4 Analysis of excellent institutions

2.4.1 Institution lists by country

Table 2.4-1: List of patenting institutions in Genetics/heredity (extract for Switzerland)

No of Patents	Institution	Town
66	SWITZERLAND/BASEL/ROCHE AG	BASEL
63	Novartis AG	4056 Basel
56	Syngenta Participations AG	4058 Basel
21	SWITZERLAND/LAUSANNE/UNIV LAUSANNE	LAUSANNE
20	SWITZERLAND/ZURICH/UNIV ZURICH	ZURICH
19	SWITZERLAND/ZURICH/SWISS FED INST TECHNOL ETHZ	ZURICH
18	SOCIETE DES PRODUITS NESTLE S.A.	1800 Vevey
15	SWITZERLAND/GENEVA/UNIV GENEVA	GENEVA
13	SWITZERLAND/BASEL/UNIV BASEL	BASEL
10	SWITZERLAND/LAUSANNE/SWISS INST EXP CANCER RES	LAUSANNE
10	Apotech R&D S.A.	1204 Geneve
9	SWITZERLAND/BASEL/FRIEDRICH MIESCHER INST	BASEL
7	Cytos Biotechnology AG	8952 Zuerich-Schlieren
7	RMF DICTAGENE S.A.	1008 Prilly
5	SWITZERLAND/BERN/UNIV BERN	BERN

The major aim of the patent analysis was to generate lists of institutions with considerable activities in patenting in the five fields analyzed. These lists provide the institution's name, its town of residence, and its number of EPO/PCT applications in the period from 1995 to 2000 (priority years). Table 2.4-1 shows an extract of Swiss applicants in the field of Genetics/heredity. All respective lists can be found in Excel format on a CD-ROM attached to the report. They can be downloaded, too, at the interface available at http://www.cwts.nl/ec-coe/cgi-bin/izite.pl?show=pub-lications.

2.4.2 Institutional differentiation: profit and non-profit institutions

In addition to compiling lists of institutions, a central target of the pilot study is to examine the extent to which technical knowledge originates from non-profit research institutions, such as universities and non-university research facilities, for example, the French CNRS, the German Max-Planck-Gesellschaft and Fraunhofer-Gesellschaft, or the Spanish CSIC. Table 2.4-2 exemplarily documents the results for Austria in Genetics/heredity.

It is not only instructive to look at country specific lists of institutions, but also to discuss the extent to which non-profit and for-profit institutions contribute to the production of technological knowledge. Figure 2.4-1 depicts for all fields the absolute and relative numbers of non-profit and for-profit institutions as they appear in the applicant field of patent documents. Quite clearly, before the cleaning/matching and implementation procedures, companies and other for-profit institutions make the lion's share of all patent applications across all fields. The picture emerging from Figure 2.4-1 tends to suggest an activity distribution with for-profit institutions engaged about twice as much in patent activity than non-profit institutions.

This finding, however, is more or less reversed after taking into account the institutional affiliation of inventors that were retrieved in the SCI. Figure 2.4-1 shows not only a substantial increase in terms of non-profit institutions' patent activity. Except for Bioinformatics, non-profit institutions even exceed for-profit institutions. Hence, non-public institutions such as universities and non-university research institutions generate technological knowledge at least to the same extent as do private companies.

Table 2.4-2: List of Austrian patenting institutions in Genetics/heredity, differentiated by non-profit and for-profit institutions

non-profit

No of Patents	Institution	Town
17	AUSTRIA/VIENNA/UNIV VIENNA	VIENNA
11	AUSTRIA/VIENNA/INST MOL PATHOL	VIENNA
5	AUSTRIA/VIENNA/BIOCENTRE VIENNA	VIENNA
3	AUSTRIA/VIENNA/TECH UNIV VIENNA	VIENNA
3	AUSTRIA/VIENNA/UNIV VET MED VIENNA	VIENNA
2	Oesterreichisches Forschungszentrum Seibersdorf Ges.m.b.H.	1010 Wien
2	AUSTRIA/VIENNA/AGR UNIV VIENNA	VIENNA
1	AUSTRIA/SALZBURG/LANDESKLINIKEN SALZBURG	SALZBURG
1	AUSTRIA/VIENNA/ST ANNA CHILDRENS HOSP	VIENNA
1	AUSTRIA/GRAZ/TECH UNIV GRAZ	GRAZ

for profit

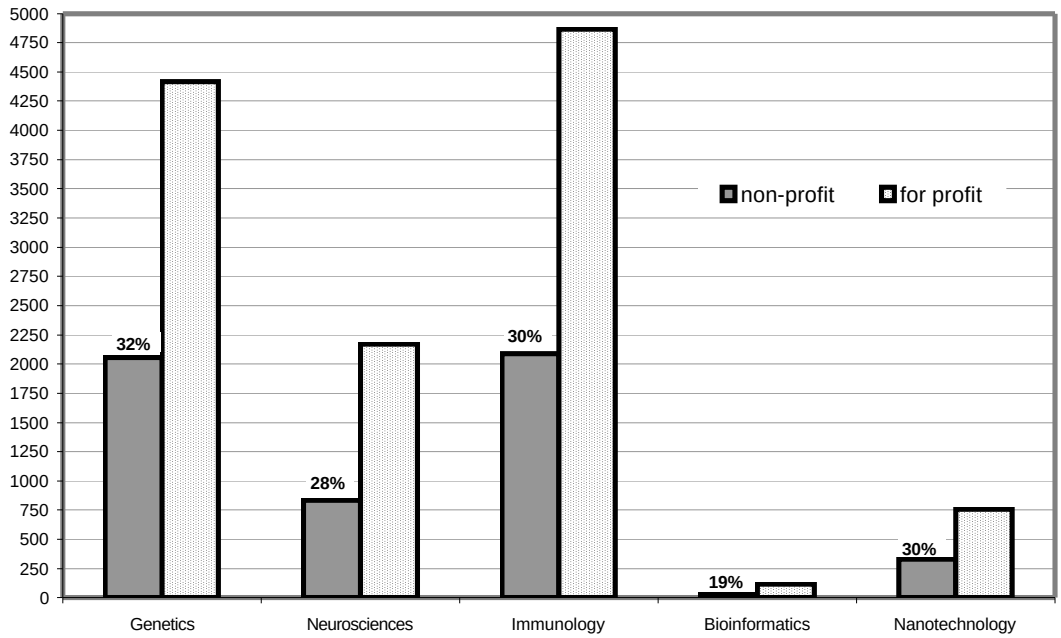
No of Patents	Institution	Town
27	Novartis-Erfindungen Verwaltungsgesellschaft m.b.H	1235 Wien
16	Syngenta Participations AG	1235 Wien
8	Baxter Aktiengesellschaft	1221 Wien
5	BIOCHEMIE GESELLSCHAFT M.B.H.	6250 Kundl
4	Austrian Nordic Biotherapeutics AG	1210 Wien
3	Polymun Scientific Immunbiologische Forschung GmbH	1190 Wien
3	AUSTRIA/VIENNA/BOEHRINGER INGELHEIM GMBH	VIENNA
2	IMMUNO Aktiengesellschaft	A-1221 Wien
1	AUSTRIA/VIENNA/VIENNALAB LABORDIAGNOSTIKA GMBH	VIENNA
1	Biomay Produktions- und Handelsgesellschaft mbH	4020 Linz
1	Biofrontera Pharmaceuticals AG	1235 Wien
1	Intercell Biomedizinische Forschungs- und Entwicklungs GmbH	1180 Vienna
1	Innovationsagentur Gesellschaft m.b.H.	1020 Wien
1	Bio Life Science Forschungs- und Entwicklungsges.m.b.H.	1010 Wien
1	DSM Fine Chemicals Austria Nfg GmbH & CoKG	4021 Linz
1	Cistem Biotechnologies GmbH	1030 Vienna
1	Bio-Products & Bio-Engineering Aktiengesellschaft	1010 Wien
1	ISIS PHARMACEUTICALS, INC.	1235 Wien

As to the interpretation of these results, it is necessary to look at the way of institutional counting in more depth which in Figure 2.4-2 is based on non-fractional counting. I. e., if a patent application is linked to more than one institution it is counted more than once. So if a patent is applied by a firm, but the inventor works at a scientific institution, the firm and the scientific institution get one count respectively. As to Figure 2.4-1, most patents have one applicant only, whereas in Figure 2.4-2, many patent applications are linked to several institutions so that the absolute number of counts increases. Referring again to the example of Genetics/heredity, the share of non-profit institutions increases from 32 to 52 percent due to the inclusion of additional institutions by the matching process. The absolute number of non-profit institutions grows from about 2050 to 5100.

However, we have to ask, whether the non-fractional counting is appropriate. In the case of an application linked to a firm and a scientific institute, it is realistic to assume that in most cases, the basic technical concept originates from the institute. According to this perspective, fractional counting is not appropriate, but a full count for the institute and no count for the firm. In some cases, the applications is linked to a firm as applicant and several scientific institutes. In this case, it seems to be appropriate to assume one count for non-profit institutions and no count for for-profit ones. When we apply this counting approach, the share of non-profit institutions in Genetics/heredity even increases to a share of about two third of

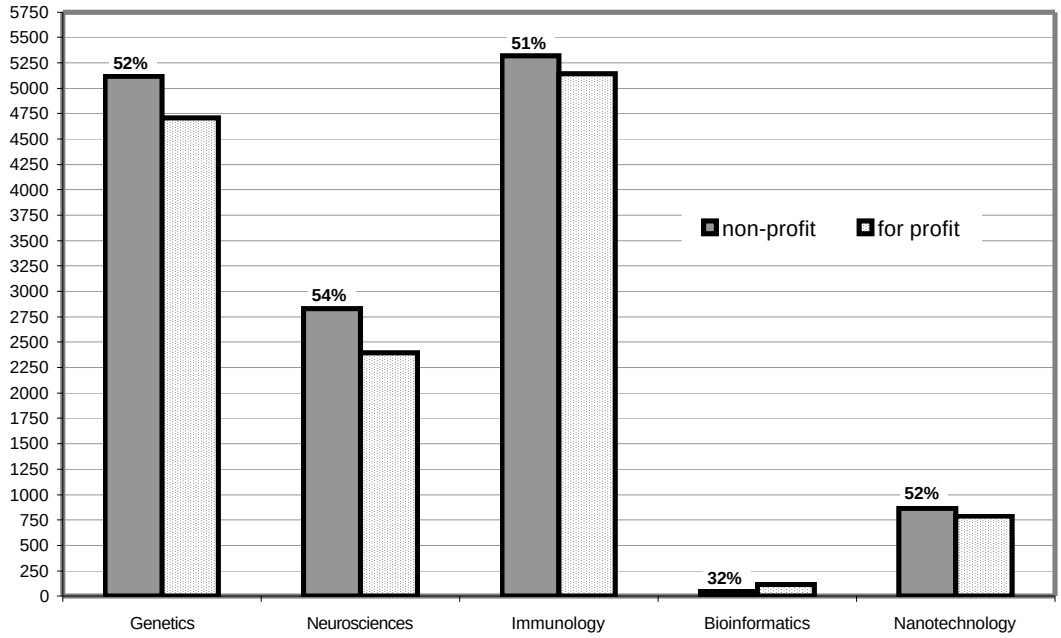
all applications. To conclude, the very high share of non-profit institutions in Figure 2.4-2 is not an artifact due to an inappropriate way of counting. It is even a minimum level due to conservative assumptions. In contrast, the level of about two third has to be considered a maximum level, as in many cases, the original idea comes from a scientific institute, but is improved and modified by the firm, so that it is useful for practical purposes. Then a fractional count with equal weights would be more realistic.

Figure 2.4-1: For-profit and non-profit institutions before cleaning/matching (all fields)



In order to check the reliability of the very high contribution of non-profit organizations, we analyzed the institutional data with US-American origin in more detail. This is feasible with a limited amount of work, as in the United States, public non-profit organizations actively take out patents and appear as patent applicants in the document records. The patents from universities, for instance, are applied by the universities, university boards, the university presidents etc., so that the university link can be identified without problems. For the period 1996 to 2000, the share of non-profit organizations within all US-American patents in Genetics/heredity appeared to be 37 percent. There might be some cases where patents with university origin are taken out by companies or individuals similar to the European situation. However, a share of non-profit organizations above 40 percent is not realistic. Thus, also in the United States, the share of non-profit organizations in the patents referring to this knowledge-based area is very high, but lower than in European countries. This might reflect the fact that in the united States, the research activities with regard to Genetics/heredity started much earlier than in Europe. Therefore, the potential to transfer scientific research results into technological application may be higher. The share of public non-profit organizations in knowledge-based fields of technology should be analyzed in more detail with reference to the life cycle. It might be an indicator for the technological maturity.

Figure 2.4-2: Absolute number of profit and non-profit institutions after cleaning/matching (all fields)



To sum up, the direct contribution of scientific institutions to the generation of new technology is very high with regard to the science-based fields considered. In all fields, the share of non-profit institutions is higher than 50 percent which is largely above average values for all technologies.[14] This striking result is already indicated by the findings before the match with levels for non-profit institutions of about 30 percent, but it is confirmed in a substantial way by the inclusion of additional information from the SCI.

2.4.3 Institutional coverage for different countries

The analysis of the shares of non-profit and for-profit institutions across fields yielded, apart from Bioinformatics, a rather uniform picture. However, one can expect considerable variation at the country level. For this reason, the analysis carried out in the preceding section is repeated, but on the level of single countries that have applied for EPO/PCT patents. The number of countries with at least one EPO/PCT application varies from field to field, covering 27 countries in Genetics/ heredity and Immunology, 25 countries in Neurosciences, and 11 countries in Bioinformatics.

Results are presented as follows: First, Figure 2.4-3 shows the absolute numbers of non-profit versus for-profit institutions in every field and every country at the outset of the study. The relevant information is here to obtain an overview about the absolute distribution of the two types of institutions across countries. Figure 2.4-3 makes clear, for instance, that France, Germany and the UK are the three most active countries in Genetics/ heredity. Further, for-profit institutions apply less often for patents in France than in Germany or the UK respectively. Second, Figure 2.4-4 shows the same distribution after the implementation of institutional information from the SCI. Again, France, Germany and the UK are the top 3 absolute performers in the field of Genetics/ heredity. However, as in France there were more non-profit institutions applications at the outset than in Germany or the UK respectively, the

[14] A major exception in the present data set is Bioinformatice with "only" 32 percent probably due to problems of field defintion.

SCI-matching procedure yielded less additional non-profit institutions for France than for the latter two countries. Third, in order to compare the relative increase in non-profit institutions from the matching procedure, Figure 2.4-5 depicts the country-specific additional shares of non-profit institutions yielded by the matching procedure. Figure 2.4-5 illustrates, for instance, that Sweden, Norway and Austria gained most, for their share of non-profit institutions increased by more than one third. For Luxembourg and France, however, the increase lies at modest 10 percent. They profited least from the SCI-matching procedure.

The respective figures for Neurosciences, Immunology, and Bio-informatics can be found in the Annex (see Figures A-3 - A-14). There can be found, too, absolute number for the first two figures.

The analysis of the institutional analysis reveals that for some countries, a substantial share of additional non-profit institutions could be identified by means of the SCI-matching procedure, while for other countries the effect tends to be negligible. Results for all fields[15] suggest that for one group of countries the methodology of this study was particularly effective. This first group consists of the following countries: Norway, Sweden, Finland and Austria with an average gain of 31 percent. In terms of additional non-profit institutions, these countries belong either to the top 5 or to the middle range of the field-specific distribution. With regard to the associated countries, Hungary can be added to this first group, too. Both in Neurosciences and Immunology a considerable number of Hungarian public research organizations could be identified through matching inventors with the SCI.

A second group with a somewhat smaller average gain of 26 percent comprises Germany, Italy and the UK. These countries constantly belong to the upper middle range of the institutional distribution. The results are not surprising for Germany and Italy, where universities had little incentives to patent for a long time, as property rights were entitled to university professors and not to research institutions. In Germany, this situation might change in the future, because property right legislation was amended in 2002. Hence, as German universities can be expected to exploit patents commercially more often, they will appear as patent applicants. Consequently, the share of additional non-profit institutions introduced through the SCI will drop.

Three smaller countries belong to a third group, whose average increase in additional non-profit institutions of about 16 percent is significantly lower than that of the second group: Switzerland, the Netherlands and Belgium. However, it appears still worthwhile for this group to apply the methodology of this study in future studies. In these countries, major universities adopted an active patent policy and already appear as formal applicants.

For the fourth group, the matching procedure had only a limited effect. France, Luxembourg and Spain yield merely 9 percent more non-profit institutions in their final sample compared to the outset of the study. In these countries, major non-university research institutions play a considerable role and hold a substantial number of patents. They have close relations to universities, so that also inventions with university origin have non-profit applicants.

[15] Except for Bioinformatics suffering from substantial data problems.

Figure 2.4-3: For-profit and non-profit institutions at the country level before cleaning/matching in Genetics/ heredity

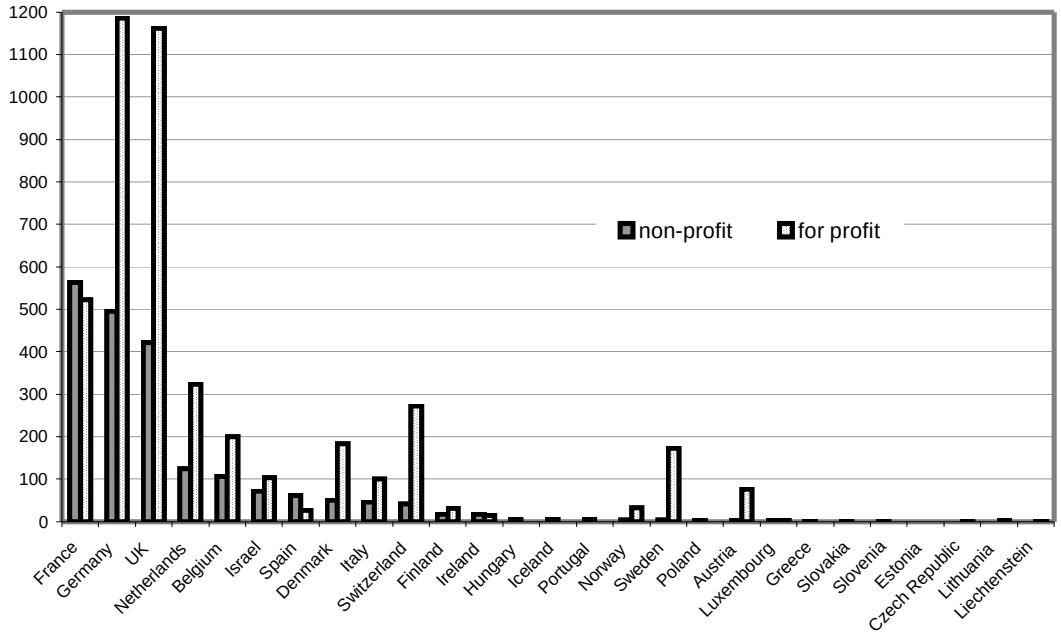


Figure 2.4-4: For-profit and non-profit institutions at the country level after cleaning/matching in Genetics/ heredity

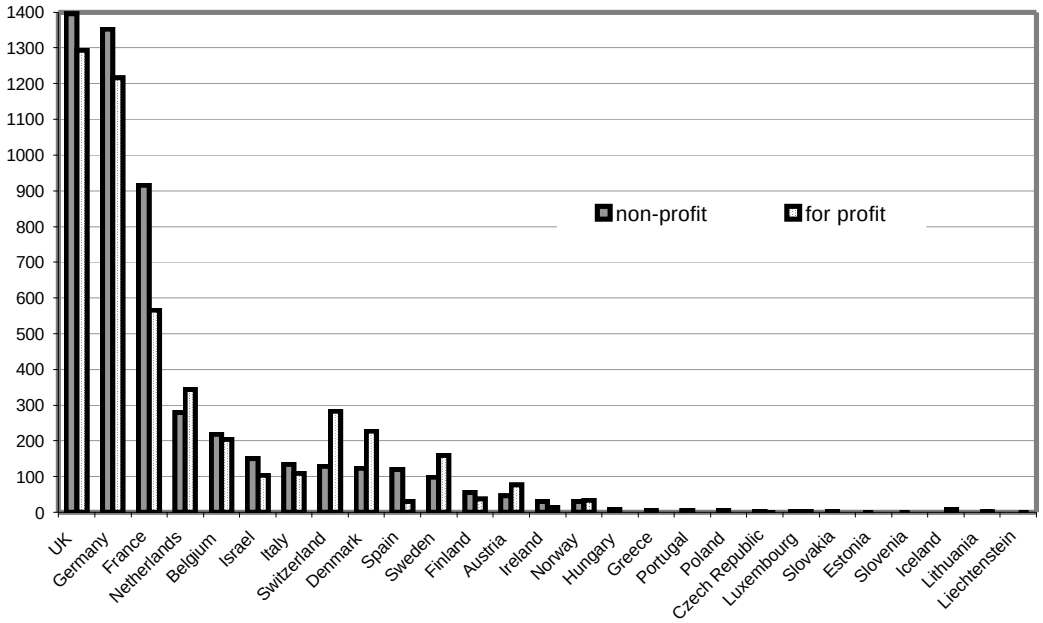
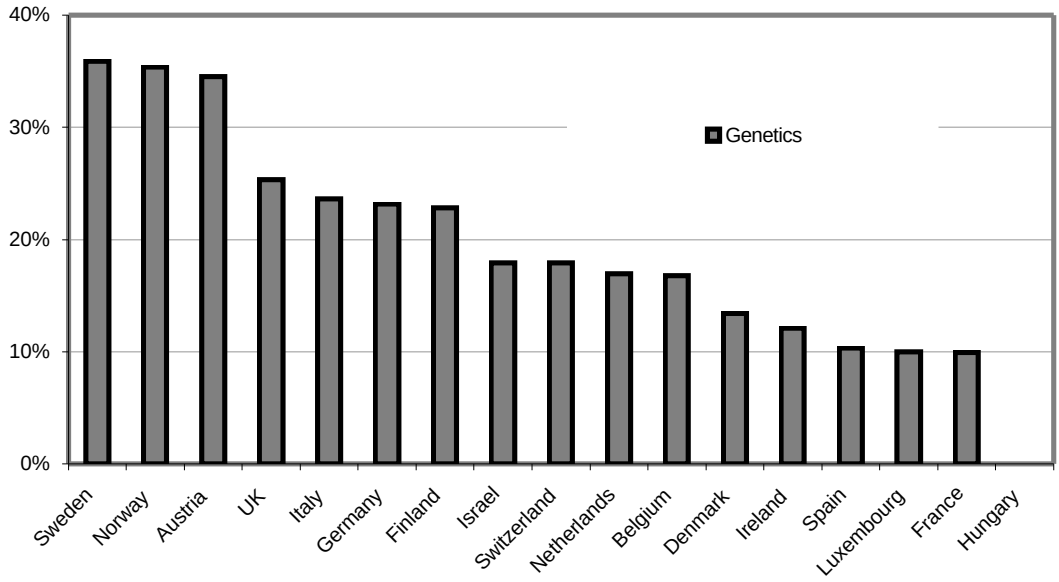


Figure 2.4-5: Additional share of non-profit institutions at the country level after cleaning/matching in Genetics/ heredity



2.4.4 Institutional analysis of the field Genetics/heredity

The preceding section has argued that non-profit institutions contribute substantially to the production of technological knowledge. The match of inventors with institutional information from SCI has impressively shown that the role of non-university research centers and universities is not solely the production of scientific papers in science journals but also of knowledge relevant for application of patents. Mapping excellence in science and technology would, therefore, be short-sighted if it took only the bibliometric, publication-linked indicators into account. The activity in patenting has to be regarded as further indicator of research excellence, namely application-oriented knowledge production.

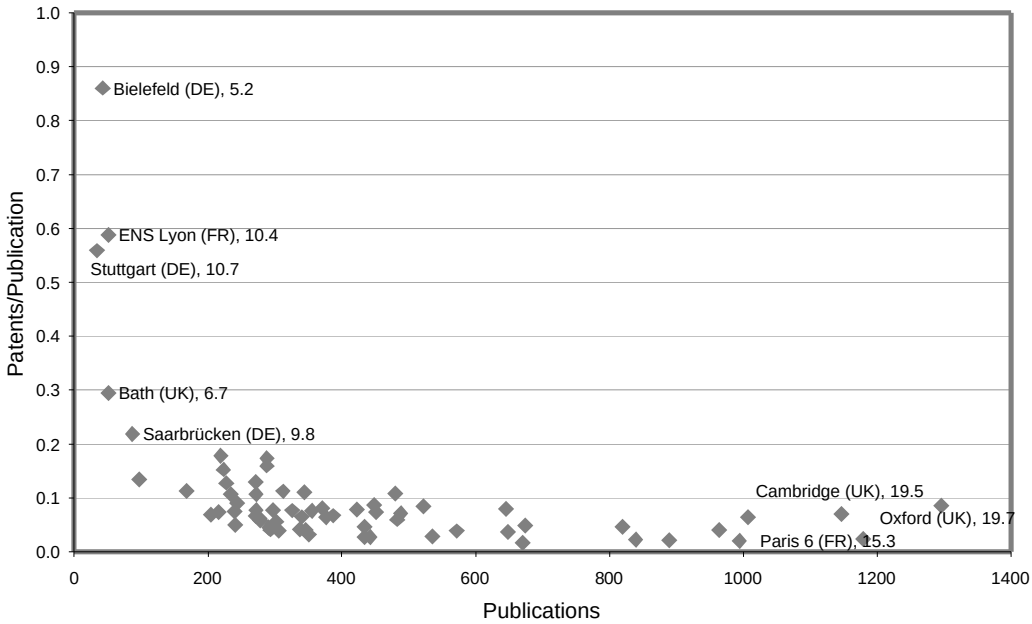
In order to demonstrate the type of non-profit institutions that would be missed, if excellent institutions were exclusively identified and examined by bibliometric indicators, a more detailed analysis of the institutional profile of this type of institutions is needed. For illustrating this aspect, universities[16] in the field of Genetics/heredity, that contributed at least 10 patents in the period of 1996-2000, were considered as to their patent and publication activity.

Figure 2.4-6 shows the result of such an analysis. On the vertical axis, the ratio of patents per publication is depicted, while on the horizontal axis, the absolute number of publications within 1996-2000 is computed. First, as can be easily seen, bibliometrically high performing universities such as Oxford, Cambridge or Paris 6 have an output of publications far above average (mean publication output: 425; median publication output: 340), but a moderate ratio of patents per publication. For instance, the University of Oxford is the most active patenting university in absolute terms. But their 111 patent applications have to be referred to their 1296

[16] As to firms, the publication activity proved to be quite erratic and not useful for systematic analysis. The missions of scientific non-university organisations are very heterogeneous, so that an appropriate interpretation of the patent and publication pattern needs further background knowledge.

publications, leading to a patent to publication ratio of 0.09. So the application-orientation of the University of Oxford has to be considered weak. The universities Oxford, Cambridge, or Paris 6 would be identified in any case by a bibliometric analysis, and they certainly belong to the most active universities in terms of publications. They also belong to the most visible universities as indicated by the high citation impact numbers in Figure 2.4-6.

Figure 2.4-6: Patent and publication production by European universities in Genetics/heredity, including citation impact coefficients[17]



Second, on the other end of the distribution, universities, such as Bielefeld, Lyon, Stuttgart, Bath or Saarbrücken, show up with a comparatively small publication output, but high relative patent activity indicated through an above average patent/publication ratio. Bielefeld produces more than 8 patents per 10 publications, Stuttgart and Lyon about 5. Furthermore, these universities are less visible in terms of citation impact than Oxford, Cambridge and Paris. Although they are still embedded into the wider scientific community (otherwise their citation impact would tend towards zero), their institutional profile appears to be distinctly application-oriented. The data provide strong evidence for the thesis that scientific institutions generally specialize either on basic or on applied research and that only few are able to be excellent in both dimensions of performance.

2.5 Conclusion of the patent analysis

Patents prove to be a powerful tool for identifying industrial enterprises actively engaged in specific fields. As companies do not publish in journals to a large extent, patents are the easiest way for showing up excellent research centers in an industrial context. However, it has

[17] The University of London is an "outlier" and is not depicted, for its publication output is so vast (three times higher than Oxford: 3964 publications) that it would lie far beyond the high performing universities of Oxford, Cambridge and Paris, thus rendering a sensible presentation of the results difficult. For London university we found 210 patents.

to be taken into account, that enterprises may follow different patent policies leading to different propensities to patent. Therefore, the number of patent applications cannot be taken as a strict ranking criterion. Nevertheless, companies actively engaged in excellent research will always apply for a relevant number of patents and get visible by patent indicators. For a more detailed analysis, it will be helpful to collect additional data such as R&D budgets for top companies identified by patents. Furthermore, the company lists should be differentiated according to large and small enterprises so that the activities of excellent small companies can be highlighted and do not disappear in comparison with large enterprises.

With regard to public non-profit institutions, in particular universities and non university research institutes, patents proved to be more relevant than expected. In all areas[18], at least 50 percent of all patent applications originate from non-profit institutes. This finding brings in a completely new perspective on the role of publicly funded research in the generation of technology, at least in research-intensive areas such as these life sciences fields. For up to now, analyses for all technologies found average shares of university patents of 4 percent (Becher et al. 1996: 42) and additional 3 percent of other public research institutions (Schmoch et al. 2000: 24ff). Against this background, the major impact of public research on the generation of technology was primarily attributed to indirect mechanisms such as the information on new trends of science or the transfer of new concepts through publications or informal meetings (cf., e.g., Salter/Martin 2001). The data of this report show that public research is directly engaged in technology generation to a considerable extent. The progress of technology development would be insufficiently recorded, if exclusively companies were considered. This finding justifies the considerable additional effort in this project linked to the identification of patents originating from public research.

A more detailed analysis of public research institutes shows that excellent performance is not exclusively reflected in high numbers of publications and referring citations. Some relevant field studies came to the result that within research areas, most institutes specialize either in basic research or in applied research (Schmoch 2003: 251). There are only few institutes which manage to be excellent in both dimensions (Laredo 1999). Therefore, several public research institutes can be identified as excellent by patent indicators which would have been overlooked by a pure analysis of publications. Due to the high patent activity of public institutions, the absolute number of patent applications on the institutional level is sufficiently high to get statistically relevant data.

With regard to public institutions, the ratio of patents to publication referring to an institution proved to be a good indicator to characterize its orientation towards applied research independently of its size. The initial thesis of largely varying patent cultures in different countries and consequently largely varying propensities to patent could not be confirmed, so that this indicator is even suitable for international comparisons. The major methodological problem linked to this indicator is the substantial effort which has to be invested in the institutional cleaning/matching of the patent and in particular the publication data, and the appropriate institutional match of the publication and patent data. As to industrial enterprises, the quotient of patents and publications is less meaningful, as their publication activities are too erratic.

For further analyses, the identification of public research institutions can be made with less work expenditure, as the extraction of institutional data from the inventor field primarily yields redundant information to the SCI data. So all working steps linked to the institutional inventor data can be left out. Furthermore, a recent OECD study came to the conclusion that at present, public research organizations are not active patent applicants, but that in many

[18] With exception of Bioinformatics suffering from substantial data problems.

OECD countries, the policies on ownership of intellectual property are being redefined in two ways. First, in several European OECD countries reviewed or modified ownership rules regarding inventions now provide incentives for public research institutions to apply for patents instead of leaving ownership rules in the hands of individual employees (mainly university professors). Apart from legal changes, there is, secondly, an increasing awareness of and support for technology transfer, especially among public research institutions (OECD 2003). As these changes have made the way free for public research institutions to pursue patenting more actively, they will appear in the applicant field of patent documents more often. Thus it will be possible, in a midterm perspective, to assess the patent activities of public research organization to a large extent by the direct exploitation of applicant lists.

In the concrete work, the appropriate definition of patent samples is decisive for all subsequent steps. Search strategies should be conceived in close co-operation of field and patent experts, and cannot be considered a mere side-effect of publication-oriented searches. The setting of this project with some meetings with broad attendance and distributed experts is not sufficiently effective. Instead of this, workshop-like face-to-face meetings with a small number of experts are suggested where a suitable delineation of the fields can be discussed and suggestions for search strings be directly tested and modified.

If the official databases of the EPO are used, it has to be taken into account that the inventor names with special letters such as umlauts are often problematic. Either the letters are completely missing or they are spelled in different versions. In the releases offered by professional hosts, these spelling problems are solved, but the download of a relevant number of records from online databases is very expensive. It seems to be more suitable to use the EPO version and apply specific computer programs for correcting these errors.

The major methodological problem of the match of author and inventor names is the fact that the SCI records only the initial of the first name and not the full first name. Therefore, the search for names in the complete database leads to a considerable number of ambiguous matches. The best approach is, according to the experiences of this project, the introduction of the following criteria:

- Match of surname and initial of first name,
- Match of countries of authors and inventors,
- Inclusion only of authors of the relevant field,
- Match of publication periods of patents and publication.

If the match is enlarged to cities in addition to countries, the output is more precise, but the number of hits is significantly reduced, as many inventors/authors do not live in the city where their institute is located. In the SCI, the match must exclusively refer to the first author, as only in these cases, he or she definitively works at the first institution. In addition, publications with only one institution can be included for all authors mentioned.

In some patent documents, the inventors have not indicated their private, but their institutional address. As these cases have no special codes, they have to be identified manually. This process is quite labor-intensive and leads to fewer institutions than the semi-automatic matching procedure in SCI. The majority of the inventor-based institutional information is redundant with the SCI information, so that for further analyses, the work should be limited to the SCI. In the context of the present project, the inventor-based data showed a high accuracy of the institutions identified by the inventor-author match in the SCI.

In any case, the results of the SCI match have to be checked in more detail with regard to their validity. Sources of uncertainty are multiple first names, which are not recorded in a systematic way neither in patents nor in SCI records, or multiple surnames, for instance, in the case of Spanish authors/inventors.

After the elimination of redundant information and mismatches, there is still the problem of multiple versions of institution names within the data derived from patents and between the data from patents and SCI publications. Therefore, the institutional lists have to be cleaned, a task which cannot be carried out by automatic tools; again, a manual check is necessary. For this study, this was done for some selected countries and fields by national experts. The rest of the cleaning/matching, being clearly the greater part of the data set, was made by staff of Fraunhofer ISI. The support of national experts proved to be very helpful, as the institutional cleaning/matching requires specific knowledge on the structure of national innovation system. Nevertheless, it was necessary to check the lists of the national experts, as they still contained some errors by oversight. This is due to the outcome of this project that the number of institutions is much higher than originally expected. With regard to large countries, the lists comprised some hundreds of institutions instead of some dozens. In any case, the enormous number of relevant institutions has to be taken into account as to cleaning/matching and matching processes.

This problem is even more severe for other approaches of mapping excellent institutions such as surveys, as all institutional information has to be collected through manuals, scientific boards and associations, experts etc. Therefore, it seems to be appropriate to start patent and publication statistics for defining data sets for refined additional analysis in order to minimize the costs of identifying relevant actors.

If the number of patents is combined with publications, an additional matching and cleaning/matching step is needed, as the representation of institutions in the SCI and in the patent lists is often different. This step is very labor-intensive, as the institutional lists derived from publications are much longer than the patent lists.

The major output of the patent analyses is formed by the institutional lists by field, country and type of institution. The separation of public and private, or not-for-profit and for-profit institutions is helpful, as the propensity to patent of public research institutes and industrial enterprises is not comparable. Both types of institutions have different cultures and different logics as to patent applications. The separation was achieved by specific codes, assigned concomitantly at the institutional cleaning/matching process.

All in all, the patent analyses proved to be very useful not only with regard to private companies, but also to public research institutes. Patents should be considered, as they reflect in an appropriate way the dimension of orientation on application and technological exploitation. Scientific excellence is important and a relevant pre-requisite for technological competitiveness. But the transfer of knowledge to technological exploitation should be analyzed as additional dimension of excellence.

3 Publication analyses

As a separate part of this study we performed an extensive bibliometric analysis of the publications data collected for each of the five fields. As mentioned before, the procedure of data collection was designed in such a way that it could be adopted for any science field within a limited period of time.

3.1 Methodology

The results we compiled in this study are based on the experience of decades of research of developing bibliometric indicators for evaluative purposes. At CWTS these indicators have been developed in two sections:

1. Performance analyses of scientific actors, mainly based on production and impact;
2. Cognitive mapping of science, mainly concerned with self-organizing structures of science.

In this study we combine the achievements of both sections and perform an integrated analysis of production, impact and cognitive orientation.

In this study we apply a so-called top-down approach. This means that we start from the entire field, and identify actors with a particular performance on the basis of the addresses in publication data. Hence, the performance of the (candidate) centers is defined by the publications identified as belonging to a certain address. This means that the actual output of a center may be larger than the output we assign, due to spelling errors, name variants etc. As a result, the quality of the address data is of major importance. We may never be able to make a perfect match, due to the unexpected errors, changing organizational systems and missing address data, but the approach should be able to provide indications of performance. The extensive experience we have with the publication address data learned us that at country, city and organization (university, companies, …) level, the address data of the ISI databases are a reasonable shape. This means that a cleaning effort within a country is possible within reasonable boundaries. At the departmental level, however, this makes no sense. There are at least four reasons for that:

1. too many variants of one name;
2. unstable organizational structure;
3. Different systems/structures in countries;
4. Missing information.

As a result, we take the main organization as the primary level for our analyses. The centers of excellence are identified at that level. However, the on-line tool on the internet page of this project, allows an analysis on departmental level.

An added aspect to the traditional results is the geographical mapping of actors. But this aspect is primarily introduced to present the results. It doesn't add any information with respect to the performance of scientific actors.

3.1.1 Bibliometric indicators

For the identification of centers of excellence, we provide a list of indicators that should cover several aspects. We stress that the possibility to identify centers of excellence from different

perspectives is vital. There is and probably never will be one definition of what excellence is, let alone that we can measure it with one indicator. The aspects of performance we attempt to measure by one or more of the indicators are:

- scientific activity
- scientific visibility
- scientific impact
- technological activity/ visibility

The indicators we provide in this study are:

1. Number of publications in 1996-2002
2. Total number of citations received, author self-citations excluded
3. Growth of activity during the entire period
4. CPP, the number of citations received per publication
5. CPPf, the average number of citations per publication normalized by traditional science areas
6. Number of publications contributed to the top-10% of most highly cited publications (multiple perspectives)
7. Number of patent applications in the entire period of time
8. Average number of patents per publication.

The activity and visibility indicator designated by the number of publications is the best-known, but also an often criticized indicator. Of course, the absolute number of publications per organization depends a great deal on the number of researchers available and on many other factors. In most bibliometric studies this indicator is used only in combination with other indicators. The same critical remarks could be made about the absolute number of citations (visibility or impact). Therefore, the CPP indicator seems much more valuable, because in that case a normalization of the number of received citations (visibility or impact) is made by the number of publications (production) as a possible measure for size.

A more dedicated indicator is the CPPf. This indicators normalizes the impact (CPP) by the field in which the research is published. It is a well-known fact that the probability of being cited in basic science is much higher than in applied science. Therefore, the impact (CPP) an actor achieves, will depend heavily on the orientation (mainly, basic vs. applied) of its research. The normalization of the CPP is based on the average impact in science fields, as defined by journal packages.

The indicator measuring the contribution within the top-10% of most highly cited publications is relatively new and correlates highly with the absolute number of publications. An interesting aspect about this indicator is that it can be measured from different perspectives. The list of most highly cited publications of one particular country differs, of course, from the one of the EU and associated states or world-wide.

The technology indicators are described in more detail in Chapter 2.

3.1.2 Cognitive mapping

The main objective of introducing the cognitive mapping into this study is to allow a more detailed view on the activity and impact (the performance) of the identified actors (the centers

of excellence). The reason for this is obvious. The science fields as defined in this study, but probably as defined in all studies like this are too broad to determine why an actor is identified as being excellent. In order to assure the results to be statistically sound, significant, we need these fields to be 'bigger' than some hundreds of publications. The cognitive structure provides the means to disaggregate without losing the substance of the entire field.

The cognitive map is a two-dimensional representation of a field, represented by the collection of publication data. From these publications, we extract noun phrases from titles and abstracts. From this huge list of noun phrases we select the ones to be used for a co-occurrence analysis on the basis of bibliometric distribution, syntactic features and (semantic) content.. The selected noun phrases are identified as keywords. The co-occurrence analysis calculates the number of times that two keywords occur in the same publication title or abstract and creates clusters of these keywords on the basis of this information.

The clusters of keywords designate sub-domains in the field. Using the clusters of keywords, we assign publications to sub-domains. Thus, sub-domains are in fact, subsets of publications from the entire collection, the field. As publications may be assigned to more than one sub-domain, we can generate a co-co-occurrence matrix of sub-domains. The cells in this N*N matrix (in which N designates the number of sub-domains), contain the number of publications overlapping in two of the N sub-domains. This matrix is the input for Multidimensional scaling (MDS). This technique puts the N elements in a two-dimensional space in such a way that sub-domains with a similar orientation in relation to all other the sub-domains, are in each other vicinity, whereas sub-domains with a different orientation are distant from each other. This two-dimensional representation is the cognitive map of the field.

In the map, sub-domains are represented by circles. The surface of these circles corresponds to their relative size (numbers of publications represented).

The cognitive map interface allows the user to indicate the contribution of an actor (organization) within each of the sub-domains. We use a color-coding to indicate the relative contribution as compared top its own average. In this way, we provide in one view the cognitive orientation, of this organization within the landscape of a particular field. Thus, we provide a tool to explore the activity and impact in more detail of any actor. A more detailed description of the methodology is in Appendix B.

3.1.3 Geographical mapping

As mentioned before, the geographical representations are invoked as a way of presenting the results of the bibliometric analyses to identify centers of excellence. The geographical information is used to present the distribution over a country of centers with a particular performance. It contributes by no means to the identification as such.

By the information we have from the organizations we evaluate in our analyses, it appeared not feasible to assign geographical information (longitude and latitude) to their exact address, which could be a postal code in combination with a number in the street. We don't have this information to our disposal from the publication data we use or we don't have unambiguous information. Therefore, the only link between the publication address data and the geographical information was the name of the city.

As input we used a database with more than three million records of cities all over the world, with the longitude and latitude information. We had to link the city names to the ones in this geographical database. We encountered a number of problems to make this link:

1. The variants of one and the same city name. The way in which city names appear in the ISI databases do not always match the variant in the geographical database (e.g.,

Göttingen, Goettingen and Gottingen). Especially in Israel this appeared to cause difficulties.

2. In some cases we found more cities with the same name in one country. In the geographical data they appear with different geocodes. In the case of Paris, this yielded a quite severe 'dislocation' of organizations from the capital somewhere in the South-East of France.
3. In some cases, region names were given instead of city names in the publication data.
4. We found misspellings of city names in the publication database.
5. In some cases there was no entry for a city in the geographical database.

We developed a strategy to be able to cope with the most obvious problems as mentioned above. This strategy consists of the following steps:

1. For every city name in both databases we created a link variant in which special characters are translated into normal ones (Ö => O, Ü => U, Ñ => N, etc.);
2. In this link variant, we translated UE, OE and AE into U, O and A;
3. In case of homonym city names within a country (geodata) we selected the one with the largest population (an additional information element)
4. We created a list of organizations with a city name that did not match the geodata after the previous steps and invoked EC country experts to change the city name to proper one. Thus we got rid of misspellings, and regions instead of cities. In some cases the experts would provide the name of a larger city in the vicinity of the missing one.
5. If we couldn't find the geographical codes, we looked them up in the Times Atlas. This was only done for the organizations with a substantial output (N>5).

With this strategy we were able to match 99.5% of the organizations involved in this study. This missing 0.5% is not available in the geographical representation, but is integrated in all other results reports.

We compiled a huge database with all bibliometric information per organization and the geographical codes. This database was used in the GIS module of SAS to compile country maps in which we indicate the cities with centers of excellence.

In the web-interface, we developed for this project, we allow users to select their country of interest (EC member states and associated states). In a second step they can determine their own indicators to be used to identify centers of excellence and the thresholds they want to use plus the field they want to explore. With these criteria they get a map in which the cities are drawn containing the organizations that meet their criteria.

From thereon, the user can list the names of those organizations by city and run a full bibliometric profile of that organization.

3.2 Feedback

However objective we claim bibliometric studies to be, we will always need expert input to verify or validate the results. First of all to determine whether we are measuring what we think we are measuring, second to clean misspellings and other 'errors', and finally to determine how the results relate to the actual situation.

In this study we had experts input at several stages during the entire process from data collection to final results. In most stages in which expert input was required, we provided electronic forms through the project webpage.

3.2.1 Field delineation

The expert input of this stage was of crucial importance for the reliability and validity of the results. We are very much aware of the fact that this is the stage that should consume the most time in the entire project.

The design of the expert input for field delineation is such that it could be applied to more than the fields involved in this project. In the back of our minds we had the requirement that the methods should be applicable to many more fields.

In main lines, in this delineation procedure we start with a core set of publications of which the experts agree that they represent a trustworthy set of core publications in the field. This may be the articles in a journal or set of journals, this may be the oeuvre of a research group, etc. etc.

These core publications are parsed and we extract noun phrases from titles and abstracts. Based on particular characteristics of the noun phrases we select candidates to be used to expand the selection of publications beyond the primary core set. These lists of candidates were available at the website for experts to do their suggestions. Furthermore, they could make new suggestions to us to expand the selection. A flow chart of this entire procedure is presented in Appendix C.

In order to use the input received in the preparatory studies of this project, we incorporated their suggested search terms as input at this stage. We did not use the input from the preparatory study where the experts reviewed the publications on a paper-to-paper basis. The reason for this is that we wanted to develop a generic tool to deal with the difficult issue of field delineation. The approach of checking individual publications does not fit into this approach.

We used the following search strategies to collect publications for the 5 different fields in this study:

- Neurosciences: the Neuroscience Citation Index (a dedicated ISI database)
- Immunology: a list of journals from the field Immunology
- Genetics: a list of journals from the field Genetics & Heredity
- Bioinformatics: a search string extracted by one of the experts from a general description of this field

For each field we compiled a feedback form with a list of candidates to be used. Experts could check the candidates that they considered useful to expand the collection. Furthermore, they could add search terms at the end of the form.

The feedback at this stage was disappointing. For genetics we received one form with suggested terms as selected from the form we provided. For immunology we received no feedback form but by discussing the procedure, two experts were able to suggest terms they would use to delineate this field. For Bioinformatics we received no input after the first suggestion of search terms. One of the experts involved was willing to get involved but we had to close the deadline for this phase of the project in September and he was not available before that time.

In addition to the suggested terms we added the terms suggested in the preparatory study for genetics and immunology. The expert had the opportunity to suggest search terms in the period July and August of 2002.

The final search strategies used in this study for the different fields are listed in Appendix D.

The data are retrieved from ISI's current contents database. Subsequently they are matched to the CWTS data system. The numbers of publications retrieved per field are listed below.

Table 3.1 Numbers of publications in 4 LS fields

Total number of retrieved publications and the number matched with the CWTS system

field	type	1996	1997	1998	1999	2000	2001
bi	Total	4160	4551	4720	4782	5379	6024
	matched	4032	4419	4560	4445	4957	5106
ge	Total	33413	36006	38318	40278	42683	43756
	matched	32082	34701	36793	36737	38960	36612
im	Total	78302	79268	79759	81156	82890	81596
	matched	72777	73871	74110	71902	73529	66698
ns	Total	82396	86174	92701	99334	106549	79180
	matched	82392	86174	92697	99334	100636	79180

3.2.2 Address data

The procedure we adopted for cleaning the address data in the project was designed in close relation with the Commission and members of the High Level Group. This procedure should be designed is such a way that the cleaning itself could be done by any person acquainted with the national science system and able to use a computer and basic desktop applications, e.g., MS-Word, Excel, Access.

By using the suggested search terms by the experts, we collected the matching publication data from the CWTS data system. For each field, we compiled a list of addresses per publication. Such an address string consists of a country, a city, a main organization (e.g., university, hospital of company name) and a department. Then by field, we compiled a list of addresses and the number if occurrences in the collected publication data.

CWTS preprocessed these address lists in the following way:

- a first cleaning on the level of country was made;
- a first cleaning of city names was performed;
- a first cleaning of main organization names was performed.

These cleaning steps are performed according to the in-house cleaning procedures. This is a hierarchical procedure, going from country names to city names to organizations in distinct steps. The stage of cleaning main organizations is the most difficult one, because different

countries and even organizations have different organizational structures. It is often difficult to decide which level is to be considered the main organization (e.g., research organizations CNRS, CSIC, and Max Planck.) Generally speaking we take the hierarchical structure as represented in 'The World of Learning' as guidance.

Within a field we compiled a list as mentioned for each of the 32 countries involved and uploaded them in an ASCI-file (CSV format) on the web-page of the project. These files were generated at two different levels of aggregation: main organization and department level. In these files we put the following information:

- Country
- City
- Organization
- (Department)
- Frequency
- ID (for internal use)

The commission asked the HLG members to provide one national expert in the field of Neurosciences, to clean the address data. They were asked to correct the address data at the organizational level. Thus they were able to clean misspelling and to unify name variants. They were told explicitly not to change the ID. This information was to be used to feed the corrected data back into our system.

Below is a list of the countries that put effort in the cleaning process. In total we received cleaned data from 19 countries. In only one case we were not able to process the data back into our system because the expert changed the ID information.

Table 3.2 Address cleaning effort by country

By field and level of aggregation (1: organization, 2: department)

	neuro	
Country	**1**	**2**
Austria		1
Belgium		
Czech republic		
Denmark		
Estonia		1
Finland		1
Germany		1
Greece		1
Hungary		1
Iceland		1
Israel		1
Lithuania		1
Luxembourg		1
Netherlands	1	1
Norway		
Slovenia		1
Spain	1	
Sweden	1	
Switzerland		1

The results show that most countries seem to have cleaned at the departmental level, but that is not true. In almost all these cases, the expert used the data at department level, but cleaned the data at the organizational level. They left the department information untouched.

As far as the procedure is concerned, we found some difficulties. First of all, the CSV format of the data files. This format was chosen because it should be compatible with standard data processing desktop applications such as MS-excel and MS-access. However, we experienced that in some countries the integration of this format required more knowledge about data processing and spreadsheet programs than expected. This issue could be tackled by introducing a telephone helpdesk.

A second issue is the fact that a lot of the expert used the department level data, but in fact only cleaned at organizational level. The drawback of this is the integration of the cleaned data of Neuroscience into the other life science fields. Because we could only use the file with data at the department level, we could only clean those data in the other fields with the similar department data. In the cases where in the other fields, other department names were sued, we can't clean the data at organizational level. If the experts would have used the file at organizational level, this link could have been made.

Finally, we found that the fact that they used these CSV files, no protection was possible. We could not prevent them from changing the ID, and thus corrupting the data.

In sum, the use of the CSV files caused some problems that could have been prevented if we would have provided an on-line cleaning form. In that case, we would have more control over the actions of the experts.

With respect to the added value of the cleaning process in general, we conclude that this is marginal, considering the table below.

Table 3.3 Number of organizations per country in top-100 most publishing within EU member and associated states using un-cleaned (u) and cleaned (c) addresses for 4 LS fields

	Bioinform.		**Genetics**		**Immunol**		**Neuro**	
country	**u**	**c**	**u**	**c**	**u**	**c**	**u**	**c**
AUSTRIA	2	1	1	1	2	2	3	2
BELGIUM	3	3	5	5	3	3	3	3
CZECH REPUBLIC								
DENMARK	3	3	2	2	2	2	2	1
FINLAND	2	2	5	4	2	2	3	3
FRANCE	10	8	10	9	10	10	8	8
GERMANY	29	31	24	25	24	26	27	27
GREECE					1	1		
HUNGARY								
IRELAND								
ISRAEL	3	3	2	2	3	3	3	3
ITALY	7	7	9	9	11	11	11	11
NETHERL.	8	8	8	8	8	8	13	13
NORWAY	1	1	1	1	1	1	1	1
POLAND								
SPAIN	2	2	2	1	3	3	2	2
SWEDEN	7	8	5	7	8	7	4	6
SWITZERL.	6	6	5	5	6	5	6	6

	Bioinform.		Genetics		Immunol		Neuro	
country	**u**	**c**	**u**	**c**	**u**	**c**	**u**	**c**
UK	17	17	21	21	16	16	14	14

In this table, we calculated the contribution of a country in the top 100 most publishing organizations within the EU15 and associated states, before and after cleaning the addresses. The results show that in most cases nothing changes. The possible changes that take place can be characterized as indicated below.

No-cleaning with a decreasing number of contributions (ND)

No-cleaning with an increasing number of contributions (NI)

Cleaning with a decreasing number of contributions (CD)

Cleaning with an increasing number of contributions (CI)

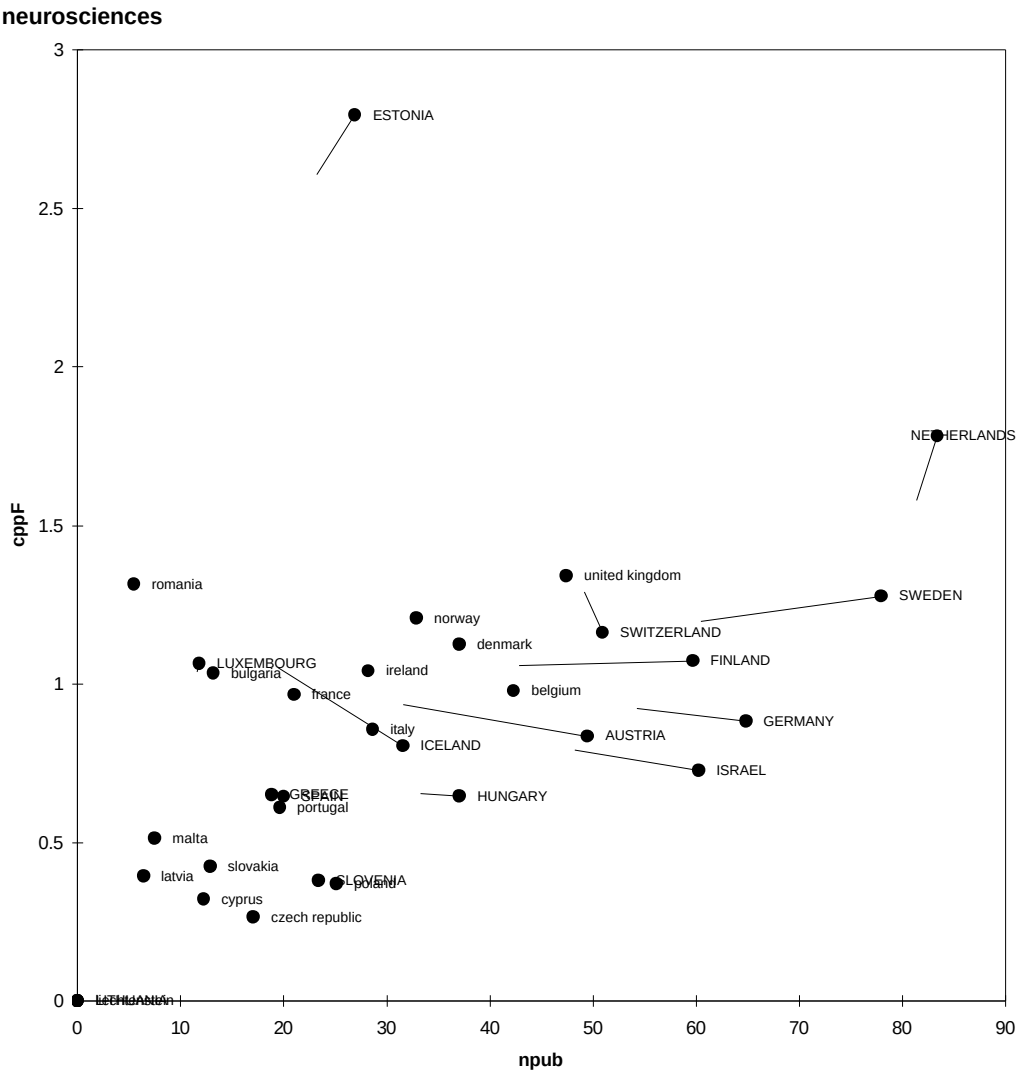
In the case of ND, the cleaning effort of other countries put one of their organizations into the top-100 (CI), lowering the position of others. In the case of NC, the cleaning effort of other countries probably merged two variants of the same organization already in the top-100 to one name (thus CD), allowing ‘newcomers’ in the top-100. From these results it is difficult to say what the effect of the cleaning effort is. Obviously, the effect of the cleaning process on a top-100 of most publishing organizations have a very direct effect. As the number of contributing organizations can easily be reduced, the results in this table don’t seem to indicate any positive or negative effect. If we explore the effect on a derived indicator, e.g., impact, we can see a more sophisticated effect.

Table 3.4 Number of organizations per country in top-100 highest impact within EU member and associated states using un-cleaned (u) and cleaned (c) addresses for 4 LS fields

	Bio		**ge**		**im**		**neuro**	
Country	***u***	***c***	***u***	***c***	***u***	***c***	***U***	***c***
AUSTRIA	1	1	2	2	3		1	1
BELGIUM	1	1	3	4	4	5	2	2
BULGARIA							1	1
CZECH REP	1	1			1	1	1	1
DENMARK	3	3			1	1	2	2
ESTONIA							1	1
FINLAND	2		11	10			1	
FRANCE	20	22	24	25	20	20	16	16
GERMANY	11	9	11	11	14	13	7	8
GREECE	1	1	1	1		1	1	1
HUNGARY	1	1	1	1			1	1
IRELAND	3	3	2	2				
ISRAEL			3	2	1	1	1	
ITALY	4	4	8	8	11	11	8	7
LITHUANIA					1	1		
NETHERL	4	4	3	3	3	3	14	18
NORWAY	3	3	3	3	1	1	1	1
POLAND	1	1						
PORTUGAL					1	1		
ROMANIA					1	1	1	1
SLOVAKIA	1	1						
SPAIN	3	3	2	3	8	8	5	5
SWEDEN	4	4	2	2	5	6	3	3
SWITZERL	13	13	2	1	2	3	4	3
UK	23	25	22	22	23	23	29	28

In more detail per country, we plotted the effect of the cleaning in Neuroscience. We calculated the average number of publications per organization as well as the average CPPf. In the chart below, we plotted the average output (X-axis) and the average CPPf (Y-axis). Only organizations with more than one publication were considered. If either of the two indicators changed, this indicated by a tail at a data point, the dot being the results for the cleaned data.

Figure 3.1 Average output vs. average impact of organizations within EU&AS countries, plus effect of cleaning effort



In most cases the situation remains exactly the same. Obviously, this is the case for countries with no cleaning effort. Furthermore, we see that most cleaned countries show a positive trend with regard to the average output. That is also what we would expect, because in most cases the cleaning of addresses means unification of variants. Thus, the number of publications remains the same, but the number of organizations decreases.

For the average impact we see that some countries benefit from the cleaning effort (Netherlands), while others seem to suffer from it (Iceland). Generally speaking however, these differences don't seem to make much difference.

3.2.3 Rankings per country

The feedback form we provided for the ranking tool, was available via the rankings per country. User who generated a ranking for a country could give their comments on:

- Organizations unexpectedly in the list;
- Organizations unexpectedly not in the list.

We assume that the user-experts were able to evaluate a national list. And by doing this, they indirectly commented the international list. However, we claim that their comments on an international list would be beyond their scope.

We did not receive any comment on the rankings through the online forms we supplied.

3.3 Results

3.3.1 Publication activity and impact (research performance)

As mentioned many times before, the issue of excellence is far too complex to decide from one single perspective what are the most excellent institutes in the EU15 and Associated states. The main problem is the fact that the perception of excellence is variable. Every user posing the question 'what are the most excellent groups?' may have a different view on the definition of this excellence. Furthermore, the thresholds a user sets to discern excellent or not, maybe be too high in certain situations (e.g., national science systems, phase of development) to be used in general or too low in other. As this pilot project was setup to test the validity and utility of the bibliometric method in general, we chose to design an interactive tool, a dynamic table, to be able to extract results from different perspectives.

In this section of the report we will illustrate the potentials of the tool with illustrative results, rather than to present final results.

As an example, we take the field of neuroscience. In the table below, we listed the most active organizations world-wide.

Table 3.5 Top ten most publishing organizations (1996-2001) world-wide

Rnk	Country	City	Organization	P	CX	P10	CPP	CPPf	PS	PN
1	UNITED KINGDOM	LONDON	UNIV LONDON	16207	57230	771	3.53	1.51	21.8%	69.4%
2	USA	BOSTON	HARVARD UNIV	8596	54017	767	6.28	2.36	16.7%	67.1%
3	USA	BETHESDA	NATL INST HLTH	7563	38671	558	5.11	1.76	20.3%	65.3%
4	USA	PHILADELPHIA	UNIV PENNSYLVANIA	5288	23408	324	4.43	1.81	20.8%	67.5%
5	USA	LOS ANGELES	UNIV CALIF	5138	19655	264	3.83	1.61	20.2%	69.6%
6	CANADA	TORONTO	UNIV TORONTO	4875	17968	221	3.69	1.60	19.7%	68.8%
7	USA	BALTIMORE	JOHNS HOPKINS UNIV	4483	24564	349	5.48	2.11	18.6%	67.5%
8	USA	PITTSBURGH	UNIV PITTSBURGH	4282	14662	189	3.42	1.53	21.2%	68.1%
9	USA	SAN FRANCISCO	UNIV CALIF	4252	27816	399	6.54	2.25	16.7%	65.1%
10	SWEDEN	STOCKHOLM	KAROLINSKA INST	4200	14844	175	3.53	1.26	25.9%	64.2%

Rnk: rank; P: number of publications; Cx: number of citations received, excluding self-cits; P10: contribution to top 10% from same perspective; CPP: citations per publication; CPPf: normalized CPP to field; PS: percentage self-cits; PN: percentage non-cited articles

By changing the perspective to the EU and Associated states, the Univ London and Karolinska are, of course, still on top, but now there is room for 8 others.

Table 3.6 Top ten most publishing organizations (1996-2001) within EU and associated states

Rnk	Country	City	Organization	P	CX	P10	CPP	CPPf	PS	PN
1	UNITED KINGDOM	LONDON	UNIV LONDON	16207	57230	849	3.53	1.51	21.8%	69.4%
2	SWEDEN	STOCKHOLM	KAROLINSKA INST	4200	14844	194	3.53	1.26	25.9%	64.2%
3	UNITED KINGDOM	OXFORD	UNIV OXFORD	3594	13637	206	3.79	1.60	22.6%	68.9%
4	FRANCE	PARIS	UNIV PARIS 06	3269	11779	152	3.60	1.28	22.0%	67.2%
5	UNITED KINGDOM	CAMBRIDGE	UNIV CAMBRIDGE	3211	12725	179	3.96	1.66	22.9%	69.7%
6	ITALY	MILAN	UNIV MILAN	2962	8329	111	2.81	1.05	29.4%	67.2%
7	GERMANY	MUNICH	LM UNIV MUNICH	2844	5991	66	2.11	0.96	28.4%	72.9%
8	GERMANY	TUBINGEN	UNIV TUBINGEN	2537	5497	62	2.17	0.98	31.3%	73.0%
9	NETHERLANDS	UTRECHT	UNIV UTRECHT	2293	5225	54	2.28	0.99	25.9%	67.5%
10	GERMANY	BERLIN	HUMBOLDT UNIV BERLIN	2081	3875	45	1.86	0.88	30.6%	74.5%

If we compile this top ten on the basis of the highest average of citations per publication, it looks completely different from the previous one. The problem here is that smaller subsets (low numbers of publications) are more likely to achieve a high CPP score. The average is not smoothed by large numbers of less cited publications. Therefore, we see here a list of organizations with a low number of publications, but with a high CPP. In fact, it seems that we are not comparing the same thing. It is almost impossible for organizations with such large output as the Univ London and Karolinska to achieve such high CPP.

Table 3.7 Top ten organizations with highest impact (CPP) within EU and associated states

Rnk	Country	City	Organization	P	CX	P10	CPP	CPPf	PS	PN
1	GERMANY	BERLIN	MAX PLANCK INST MOL GENET	5	520	3	104.00	14.31	17.7%	20.0%
2	UNITED KINGDOM	NEWCASTLE UPON TYNE	MRC-NEUROCHEM PATHOL UNIT	6	242	2	40.33	12.59	24.1%	50.0%
3	FRANCE	MARSEILLE	DEV BIOL INST MARSEILLE	5	188	2	37.60	5.64	13.4%	40.0%
4	SWITZERLAND	GENEVA	GENEVA BIOMED RES INST	6	200	3	33.33	4.13	8.3%	50.0%
5	SPAIN	MADRID	CSIC-INST RAMON Y CAJAL	5	145	2	29.00	5.82	11.6%	20.0%
6	FRANCE	LYON	BIOMERIEUX PIERRE FABRE	5	126	2	25.20	2.93	20.8%	20.0%
7	FRANCE	LABEGE	SANOFI SYNTHELABO	21	525	6	25.00	4.12	12.4%	28.6%
8	UNITED	LONDON	PARKINSONS	5	125	3	25.00	4.02	26.5%	40.0%

	KINGDOM		DIS SOC							
9	FRANCE	SOPHIA ANTIPOLIS	CNRS-INST PHARMACOL MOLEC & CELLULAIRE	5	124	3	24.80	2.83	22.0%	0.0%
10	FRANCE	PARIS	CHNO XV XX	6	147	1	24.50	3.57	10.9%	50.0%

In order to be able to make a more 'fair' comparison, we added the possibility to set a threshold of numbers of publications and then to look at the top CPP organizations.

Table 3.8 Top ten organizations with highest impact (CPP) within EU and associated states, organizations with 50 publications or more only

Rnk	**Country**	**City**	**Organization**	**P**	**CX**	**P10**	**CPP**	**CPPf**	**PS**	**PN**
1	SWITZERLAND	GENEVA	GLAXOSMITHKLINE	88	2140	38	24.32	3.19	15.5%	18.2%
2	FRANCE	ILLKIRCH	INST GENET & BIOL MOL & CELLULAIRE	50	653	6	13.06	3.03	17.7%	60.0%
3	UNITED KINGDOM	BECKENHAM	GLAXO WELLCOME SMITHKLINE BEECHAM	53	671	12	12.66	2.59	13.2%	30.2%
4	FRANCE	ILLKIRCH	UNIV STRASBOURG 1	209	2424	34	11.60	2.71	18.3%	61.2%
5	UNITED KINGDOM	CAMBRIDGE	MRC-CAMBRIDGE CTR BRAIN REPAIR	64	735	10	11.48	2.08	20.2%	35.9%
6	UNITED KINGDOM	LONDON	NATL INST MED RES	305	3493	46	11.45	2.63	15.6%	62.3%
7	UNITED KINGDOM	CAMBRIDGE	MRC-MOL BIOL LAB	237	2592	37	10.94	3.21	21.0%	67.5%
8	GERMANY	HEIDELBERG	MAX PLANCK SOC	203	2144	37	10.56	3.83	12.1%	70.9%
9	FRANCE	EVRY	GENETHON SA	86	879	13	10.22	2.19	20.2%	46.5%
10	FRANCE	NOGENT SUR MARNE	COLL FRANCE	57	573	12	10.05	1.22	17.1%	43.9%

By creating the opportunity to combine two indicators, we allow the user create a particular context within which he can identify centers of excellence.

An other context that can be used to identify excellence is the national perspective. This is provided by the option to rank within the context of one country. In the sample below, we present the ranking within the Netherlands.

Table 3.9 Top ten organizations with highest impact (CPP) within the Netherlands, organizations with 50 publications or more only

Rnk	**Country**	**City**	**Organization**	**P**	**CX**	**P10**	**CPP**	**CPPf**	**PS**	**PN**
1	NETHERLANDS	AMSTERDAM	HOSP SLOTERVAART AMSTERDAM	61	406	8	6.66	2.54	20.5%	54.1%
2	NETHERLANDS	AMSTERDAM	NETHERL CANC INST	122	794	13	6.51	2.09	19.0%	59.0%
3	NETHERLANDS	ENSCHEDE	HOSP MED SPECTRUM TWENTE	66	426	4	6.45	2.73	21.5%	74.2%
4	NETHERLANDS	AMSTERDAM	KNAW NETHERL OPHTHALM RES INST	148	784	12	5.30	1.77	17.8%	62.8%
5	NETHERLANDS	ROTTERDAM	ERASMUS UNIV	1412	6924	88	4.90	1.84	21.2%	65.0%
6	NETHERLANDS	OSS	AKZO NOBEL	89	433	7	4.87	1.66	11.1%	53.9%

			ORGANON							
7	NETHERLANDS	THE HAGUE	HOSP WESTEINDE	52	239	2	4.60	1.24	20.6%	55.8%
8	NETHERLANDS	AMSTERDAM	HOSP ST LUCAS ANDREAS AMSTERDAM	52	238	4	4.58	1.67	15.3%	57.7%
9	NETHERLANDS	LEIDEN	TNO	99	404	6	4.08	1.47	20.2%	66.7%
10	NETHERLANDS	AMSTERDAM	KNAW NETHERL INST BRAIN RES	367	1449	23	3.95	1.16	27.3%	59.9%

From the results in this table, we see that smaller organizations with a significant output in 5 years can be discerned, and not only the bigger organizations. It seems, however, that in this case the specialized research centers are somewhat in favor. Still, we also detect bigger organizations which are less specialized.

3.3.2 Cognitive maps of the fields

For the 4 life science fields we compiled the following maps. In the maps, each circle designates a sub-domain with a particular field. A sub-domain is determined by a cluster of keywords. In the map a sub-domain is labeled with the most frequent keyword within. The surface of a circle is determined by the number of publications represented in a sub-domain. The position of the sub-domains depends on their cognitive orientation. The closer two sub-domains are related the closer they are in the map.

These maps were used to characterize the activity and impact of actors in the individual fields and thus fully integrated in the performance profile interface.

Figure 3-1: Cognitive map of Bioinformatics

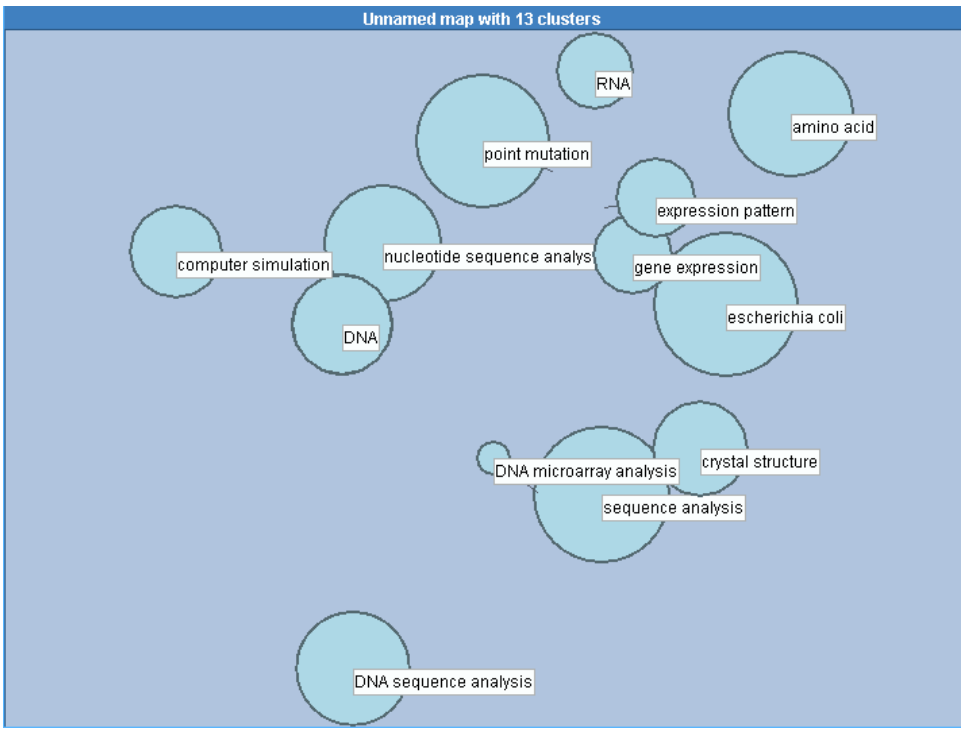


Figure 3-2: Cognitive map of Genetics

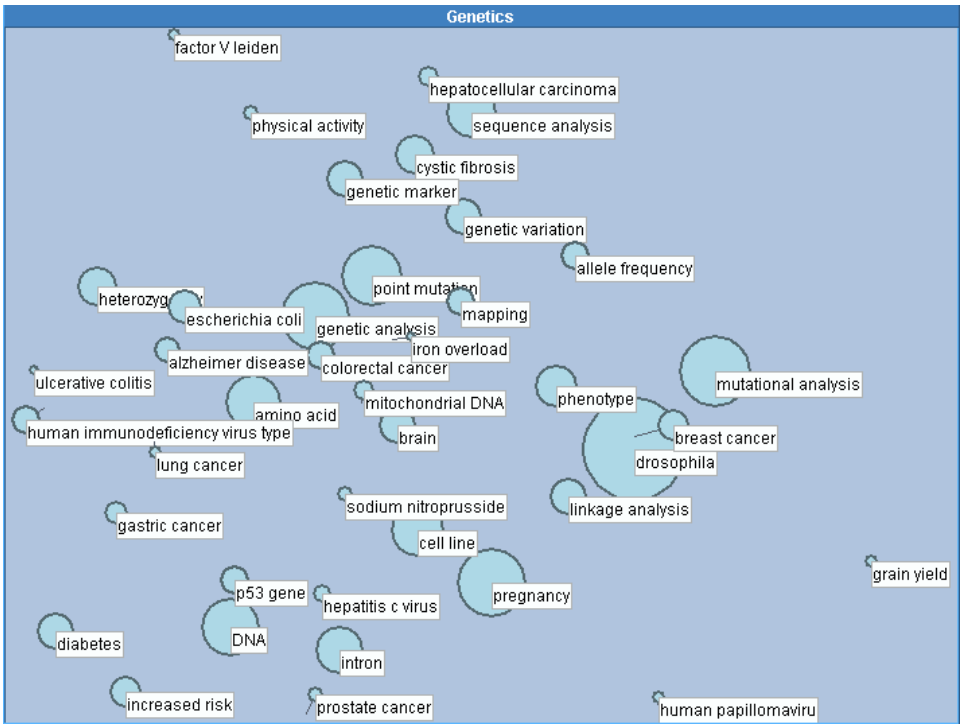


Figure 3-3: Cognitive map of Immunology

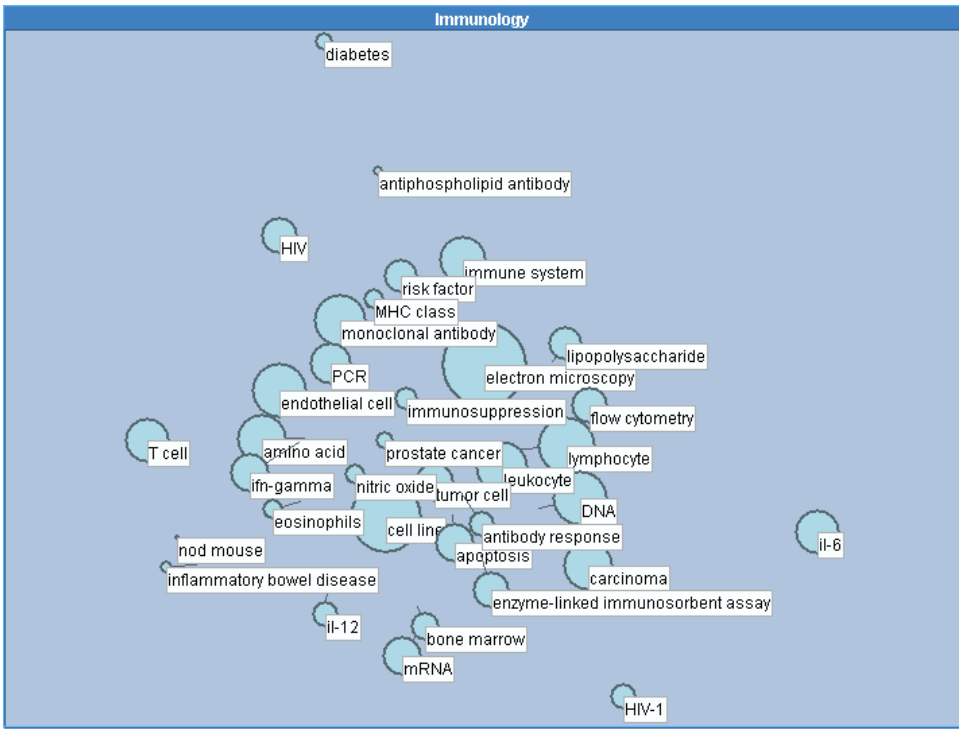
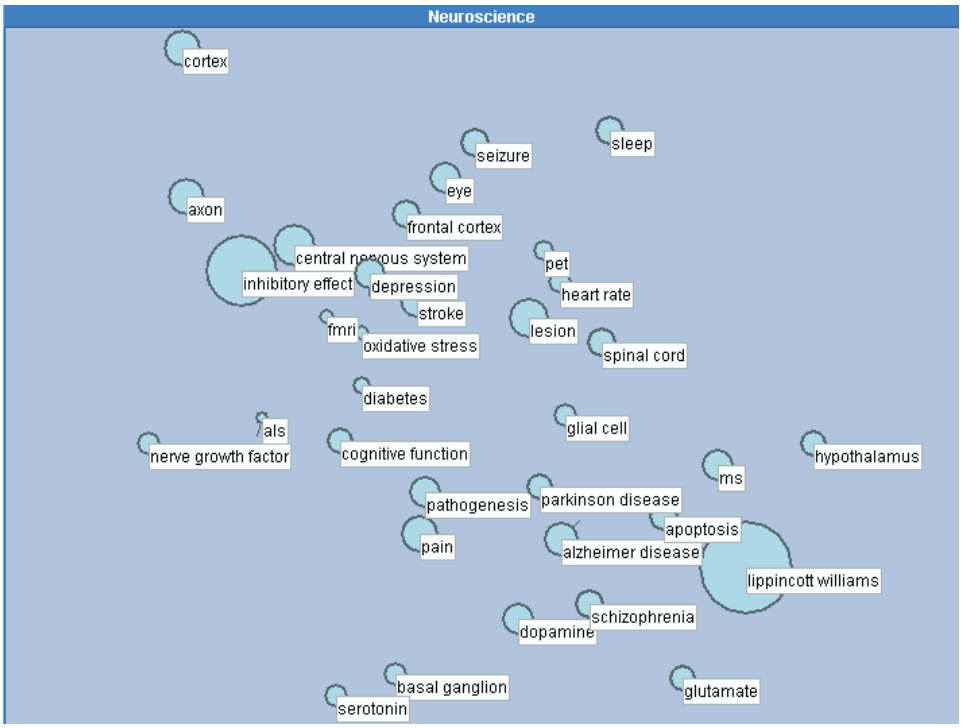


Figure 3-4: Cognitive map of Neuroscience

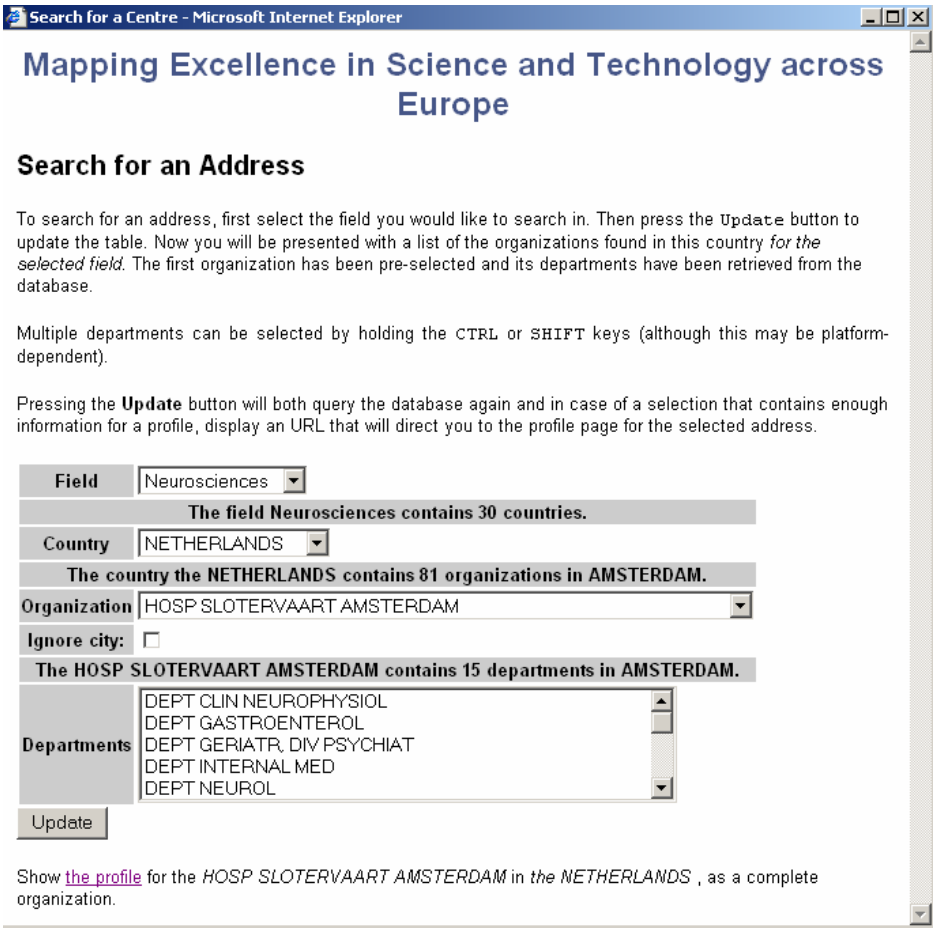


3.3.3 Integrated results

As mentioned before, the results depend heavily on the question of what the definition of excellence is. And of course what we wish to define as a center. In this study we take the main organization as a starting entity. But as we can see from the table above, they may be specialized research institutes or large universities. From the number of publications in the table we already discern large differences. In large fields such as neuroscience, with around 90,000 publications per year, this plays an important role. To illustrate this, we will run a full profile of two organizations from the table: the Slotervaart Hospital and the Erasmus University. We will demonstrate how the cognitive map can help to characterize the differences between these organizations. This should shed more light on the matter of how to decide what a center of excellence is.

A full bibliometric profile of an organization can be retrieved by clicking the name of the organization in the ranking. For the Slotervaart Hospital, the form to run a profile is shown in the figure below.

Figure 3-5Form to run a complete bibliometric profile



This form allows the user to run a profile of the entire organization. There is an option to ignore the city name in order to be able to run a profile of organizations that have locations in different cities. Furthermore it is possible to run a profile for a part of that organization, by selecting a limited set of departments within. We will address the issue of departments later.

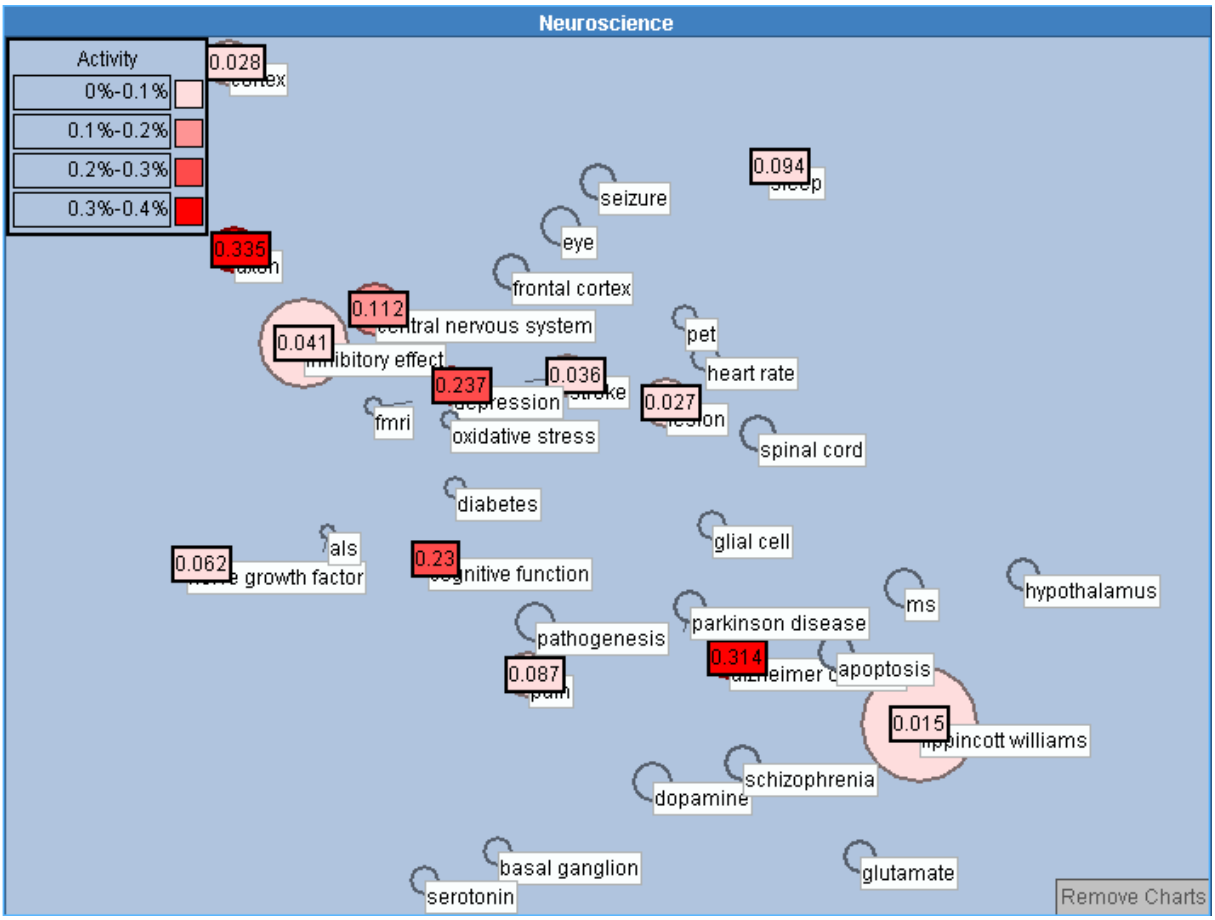
A complete bibliometric profile looks like the Figure below.

Figure 3-6: Bibliometric profile of an organization

Apart from the known indicators, this profile presents the distribution of publications over clusters that are identified in the cognitive map of the field. By clicking the button [show the map] this map is shown. On this map, we can characterize the activity of this organization by showing the activity distribution. This is shown for Slotervaart Hospital on the figure below.

Here we see the entire map of the field and in colors is indicated what the contribution of this actor is. In the left upper corner a explanation is given of the colors. The deeper red a circle, the higher the relative contribution is.

Figure 3-7: Activity distribution of the Slotervaart Hospital in Neuroscience



If we visualize the activity distribution of the Erasmus University, we see a completely different pattern. Being such a large organization their activity covers the entire map.

As seen from the ranking, both organizations show a good performance with respect to impact. The Slotervaart Hospital, however, is much smaller and focuses clearly on the area of *cognitive function*, *axon*, *alzheimer* and *depression*, while the Erasmus is present in the entire field as can be seen in the chart in Annex F.

This leads us to the possibility to select parts from a bigger organization to run a profile. This possibility is created to allow users to use their own definition of what a center is, and to allow cleaning of the departmental address data for individual cases. Thus, we provide the user to clean address data on he spot for individual use, rather than to force national experts to clean the address data for the whole country. The results for one department in the Erasmus University are in Annex F.

This profile shows clearly the focus of this department of the Erasmus. Their focus is *basal ganglion*, *pathogenesis* and *seizure*. This shows how one the one hand that it is possible to profile parts of a large organization by using (variants of) a department name, and on the other that it is possible to compare entire organizations with parts of larger organizations. This enables the user to make his own choice to compare like with like. These comparable entities can be composed on the basis of the user's own knowledge of the science organization in his own country.

3.4 Conclusions of the publication analyses

3.4.1 Field delineation

We designed the procedure for field delineation in such a way that it could be applied to many more than these 5 fields, with non extra effort from the analyst perspective. The idea behind the design was that we allow expert interference but as little as possible. We don't want them to go through huge lists of publications to decide which papers belong to field and which do not.

The generic character of this design has shown different drawbacks at this stage. The main problem is the fact that we had to design something in a short period of time. We were not able to process the comments that came to the first version of the procedure. Moreover, we had to collect the data in a relatively short period of time. In at least one case, this led to the situation that one of the two experts was not able to comment on the preliminary results within the set period of 6 weeks. In this particular case, the lack of time, consensus and input from experts led to a selection of publications that seems completely useless. Therefore, the results are difficult, if not impossible, to interpret and to validate.

In a technical sense, the internet form to add or remove search terms from the applied delineation, appeared useful for some of the experts. They used the form to add search terms to the existing list. Furthermore, one of the experts sent his own search strategy to add to the existing search strategy not using the form. Generally speaking, however, we received a limited response to our request to help delineating the field. The design of the form should be revised considerably before it can be used.

As indicated in the preparatory studies as well, there is a problem with regard to reaching consensus among experts about the delineation of their own field. This makes it difficult to decide what the field looks like, and to decide what search strategy could be used to collect the relevant publication data. One of the possibilities to cope with this is to allow users to create their own definition of the field, via a web-interface and to compile the results after this stage of delineation. In order to establish this 'open' delineation, all subsequent stages should be highly automated. In such a way there is no limited number of fields to be considered, but a field is defined by the search strategy compiled by an expert or other user. A field is no longer a field in the traditional sense but can also be a theme (e.g., Alzheimer's disease, prostate cancer, CO_2 emission).

3.4.2 Bibliometric indicators

The interactive ranking tools developed in this project allow the user to identify centers of excellence on the basis of different indicators. Each indicator yields a different perspective from which to address the complex concept of excellence. It depends on the kind of excellence one is looking for to determine which indicator or combination of indicators is preferably used.

We think that particularly the option to combine indicators is a major improvement. The possibility to change thresholds for individual indicators while leaving an other unchanged, immediately shows the effect of using certain thresholds and the effect of using certain

indicators. As excellence is a complex concept, the identification of excellent groups could be helped by complex tools.

As for the different indicators are concerned, we already know the strengths and weaknesses of the production (P) and the impact (CPP). The latter indicator is particularly valuable in combination with a field average (CPP/FCSm). Still, we find that the CPP is biased towards smaller numbers. This means that smaller institutes tend to reach higher impact. And also variants of institute names with small numbers of publications (or only one if there is a uncorrected spelling error) are more likely to reach high impact scores (CPP and CPP/FCSm). Here the combination with P is available to enable the user to compare organizations (institutes) with a similar size. By choosing a P of at least 100, and ranking on the basis of impact, a much fairer comparison is made.

The Top 10% indicator correlates with the production indicator (P). The higher production P, the higher the top 10% contribution is. This indicator needs more sophistication by normalization to P. This is suggested by van Leeuwen et al. (2003) and should be integrated in future studies.

The growth factor is highly biased towards smaller entities and should be investigated in more detail to determine its value added for this kind of studies.

3.4.3 Address data

There has been a lot of discussion on the level of aggregation to be used to identify centers of excellence. It has been argued that a center should be a group, a team. Others consider a center much larger. Their perception of a center would be an organization (i.e., a university, a research institute or company) or intermediate levels (faculty at a university). Although, we may never reach consensus on this, the bibliometric approach should be able to provide information at different levels. As we are working with the address data in the publication databases, we had to deal with the problem of name variants, spelling errors in the address data at all levels. Although the CWTS data system performs a thorough cleaning of address data at the level of country, city and organization, there is still a substantial amount of ‘mistakes’. The national experts were involved via the High Level Group to clean their national publication addresses down to the level of organization. At the departmental level the job would be too much time-consuming. Even for the bigger countries, this effort was done reasonably fast. In some countries, the complexity of the national science system and the quality of the input address data was the reason that no cleaning effort was undertaken.

As the cleaning process of CWTS at the first stage seemed to be to rigorous in some cases, we need to reconsider the structure we applied in our cleaned data. For instance, the ‘legal’ entity of the University of London is too big. It has been argued that we should use one level down as an organization (i.e., colleges etc.). Furthermore, the level of national research institutes (CNRS, Max Planck Gesellschaft, CSIC) should be disaggregated to smaller entities.

Another problem was the procedure we used for cleaning. We provide comma-delimited files to be used in any database-like desktop application (excel, access, etc.) And although we provide strict rules of what the experts should do or not, the procedure allowed the user to change the data in such a way that we had to reprocess the data to integrate into our system again. Furthermore, we noticed that some settings of personal computers may cause difficulties to read and process the data in its basic format. A solution could be that the cleaning process is done via a web-based form. In this way we keep full control of what the user is allowed to do or not.

We concluded during the run of the project that the data on departmental level is such that it made no sense to clean them. We had to deal with name variants, spelling variants, spelling mistakes, structural changes at the organizational level and complex structures that are not represented in the address data. It seems that the only robust and therefore useful structure is at the organizational level (universities, companies, research institutes). The only possibility we see at present to go down a level, is by using the author names. By using this information, we should be able to identify research groups around one or more leaders with an organization. As we are dealing with separate research fields, we don't have to worry too much about homonym author names (i.e., one name referring to multiple persons). We would need to investigate in more detail this possibility. One of the problems will be to uniquely assign an author to an organization. As publications contain more than one address, and from the ISI data we only know that the first author is linked to the first address, we cannot simply link publications to organizations, to authors. A second issue to deal with is the mobility of authors and the multiple assignments of authors.

Finally, in the final version of the results we encountered the discrepancies of publication address cleaning on the one hand and patent address cleaning on the other. As these efforts took place approximately at the same time, it was not possible to assure a complete compatibility. It appeared that in some cases, the patent experts used a different structure than the publication experts. For instance, in Leiden, the Netherlands the University hospital was assigned to the Leiden University in the publication data, whereas in patents they were cleaned to separate entities. In the final results, where the patent data was integrated into the publication data this may have led to multiple entities. It is therefore, advisable to integrate the patent addresses into the publication addresses from the start and to distribute the address data to cleaning experts and to process the results from one location.

3.4.4 Geographical mapping

Although we had to broaden our scope as bibliometricians to create the geographical maps to plot the identified cents of excellence, we manage to do so. This could not have been done without the database, we had to purchase with geographical codes fro 3 million city entries world-wide. Using the SAS geographical country maps we could plot most of the entries via an interactive tool, which allows the user to combine indicators and to change thresholds.

In order to reach one hundred percent coverage of the cities, we need to look up by hand the geographical codes for the missing cities in the future.

3.4.5 Cognitive mapping

The cognitive maps have the added value of putting the activity or impact of a research entity into a cognitive perspective. This enables the user to determine what the specific expertise of the identified center is. As these cognitive maps are generated with an almost completely automated process, we need a good feedback procedure to evaluate the cognitive map before they are used to map a field. The map should be 'recognized' otherwise the structure may not make sense to the user. In this study we were not able to make much progress in this respect. We could apply a basic methodology to generate the maps. Unfortunately this was established too late for most fields for experts to evaluate. If we are developing a tool to allow an 'open' delineation of a field, this aspect becomes even more important.

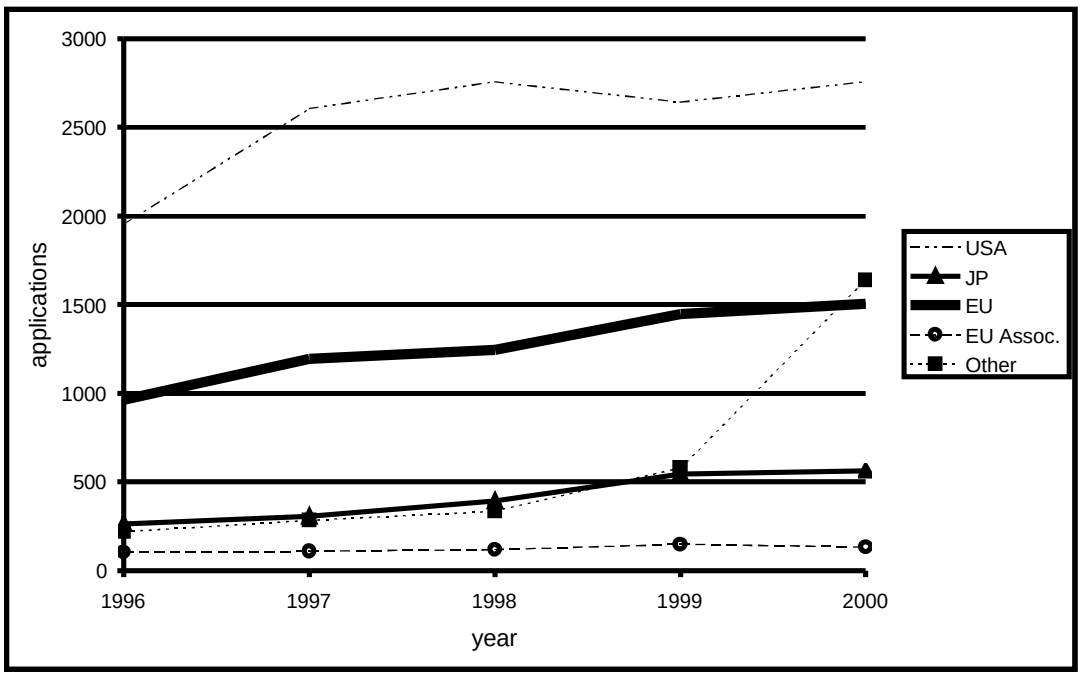
4 Macro analyses

4.1 Patents

The major aim of this project is to identify excellent institutions in different fields. The details of the referring analyses are documented in chapters 2 and 3. In this section, some results of macro analyses are additionally presented, as they give interesting insights into the overall structures within the EU and associated countries.

A standard macro approach is to look at time series on the national or regional level. Figure 4.1-1 shows a comparison of the patent applications in the years of 1996 to 2000 for the area of Genetics/heredity. In this short period, the number of referring patent applications has increased substantially. The number of the EU countries grew by 50 percent, of the United States by 40 percent. The world-wide patent activities are dominated by the United States which in 2000, applied for about 80 percent more patents then all EU countries taken together. The Japanese patent applications represent about one third of the EU applications and doubled in the period considered. A striking point is the enormous increase of the patents of „other countries“ by 200 percent between 1999 and 2000. This is primarily due to patent applications from China, to a lower extent also from Australia and South Korea. The number of applications from the associated countries is rather modest.

Figure 4.1-1: Trends of EPO and PCT patent applications in Genetics/heredity

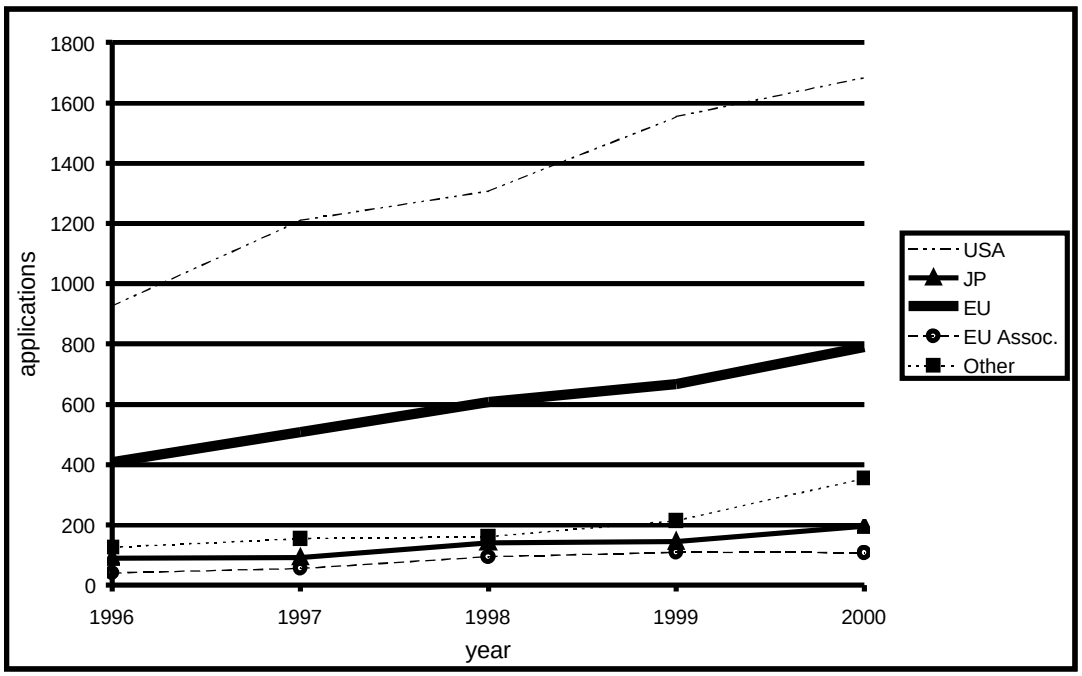


Source: EUREG, computation of Fraunhofer ISI

In the case of Neurosciences, the growth of patent applications is even stronger than in Genetics/heredity, 80 percent for the United States, 90 percent for the EU, 115 percent for Japan, and 180 percent for other countries, 160 percent for the associated countries (Figure

4.1-2). In Neurosciences, the position of Japan, the associated countries, and other countries is modest compared to the EU and the United States. The dominance of the United States is more distinct compared to genetics/heredity with two times as many applications at the EU.

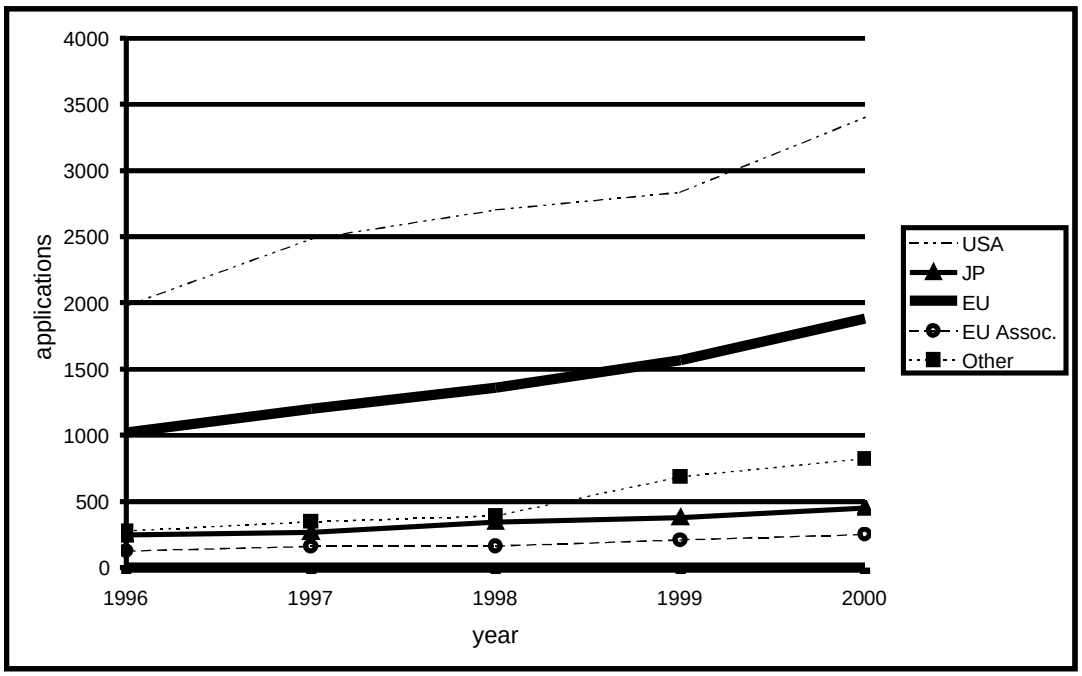
Figure 4.1-2: Trends of EPO and PCT patent applications in Neurosciences



Source: EUREG, computation of Fraunhofer ISI

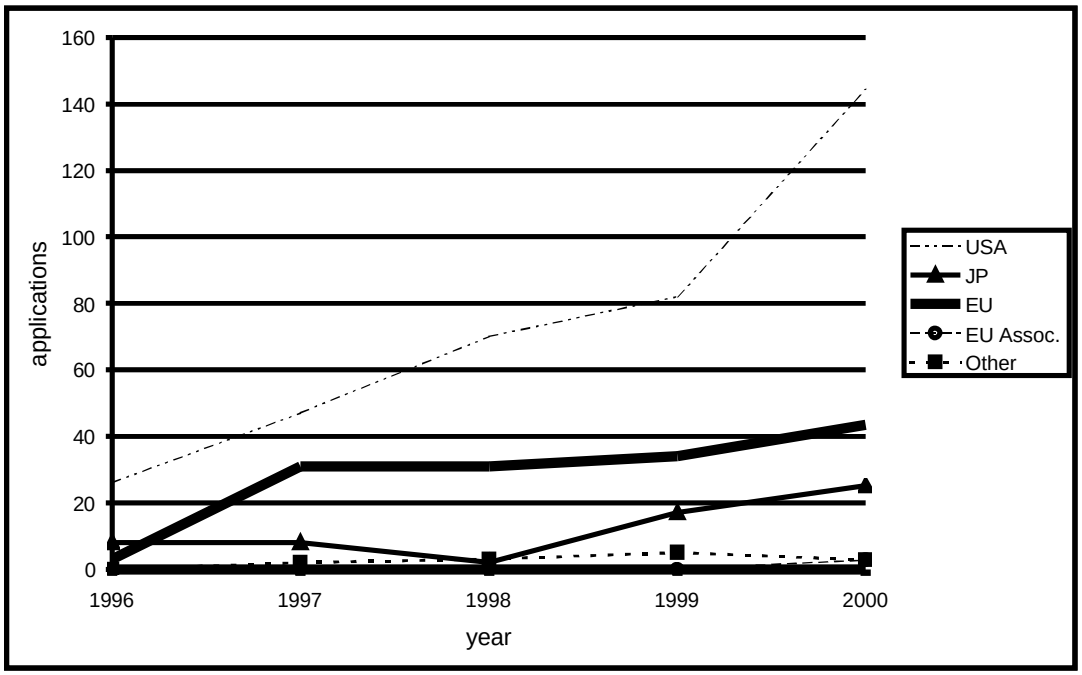
The trends in Immunology shows characteristics similar to Genetics/heredity and Neurology, in particular a strong increase in a short period and a dominant position of the United States, whereas the level of Japan, the associated countries and other countries is modest (Figure 4.1-3). Again, the other countries display a clear uptake in the present situation, but they do not reach the level of the EU like in Genetics/heredity.

Figure 4.1-3: Trends of EPO and PCT patent applications in Immunology



Source: EUREG, computation of Fraunhofer ISI

Figure 4.1-4: Trends of EPO and PCT patent applications in Bioinformatics



Source: EUREG, computation of Fraunhofer ISI

All in all, the general structures in the life sciences are coherent. However, the situation in Bioinformatics is quite different. The activities of the associated countries and the other countries are very low, but Japan appears to be relatively stronger than in the other three areas

(Figure 4.1-4). The United States have about 140 percent more patents than the EU until 1999, and show in 2000 a dramatic increase, so that they presently dominate the EU by 230 percent. This result for Bioinformatics should be interpreted with caution. It can be assumed that the data comprise a large sample of all patent applications registered in this area. However, the propensity to patent software–based inventions may considerably vary between different countries. The high number of patents with US origin may be due to the broader patentability of software at the United States Patent and Trademark Office (USPTO), encouraging US investors to apply for software patents also on the international level.

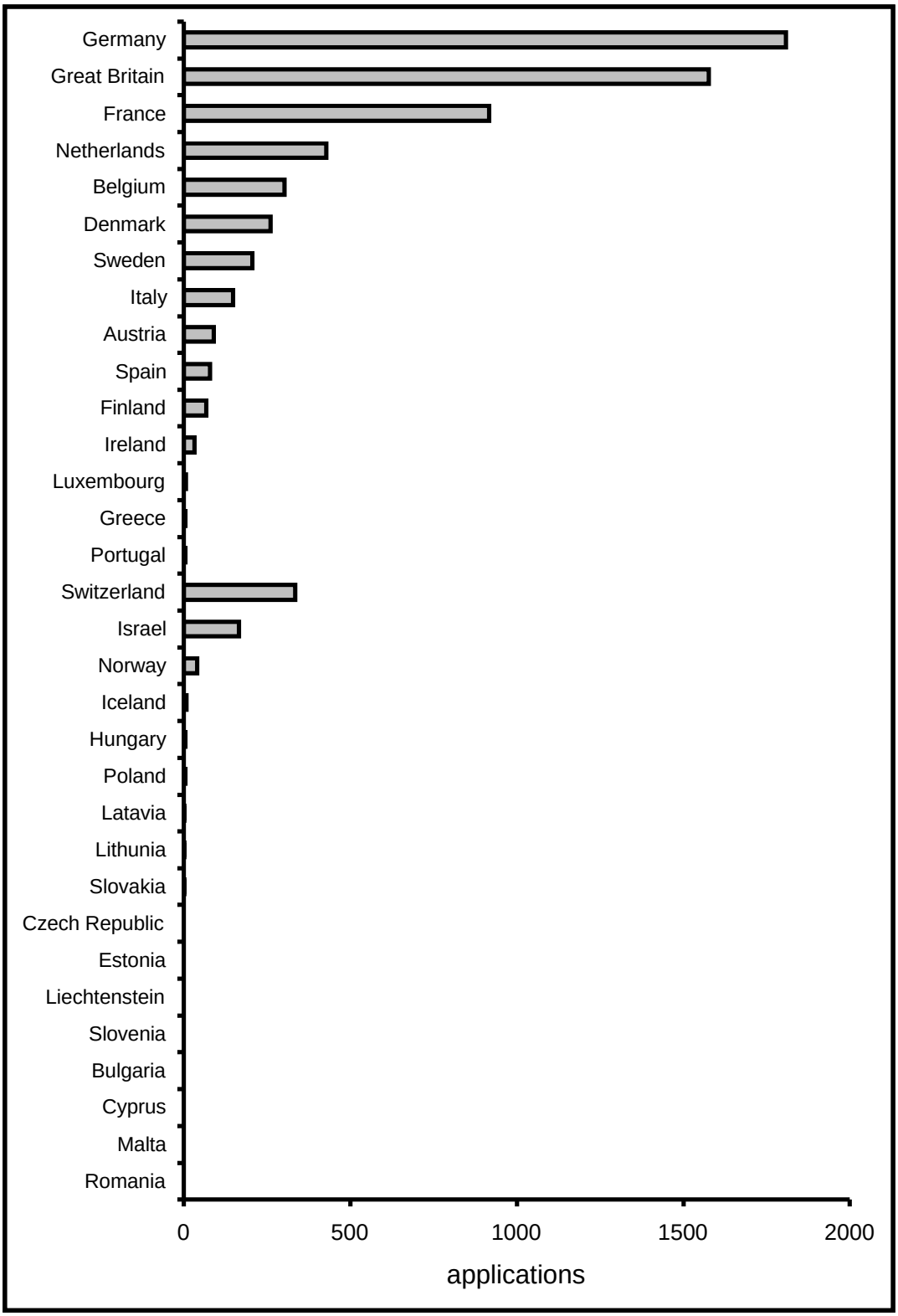
A further interesting perspective is to look at the patents within the EU and the associated countries. In absolute terms, the EU patents in Genetics/heredity are dominated by about eight countries (Figure 4.1-6). As to the associated countries, Switzerland, Israel, and Norway have a relevant impact, whereas most other countries have very few or no patents at all. It is not surprising that the larger countries take out the majority of patents, so that it is more interesting to control the data for country size. It suggests itself to use R&D budgets as reference for patents, but for many countries these data are not available. Therefore we refer to GDP which represents the size as well as the economic strength of a country[19]. With reference to GDP, Denmark shows up to be the most active country in the EU, whereas Germany has a medium position in this perspective (figure 4.1-7). As to the associated countries, Switzerland, Israel and Norway are still leading, but the productivity of Baltic countries, for instance, gets visible in a better way[20]. This form of representation allows for a first screening with regard to centers of excellence. The good performance of countries such as Denmark or Switzerland, or the still relevant activity of the Baltic countries can be interpreted as a strong indication that some centers of excellence can be found in these countries. So the data advise to look at the institutional structure of these countries in more detail. The analogous data for Neurosciences, recorded in Figure 4.1-8, show high patent activities in Sweden and again in Denmark. The performance of Israel is surprisingly high and is above EU countries' average.

The patents with reference to GDP reveal a similar structure in Immunology as in the other life science areas Genetics/heredity and Neurosciences (Figure 4.1-9). However, the structure in Bioinformatics appears to be completely different (Figure 4.1-10). Within the EU, Sweden and Great Britain hold dominant positions, whereas the associated countries do not show up at all. The reasons for this specific situation are discussed further above.

[19] The GDP data are in million Euro in purchasing power standards (PPS) of the year 2000. The data were taken from the following cources: Eurostat (2002), Worldbank (2002), Statistisches Bundesamt (2002), Fischer Weltalmanach (2002)).

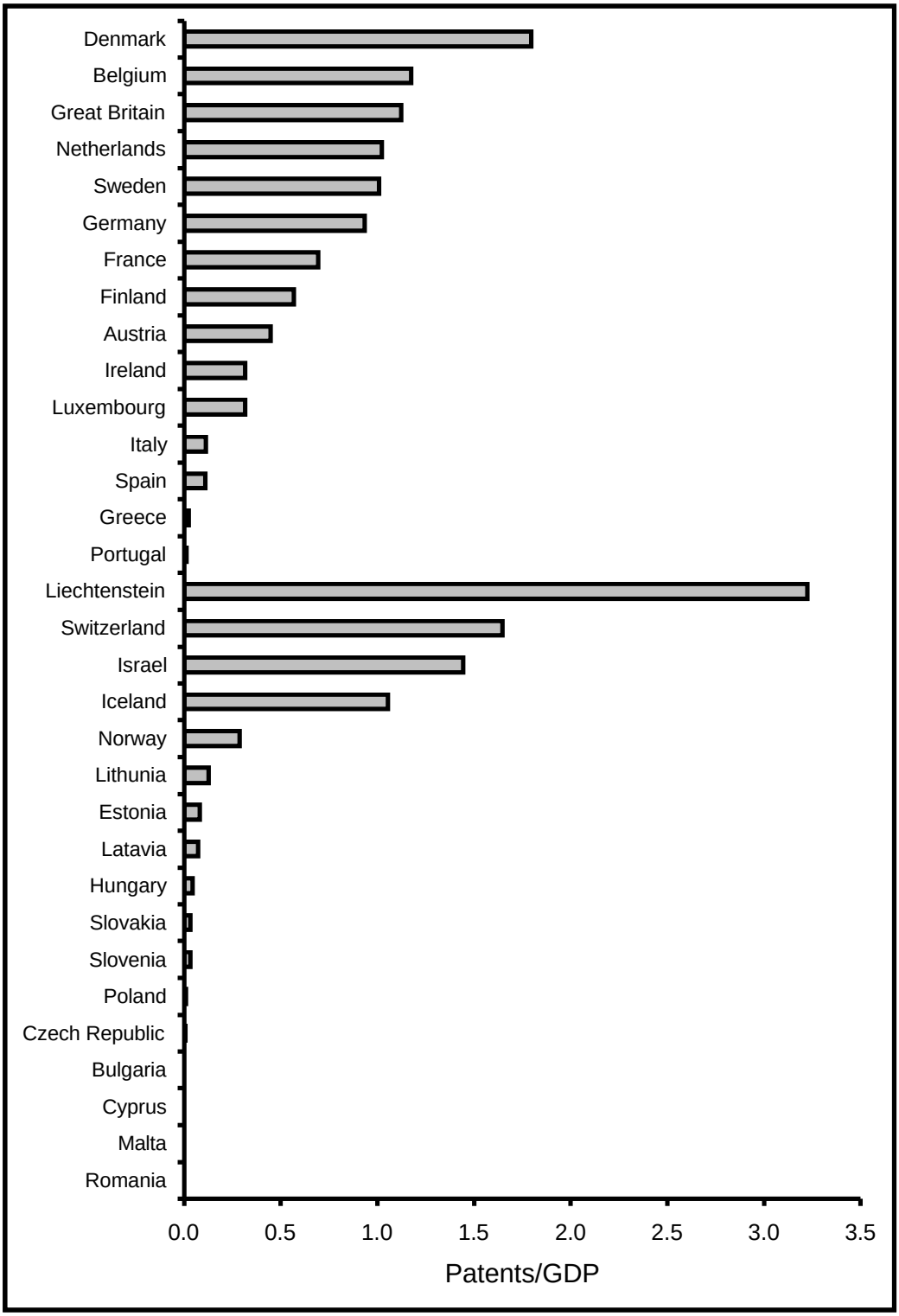
[20] The results for Liechtenstein and Ireland were deleted due to extreme effects of quotients of low figures.

Figure 4.1-6: PCT and EPO applications of selected countries in Genetics/heredity in the priority period 1996 to 2000



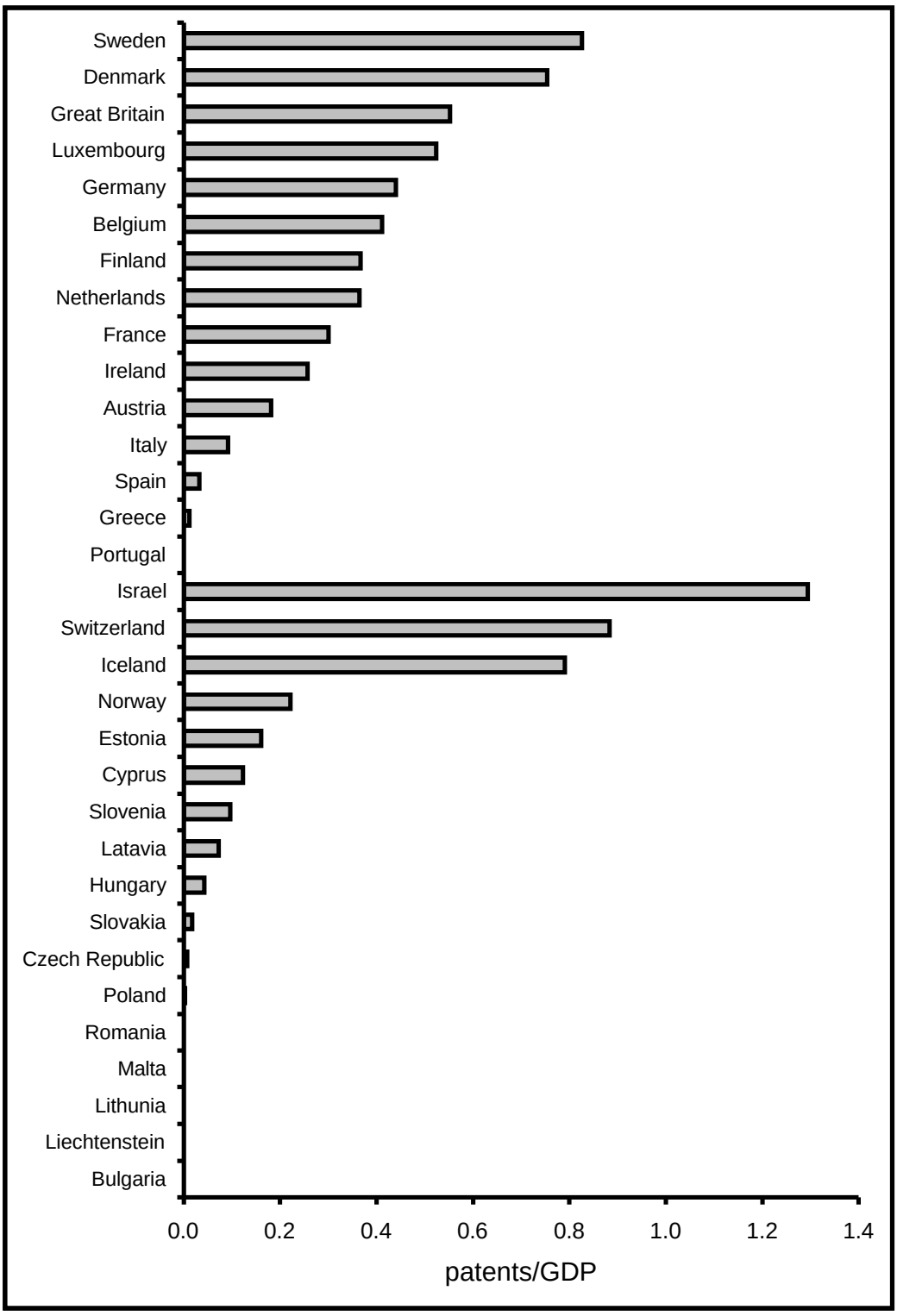
Source: EUREG, computation of Fraunhofer ISI

Figure 4.1-7: PCT and EPO applications with reference to GDP for selected countries in Genetics/heredity in the priority period 1996 to 2000 (GDP in billion € in purchasing power standards 2000)



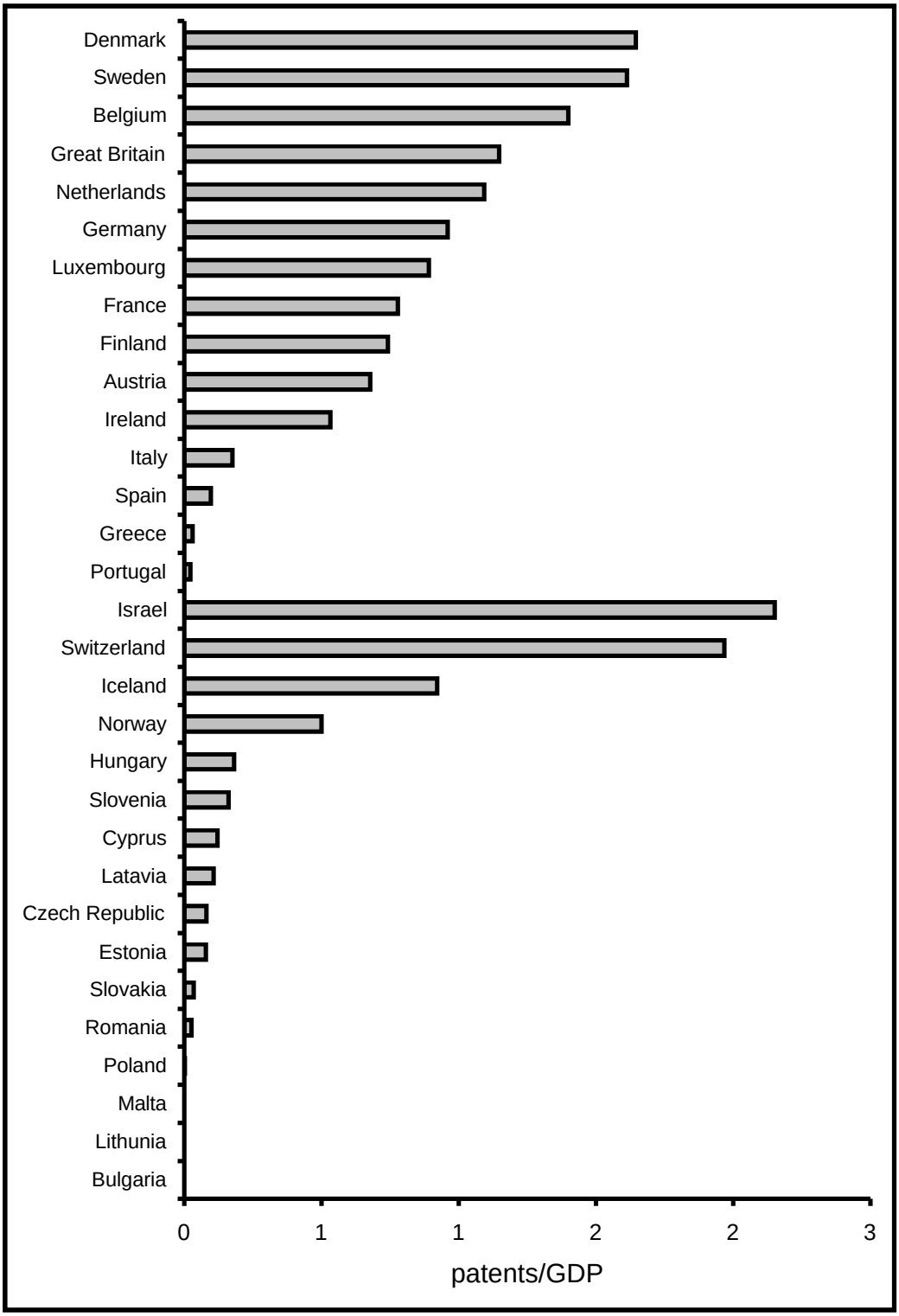
Source: EUREG, computation of Fraunhofer ISI

Figure 4.1-8: PCT and EPO applications with reference to GDP for selected countries in Neurosciences in the priority period 1996 to 2000 (GDP in billion € in purchasing power standards 2000)



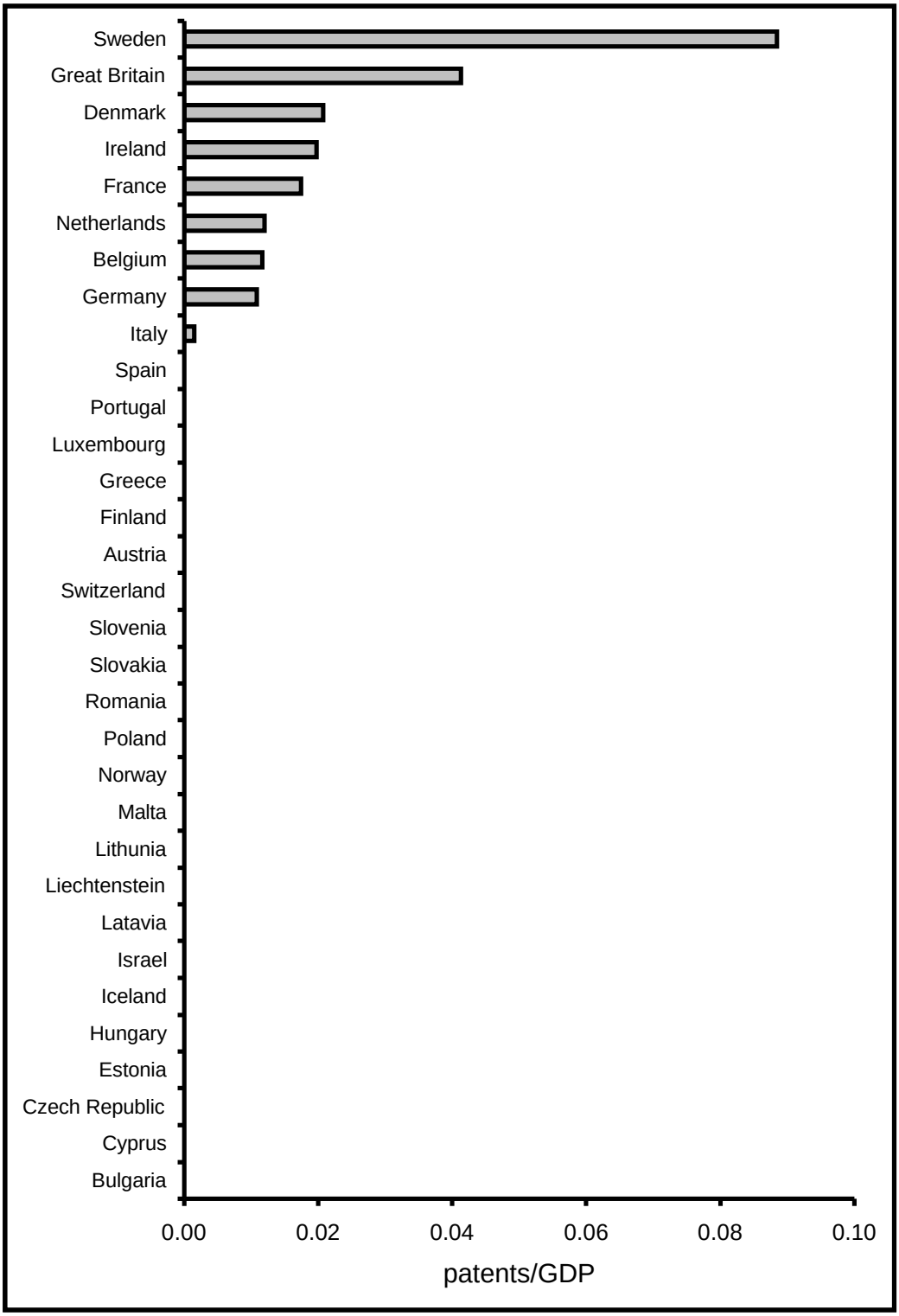
Source: EUREG, computation of Fraunhofer ISI

Figure 4.1-9: PCT and EPO applications with reference to GDP for selected countries in Immunology in the priority period 1996 to 2000 (GDP in billion € in purchasing power standards 2000)



Source: EUREG, computation of Fraunhofer ISI

Figure 4.1-10: PCT and EPO applications with reference to GDP for selected countries in Bioinformatics in the priority period 1996 to 2000 (GDP in billion € in purchasing power standards 2000)

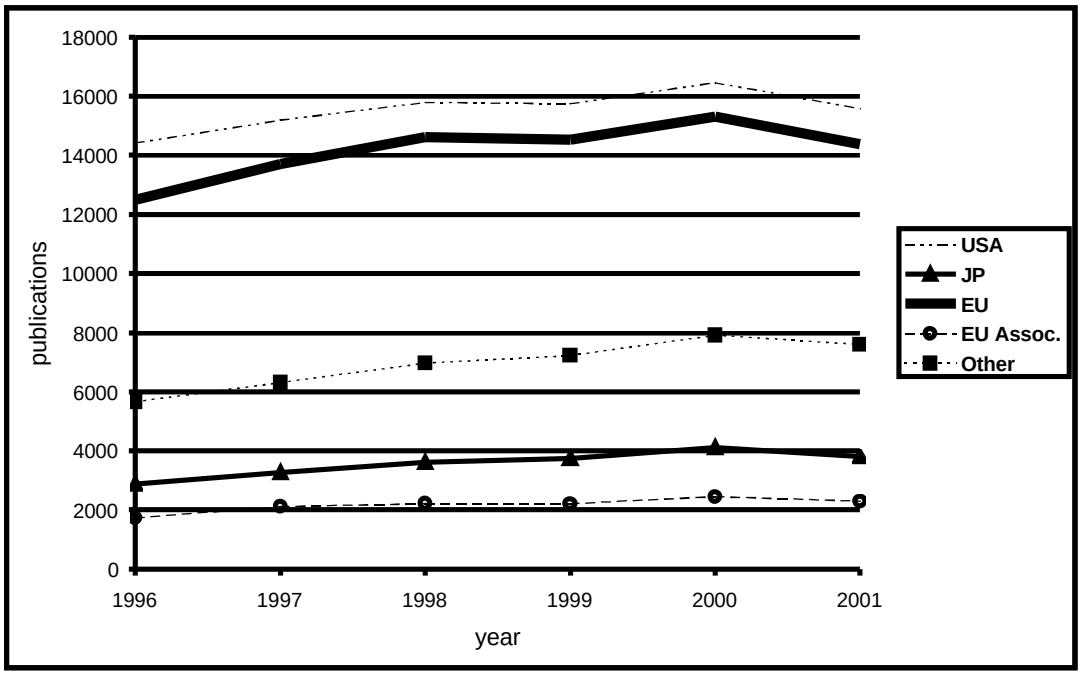


Source: EUREG, computation of Fraunhofer ISI

4.2 Publications

The trends of publications considerably differ from those of patent applications. In the case of Genetics/heredity, the United States display the majority of publications, similar to patents, but the relative level of the EU countries is much higher (Figure 4.2-1). Furthermore, the other countries have a distinctly higher number of publications than Japan, whereas in terms of patents, they had a comparable level.

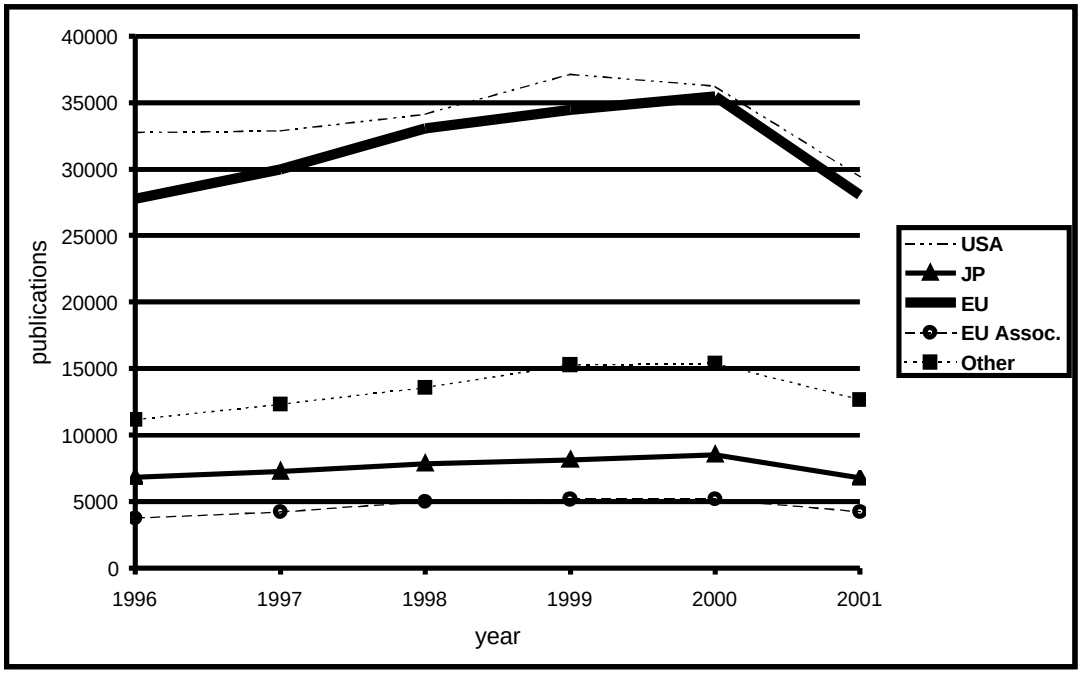
Figure 4.2-1: Trends of publications in Genetics/heredity



Source: SCI, computation of CWTS

In Neurosciences, the position of the EU countries is again stronger in terms of publications them in terms of patents (Figure 4.2-2). A remarkable point is the distinct decrease of publications in the last year of observation, whereas the patents are still increasing. The last year of the publication analysis refers to 2001, which is the year of publication. Due to publication delays, this is largely equivalent to the year 2000 of submission of the articles. Therefore the priority year 200o and the publication year 2001 are directly comparable.

Figure 4.2-2: Trends of publications in Neurosciences

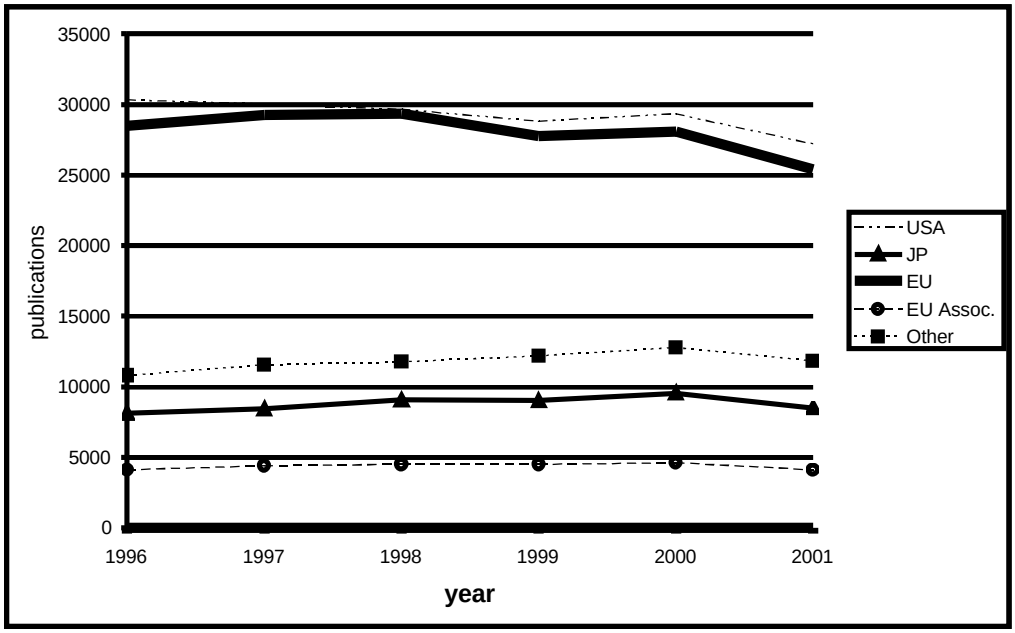


Source: SCI, computation of CWTS

In Immunology, the generally decreasing activities in publications show up in an even more distinct way (Figure 4.2-3), whereas patens are still increasing. This phenomenon obviously indicates that science in this area has achieved a high absolute level, but is largely saturated, whereas the transformation into technical application is still in an early phase. This situation can be illustrated more clearly by relating the number of patents to that of publications (Figure 4.2-4). According to this, the technological application is increasing with reference to science for all countries, but the United States display the highest relation. This outcome might be interpreted as a stronger orientation of the United States on technology.

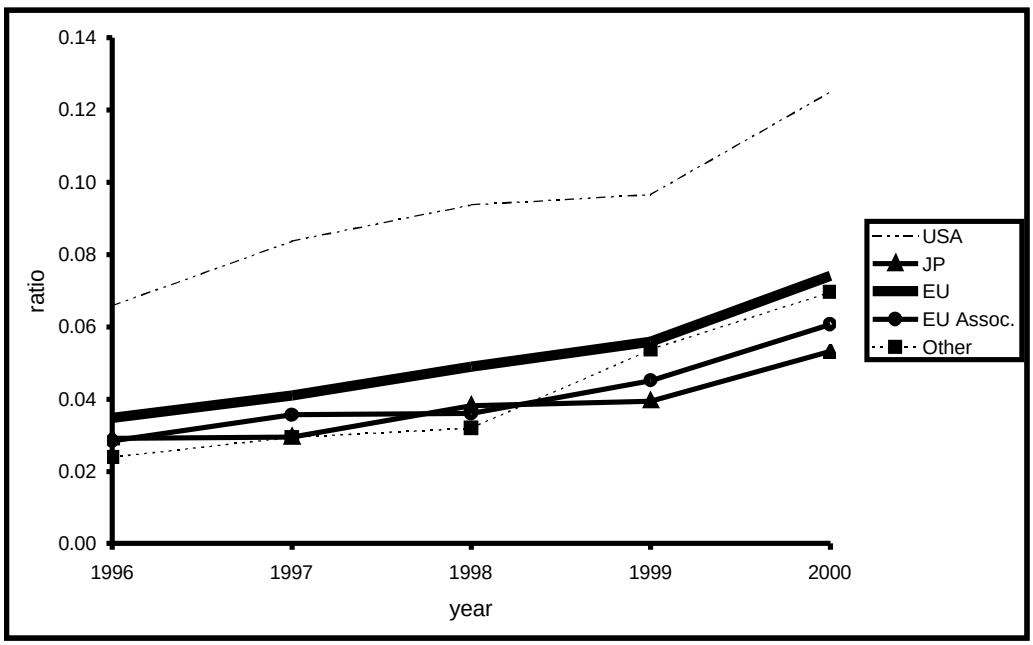
Looking at Bioinformatics, we see an increase for all countries (Figure 4.2-5), so this area is obviously in an earlier scientific state than the other areas of life sciences. However, due to the limitations of the data, described in section 4.1, it is not useful to look at the relation of patents and publications. As the trends in patents and publications increase simultaneously, the ration of patents to publications is more stable than for Immunology (Figure 4.2-7). But again, their orientation of the United States on Technology is the strongest one among the countries compared.

Figure 4.2-3: Trends of publications in Immunology



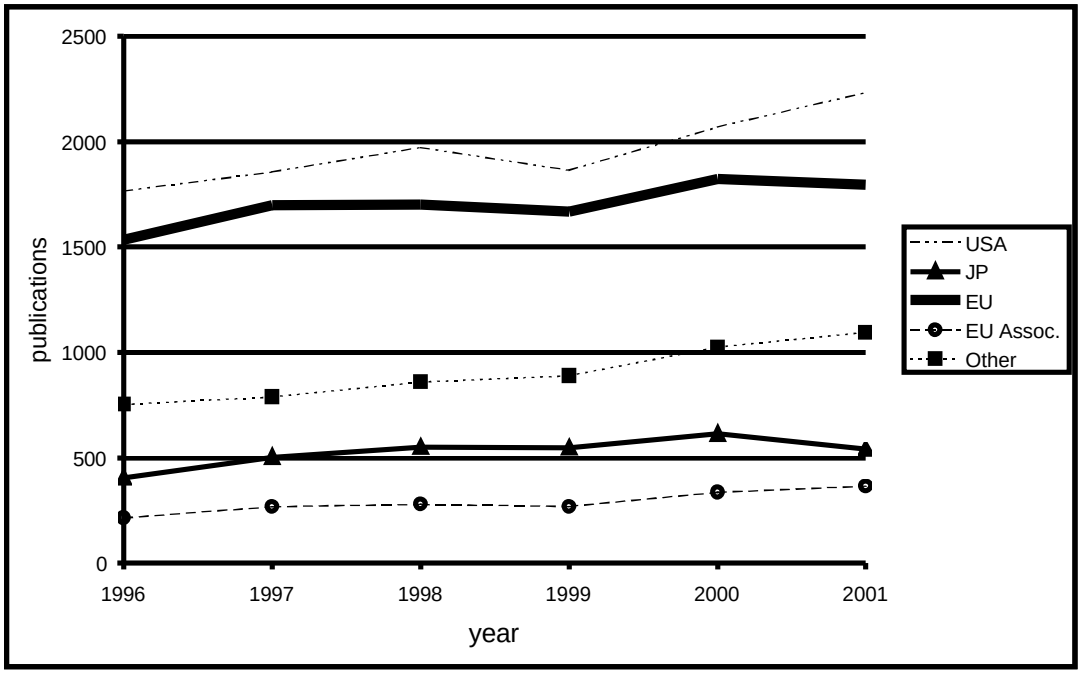
Source: SCI, computation of CWTS

Figure 4.2-4: Trends of relation of patent applications to publications in Immunology



Source: SCI, computations of CWTS and Fraunhofer ISI

Figure 4.2-5: Trends of publications in Bioinformatics

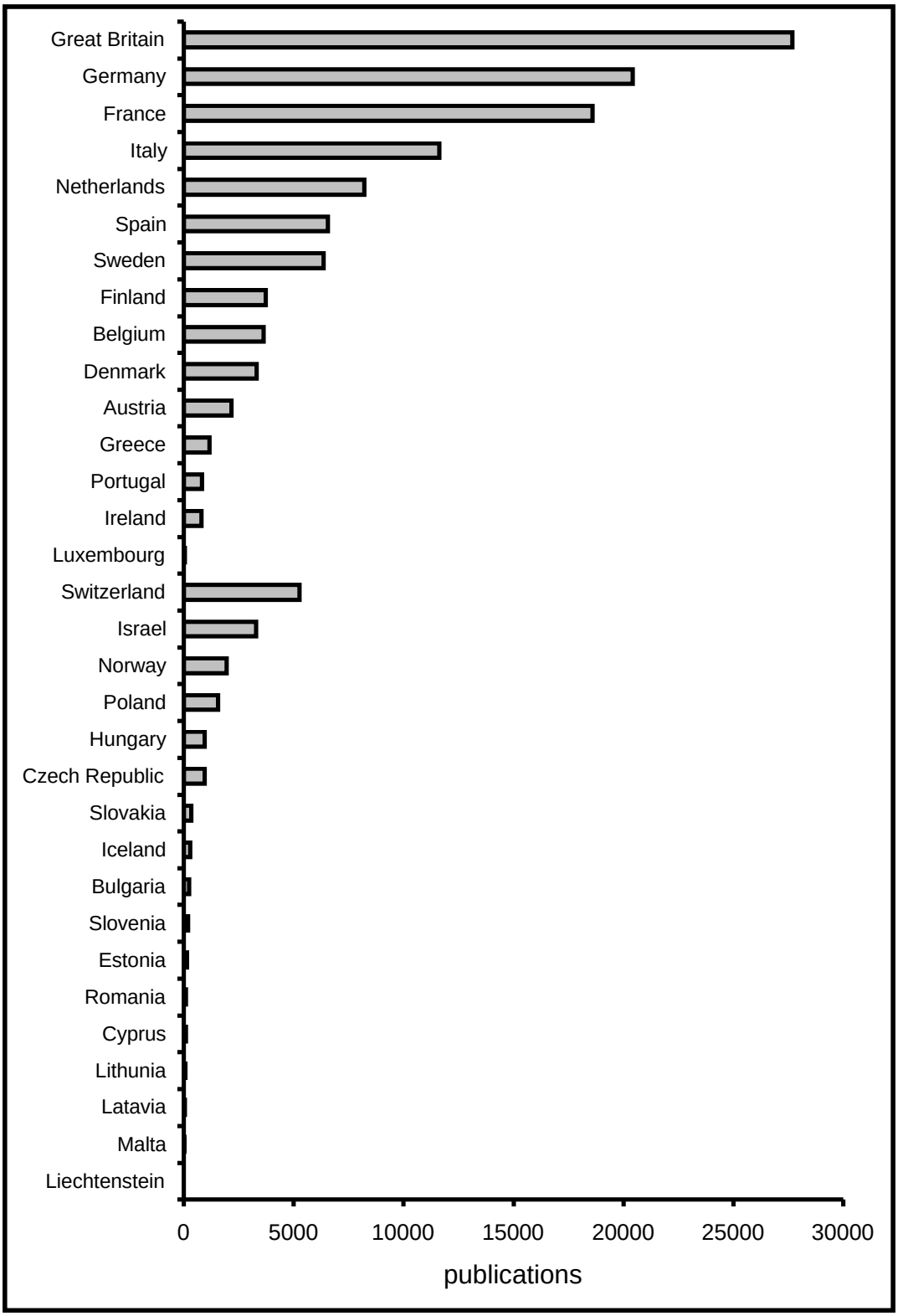


Source: SCI, computation of CWTS

In terms of absolute publication numbers for the whole observation period, Great Britain appears to be on a higher level than Germany in genetics/heredity (Figure 4.2-8). This structure is not particular for this field, but is generally characteristic for the SCI wherein a significant language bias towards English-language countries can be shown in contrast to a less strong position of countries with a large own language area such as Germany, France, Italy, or Japan (Grupp et al. 2001). Compared to the situation in patents (Figure 4.1-6), the publication activity of the candidate countries Poland, Hungary, and Czech Republic is remarkable. With reference to GDP, Sweden and Finland are on the top of the EU countries. The position of Great Britain is still exaggerated, but less extremely than in absolute terms (Figure 4.2-9). As to the associate countries, it has to be taken into account that quotients of low figures may lead to high values, e.g. in the cases of Bulgaria or Estonia.

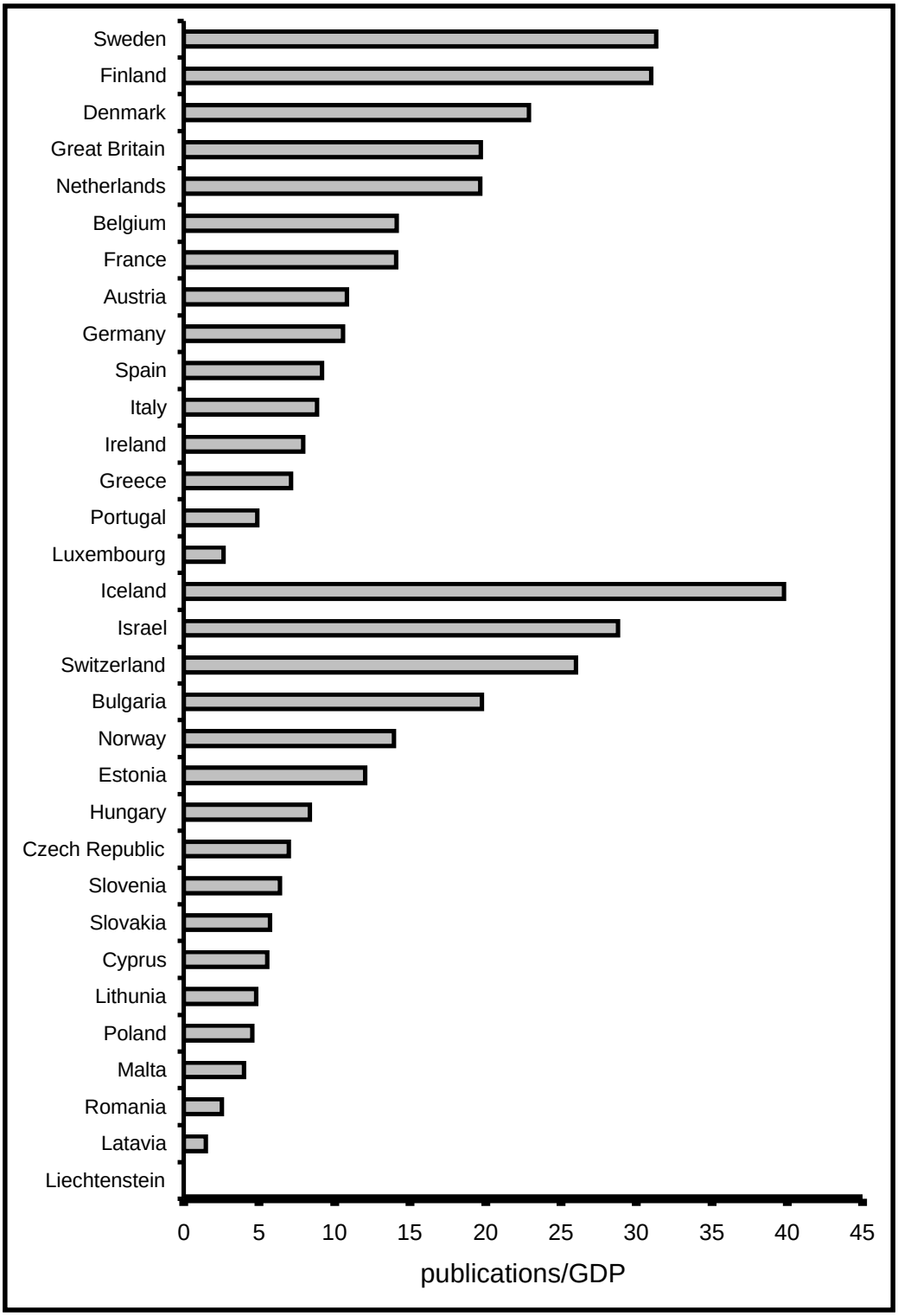
The strong position of Sweden and Finland is also confirmed for neurosciences and immunology (figures 4.2-10 and 4.2-11). In the case of publications, the structures in bioinformatics show a stronger affinity to the other life science areas than in the case of patents (figure 4.2-12).

Figure 4.2-8: SCI publications for selected countries in Genetics/heredity in the period 1995 to 2000 (GDP in billion € in purchasing power standards 2000)



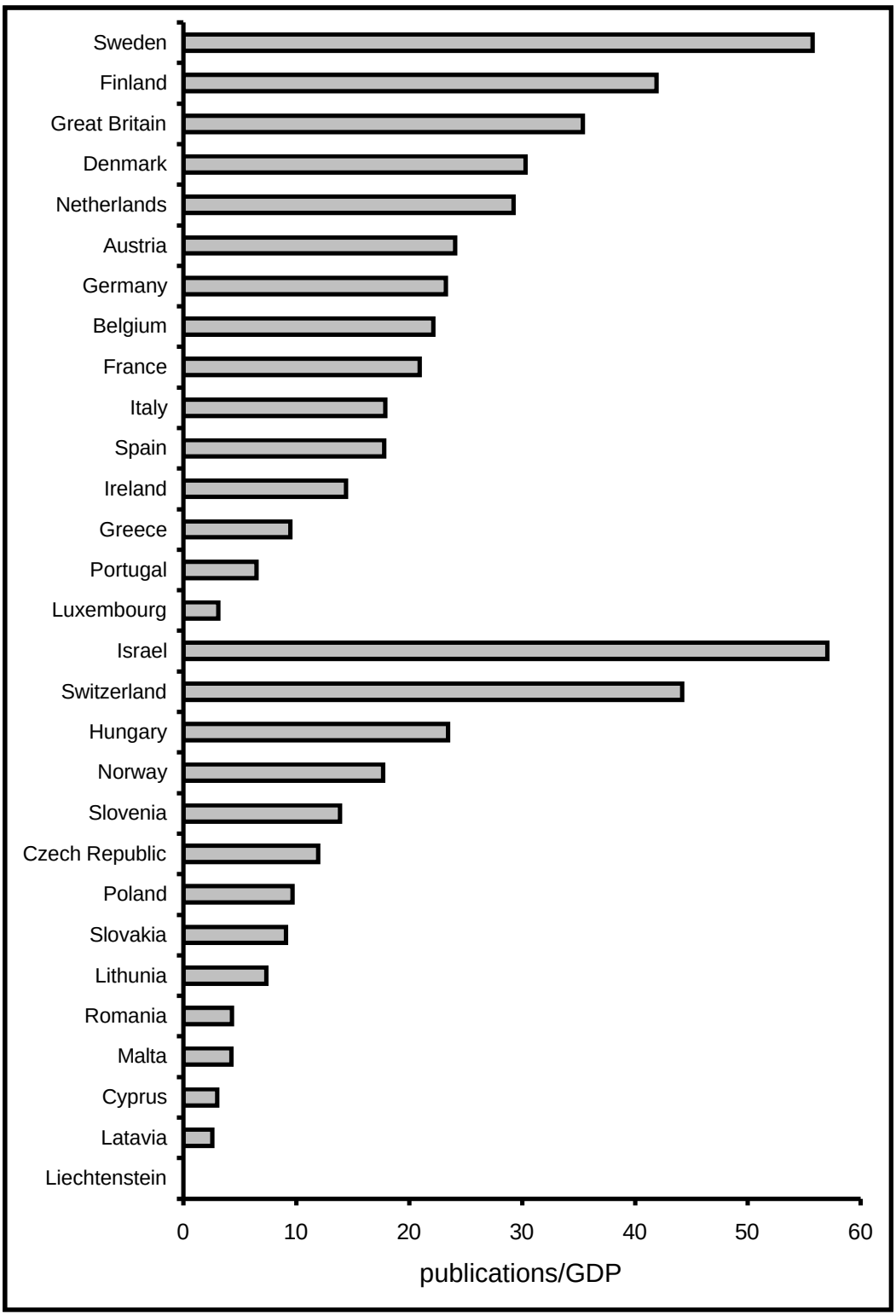
Source: SCI, computation of Fraunhofer ISI and CWTS

Figure 4.2-9 SCI publications with reference to GDP for selected countries in Genetics/heredity in the period 1995 to 2000 (GDP in billion € in purchasing power standards 2000)



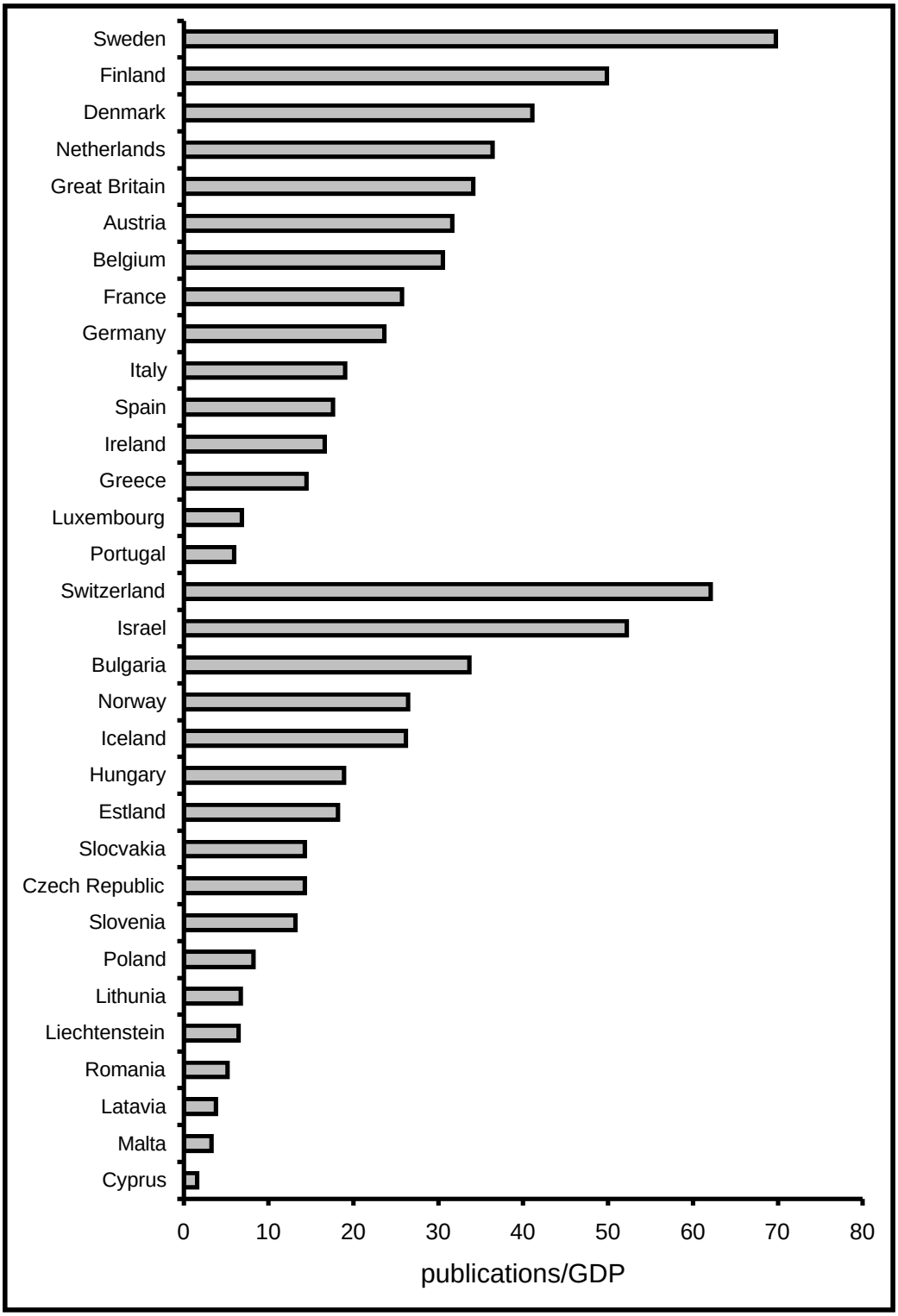
Source: SCI, computation of Fraunhofer ISI and CWTS

Figure 4.2-10: SCI publications with reference to GDP for selected countries in Neurosciences in the period 1995 to 2000 (GDP in billion € in purchasing power standards 2000)



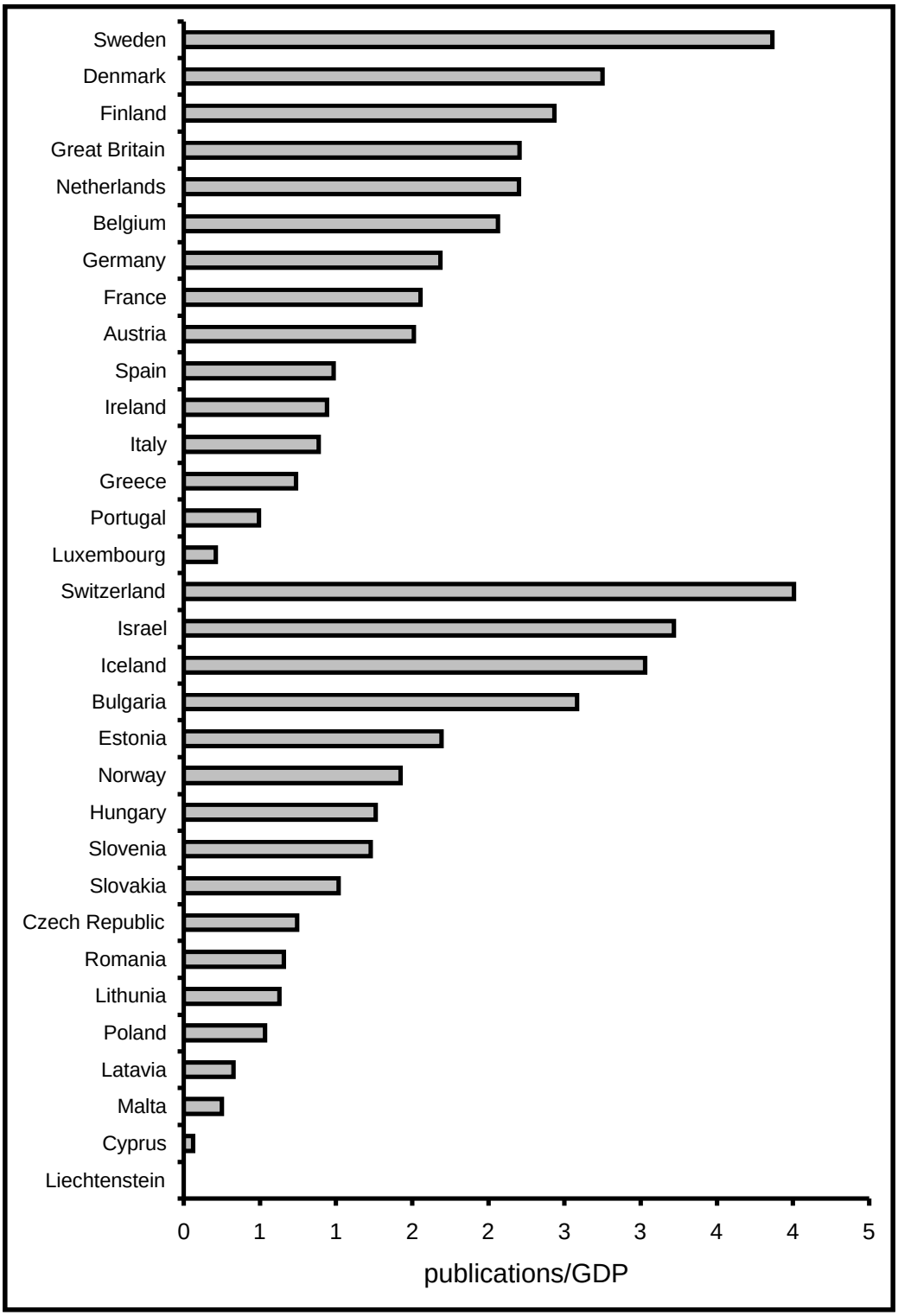
Source: SCI, computation of Fraunhofer ISI and CWTS

Figure 4.2-11: SCI publications with reference to GDP for selected countries in Immunology in the period 1995 to 2000 (GDP in billion € in purchasing power standards 2000)



Source: SCI, computation of Fraunhofer ISI and CWTS

Figure 4.2-12: SCI publications with reference to GDP for selected countries in Bioinformatics in the period 1995 to 2000 (GDP in billion € in purchasing power standards 2000)



Source: SCI, computation of Fraunhofer ISI and CWTS

4.3 Integrated analysis of publications and patents

In the two preceding sections, the performance of countries in terms of publications and patents has been discussed. This way of separate presentation does not indicate, however, whether there is any linkage between publication and patent performance, thus between scientific research and technological exploitation. The high share of public institutions in patents applications supports the hypothesis of a close link between science and technology. But the relevant publication activity of smaller countries without visible patent activity is in favor of the view that the linkage between science and technology is less strict, at least if applied to a larger set of countries.

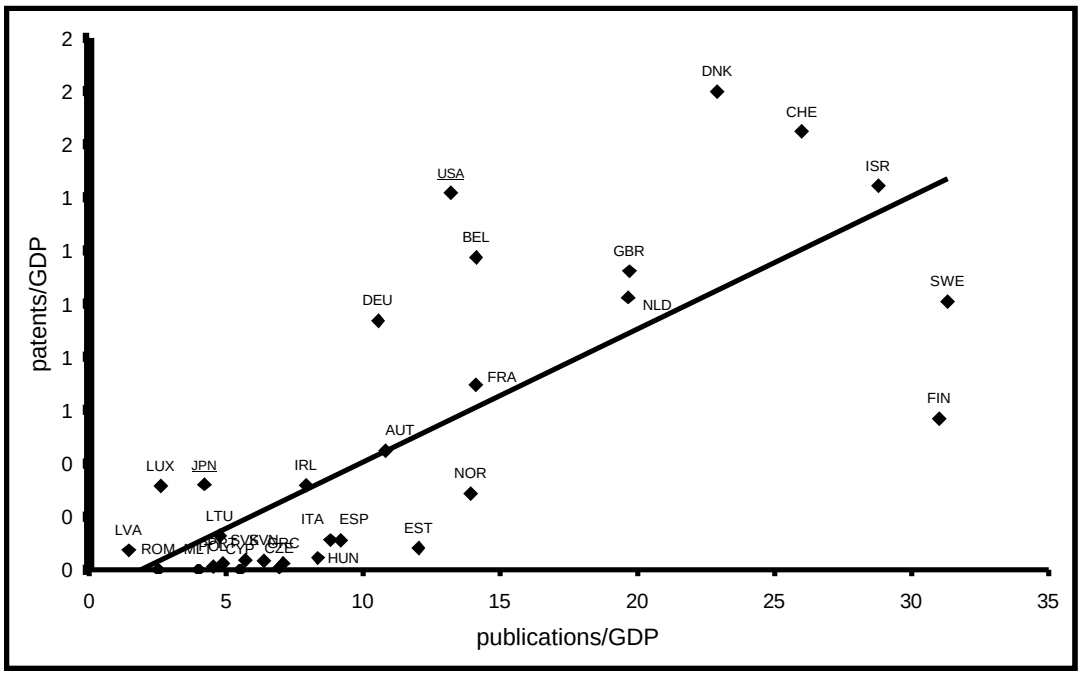
In this section, the interplay between publications and patents will be examined. The analysis does not refer to absolute values, because differential country size would automatically result in a high, but meaningless correlation coefficient, as large countries have many publications and patents and small countries have comparatively fewer publications and patents respectively. Therefore, the data are referred to GDP as a rough proxy to control for country size. The calculations are conducted in the following way:

- All EU and associated countries are included in the data set.
- Extreme outliers due to quotients with low figures are excluded.
- A regression analysis is calculated on the basis of the remaining countries.
- The United States and Japan are included in the graphs for illustration purposes, but not considered in the regression analysis.

As to the area of genetics/heredity, the plot of patents and publication (controlled for GDP) shows many countries with low patent and publication activities due to the inclusion of associated countries with less economic power (figure 4.3-1). Most of these weaker countries produce publications, but virtually no patents. Above a certain threshold of publications, also patents can be observed. This is to say, that countries with a high level of publications (controlled for GDP) display, too, high patent activities. Nevertheless, Finland has a strong orientation on publications, whereas Denmark or Switzerland show a focus on patents and thus technological exploitation.

The correlation was tested on the basis of various regression approaches. The linear and the quadratic trends proved to be the most appropriate ones. All in all, the correlation between the publication and patent intensities are high and highly significant. With R=0.80 in the case of a linear regression and R=0.82 in the case of a quadratic relation (Figure 4.3-1). It is interesting to note the strong orientation of the United States on technology which may be due to the fact that they are engaged in this area since a longer time than European countries. With an average relation of 0,050 between patent applications and publications for EU and associated countries, the orientation of genetics/heredity on technology is quite high, at least compared to the other areas considered in this study (Table 4.3-1).

Figure 4.3-1: Patent applications and publications in the area of Genetics/heredity for selected countries with reference to GDP(GDP in billion € in purchasing power standards 2000)



Source: SCI, EUREG, computation of Fraunhofer ISI and CWTS

Table 4.3-1: Regression coefficients between patents and publications with reference to GDP and relations of patents to publications for different areas

Area	linear regression	quadratic regression	patents/ publications
Genetics	0.80	0.82	0.050
Neurosciences	0.89	0.90	0.014
Immunology	0.88	0.88	0.028
Bioinformatics	0.53	0.55	0.008

Source: SCI, EUREG, computation of Fraunhofer ISI and CWTS

As to Neurosciences, the relation between publication and patent intensities is even closer than in genetics with R=0.89 for a linear regression. With a relation of patents to publications of 0.014 the orientation of the area on technology is distinctly weaker than in genetics, probably due to a stronger focus on medical research without a potential of technology application (Figure 4.3-2).

In Immunology, similar pattern as in the other two life science areas appear: The correlation coefficient R=0.88 (linear regression) is very high. Israel, Switzerland, Denmark, Belgium, Germany and the United States have an orientation on technology; Finland, Sweden, Austria, or Norway a stronger focus on publications (science) (Figure 4.3-3). In all cases, there is obviously a distinct impact of the scientific activity on the technological output (or vice versa). This outcome may be explained by the high direct participation of public research

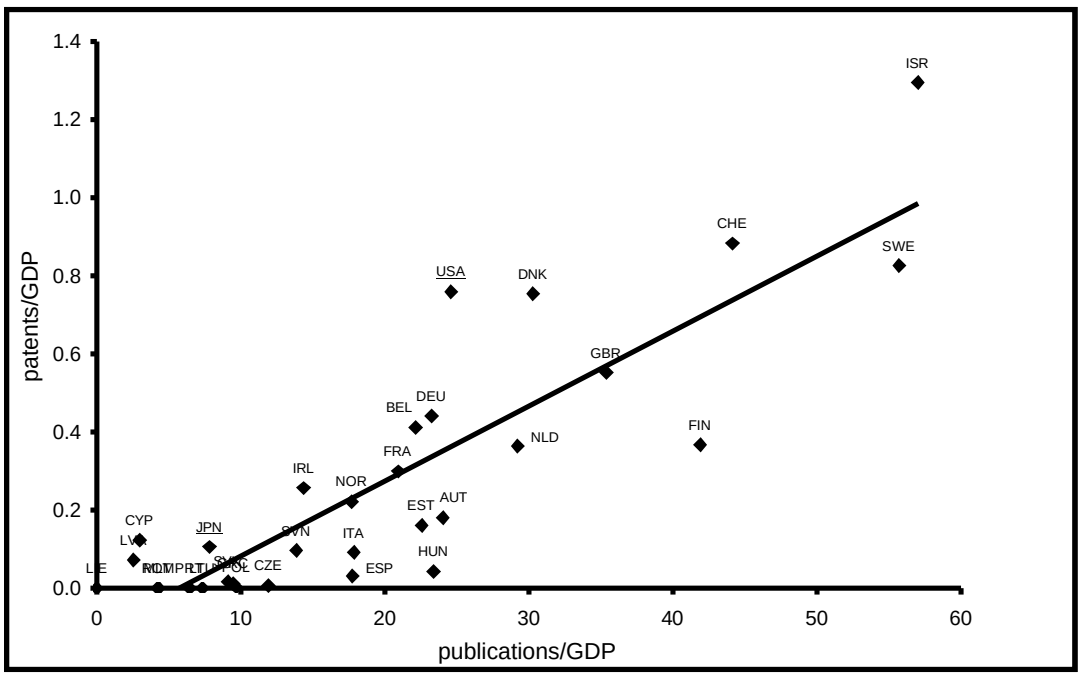
institutions in patents. But the very high correlation is, nevertheless, surprising, as the computation refers to intensities and not absolute patent and publication numbers.

The structures are less clear in the case of Bioinformatics. Sweden, Great Britain or the United States display relatively high patent intensities, whereas Finland, Israel or Switzerland have high publication intensities, but no patents (Figure 4.4-4). This may be due to different national cultures with regard to software patents, or an insufficient match of the search strategies of patents and publications compared to the other life science areas. In consequence, the correlation coefficient R=0.53 (linear regression) is relatively low and might be linked to other factors, in particular the general engagement of the countries in science and technology and not the specific structures in Bioinformatics. Due to the restrictions in the patenting of software and the early stage of Bioinformatics, the relation of patents to publication is very low with 0.008.

With a relation of 0.028 for patents and publications, the general orientation of the area of immunology appears to be on technology. For an adequate interpretation of this indicator, it has to be taken into account, that in the life sciences, a considerable share of publications refers to pure medical knowledge without a potential of technological exploitation.

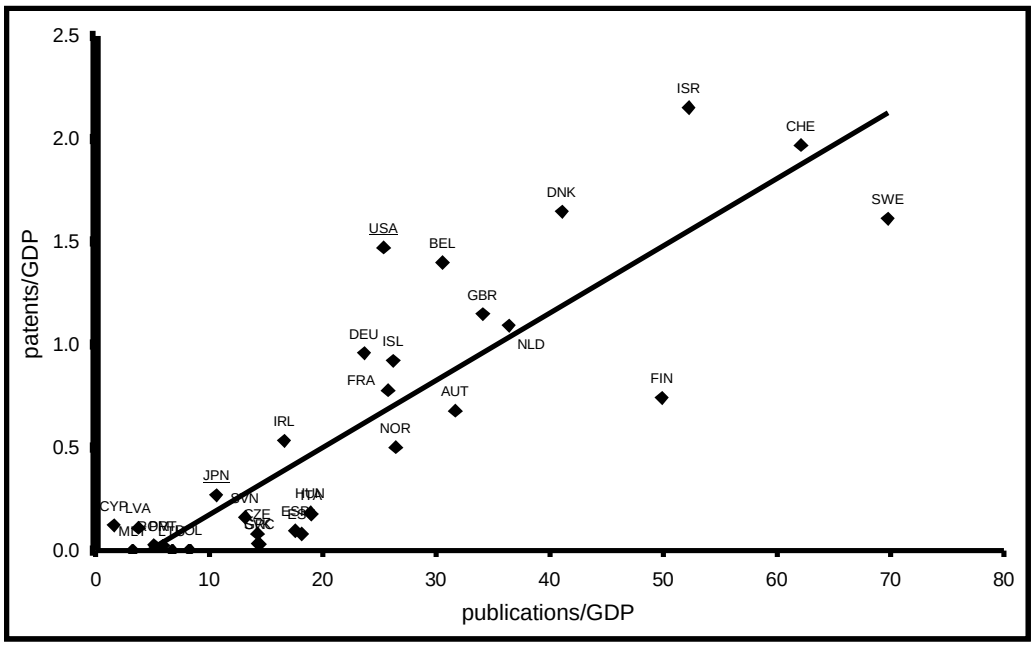
All in all, we can see a close correlation between the patent and publication intensities in the considered, research-intensive areas. Even if the technological competitiveness is considered as major political aim, the adequate support of the scientific basis is important. Against this background, the identification of centers of excellence in a scientific perspective proves to be a relevant approach, but it should be complemented by a technological analysis.

Figure 4.3-2: Patent applications and publications in the area of Neurosciences for selected countries with reference to GDP(GDP in billion € in purchasing power standards 2000)



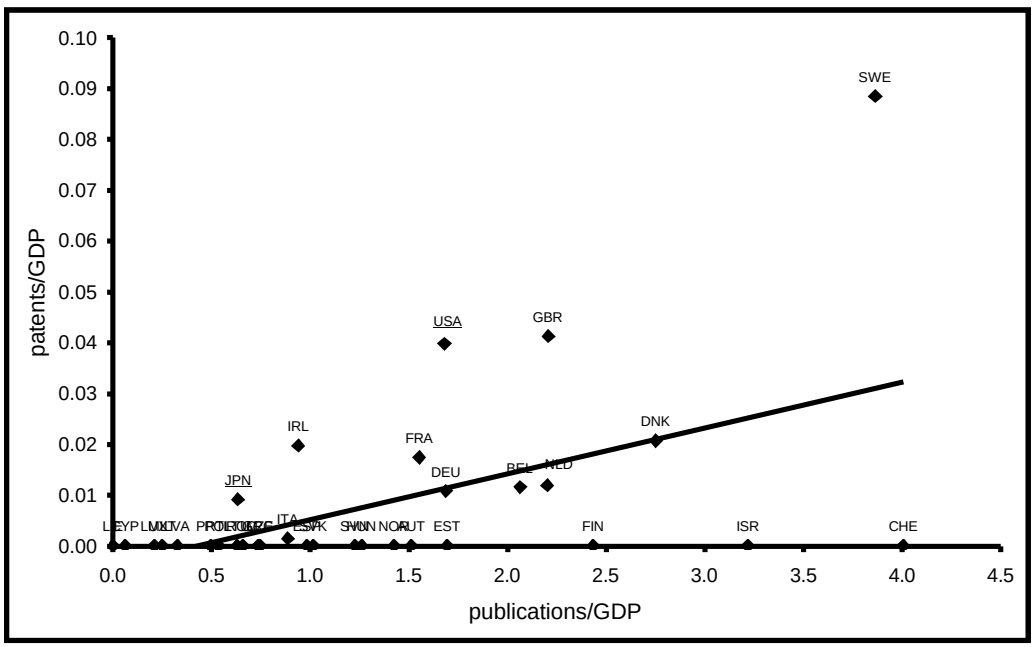
Source: SCI, EUREG, computation of Fraunhofer ISI and CWTS

Figure 4.3-3: Patent applications and publications in the area of Immunology for selected countries with reference to GDP(GDP in billion € in purchasing power standards 2000)



Source: SCI, EUREG, computation of Fraunhofer ISI and CWTS

Figure 4.3-4: Patent applications and publications in the area of Bioinformatics for selected countries with reference to GDP(GDP in billion € in purchasing power standards 2000)



Source: SCI, EUREG, computation of Fraunhofer ISI and CWTS

5 Conclusions and perspectives

In this study the main objective was to identify centers of excellence in selected fields of science. The question rises whether we were able to do that with the methodology used. The answer from our perspective to this question is yes. We are able to identify centers of excellence if we have adequate input from field experts. Still, in the course of this project we encountered some problems that should be solved in order to carry out the identification in a cost-effective way and with reliable results.

Expert input is crucial for collecting the proper publication and patent data to be used as the basis for the analyses and particularly the delineation of fields. At present it is not possible to collect these data without experts who are able to compile an effective search strategy to retrieve the relevant data. These experts should not only 'know' the field but also have knowledge about search strategies and their use. These experts can be supported by a search interface to see the effect of specific search strings.

In view of the difficulties we encountered with the input of experts in order to delineate the fields, we wonder whether this is the best approach to meet the objective. In practice, there are two types of 'fields'. One concerns a 'known' field, i.e., a field established for years and years (e.g., genetics, immunology, and neuroscience). For these fields, delineation is much easier because experts know the journals that cover the core of the field and they have ample experience using search terms. The second type is the 'developing' field (e.g., bioinformatics). In these fields there appears to be much less consensus among experts about what should and should not be covered, there are hardly or no journals specifically for that field, and the experts have a rather limited experience as to what search terms to use for collecting relevant data. Furthermore, the results for the 'known' fields are much easier to validate than the results for the 'developing' fields. As the results should in some way refer to what the field experts expect, validation appears to be considerably easier for the 'known' fields.

It should also be noted that field delineation on the basis of patents differs from publications, because patent databases contain a well-developed classification system (the International Patent Classification, IPC). This classification creates an additional and powerful facility to collect the proper data. The publication databases essentially lack such an overall generic scheme.

The above observations lead to the conclusion that implementation of our approach as described in this report on a larger scale (i.e., applying it to hundreds of 'fields') is not feasible, simply because we expect that it is impossible to get experts involved on such a large scale in a reasonable way, without losing control over the results. Moreover, we know that the science landscape is changing, and that particularly new and developing fields will attract interest to identify centers of excellence. But precisely in these developing fields delineation of he field on the basis of expert input is problematic, as discussed above. However, there are good prospects to deal with this delineation issue, but this has to be investigated in more detail. In principle it should be possible to start with a limited set of publications and to enlarge this set on the basis of co-citation relations, similar keyword patterns and other bibliometric characteristics.

With respect to the use of address data in publication and patent data, we conclude that they may be used at the level of 'main organization' (university, company, research institute) in most member states of the EU and associated states. At that level, cleaning of data by national experts is certainly feasible. It seems however, that problems with cleaning are not the same

in every country. Apart from the size of the country, it is well known that the science system in countries like France (particularly, the 'interwoveness' of the CNRS) differs considerably from the system in the Netherlands. The complexity of the system in France makes it almost impossible, also for national experts, to clean the data, even on the level of organization. Cleaning of these address data would be easier in a 'bottom-up approach'. This means that beforehand a limited list of organizations has to be compiled within each country. Then the address data could be cleaned using this basic list of organizations.

With respect to linking patent and publication indicators, we have made in this project a huge step forward as we were able to identify inventors as authors in the same field. Hence, we were able to identify the 'research address' of inventors and thus to build indicators for institutions having both patent and publication data. This enables us to find 'bridges' between scientific and technological performance within an R&D field.

In this project, we created a tool for different users to enter the fields chosen for this study. The design of this tool had to be flexible enough to be used by different types of users. Because of the variety of users (from scientific experts to policy makers), we are not yet completely able to determine whether the requirements of all users are satisfied. Still in view of the purposes of this study we are convinced that we indeed have. The tool enables users to determine their own criteria and thresholds to identify research entities of a certain productivity or impact. In particular, the possibility to combine different indicators enhances the utility of the tool for the different user groups considerably. On a large scale we were able to combine patent and publication indicators, which can be considered as a major step forward to explore the multiple aspects of excellence.

With respect to the size of research entities, we were within the scope of this project not able to go a step below the level of 'main organization', e.g., from university to department. It appeared that the quality of address data in publications on the level of departments and (if available) faculty, was so low, that we do not provide results on department level systematically. Moreover, the cleaning efforts for experts in the different national science systems would be huge. Especially in larger countries like Germany, France and the UK, we could not ask to clean the address data at any lower level than the main organization.

It should be noted that the activity and performance of these organizations are only measured within the field. The name of the organization as mentioned in the tables and rankings do not refer to the entire organization but only for the part active in a particular field.

Apart from these data problems, we mention the debate on the validity of performance indicators on the level of departments. For some purposes and within particular contexts, the entity to focus on should be even below the departments. In these cases the 'group' seems more appropriate. In this study we were not able explore this, but we have ideas as to how to deal with this. We suggest that combination of author names and organization name could be used effectively to define groups. A combination of groups may be used to define a department or even a faculty.

Still, as mentioned above, we were able to provide information on research in a specific field in an efficient interactive tool, enabling the user to use his/her own criteria and thresholds to identify research entities at the level of organization, with a particular performance. Moreover, we provide the tool at different levels of aggregation (world, EU, and national level).

The geographical interface can be used to localize the identified organizations. This enables a specific user to search for entities together with the information of its geographical position.

References

World of Learning (2000). 50th edition. London: Europa Publicatons Unlimited.

Airaghi, A., J. Viana Baptista, N.E. Bush, L. Georghiov, M.J. Ledoux, A.F.J. van Raan, and S. Kuhlmann. 'Options and Limits for Assessing the Socio-Economic Impact of European RTD Programmes', ETAN Working Paper, EUR 18884, Luxembourg: Office for Official Publications of the European Communities, 1999. (ISBN 92-828-3721-1).

Baratta, M. v. (ed.) (2001): Fischer Weltalmanach 2002. Frankfurt a.M.: Fischer Taschenbuch Verlag.

Barré, E. / Laville, F. / Schmoch, U. (2001): Indicateurs brevets. Les grands ensembles régionaux. La triade et les pays de l'UE. Report to the European Commission. Paris: OST.

Becher, G. / Gering, T. / Lang, O. / Schmoch, U. (1996): Patentwesen an Hochschulen, Bonn: BMBF.

Blind, K. / Edler, J. / Nack, R. / Straus, J. (2003): Softwarepatente. Eine empirische Analyze aus ökonomischer und juristischer Sicht. Heidelerg: Physica-Verlag.

Commission of the European Communities (CEC) (2001): How to map excellence in research and technological development in Europe. Commission staff working paper SEC(2001)434. Brussels: CEC.

Dybkaer, R. / Bauin, S. (2002): Final report on the preparation exercise of mapping of excellence in research and technological development in Europe in life sciences. Report to the European Commission. Brussels: European Commission.

European Patent Office (EPO) (2002): Annual Report 2002. Munich: EPO.

Eurostat (2002): R&D and Innovation Statistics in Candidate Countries and the Russian Federation. Data 1990-99. European Communities. Luxembourg

Hoffmann-La Roche AG / Urban & Schwarzenberg (ed.) (1998): Roche Lexikon Medizin. München: Urban & Schwarzenberg.

Horrobin, D.F. (1990). The philosophical basis of peer review and the suppression of innovation. *Journal of the American Medical Association (JAMA)* 263, 1438-1441.

Laredo, P. (1999): The development of a reproducible method for the characterisation of a large set of research collectives. Report to the European Commission. Armines: CSI

Meyer, M. / Perrson, O. / Power, Y. (2001):. Mapping excellence in nanotechnologies. Preparatory study (Nanotechnology expert group and Eurotech data). Report to the European Commission. Brussels: European Commission.

Moed, H.F. and Th.N. van Leeuwen (1995). Improving the accuracy of the Institute for Scientific Information's Journal Impact Factors. *Journal of the American Society for Information Science* (JASIS) 46, 461-467.

Moed, H.F. and Th.N. van Leeuwen (1996). Impact Factors Can Mislead. *Nature* 381, 186.

Moed, H.F., De Bruin, R.E., and Th.N. Van Leeuwen (1995). New Bibliometric Tools for the Assessment of National Research Performance: Database Description, Overview of Indicators and First Applications. *Scientometrics* 33, 381-422.

Moxham, H. and J. Anderson (1992). Peer review. A view from the inside. *Science and Technology Policy* 7-15.

Noyons, E.C.M. and A.F.J. van Raan (1998). Monitoring Scientific Developments from a Dynamic Perspective: Self-Organized Structuring to Map Neural Network Research. *Journal of the American Society for Information Science and Technology (JASIST)*, 49, 68-81.

Noyons, E.C.M., M. Luwel and H.F. Moed (1999). Combining Mapping and Citation Analysis for Evaluative Bibliometric Purpose. A Bibliometric Study on Recent Development in Micro-Electronics. *Journal of the American Society for Information Science and Technology (JASIST)*, 50, 115-131.

Noyons, E.C.M. (1999). *Bibliometric mapping as a science policy and research management tool*. Thesis Leiden University. Leiden: DSWO Press (ISBN 90-6695-152-4).

Noyons, E.C.M., R.K. Buter and A.F.J. van Raan (2000). *Mapping the field of Neuroscience.* Electronic version with interactive facilities available via www.cwts.leidenuniv.nl .

Organization for Economic Co-operation and Development (OECD) (2003): Turning Science into Business: Patenting and licensing at public research organizations. DISTI/STP(2003)22. Paris: OECD.

Salter, J., Martin, B. (2001): The economic benefits of publicly funded basic research: a critical review, in: Research Polocy, vol. 30, pp. 509-532.

Schmoch, U. (1999a): Eigenen sich Patente als Innovationsindikatoren?, in: Boch, R. (ed.), Patentschutz und Innovation in Geschichte und Gegenwart, Frankfurt a. M., Berlin, Bern u.a.: Peter Lang, pp. 113-126.

Schmoch, U. (1999b): Impact of International Patent Applications on Patent Indicators, in: Research Evaluation, vol. 8, no. 2, pp. 119-131.

Schmoch, U. (2003): Akademische Forschung und industrielle Forschung. Pertspektiven der Uinteraktion. Frankfurt a.M. / New York: Campus-Verlag.

Schmoch, U., Licht, G., Reinhard, M. (ed.) (2000): Wissens- und Technologietranfer in Deutschland. Stuttgart: Fraunhofer-IRB Verlag.

Statistisches Bundesamt (ed.) (2002): Statistisches Jahrbuch 2002 für das Ausland. Wiesbaden.

The Worldbank Group (2002):World Development Indicators 2002. Access via: www.worldbank.org/data/wdi2002/Tables/Table1-6.pdf.

Van Leeuwen, T.N., M.S. Visser, H.F. Moed, A.J. Nederhof, and A.F.J. van Raan (2003). The Holy Grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence. *Scientometrics*, 57, 257-280.

van Raan, A.F.J. (1996). Advanced Bibliometric Methods as Quantitative Core of Peer Review Based Evaluation and Foresight Exercises. *Scientometrics* 36, 397-420.

van Raan, A.F.J. (2000a). The Pandora's Box of Citation Analysis: Measuring Scientific Excellence, the Last Evil? In: B. Cronin and H. Barskt Atkins (eds.). *The Web of Knowledge. A Festschrift in Honor of Eugene Garfield*. Ch. 15, p. 301-319. Medford (New Jersey): ASIS Monograph Series, 2000 (ISBN 1-57387-099-4).

van Raan, A.F.J. (2000b). The Interdisciplinary Nature of Science. Theoretical Framework and Bibliometric-Empirical Approach. In: P. Weingart and N. Stehr (Eds.). *Practicing Interdisciplinarity*. Toronto: University of Toronto Press (ISBN 0-8020-8139-8).

Van Raan, A.F.J. and Th. N. van Leeuwen (2001). *Identifying the Fields for Mapping RTD Excellence in Life Sciences, First Approach*. Brussels: Report to the European Commission, available on website http://europa.eu.int/comm/research/era/pdf/lifesciences-mapexcel.pdf .

Van Raan, A.F.J and E.C.M. Noyons (2002). Discovery of patterns of scientific and technological development and knowledge transfer. In: W. Adamczak and A. Nase (eds.): *Gaining Insight from Research Information*. Proceedings of the 6th International Conference on Current Research Information Systems, University of Kassel, August 29-31, 2002. Kassel: University Press. Page 105- 112 (ISBN 3-933146-844).

World Health Organization (WHO) (1993): International Statistical Classification of Diseases and Relatated Health Problems (ICD), 10th revision, reprinted with corrections 1996. Geneva.

World Intellectual Property Organization (WIPO) (2000): International Patent Classification. Seventh edition. München: Carl Heymanns Verlag.

Annexes

Appendix A: Bibliometric indicators

A1: Details of the bibliometric methodology

Research output is defined as the number of articles of the institute, as far as covered by the Science Citation Index (SCI), the Social Science Citation Index (SSCI), or the Arts & Humanities Citation Index (AHCI). As 'article' we consider the following publication-types: normal articles (including proceedings papers published in journals), letters, notes, and reviews (but not meeting abstracts, obituaries, corrections, editorials, etc.). We developed software to calculate a set of standardized, basic indicators.

To discuss this set of indicators, we take the results of our recent analysis of a German medical research institute as an example (time period 1992 – 2000). Table 1 shows in the first column the number of papers published, ***P***, which is also a first but good indication of the size of an institute. This number is about 250 per year. In the second column we find the total number of citations, ***C***, received by ***P*** in the indicated time period, *and corrected for self-citations.*

The analytic scheme is as follows. We take the last sub-period 1996-2000 as an example. For papers published in 1996, citations are counted during the period 1996-2000, for 1997 papers citations in 1997-2000, and so on. There is ample empirical evidence that in the natural and life sciences -basic as well as applied- the average 'peak' in the number of citations is in the third or fourth year after publication (Moed et al 1995). Therefore a ('moving' and partially overlapping) five-year analysis period is appropriate for impact assessment.

The third and fourth indicators are the average number of citations per publication ***(CPP)***, again without self-citations, and the percentage of not-cited papers, ***% Pnc***. We stress that this percentage of non-cited papers concerns, like all other indicators, the given time period. It is very well possible that publications not cited within such a block will be cited after a longer time. This is clearly visible when comparing this indicator for the five-year periods (e.g., 1996-2000: 30%) with that of the whole (that is, longer) period (1992-2000: 21%). The values found for this medical research institute are quite normal.

How do we know that a certain volume of citations, or a certain citation-per-publication value is low or high? Therefore it is crucial to make a comparison with (or normalization to) a well-chosen international reference value, and to establish a reliable measure of *relative, internationally field-normalized impact*. Furthermore, as overall, worldwide citation rates are increasing, it is also necessary to normalize the measured impact of an institute (***CPP***) to international reference values.

First, we calculate the average citation rate of all papers (world-wide) in the journals in which the institute has published (***JCSm***, the **m**ean **J**ournal **C**itation **S**core of the institute's 'journal set'). Thus, this indicator ***JCSm*** defines a worldwide reference level for the citation rate of the institute. It is calculated in the same way as ***CPP***, but now for all publications in a set of journals (see van Raan 1996). With help of the ratio ***CPP/JCSm*** (5^{th} indicator) we observe whether the measured impact is *above* or *below* international average.

Table 1: Bibliometric analysis of a medical research institute 1992 – 2000

period	*P*	*C*	*CPP*	*%Pnc*	*CPP/ JCSm*	*CPP/ FCSm*	*CPP/ D-FCSm*	*JCSm/ FCSm*	*%SCit*
1992 - 00	**2,245**	**43,665**	**19.45**	**21**	**1.26**	**1.95**	**1.85**	**1.55**	**18**
1992 – 96	1,080	11,151	10.33	36	1.27	2.02	1.95	1.58	22
1993 – 97	1,198	12,794	10.68	34	1.24	2.03	1.92	1.63	21
1994 – 98	1,261	12,217	9.69	32	1.19	1.85	1.72	1.55	22
1995 – 99	1,350	13,709	10.15	31	1.21	1.89	1.76	1.56	21
1996 – 00	1,410	14,815	10.51	30	1.20	1.91	1.76	1.59	21

Comparison of the institute's citation rate ***(CPP)*** with the average citation rate of its journal set (***JCSm***) introduces a specific problem related to journal status. For instance, if the institute publishes in prestigious (high impact) journals, and another institute in rather mediocre journals, the citation rate of articles published by both groups may be equal *relative to* the average citation rate of their respective journal sets. But the first group evidently performs better than the second. Therefore, we developed a second international reference level, a *field*-based world average ***FCSm***. This indicator is based on the citation rate of *all* papers (world-wide) published in *all* journals of the field(s)[21] in which the institute is active, and not only the journals in which the institute's researchers publish their papers. For a publication in a less prestigious journal one may have a (relatively) high ***CPP/JCSm*** but a lower ***CPP/FCSm,*** and for a publication in a more prestigious journal one may expect a higher ***CPP/FCSm***, as publications in a prestigious journal will have generally have an impact above the field-specific average.

We use the same procedure as the one we applied in the calculation of ***JCSm***. A novel and unique aspect of our comparison with both worldwide reference indicators is that we take into account the type of paper (e.g., letters, normal article, review) *as well as* the specific years in which the papers were published. This is absolutely necessary, as the average impact of journals may have considerable annual fluctuations and large differences per article type, see Moed and Van Leeuwen 1995, 1996).

Often an institute is active in more than one field. In such cases we calculate a weighted average value, the weights being determined by the total number of papers published by the institute in each field. For instance, if the institute publishes in journals belonging to genetics and heredity, as well as to cell biology, then the ***FCSm*** of this institute will be based on both field averages. Thus, indicator ***FCSm*** represents a *world average*[22] in a specific (combination of) field(s). It is also possible to calculate ***FCSm*** for a specific country or for the European Union. The example discussed in this paper concerns a German medical research institute and for this institute we calculated the Germany-specific ***FCSm***-value, ***D-FCSm***.

As in the case of ***CPP/JCSm***, if the ratio ***CPP/FCSm*** (6th indicator) is above 1.0, the impact of the institute's papers exceeds the field-based (i.e., *all* journals in the field) world average. We observe in Table 1 that the ***CPP/JCSm*** is 1.20, ***CPP/FCSm*** 1.91 and ***CPP/D-FCSm*** (7th indicator) is 1.76 in the last period 1996 – 2000. These results show that the institute is performing well above international average. The ratio ***JCSm/FCSm*** (8th indicator) is also an interesting indicator. Is it above 1.0, the mean citation score of the institute's journal set exceeds the mean citation score of all papers published in the field(s) to which the journals

[21] We use the definition of fields based on a classification of scientific journals into *categories* developed by ISI. Although this classification is not perfect, it is at present the most suitable classification available to us in terms of an automated procedure within our data-system.

[22] About 80 percent of all SCI-covered papers is authored by scientists from the United States, Western Europe, Japan, Canada, and Australia. Therefore, our 'world average' is dominated by the Western world.

belong. For the institute this ratio is around 1.59. This means that the institute publishes in journals with, generally, a high impact. The last (9^{th}) indicator shows the percentages of self-citations (***%Scit***). About thirty percent is normal, so the self-citation rates for this institute are certainly not high (about 20%).

We regard the *internationally standardized impact indicator* ***CPP/FCSm*** as our 'crown' indicator. This indicator enables us to observe immediately whether the performance of a research group or institute is significantly far below (indicator value < 0.5), below (indicator value 0.5 - 0.8), around (0.8 - 1.2), above (1.2 – 1.5), or far above (>1.5) the international (western world dominated) impact standard of the field. We stress that in the measurement of scientific impact one has to take into account the *aggregation level of the entity* under study. The higher the aggregation level, the larger the volume in publications and the more difficult it is to have an impact significantly above the international level. Based on our long-standing experiences, we can say the following. At the 'meso-level' (e.g., a large institute), a ***CPP/FCSm*** value above 1.2 means that the institute's impact as a whole is significantly above (western-) world average.

Particularly with a ***CPP/FCSm*** value above 1.5, such as in our example, the institute can be considered as scientifically strong, with a high probability to find very good to excellent groups. Thus, the next step in a research performance analysis is a breakdown of the institution into smaller units, i.e., research groups and/or programs. Therefore the bibliometric analysis has to be applied on the basis of institutional input data on personnel and composition of groups.

Then, the bibliometric algorithms can be repeated efficiently on the lowest but most important aggregation level, that of the research group or research program. In most cases the volume of publications at this level is between 10 and 20 per year. At the group level a ***CPP/FCSm*** value above 2 indicates a very strong group, and above 3 the groups can be, generally, considered as excellent and comparable to top-groups at the best US universities. If the threshold value for the ***CPP/FCSm*** indicator is set at 3.0, we filter out the excellent groups with high probability.

As an additional indicator of scientific excellence, we determines for the target entity the number of publications within the *top-10%* of the worldwide impact distribution of the field concerned (***P10***)

For all above indicators we also perform a breakdown into types of *scientific co-operation* according to the publication addresses: work by only the unit itself; in a national collaboration; or in an international collaboration.

A further important part of our bibliometric methodology is the *breakdown* of the institute's output *into* research fields. This provides a clear impression of the research scope or 'profile' of the institute. Such a *spectral analysis* of the output is based on the simple fact that the researchers publish in journals of many different fields. Our example, the German medical research institute, is a center for broad, medical science oriented, molecular research. The researchers of this institute are working in a typical interdisciplinary environment. The institute's publications are published in a wide range of fields: biochemistry and molecular biology, genetics and heredity, oncology, cell biology, and so on.

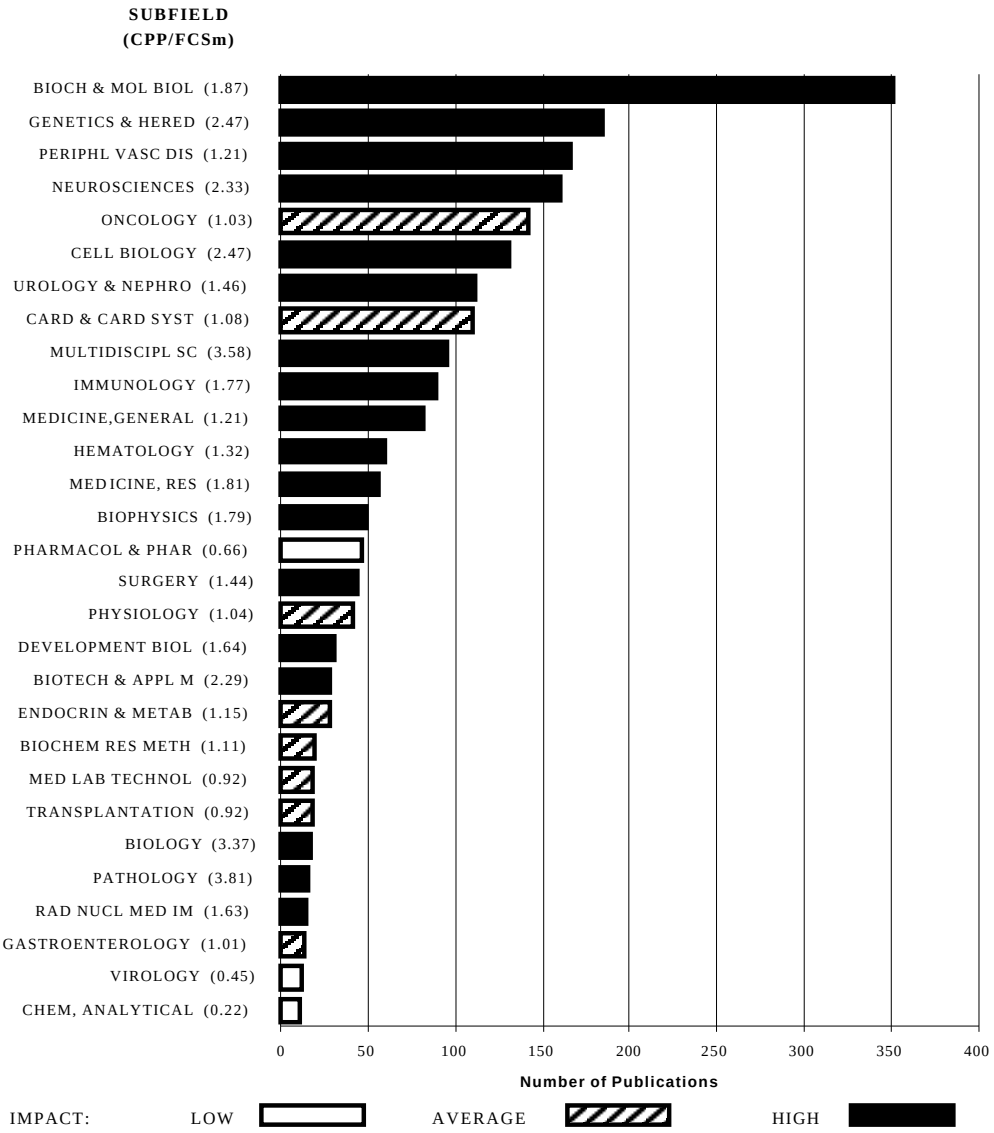
By ranking fields according to their size (in terms of numbers of publications) in a graphical display, we construct the research profile of the institute. Furthermore, we provide the *impact* of the institute's research in these different fields with help of ***CPP/FCSm*** as impact indicator normalized for *each the fields separately*. Figure 1 shows the results of this *bibliometric*

spectroscopy. Thus it becomes immediately visible in which fields within its interdisciplinary research profile the institute has a high (or lower) performance (van Raan 2000b).

In Fig. 1 we observe the scientific strength of the target institute: its performance in the top-four fields is high to very high. If we find a smaller field with a relatively low impact (i.e., a field in the lower part, the 'tail' of the profile), this does not necessarily means that the (few) publications of the institute in this particular field are 'bad'. Often these small fields in a profile are those that are quite 'remote' from the institute's core fields. They are, so to say, peripheral fields. In such a case, the group's researchers may not belong to the dominating international research community of those fields, and as the consequence their work will be not be cited as frequently as the work of these dominating ('card holding') community members.

In a similar way, we construct a profile of *the citing publications* into fields of science, i.e., the *'users'* of scientific results (as far as represented by citing publications). This 'knowledge users' profile is a powerful indicator of *who* is using *which* research results, *where* (in which fields) and *when*. Thus it analyses *knowledge diffusion and knowledge use* and it indicates further interdisciplinary 'bridges', potential collaboration, and possible 'markets' in the case of applied research.

Figure 1: Research profile of medical research institute, 1992 -2000



A2. Timeliness of the bibliometric method

Example of 'process time' (in terms of age) of publication, references, citations

Publication in Physical Review Letters, vol. 88, page 138701, year of publication 2002 (April, 1)

Cited Articles:

S. H. Strogatz, Nature **410**, 268 (2001).
R. Albert and A. -L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
S. N. Dorogovtsev and J. F. F. Mendes, Adv. Phys. (to be published).
R. Albert, H. Jeong, and A. -L. Barabási, Nature **401**, 130 (1999).
B. A. Huberman and L . A. Adamic, Nature **401**, 131 (1999);
R. Kumar *et al.*, in *Proc. of the 25th Int. Conf. on Very Large Databases* (Morgan Kaufmann Publ., San Francisco, 1999), p. 639;
A. Broder *et al.*, Comput. Netw. **33**, 309 (2000);
P. L. Krapivsky, S. Redner, and F. Leyvr az, Phys. Rev. Lett. **85**, 4629 (2000);
S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, . Phys. Rev. Lett. **85**, 4633 (2000);
A. Vazquez, Europhys. Lett. **54**, 430 (2001).
M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999);
G. Caldarelli, R. Marchetti, and L. Pietronero, Europhys. Lett. **52**, 386 (2000);
A. Medina, I. Matta, and J. Byers, Comput. Commun. Rev. **30**, 18 (2000);
R. Pastor -Satorras, A. Vazquez, and A. Vespignani, arXiv:cond -mat/0105161;
L. A. Adamic *et al.*, Phys. Rev. E **64**, 046135 (2001).
F. B. Cohen, *A Short Course on Computer Viruses* (Wiley, New York, 1994);
R. Pastor -Satorras and A. Vespignagni, Phys. Rev. Lett. **86**, 3200 (2001);
R. Pastor -Satorras and A. Vespignagni, Phys. Rev. E **63**, 066117 (2001);
F. Liljeros, C. R. Edling, L. A. Nunes Amaral, H. E. Stanley, and Y. Åberg, Nature **411**, 907 (2001).
A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
Y. Ijiri and H. A. Simon, *Skew Distributions and the Sizes of Business Firms* (North -Holland, Amsterdam, 1977).
G. Bianconi and A. -L. Barabasi, Europhys. Lett. **54**, 436 (2001).
A. F. J. Van Raan, Scientometrics **47**, 347 (2000).
L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, Proc. Natl. Acad. Sci. U.S.A. **97**, 11 149 (2000).
H. A. Simon, *Models of Bounded Rationality: Empirically Grounded Economic Reason* (MIT Press, Cambridge, 1997).

The first citing articles are in the same year as the cited publication, examples:

Marc Barthélemy *et al.*, Phys. Rev. E **66**, 056110 (2002)
Petter Holme, Phys. Rev. E **66**, 036119 (2002)
Holger Ebel *et al.*, Phys. Rev. E **66**, 035103 (2002)
Haijun Zhou, Phys. Rev. E **66**, 016125 (2002)

Appendix B: Cognitive mapping methodology

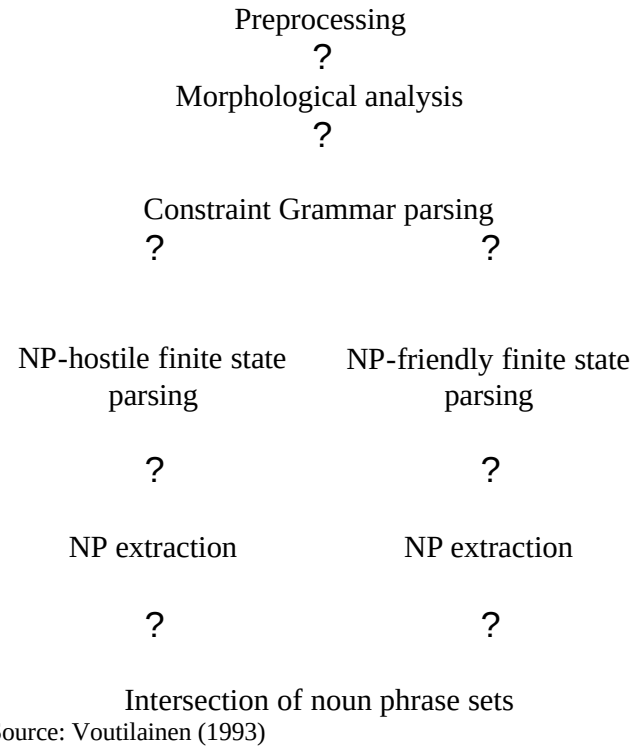
The main aspects of the general methodology of cognitive bibliometric mapping are outlined. They concern the selection of field keywords, the clustering of these keywords (identification of sub-domains), and the two-dimensional scaling of the identified field sub-domains.

Field keyword selection

The selection of keywords to structure the bibliographic database, representin g the field under study, is a procedure based on four word characteristics:

- lexical features;
- linguistic characteristics;
- bibliometric distribution;
- semantic scope.

First of all, only noun phrases (NPs) can become field keywords. In order to identify noun phrases in English texts, we use a 'noun phrase extractor' developed by Connexor OY in Finland (Englite). The process from text to NPs is described in the scheme below.

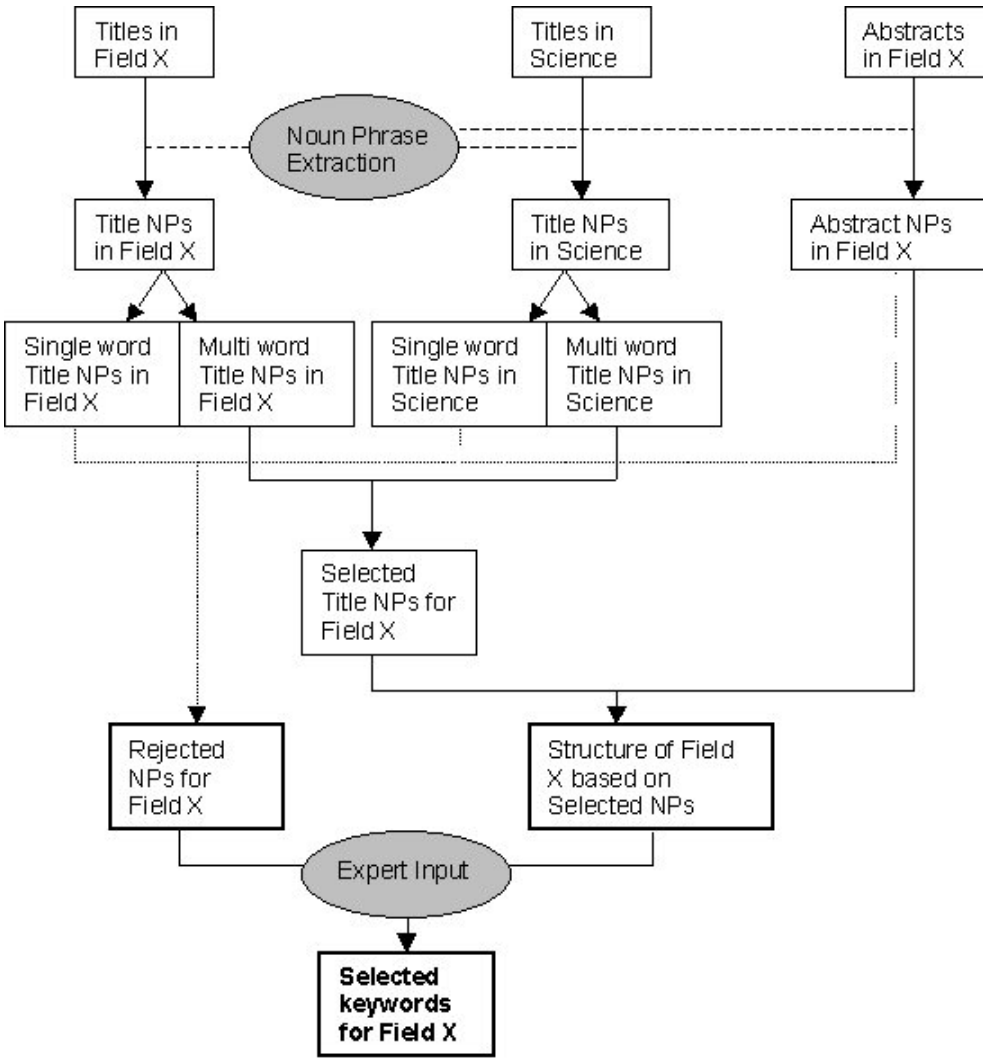


Source: Voutilainen (1993)

Englite system flowchart

The identified NPs are divided into two groups: the single word NPs (SWNP) and the multiword NPs (MWNP). At first, only a MWNP becomes a field keyword. From the list of MWNPs a list of phrases is withdrawn because they are used in text primari ly for other reasons than to describe contents of research. Their semantic scope is outside scientific research. In the near future, we will be able to rule out such phrases on the basis of their bibliometric distribution within science.

Furthermore, some minor unification is conducted to words and phrases for efficiency and esthetical reasons. It concerns unification of plural to singular form not identified by Englite, and unification of full terms to acronyms and abbreviations. In each project this list is adjusted, because word unification in one field can have unwanted effects in another.



Field keyword selection flowchart

The selection of field keywords from the list of candidates is presently established on the basis of their bibliometric distri bution, and (if possible) the input of a field expert. For

each MWNP, we count the number of appearances in titles on the one hand, and in abstracts on the other within the field under study, as well as the number of appearances in titles in science as a whole. A combination of these three figures, indicates both the specificity of the NP within the field and its 'centrality' within the field. The expert input is directed at two sources of information:

- the excluded list of single word noun phrases (SWNPs) ;
- the preliminary list of selected field keywords within its cognitive context (see next section).

By using an 'on-line' feedback form, the field expert is able to remove preliminary selected keywords or to add preliminary excluded NPs from the two lists . A flowchart of the selection procedure is depicted below.

Keyword clustering and sub-domain identification

In order to identify sub-domains within a field, the selected keywords are clustered into groups on the basis of their similar cognitive orientati on. At first, this cluster structure is used to provide a cognitive context for each preliminary selected keyword. In the final stage, this cluster structure is used to delineate sub-domains within the field.

The clusters are identified by a cluster analys is on a normalized co-occurrence matrix. The matrix is composed by the keyword co-occurrences in the set of publications defining the field in a certain period of time. It will depend on the aims of a project, which period of time this is. The co-occurrence matrix looks like the example below, where each cell contains the number of times that the row and the column item appear together in a publication.

	keyword #1	keyword #2	keyword #3	…
keyword #1	100	30	80	…
keyword #2	30	50	0	…
keyword #3	80	0	100	…
…	…	…	…	…

Keyword co-occurrence matrix sample

In the above sample, *keyword #1* appears in 100 publications. In 30 publications it co-occurs with *keyword #2*, which appears in 50 publications.

The 'raw data' matrix is normalized in such a way that the similarity of keywords is no longer based on the pair-wise co-occurrences, but rather on the cognitive orientation of two keywords in relation to all other keywords. The similarity is thus calculated on the co-occurrence profiles of *keyword #1* and *keyword #2* with all other keywords. In other words the vector of *#1* and *#2*, as defined by the co-occurrences with other keywords, is compared. The similarity is calculated on the based of the cosine index (the cosine of vectors):

$$sim(x,y) = \frac{\sum_i (x_i y_i)}{\sqrt{((\sum_i x_i^2)(\sum_i y_i^2))}}$$

Cosine of co-occurrence vectors

The recalculated matrix is input for a cluster analysis. In most cases, we use a hierarchical cluster algorithm with complete linkage.

The number of clusters to be formed is determined by combining three criteria provided by SAS ® (local peak of the cubic clustering criterion (CCC) and the Pseudo F statistic, together with a low Pseudo T^2 and a much higher Pseudo T^2 at the next cluster fusion). Of course, the determination of the number of clusters is related to the issue addressed in the mapping study. If a very coarse structuring of a field is required, a high number of clusters seems not appropriate. It should be noted that any number of clusters represents a structure of the field. In other words: both a structure based on 5 clusters as well as a structure based on 50 clusters is able to represent a field. In the former case, however, certain details provided by the latter may are not revealed.

The identified clusters of keywords represent field sub-domains. These sub-domains are labeled with a name by the four most frequent keywords in a cluster.

Mapping sub-domains by MDS

In order to build a map of the field, the sub-domains are positioned in a two-dimensional space. Each sub-domain represents publications on the basis of keyword occurrence. If any of the keywords is in a publication, it will be attached to the sub-domain to which the keywords belong. Thus, publications may be attached to more than one sub-domain. The overlap between sub-domains can be used to create a co-occurrence matrix (c.f., keyword co-occurrence matrix in the previous section). A normalization of the sub-domain co-occurrences is performed is established by a cosine similarity matrix (c.f., previous section). The sub-domains are positioned in two dimensions by multidimensional scaling (MDS: ordinal). The resulting field map renders the cognitive similarity of sub-domains as measured by the distance between them. The distance is determined by the cognitive orientation of sub-domains in relation to all other, in such a way, that sub-domains with a similar cognitive profile are in each other's vicinity, and sub-domains with different orientation are distant from each other.

The map provides information about the number of publications represented by a sub-domain as well. The size of a sub-domain (the surface of a circle) indicates the share of publications represented in relation to the full number in the whole field.

Finally, the pair-wise cognitive relation between two individual sub-domains is indicated by a connecting line. As the whole map is a representation of similarity between sub-domains (taking into account all co-occurrence relations), connecting lines enhance the structure by emphasizing pair-wise relations. A pair-wise relation is measured by a normalized similarity of the Salton index.

$$sim(x,y) = \frac{C_{xy}}{\sqrt{C_x C_y}}$$

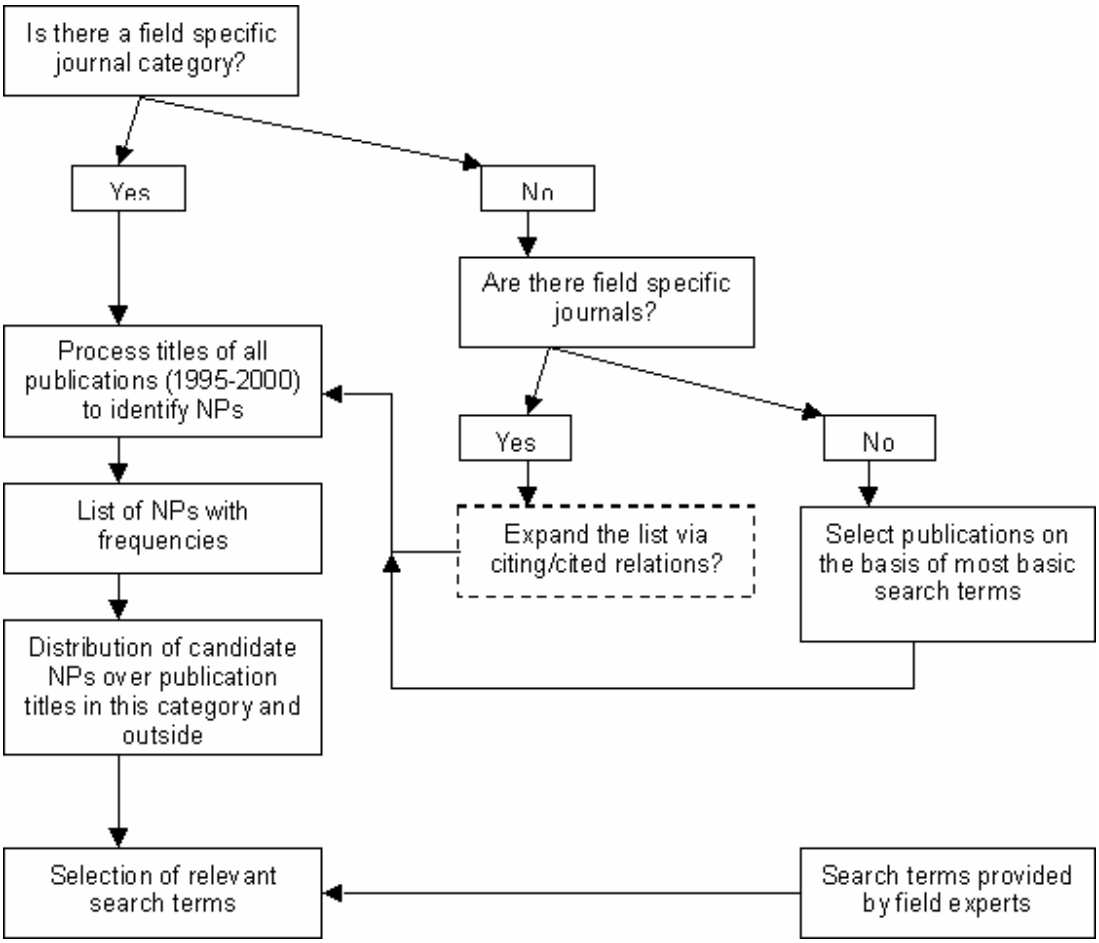
Where:
C_{xy} is the number of co -occurrences of x and y;
C_x is the number of occurrences of x; and
C_y is the number of occurrences of y.

Salton index

References

Voutilainen, A. (1993), 'NPtool, a Detector of English Noun Phrases'. In: *Proceedings of the Workshop on Very Large Corpera 1993* . Ohio State University, Columbus Ohio.

Appendix C: Flow chart of the field delineation procedure



Note: Dashed boxes are optional

Appendix D: Field delineations

Bioinformatics

Search terms:

bioinform* OR
comput* and biol* OR
(datamining or data mining) and bio* data* OR
sequen* analys* OR
(high throughput or large scale) and bio* data* OR
(microarray* or chip*) and analys* OR
rational drug design OR
model* near protein structur* OR
genome data OR
macromolecul* sequen* or macromolec* structur* OR
comput* and (genom* or proteom*)

Genetics & Heredity

Journals:

ACTA GENETICAE MEDICAE ET GEMELLOLOGIAE
AMERICAN JOURNAL OF HUMAN GENETICS
AMERICAN JOURNAL OF MEDICAL GENETICS
ANNALES DE GE NETIQUE
ANNALS OF HUMAN GENETICS
ANNUAL REVIEW OF GENETICS
ATTI ASSOCIAZIONE GENETICA ITALIANA
CANADIAN JOURNAL OF GENETICS AND CYTOLOGY
CARYOLOGIA
CLINICAL GENETICS
CURRENT GENETICS
GENE
GENETICA
GENETICA POLONICA
GENETICAL RESEARCH
GENETICS
GENETIKA
HEREDITAS
HEREDITY
HUMAN GENETICS
HUMAN HEREDITY
JAPANESE JOURNAL OF GENETICS
JAPANESE JOURNAL OF HUMAN GENETICS
JOURNAL DE GENETIQUE HUMAINE
JOURNAL OF HEREDITY
JOURNAL OF MEDICAL GENETICS
MUTATION RESEARCH
PLASMID
REVISTA BRASILEIRA DE GENETICA
ADVANCES IN G ENETICS
ADVANCES IN HUMAN GENETICS
CHEMICAL MUTAGENS -PRINCIPLES AND METHODS FOR THEIR DETECTION
TSITOLOGIYA I GENETIKA
PROGRESS IN MEDICAL GENETICS
JOURNAL OF MOLECULAR AND APPLIED GENETICS
TRENDS IN GENETICS
MUTAGENESIS
JOURNAL OF GENETICS
NATURE GENETICS
ADVANCES IN GENETICS INCORPORATING MOLECULAR GENETIC MEDICINE

REPRODUCTIVE AND GENETIC ENGINEERING -JOURNAL OF
INTERNATIONAL FEMINIST ANALYSIS
DIVERSITY
INTERNATIONAL JOURNAL OF GENOME RESEARCH
GENE GEOGRAPHY
CLINICAL DYSMORPHOLOGY
BRAZILIAN JOURNAL OF GEN ETICS
HUMAN MUTATION
JOURNAL OF HUMAN GENETICS
GENETICS IN MEDICINE
RUSSIAN JOURNAL OF GENETICS
DNA RESEARCH
ANNUAL REVIEW OF GENOMICS AND HUMAN GENETICS
NATURE REVIEWS GENETICS

Search terms

mutation* OR
linkage OR
polymorphism OR
physical mapping OR
quantitative trait OR
association analysis OR
snp OR
haplotype* OR
allel* OR
transmission disequilibrium test OR
genetic anticipation OR
mosaicism OR
trinucleotide repeat OR
splicing OR
genetic predisposition OR
genotype phenotype OR
linkage analysis OR
heterozygosity OR
autosomal* OR
mendelian inheritance OR
microsatellite* OR
genotyp* OR
odds ratio OR
risk ratio OR
sib pair analysis

Immunology

Journals

ACTA PATHOLOGICA ET MICROBIOLOGICA SCANDINAVICA SECTION C - IMMUNOLOGY
ANNALES D IMMUNOLOGIE
CLINICAL AND EXPERIMENTAL IMMUNOLOGY
EUROPEAN JOURNAL OF IMMUNOLOGY
HUMAN IMMUNOLOGY
IMMUNOBIOLOGY
IMMUNOLOGICAL COMMUNICATIONS
IMMUNOLOGICAL REVIEWS
IMMUNOLOGY
IMMUNOLOGY LETTERS
IMMUNOLOGY TODAY
JOURNAL OF CLINICAL & LABORATORY IMMUNOLOGY
JOURNAL OF IMMUNOLOGY
JOURNAL OF THE RETICULOENDOTHELIAL SOCIETY
SCANDINAVIAN JOURNAL OF IMMUNOLOGY
THYMUS
ADVANCES IN IMMUNOLOGY
AFRICAN JOURNAL OF CLINICAL AND EXPERIMENTAL IMMUNOLOGY
CRC CRITICAL REVIEWS IN IMMUNOLOGY
JOURNAL OF CLINICAL IMMUNOLOGY
ACTA PATHOLOGICA MICROBIOLOGICA ET IMMUNOLOGICA SCANDINAVICA SECTION C -IMMUNOLOGY
COMPREHENSIVE IMMUNOLOGY
EOS-RIVISTA DI IMMUNOLOGIA ED IMMUNOFARMACOLOGIA
INMUNOLOGIA
SURVEY OF IMMUNOLOGIC RESEARCH
ANNUAL REVIEW OF IMMUNOLOGY
CONTEMPORARY TOPICS IN MOLECULAR IMMUNOLOGY
JOURNAL OF MOLECU LAR AND CELLULAR IMMUNOLOGY
LYMPHOKINE RESEARCH
CONTEMPORARY TOPICS IN IMMUNOBIOLOGY
ANNALES DE L INSTITUT PASTEUR -IMMUNOLOGY
IMMUNOLOGICAL INVESTIGATIONS
AIDS RESEARCH
IMMUNOLOGIC RESEARCH
INTERNATIONAL JOURNAL OF IMMUNOTHERAPY
DIAGNOSTIC AND CLINICAL IMM UNOLOGY
ISI ATLAS OF SCIENCE -IMMUNOLOGY
CRITICAL REVIEWS IN IMMUNOLOGY
CURRENT OPINION IN IMMUNOLOGY
JOURNAL OF AUTOIMMUNITY
RESEARCH IN IMMUNOLOGY

CHEMICAL IMMUNOLOGY
INTERNATIONAL IMMUNOLOGY
JOURNAL OF IMMUNOLOGICAL RESEARCH
DEVELOPMENTAL IMMUNOLOGY
FUNDAMENTAL AND CLINICAL IMMUNOLOGY
IMMUNITY
PEDIATRIC AIDS AND HIV INFECTION -FETUS TO ADOLESCENT
JOURNAL OF CLINICAL IMMUNOASSAY
AUTOIMMUNITY
SEMINARS IN IMMUNOLOGY
IMMUNOLOGIST
CLINICAL IMMUNOLOGY
NATURE IMMUNOLOGY

Search terms:

immun OR
antibod* OR
antigen* OR
cytokine* OR
lymphocyt* OR
vaccin* OR
MHC OR
histocomp* OR
thymus* OR
thymocyte* OR
interleukin* OR
t cell* OR
b cell* OR
nk cell* OR
nkt cell* OR
((apc OR antigen presenting cell*) NOT NSAID) OR
il-1 OR il-2 or ... (etc)

Appendix E: Patent analyses

Table A-1: Search strategy for patent applications in Genetics/heredity
Source: Fraunhofer, ISI

S (C12N015 or A01H001 or A61K048 or A01K067 -027 or A01K067 -033 or C12N005 -10 or C12N005 -12 or C12N005 -14 or C12N005 -16 or C12N005 -18 or C12N005 -20 or C12N005 -22 or C12N005 -24 or C12N005 -26 or C12N005 -28)/IC
s (C12N001 -11 or C12N001 -13 or C12N001 -15 or C12N001 -19 or C12N001 -21 or C12N007 -01)/IC
S (GENE OR GENES OR GENETIC? OR GENOME O R GENOMIC?) AND (C07 H021 OR C12Q001 -68)/IC
S A61P037/IC AND ((A61K03 5 NOT (A61K035 -0! OR A61K035 -10)) OR A61K038 OR A61K039 OR A61K048 O R A61K049 OR A61K051)/IC
S L1 -L4

Note:

#	truncation up to one character (0 or 1)
!	truncation of exactly one character (1)
?	unlimited truncation (0 or any number)

Table A-2: Content of IPC classes and sub -classes used in the definition of patent samples
Source: Fraunhofer ISI, WIPO 2000

A01H	new plants
A01K	new breeds of animals
A01N	biocides
A23L	preparation of food
A61B	medical diagnosis, surgery
A61F	treatment of eyes, ears etc.
A61H	physical therapy
A61M	media introduction into the body
A61N	electrotherapy
A61K	preparations for medical purposes
A61L	disinfection
A61P	therapeutic activity of preparations for medical purposes
B0	separating, mixing
B82B	Nanotechnology
C0	organic and inorganic chemistry
C02F	treatment of water
C07H	sugars
C12M	apparatus for enzymology or microbiology
C12N	micro -organisms, genetic engineering (C12N015)
C12P	enzyme -based processes

C12Q measuring processes based on micro -organisms etc.
C12R indexing scheme for micro -organisms etc.
C12S specific processes using enzymes or micro -organisms
G01 measuring, testing
G01N analyzing physical, chemical or biological properties of materials
G02 optics
G03 photography
G03G electrography
G06 computing
G11 information storage
G12B details of instruments
H01L semiconductors
H01J electric discharge tubes

Table A-3: Search strategy for patent applications in Neurosciences
Source: Fraunhofer, ISI

S (NEUR? OR NERVE? OR NERVOUS OR BRAIN? OR MENING? OR MYELITIS OR CEREBR? OR INTRACERE BR? OR BSE OR ENCEPH ALO? OR AXON OR DEND RIDE OR CORTEX OR MYELIN? OR DEMYELIN? OR SYNAPS E)/BI NOT (NEURAMIN? OR NEUROSPORA?)/BI
S (A61P025 and ((A61K035 not (A61K035 -0! or A61K035 -10)) or A61K038 OR A61K039 OR A61K048 OR A61K049 OR A61K051))/IC
S (?PLEGIA) OR ALZHEIMER# OR (AMAUROSIS(W)FUGAX) OR ARACHNOID? OR TELANGIECTASIA OR (ATYPICAL(W)FACIAL(W)PAIN) OR (AUTOSOMAL(W)RECESSIVE) OR BALO OR (BASAL(W)GANGLIA) OR (BELL? (W) PALSY) OR BE NEDIKT OR BERNARD OR (BONE (W) MARROW)
S (BRACHIAL(W)PLEXUS) OR (BRIDGE(2W)VAROLIUS) OR (BULBAR (W) PALS?) OR (CARPAL(W)TUNNEL) OR (CARTOID (W) ARTERY) OR (CARTOID(W)SINUS(W)SYNCOPE#) OR (CAUDA(W)EQUINA) OR CAUSALGIA OR MYELINOLYSIS OR (CERVICAL (W) ROOT) OR (CHARCOT (W) MARIE)
S CLAUDE OR CREUTZFE LD# OR (HEMIFACIAL(W)SPASM) OR (CONCENTRIC(W)SCLEROSIS) OR ATAXIA OR (DEJERINE (W) SOTTAS) OR (DENNY (W) DROWN) OR DEVIC OR (DIFFUSE (W) SCLEROSIS) OR DEMYELINATION OR (DURA (W) MATER) OR (DYSTONIA)
S (EPIDURAL) OR (EPILEPS?) OR (ENCEPHALITIS) OR (ENCEPHALOMYELITIS)
S (EXTRADURAL) OR (EXTRAPYRAMIDAL) OR (FACIAL (W) MYOKYMIA) OR (FAMILIAL (W) DYSAUTONOMIA) OR (FAZIO (W) LONDE) OR (FOVILLE) OR (GANGLIONITIS)
S (GLOBUS (W) PALLIDUS) OR NEURALGIA OR (GRAND (W) MAL (W) SEIZURE#) OR (GUILLAIN (W) BARRE) OR (HAEMATOMYELIA) OR (HAEMORRHAGIC (W) LEUKOENCEPHALITIS) OR (HALLERVORDEN (W) SPATZ) OR (HEMIPARKINSON?) OR (HORNER?) OR HUNTINGTON? OR HURST OR (HYDROCEPHALUS) OR (INTRACRANIAL)
S (INTRASPINAL) OR (KOZHEVNIKOF) OR (LATERAL (W) SCLEROSIS) OR (LENNOX (W) GASTAUT)

S (LEPTOMENINX) OR (LOUIS (W) BAR) OR (LUMBOSACRAL (W) ROOT) OR (MEDULLA (W) OBLOGONTA) OR (MELKERSSON?) OR (MENINX) OR (MERALGIA (W) PARAESTHETICA) OR (MIGRAIN?)
S (MILLARD (W) GUBLER) OR (MONONEURITIS) OR (MONONEUROPATH?) OR (MORVAN?)
S (MULTIPLE (W) SCLEROSIS) OR (MYONEURAL) OR (NELATON?) OR (PARALY?) OR (PARKINSON?) OR (PARSONAGE (W) ALDREN (W) TURNER) OR (PHANTOM (W) LIMB)
S (PIA (W) MATER) OR (PICK? (W) DISEASE#) OR (PIGMENTARY (W) PALLIDAL) OR (PLEXUS (W) DISORDER?) OR (POLYNEURITIS) OR (POLYNEUROPATH?) OR (PORENCEPHALIC) OR (LACUNAR)
S (REFUSUM? (W) DISEASE#) OR (RESTLESS (W) LEGS) OR (REYE?) OR (RILEY (W) DAY) OR (ROUSSY (W) LEVY) OR (SCHILDER?) OR (SHY (W) DRAGER) OR (SPASTIC) OR (SPASMODIC?) OR (S PINAL (W) CORD) OR (SPINAL (W) MUSCULAR (W) ATHROP?) OR (EPILEPTICUS)
S (STEELE (W) RICHARDSON (W) OLSZEWSKI) OR (STIFF (W) MAN) OR (STRIATONIGRAL) OR (PAROXYSMAL (W) FACIAL) OR (SYRINGOBULBIA) OR (SYRINGOMYELIA) OR (TARSAL (W) TUNNEL)
S (?THALAMUS) OR (THORAIC (W) ROOT) OR (TIC (W) DOULOUREUX) OR (AMNESIA)
S MYELOPATH? OR (VERTEBRO (W) BASILAR (W) ARTERY) OR (WALLENBERG) OR (WEBER) OR (WERDNIG (W) HOFFMAN) OR (WEST? (W) SYNDROM?) OR (WITTMAACK (W) EKBORN) OR (DIABETIC (W) AMYOTHROPH?)
S (DYSTROPHIA (W) MYOTONICA) OR (EATON (W) LAMBERT) OR (EULENBERG) OR (ISAACS) OR (MYASTHENIC)
S (MYASTHENIA) OR (MYONEURAL) OR (MYOTONIA) OR (PARAMYOTONIA (W) CONGENITA) OR (PSEUDOMYOTONIA) OR (STEINERT) OR (THYROTOXICOSIS) OR TREMOR OR HEADACHE O R TRIGEMIN? OR VAGUS OR POLIOMYELITIS OR (HEINE(W)MEDIN)
S L1 -L19
S L20 AND (A01K OR A 23L OR A61B OR A61F OR A61H OR A61M OR A 61N OR (A61K035 NOT (A61K03 5-0! OR A61K035 -10)) OR A61K038 OR A 61K039 OR A61K048 OR A61K049 OR A61K051 O R C12M OR C12N OR C1 2P OR C12Q OR C12R O R C12S OR G01N)/IC

Note:

#	truncation up to one character (0 or 1)
!	truncation of exactly one character (1)
?	unlimited truncation (0 or any number)
W	directly adjacent terms in order specified

Table A-4: Search strategy for patent applicati ons in Immunology
Source: Fraunhofer, ISI

S (IMMUN? or (?ALLERG?) OR (ADA) OR (ADENOSINE (W) DEAMINASE) OR (AGAMMAGLOBULINEMIA?) OR (AIDS(2W)(VIRUS OR INFECT? OR COMPL EX)) OR CYTOKIN? OR LYMPHOCY TE? OR LYMPHOKIN? OR INTERFERON? OR INTERLEUKIN?)/BI

S ((ALLOGENEIC (W) TRANSPLANTATION#) OR (ALYMPHOPLASIA?) OR (ANTI (W) PHOSPHOLIPID) OR (ANTIBODY (W) DEFECT#) OR (ANTIPHOSPHOLIPID) OR (ATAXIA (W) TELANGIECTASIA?))/BI
S ((AUTOIMMUN?) OR (BARE (W) LYMPHOCYTE) OR (BASEDOW) OR (B (W) CELL#) OR (BIOTIN (W) DEPENDENT (W) CARBOXYLASE) OR (BRUTON))/BI
S (COMPLEMENT(W)DEFE CT?) OR (COMPLEMENT(W)INHIBIT?)
S ((CHEDIAK (W) HIGASHI) OR (CRYOGLOBULINEMIA?) OR (CV (W) AGAMMA#) OR (CVI) OR (CYTOKINE#) OR (GEORGE#) OR (DIABETES (W) MELLITUS) OR (DIABETES (W) TYPE (W) 1) OR (DIAB ETES (W) TYPE (W) I) OR (DYSGAMMAGLOBULINEMIA?))/BI
S ((EATON (W) LAMBERT) OR (ENDOCRINE (W) OPHTHALMOPATH?) OR (EPISODIC (W) LYMPHOCYTOPENIA?) OR (EPSTEIN (W) BARR) OR (GRAFT (W) REJECTION#) OR (GRANULOMATOUS) OR (GRAVE## (W) DISEASE#) OR (GRAVE## (W) THYROIDITIS?))/BI
S ((HEMOLYTIC (W) ANEMIA) OR (HAEMOLYTIC (W) ANAEMIA) OR (HASHIMOTO#) OR (HEREDITARY (W) ANGIONEUROTIC) OR (HIOB (W) SYNDROM?))/BI
S ((HISTOCOMPATIBILITY (W) ANTIGEN#) OR (HIV) OR (HLA (W) SYSTEM) OR (HOST (W) VERSUS (W) GRAFT) OR (HYPERGAMM AGLOBULINEMIA?) OR (HYPERIMMUNOGLOBULINEMIA?) OR (HYPOALLERGENIC) OR (HYPOGAMMAGLOBULINEMIA?) OR (IGA) OR (IGE) OR (IGM))/BI
S ((INSULIN (W) DEPENDENT (W) DIABETES?) OR (JUVENILE (W) DIABETES?) OR (KAPPA (W) LIGHT (W) CHAIN) OR (KILLER (W) CELL#) OR (KUSSM AUL (W) MEYER))/BI
S ((LAMBERT (W) EATON) OR (LAZY (W) LEUCOCYTE) OR (LFA1) OR (LOUIS (W) BAR) OR (LYMPHOKINE#))/BI
S ((MCTD) OR (MHC) OR (MHS) OR (MONOCLONAL (W) GAMMOPATH?) OR (MUKOKUTANE (W) CANDIDIASIS?) OR (MYASTHENIA (W) GRAVIS?) OR (MYELOPEROXIDASE (W) DEFECT#) OR (NEZELOF##) OR (PHACOGENIC (W) UVEITIS?))/BI
S ((PHAGOCYTE#) OR (PHARYNGEAL (W) POUCH) OR (PNP) OR (POLYARTERITIS (W) NODOSA?) OR (POLYCLONAL (W) GAMMOPATH?) OR (POLYMYOSITIS?) OR (PRIMARY (W) BILIARY (W) CIRRHOSIS?) OR (PURINE (W) NUCLEOSI DE (W) PHOSPHORYLASE#) OR (RETICULAR (W) DYSGENESIS?))/BI
S ((ROOKE#) OR (SCID) OR (SHARP (W) SYNDROM?) OR (SYMPATHETIC (W) OPHTHALMIA?) OR (SYNGENEIC (W) TRANSPLANTATION#) OR (T (W) CELL#) OR (TRANSPLANT (W) REJECTION#) OR (TRANSPLANTATION (W) IMMUNIT?))/ BI
S ((WALDENSTROEM) OR (WISKOTT (W) ALDRICH) OR (XENOGENEIC (W) TRANSPLANTATION#) OR (X -LINKED (W) LYMPHOPROLIFERATIVE) or (MULTIPLE (W) SCLEROSIS) OR ADDISON? OR CRO HN? OR GOODPASTURE? OR GRAVES? OR HASHIMOTO? OR MYASTH ENIA(W)GRAVIS)/BI
S ((AGAMMAGLOBUL INAEMIA?) OR (CRYOGLOBULINAEMIA?) OR (DYSGAMMAGLOBULINAEMIA?) OR (HYPERGAMMAGLOBULINAEMIA?) OR (HYPERIMMUNOGLOBULINAEMIA?) OR (HYPOGAMMAGLOBULINAEMIA?) OR (LFA(W)1) OR (MUCOCUTANE (W) CANDIDIASIS?))/BI
S (PEMPHIGUS(W)VULGARIS OR PERNICIOUS(W)ANE MIA OR PSOR IASIS OR SCLERODERMA OR SJOER GREN? OR LUPUS(W)ERY THEMATOSUS OR ANTIBO D? OR ANTIGEN? OR HISTO COMPATIBL? OR VACCIN ?)/BI
S L1-L16
S L17 AND (A01K OR A 23L OR A61B OR A61F OR A61H OR A61M OR A 61N OR (A61K035 NOT (A61K03 5-0! OR A61K035 -10)) OR A61K038 OR A 61K039 OR A61K048 OR

A61K049 OR A61K051 O R C12M OR C12N OR C1 2P OR C12Q OR C12R O R C12S OR G01N)/IC
S (A61P037 AND ((A61 K035 NOT (A61K035 -0! OR A61K035 -10)) OR A61K038 OR A61K039 OR A61K048 O R A61K049 OR A61K051))/IC
S (A61K038 -19 OR A61K038 -20 OR A61K038 -21 OR A61K039 OR A61K049 -16 OR A61K051 -10 OR C12N015 -13 OR C12N015 -19 OR C12N015 -20 OR C12N015 -21 OR C12N015 -22 OR C12N015 -23 OR C12N015 -24 OR C12N015 -25 OR C12N015 -26)/IC
S L18 -L20

Note:

#	truncation up to one character (0 or 1)
!	truncation of exactly one ch aracter (1)
?	unlimited truncation (0 or any number)
W	directly adjacent terms in order specified

Table A-5: Search strategy for patent applications in Bioinformatics
Source: Fraunhofer, ISI

S G06?/IC
S (L1 AND (C12M OR C 12N OR C 12P OR C12S OR A61K -039 OR A61K -048 OR C02F -003)/IC) NOT G01#/IC
S BIOCOMPUT? OR BIOI NFORMAT?
S L1 AND ((DRUG#(2A) (DISCOVER? OR TARGET # OR DEVELOPMENT# OR CANDIDATE# OR DESIGN)) OR (HIGH(W)THROUG HPUT(W)SCREENING) OR ((GENOME OR DNA)(W)S EQUENC?) OR ((COMBI NATOR? OR COMBINATIONAL)(W)CHE MISTRY) OR (GENE(2A) PREDICT?))
S L1 AND ((GENOM? OR GENE#)(2A)(SCREEN? OR COMPARISON#) OR (COMBINATI ONAL(W)LIBRAR?))
S L1 AND ((METABOLIC (W)PATHWAY#) OR (BIO MOLECUL?(2A)SEQUENC?) OR (GENE# OR GENOM?)(2A)(ANALYS? OR DATA OR STR UCTURE#) OR PROTEIN#(2A)(STRUCTU RE# OR MOTIF#))
S L1 AND (((GENE# OR GENOME)(2A)ANNOTAT?) OR (BASE(2A)SEQUEN C?) OR (GENE#(2A)EXPRESSION) OR GENOMICS OR PRO TEOMICS OR (BIOCHEM?(W)PATHWAY#) OR (MODELLING(2A)P ROTEIN#))
S L1 AND (MACROMOLEC ULAR(W)(SEQUENCE# OR STRUCTURE#) OR PREDICT?(2A)MUTATION # OR BIOLOGICAL(W)DA TABASE# OR MOLECULAR(W)(EVOLUTI ON OR DYNAMICS) OR C ELL?(W)FUNCTION#)
s L2 -L8

Note:

#	truncation up to one character (0 or 1)
!	truncation of exactly one character (1)
?	unlimited truncation (0 or any number)
W	directly adjacent terms in order specified
2A	adjacent terms in any order, separated by up to 2 words

Table A-7: Matched inventor names in the SCI (all fields, total numbers)

	Inventor names sent to CWTS	Inventor names matched by CWTS
Genetics	18930	6167
Neurosciences	8541	5458
Immunology	20324	7293
Bioinformatics	374	80
Nanotechnology	3348	1303

Table A-8: Non-matches and matches (inventor -institution pairs) across all fields

	Non-matches	Matches	Total CWTS matches
Genetics	1267	4900	6167
Neurosciences	2056	3402	5458
Immunology	1730	5563	7293
Bioinformatics	23	57	80
Nanotechnology	184	1119	1303

Table A-9: Non-matches (inventor -institution pairs) across fields and countries

	Genetics	**Neurosciences**	**Immunology**	**Biocomputing**	**Nanotechnology**
UK	862	1337	1085	23	35
Germany	148	337	236	0	115
Denmark	102	118	123	0	2
Netherlands	42	48	78	0	4
France	35	69	79	0	12
Switzerland	23	21	29	0	6
Sweden	20	31	28	0	3
Spain	14	30	4	0	0
Belgium	7	6	17	0	1
Israel	5	13	21	0	1
Austria	5	1	11	0	0
Iceland	2	0	0	0	0
Italy	1	25	11	0	0
Finland	1	12	3	0	1
Ireland	0	5	3	0	4
Norway	0	3	2	0	0

Table A-10: Full and partial matches (inventor -institution pairs) across all fields

	Full match	Partial match	All matches
Genetics	3955	945	4900
Neurosciences	2812	590	3402
Immunology	4631	932	5563
Bioinformatics	57	0	57
Nanotechnology	1033	86	1119

Table A-11: Partial matches (inventor-institution pairs) across all fields

	Genetics	Neurosciences	Immunology	Nanotechnology
UK	412	284	348	22
Germany	258	114	241	36
France	77	42	47	3
Netherlands	51	10	48	9
Denmark	40	33	43	3
Austria	21	4	14	3
Belgium	20	31	29	2
Switzerland	15	14	51	2
Sweden	12	23	71	11
Italy	12	19	9	2
Ireland	10	6	4	1
Finland	8	12	7	0
Spain	6	5	2	0
Israel	4	6	18	3
Portugal	1	0	0	0
Norway	0	1	4	0
Slovakia	0	1	0	0

Table A-12: Results of cleaning procedure; three sources of institutional information (Genetics), absolute numbers

	Applicant field	Inventor field	Full or partial match with SCI
before cleaning; N=15427	7215	3312	4900
after cleaning; N=9826	6084	666	3076

Table A-13: Results of cleaning procedure; three sources of institutional information (all other fields), absolute numbers

	Applicant field	Inventor field	Full or partial match with SCI
Neurosciences (N=5228)	2844	180	2204
Immunology (N=10459)	6841	157	3461
Bioinformatics (N=173)	129	4	40
Nanotechnology (N=1657)	1043	22	592

Figure A-1: Letter to country experts

Dear colleague,

In the project "Mapping R&D Excellence in E urope", various lists of institutions were generated by database analysis. For achieving reliable results, it is necessary to check the accuracy of these lists. The European Commission transferred us your name as contact point for Belgium with regard to th e fields of nanotechnology and neuroscience.

Please find enclosed an institutional list referring to these fields and your country; these lists are the result of patent analyses. Please send them back to us until

30 September 2002.

We know that this is a tight deadline. But the lists of the patent analysis are much shorter than those of the publication analysis, so that the necessary amount of work is limited. In separate files, we describe what shall be done for the institutional a djustment.

Sincerely yours,

Dr. U. Schmoch
Fraunhofer ISI
Karlsruhe, Germany

Figure A-2: Instruction for country experts (attached to letter)

Institutional Adjustment for Patent Analysis

In the project "Mapping R&D Excellence in Europe", our institute performed a patent analysis. One output of this work is a list of institutions with patent activities, broken down by country. For a final exploitation, these lists need further adjustment, and we would like to ask you for support in this matter.

The aim of the adjustment is to bring together different institutional names referring to the same institution. It is possible that one institution appears with different names due to

- spelling errors
- the use of different sources of information
- the inconsistent self-description of an organization.

For instance, the exploitation of different sources of information sometimes has the effect that the organizational name is presented in its native language and in English in parallel.

How you can you help us

Please find enclosed the file Nano_CH_demo.xls. In the first column (column A), the names of relevant institutions, in this case Swiss institutions, are registered in alphabetical order. In the second column, the apparent town of residence – as matched from a different database – is given. In the third column (column C), the "institution number" is recorded.

We ask you to fill in additional information on institutional linkages in the fourth column (column D). For example, ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE (EPFL) (institution number 9) appears a second time as SWITZERLAND/LAUSANNE/ SWISS FED INST TECHNOL EPFL (institution number 24). Please indicate in column D for both cases the institution number of the version which you think is the relevant one. In the example, we chose number 24, the English version. The case of EPFL is obvious. In other cases, however, we need your support, as we do not know the national structures in detail.

It is not necessary to fill in column D in all cases. This should only be done for institutions with different names. For the project, **only the main organization** is of interest.

- In the case of university institutes or centers, only the university is of interest.
- In the case of large companies, the affiliations have to be linked to the parent company.
- As an exception, the different institutes of non-university research organizations, the institute/centre is considered the main organization, for instance, in the case of the Fraunhofer Institute for Production Technology and Applied Material Research (IFAM), the IFAM is the relevant unit, not the parent organization Fraunhofer Society.

In the case of names in different languages, please refer to the English version, b ecause it is easier to link it to the Science Citation Index. This will be done in subs equent steps of the project.

Finally, we ask you to indicate the type of institution in the fifth column (column E). Please differentiate between the following types of institution:

1 Non-profit research institution (e.g. university, public ly funded research centre)
2 Patent, license or transfer organization acting on behalf of an institution of type 1
3 For-profit research institution
4 Enterprise

If you have any questions, please send a message to
Dr. Ulrich Schmoch
E-Mail: us@isi.fhg.de

We will answer as soon as possible. After checking the files, please send them back until 30 September 2002 to the e -mail address mentioned.

Thank you for your support!

Table A-14: Example sheet "Switzerland" sent to country experts

Institution	Town	No	Relev. no	Type
ABB RESEARCH LTD.	8050 Zuerich	1	1	4
Alcan Technology & Management AG	8212 Neuhausen am Rheinfall	2	2	4
Andromis S.A.	1207 Genève	3	3	4
Barth Fruit AG	4010 Basel	4	4	4
Ciba Specialty Chemicals Holding Inc.	4057 Basel	5	5	4
CONTRAVES SPACE AG	8052 Zuerich	6	6	4
DEBIO RECHERCHE PHARMACEUTIQUE S.A.	CH-1920 Martigny	7	7	4
Eco 2? SA	6805 Mezzovico	8	8	4
ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE (EPFL)	1015 Lausanne	9	24	1
Eidgenoessische Materialpruefungs- und Forschungsanstalt Empa Thun	3602 Thun	10	10	1
Emil Flachsmann AG	CH-8820 Waedenswil	11	11	4
F. HOFFMANN-LA ROCHE AG	4070 Basel	12	21	4
Givaudan SA	1214 Vernier-Genève	13	13	4
Incoat GmbH	8224 Loehningen	14	14	4
METAUX PRECIEUX SA METALOR	2009 Neuchâtel	15	15	4
Mettler-Toledo GmbH	8606 Greifensee	16	16	4
Microchemical Systems S.A.	2035 Corcelles	17	17	4
Physik-Institut	8057 Zuerich	18	18	1
Schibli Engineering GmbH	4503 Solothurn	19	19	4
SWITZERLAND/BASEL/NOVARTIS PHARMA AG	BASEL	20	20	4
SWITZERLAND/BASEL/ROCHE AG	BASEL	21	21	4
SWITZERLAND/BASEL/UNIV BASEL	BASEL	22	22	1
SWITZERLAND/GENEVA/UNIV GENEVA	GENEVA	23	23	1
SWITZERLAND/LAUSANNE/SWISS FED INST TECHNOL EPFL	LAUSANNE	24	24	1
SWITZERLAND/LAUSANNE/UNIV LAUSANNE	LAUSANNE	25	25	1
SWITZERLAND/NEUCHATEL/CSEM SA	NEUCHATEL	26	26	4
SWITZERLAND/NEUCHATEL/UNIV NEUCHATEL	NEUCHATEL	27	27	1
SWITZERLAND/RUSCHLIKON/IBM CORP	RUSCHLIKON	28	28	4
SWITZERLAND/VILLIGEN/PAUL SCHERRER INST	VILLIGEN	29	29	1
SWITZERLAND/ZURICH/SWISS FED INST TECHNOL ETHZ	ZURICH	30	30	1
SWITZERLAND/ZURICH/UNIV ZURICH	ZURICH	31	31	1
Syngenta Participations AG	4058 Basel	32	32	4
Tec-Sem AG	8274 Taegerwilen	33	33	4
Tetra Laval Holdings & Finance S.A.	1009 Pully	34	34	4
Tetra Laval Holdings & Finance SA	1009 Pully	35	34	4
Unaxis Trading AG	9477 Truebbach	36	36	4
UNIVERSITE DE NEUCHATEL	2000 Neuchatel	37	27	1
Vantico AG	4057 Basel	38	38	4
VESIFACT AG	6340 Baar	39	39	4
VESIFACT AG	6342 Baar 2	40	39	4
White Spot AG	6342 Baar	41	41	4
ZW Biomedical Research AG	3008 Bern	42	42	4

Annex Type

1 = Non-profit research institution (e.g. university, publicly funded research centre)

2 = Patent, license or transfer organisation acting on behalf of an institution of type 1

3 = For-profit research institution

4 = Enterprise

Table A-15: Absolute number of profit and non-profit institutions before cleaning (all fields)

	Genetics	Neurosciences	Immunology	Bioinformatics	Nanotechnology
non-profit	2053	837	2089	27	330
for profit	4414	2170	4866	118	757

Table A-16: Absolute number of profit and non-profit institutions after cleaning (all fields)

	Genetics	Neurosciences	Immunology	Bioinformatics	Nanotechnology
non-profit	5116	2834	5318	56	867
for profit	4710	2394	5141	117	790

Table A-17: Absolute number of profit and non-profit institutions before and after cleaning (Genetics)

before	non-profit	for profit
France	563	523
Germany	495	1186
UK	421	1162
Netherlands	125	324
Belgium	107	200
Israel	72	103
Spain	62	27
Denmark	51	184
Italy	46	101
Switzerland	42	273
Finland	18	31
Ireland	18	15
Hungary	5	0
Iceland	5	0
Portugal	5	0
Norway	4	33
Sweden	4	173
Poland	3	0
Austria	2	76
Luxembourg	2	3
Greece	1	0
Slovakia	1	0
Slovenia	1	0
Estonia	0	0
Czech Republic	0	1
Lithuania	0	2
Liechtenstein	0	1

after	non-profit	for profit
UK	1396	1293
Germany	1352	1217
France	916	566
Netherlands	280	345
Belgium	219	205
Israel	150	104
Italy	134	110
Switzerland	129	284
Denmark	123	227
Spain	120	30
Sweden	98	159
Finland	56	38
Austria	46	78
Ireland	30	15
Norway	30	35
Hungary	9	0
Greece	6	0
Portugal	6	0
Poland	5	0
Czech Republic	3	1
Luxembourg	3	3
Slovakia	3	0
Estonia	1	0
Slovenia	1	0
Iceland	0	8
Lithuania	0	2
Liechtenstein	0	1

Table A-18: Absolute number of profit and non-profit institutions before and after cleaning (Neurosciences)

before	non-profit	for profit
France	224	226
UK	192	578
Germany	147	563
Netherlands	70	88
Israel	59	79
Belgium	41	68
Switzerland	22	137
Italy	20	83
Spain	17	6
Sweden	13	131
Denmark	11	89
Finland	8	26
Norway	5	22
Luxembourg	3	7
Slovenia	2	1
Ireland	1	20
Hungary	1	2
Poland	1	0
Austria	0	34
Iceland	0	5
Liechtenstein	0	2
Cyprus	0	1
Estonia	0	1
Greece	0	1
Slovakia	0	0

after	non-profit	for profit
UK	805	623
Germany	729	600
France	432	279
Israel	155	79
Italy	127	95
Netherlands	113	91
Sweden	93	179
Switzerland	85	143
Belgium	82	79
Denmark	64	93
Finland	49	26
Spain	38	6
Norway	16	25
Hungary	11	2
Austria	9	36
Ireland	9	20
Slovakia	5	0
Poland	4	0
Luxembourg	3	7
Iceland	2	6
Slovenia	2	1
Estonia	1	1
Liechtenstein	0	2
Cyprus	0	1
Greece	0	1

Table A-19: Absolute number of profit and non-profit institutions before and after cleaning (Immunology)

before	non-profit	for profit
France	560	634
UK	442	1153
Germany	400	1194
Netherlands	202	277
Israel	135	117
Belgium	108	254
Italy	58	158
Switzerland	52	320
Spain	36	23
Sweden	30	253
Denmark	19	193
Finland	10	54
Ireland	9	37
Norway	6	55
Portugal	5	1
Luxembourg	4	11
Czech Republic	4	8
Hungary	2	9
Slovenia	2	3
Greece	2	0
Austria	1	108
Poland	1	0
Slovakia	1	0
Iceland	0	4
Liechtenstein	0	4
Cyprus	0	2
Estonia	0	1

before	non-profit	for profit
France	560	634
UK	442	1153
Germany	400	1194
Netherlands	202	277
Israel	135	117
Belgium	108	254
Italy	58	158
Switzerland	52	320
Spain	36	23
Sweden	30	253
Denmark	19	193
Finland	10	54
Ireland	9	37
Norway	6	55
Portugal	5	1
Luxembourg	4	11
Czech Republic	4	8
Hungary	2	9
Slovenia	2	3
Greece	2	0
Austria	1	108
Poland	1	0
Slovakia	1	0
Iceland	0	4
Liechtenstein	0	4
Cyprus	0	2
Estonia	0	1

Table A-20: Absolute number of profit and non-profit institutions before and after cleaning (Bioinformatics)

before	non-profit	for profit
UK	8	50
Germany	6	13
Sweden	4	18
France	3	20
Switzerland	3	3
Israel	2	1
Spain	1	0
Netherlands	0	5
Belgium	0	3
Denmark	0	3
Italy	0	2

after	non-profit	for profit
UK	26	48
Germany	10	13
France	4	20
Sweden	4	18
Switzerland	3	3
Israel	3	1
Denmark	2	4
Belgium	2	3
Italy	1	2
Spain	1	0
Netherlands	0	5

Figure A-3: For- profit and non-profit institutions at the country level before cleaning/matching in Neurosciences

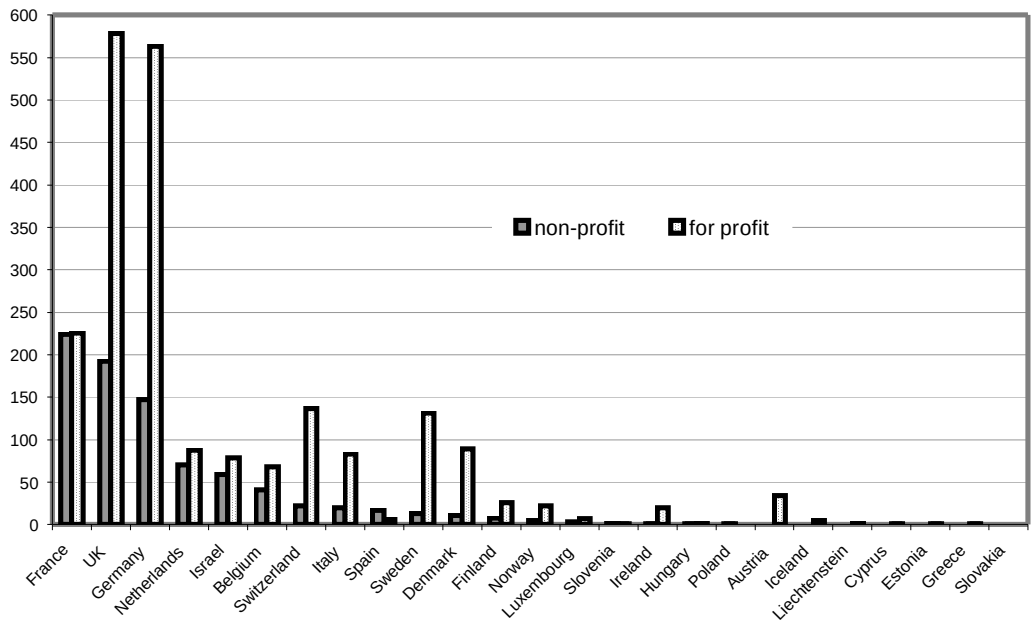


Figure A-4: For- profit and non-profit institutions at the country level after cleaning/matching in Neurosciences

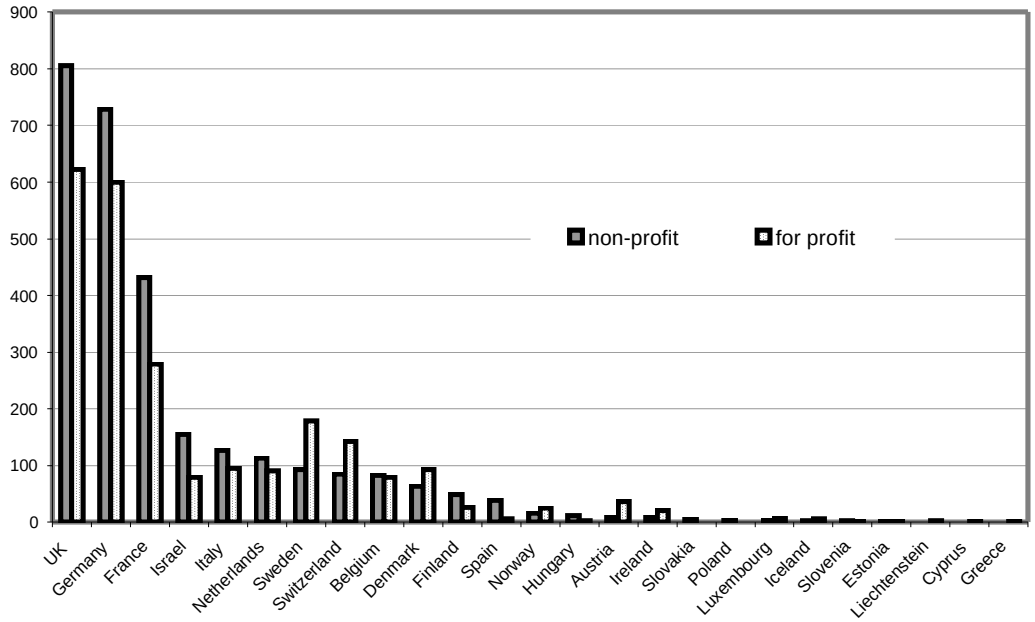


Figure A-5: Additional share of non-profit institutions at the country level after cleaning/matching in Neurosciences

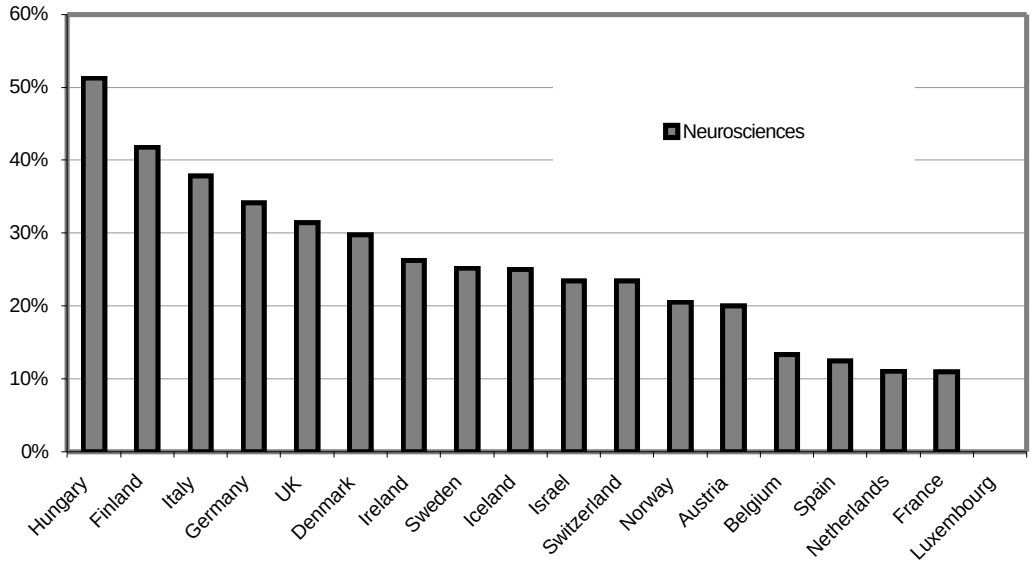


Figure A-6: For- profit and non-profit institutions at the country level before cleaning/matching in Immunology

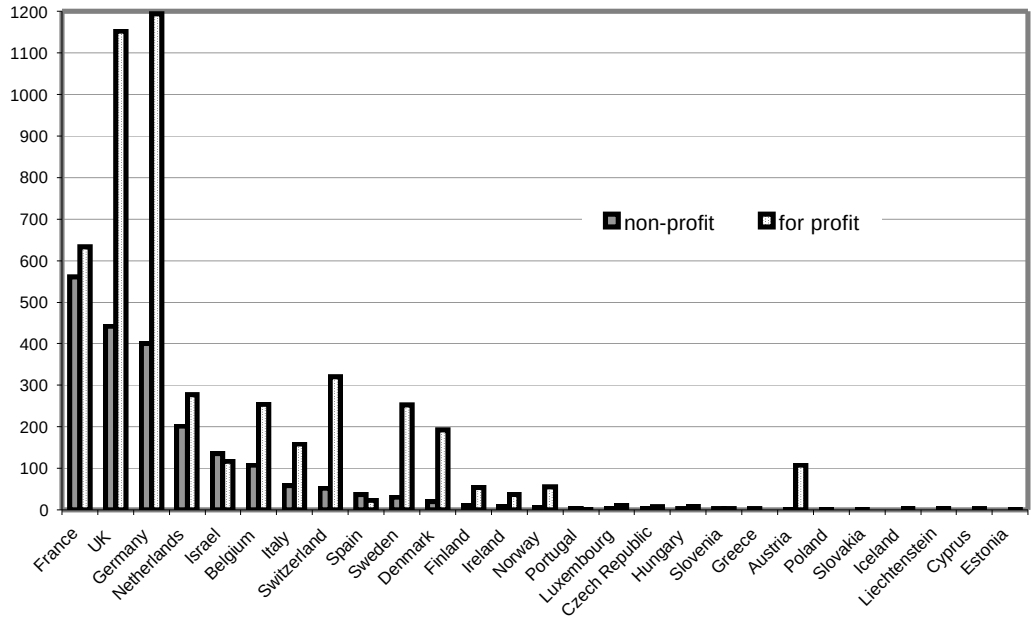


Figure A-7: For-profit and non-profit institutions at the country level after cleaning/matching in Immunology

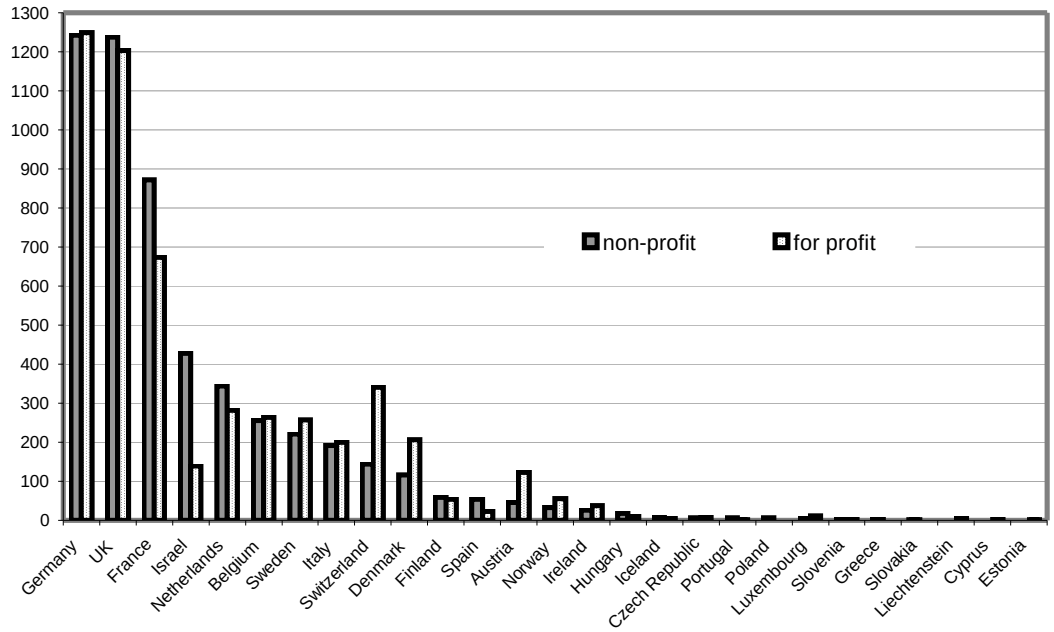


Figure A-8: Additional share of non-profit institutions at the country level after cleaning/matching in Immunology

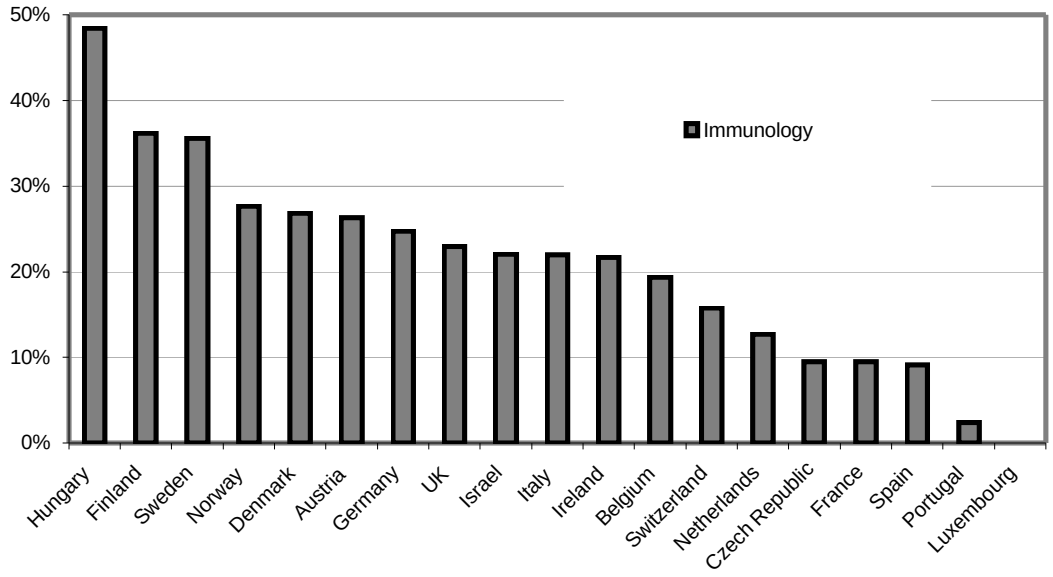


Figure A-9: For- profit and non-profit institutions at the country level before cleaning/matching in Bioinformatics

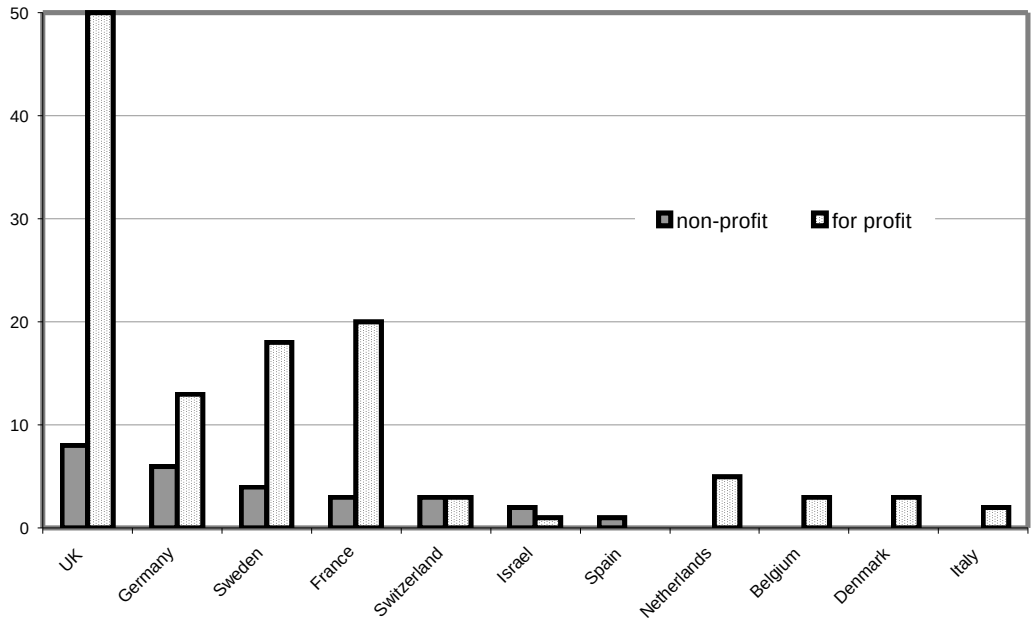


Figure A-10: For-profit and non-profit institutions at the country level after cleaning/matching in Bioinformatics

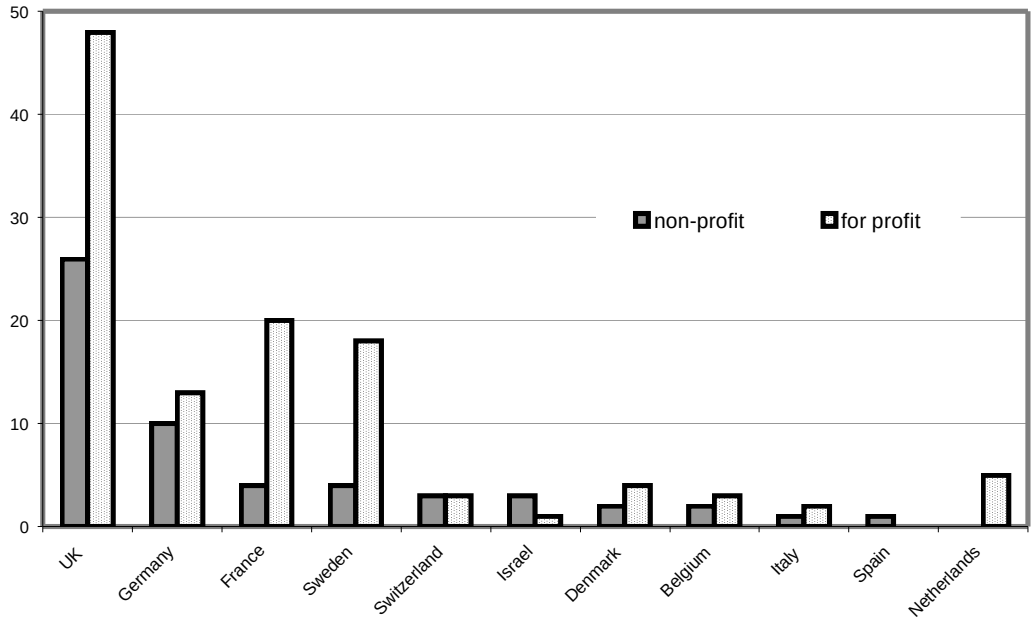
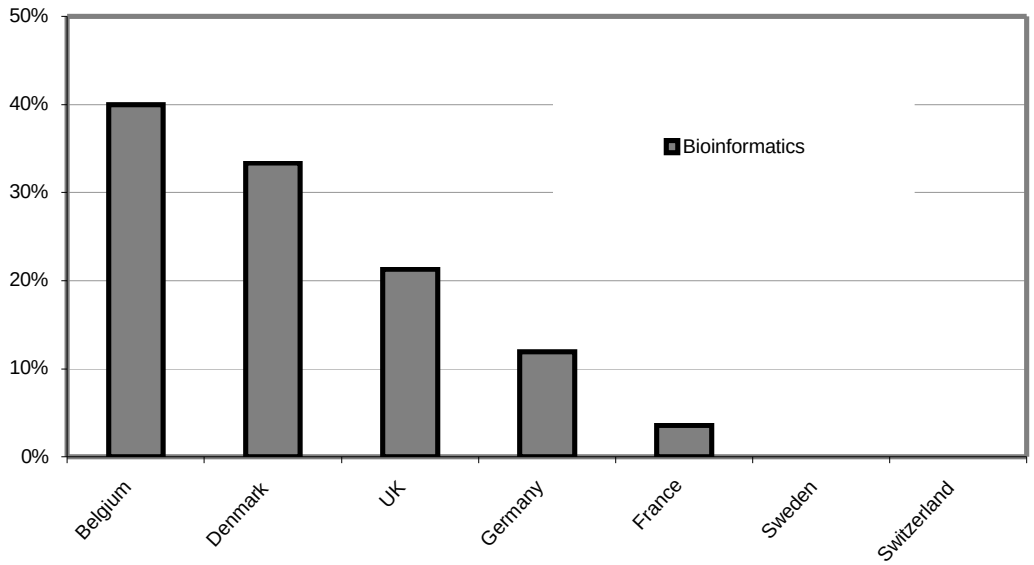


Figure A-11: Additional share of non-profit institutions at the country level after cleaning/matching in Bioinformatics



Annex F: User interface

Cognitive map of the field and activity or impact indicated.

Figure 5-1: Activity distribution of the Erasmus University in Neuroscience

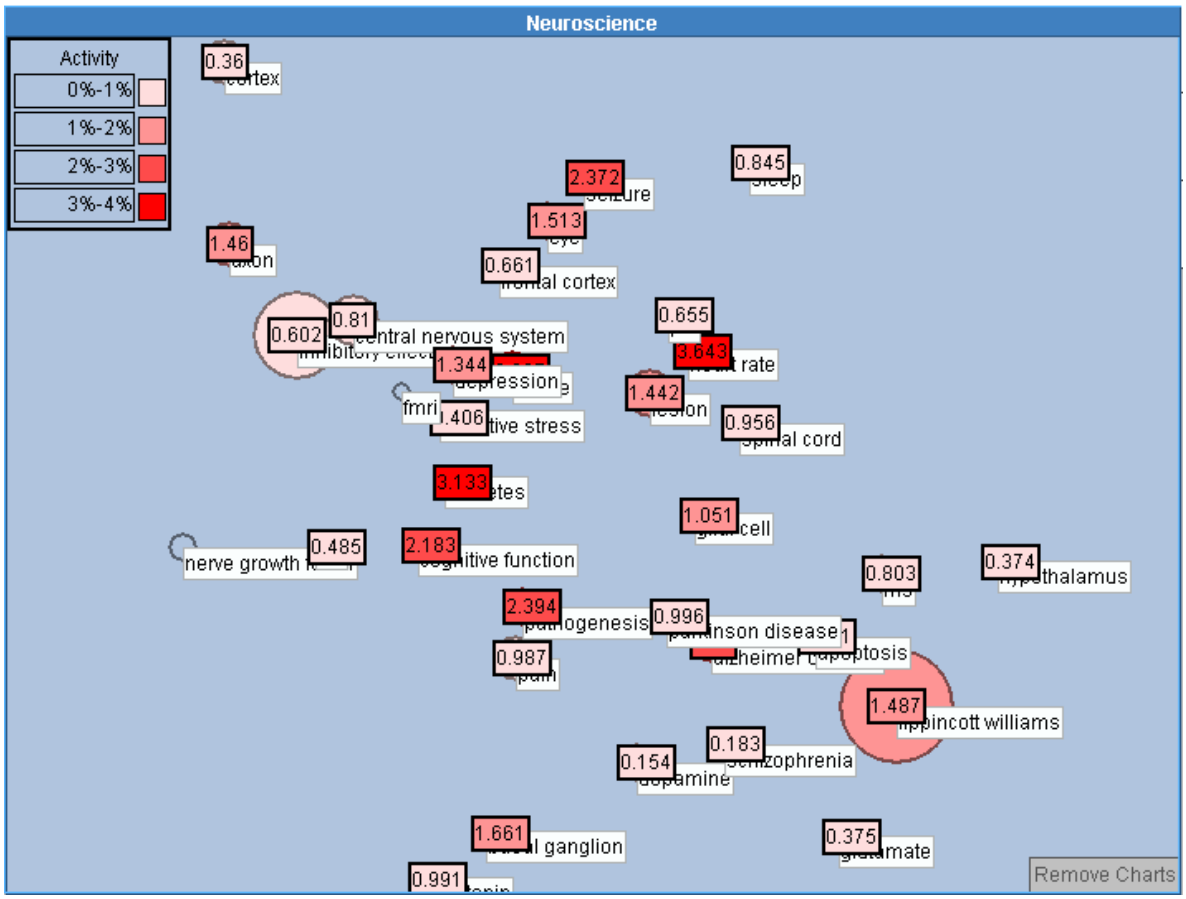
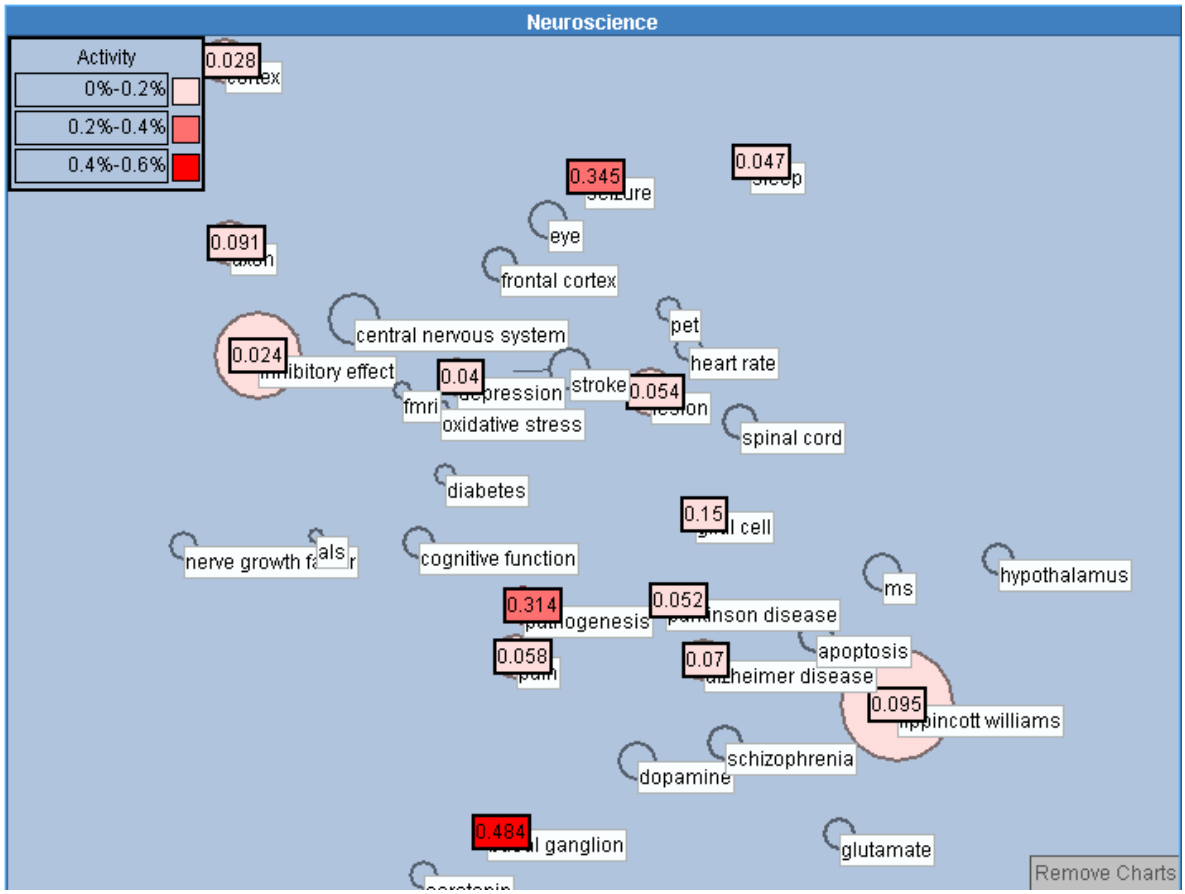


Figure 5-2: Activity profile for the Clinical Genetics department of the Erasmus University in neuroscience



Department name variants:

- ROTTERDAM, AFD KLIN GENET
- ROTTERDAM, DEPT CLIN GENET
- LEIDEN, DEPT CLIN GENET
- ROTTERDAM, DEPT CLIN GENET EE2422
- ROTTERDAM, DEPT CLIN GENET EE2422, MGC
- ROTTERDAM, DEPT CLIN GENET, GENET EPIDEMIOL UNIT
- ROTTERDAM, DEPT CLIN GENET, MGC