EPIP WORKSHOP
"Patent Data for Economic Analysis"
24-25 February 2006
Bocconi University, Milan

1. Objectives and main outcomes of the workshop

The EPIP Workshop "Patent Data for Economic Analysis" was held in Milan, Bocconi University, on February 24-25, 2006. It gathered a group of highly experienced scholars in the economic analysis of patent data, along with more junior researchers, students, and representatives of statistical and other institutions. The workshop presented existing research and data collections about patents. This included new patent data, updates of existing patent data, new collections of patent data, links between patent data and other datasets (firms, industries, regions).

The goal of the workshop was to contribute to data collections, by helping to coordinate projects, to reduce duplicative efforts, find synergies, exchange methodologies of data production and analysis.

The workshop had a largely European focus. However, US and Japanese researchers were invited for a broader coordination and exchange of data and methodologies.

The workshop also aimed at coordinating with Patent institutions, and particularly the EPO. In this respect, the EPO contribution to the production and elaboration of patent statistics was an underlying asset of this workshop. EPO will also be particularly important in future work coming out of this meeting. Among other things, it can help researchers better understand the nature of the patenting process, which can be crucial for a better interpretation of the data, the way to collect them, and more generally for assessing how these data can be used.

In terms of goals achieved, the workshop was one of the first attempts to confront and pull together the practical experiences of economic researchers involved in the construction of patent datasets on a large basis. The Program was designed to provide enough time for each presentation along with an extensive discussion at the end of each session. This enabled the speakers to enter into details about the nature of their datasets and the problems encountered in constructing them, and to compare their methodologies and know-how to address these problems. This produced very wide and lively interaction and exchange of information.

As a result, the workshop contributed certainly to improve the efficiency of the organization of patent datasets among the participating researchers. They now have a better understanding of each other's data effort and collection and several contacts and potential exchanges were carried out. Among other things, it was agreed to organize new data collections to acquire new data about patents from questionnaire surveys. All in all, the workshop was a landmark event for new and better organized data collections in this field. It has been agreed that it will be useful that another follow-up technical workshop on these same issues be organised by EPIP in early 2007.

2. Summary of the presentations

Recent developments in patents statistics and data bases at EPO and OECD

Dominique Guellec, OECD

This presentation illustrated recent developments about patent data carried out at the EPO and OECD. It focused on the Patstat dataset and the Triadic patent data.

Patstat is a new dataset designed for statistical purposes, for compiling indicators and for conducting analytical work (policy, academic). It includes patent data from 73 offices world wide and post-grant patent data from about 40 offices. The first complete version will be released in April 2006 and it will be available to any user committing to non commercial use and no further dissemination of the data.

Patstat will represent a key tool for analysts or researchers using patent data. So far, the unavailability of harmonised and cleaned patent datasets obliged users to create their own database by extracting data published by different patent offices. However, several problems have emerged such as the high costs of database creation, the duplication of costs, the uneven quality, the absence of standardisation, and the

lack of transparency. As a single source of data, Patstat will contribute to solve several of these crucial problems faced in patent data analysis.

Patstat was designed to incorporate current and future efforts on patent data harmonisation (e.g. cleaning of applicant names at EPO and USPTO currently sponsored by Eurostat) and additional tables of patent related indicators (e.g. families, citations, procedural data). Moreover, the role of Patstat users will be very important because Patstat can adapt to the needs expressed by the users. In addition, the users can check the quality by reporting "bugs" to the EPO or by making specific harmonisation efforts (e.g. cleaning names for non western companies, cleaning SME names, consolidating groups of enterprises).

A second important development at OECD is the creation of a dataset on Triadic patent families. The OECD defines a Triadic family as a set of applications at the EPO and JPO along with grants by the USPTO that share one or more priorities (protecting the same invention). There are two main advantages of data on triadic families. First, they reduce the heterogeneity of the value of patents. Members of triadic families are found to be more cited than other patents, to have more claims etc. Second, they reduce the cross country biases in the count of patents in different offices because families are measured on a more neutral ground than applications filed in a single jurisdiction.

The NBER Patent Data Project: Past data uses and future plans

Bronwyn Hall, Berkeley

This presentation focused on the currently available NBER patent data, the uses of this dataset and the new Patent Data project (PDP) at NBER.

The NBER Patent database covers about 3 million U.S. patents granted between January 1963 and December 1999 (now updated to 2002) and all citations made to these patents between 1975 and 1999 (over 16 million). It includes several bibliographic patent information (patent numbers, date, first inventors, assignees, main US and IPC patent classes), citations data (number of forward and backward citations, generality and originality measures based on citations) and the match between assignees and the Compustat dataset on firms traded in the U.S. stock market.

Since its development and public availability, the NBER patent dataset has been used very intensively in more than 100 significant research projects, of which at least one quarter carried out outside the US. It also produced about 100 published papers and about 50 doctoral dissertations in different fields (mainly economics and management, but also a few in finance, law, public policy and other areas).

The NBER is currently carrying out the PDP project, aimed at updating and extending the publicly available USPTO data. The new database is designed according to the principles of public accessibility through xml based tools, modularity, and openness towards the user community. The principle of openness is based on the development of an open source-like environment that allows others to link their data to the patent data, to provide attributions and citations so that contributors are recognized, and to benefit from annotations by users (e.g., error correction, identification of SW or gene patents, etc.).

The tasks of the new PDP project include the update of existing data to 2007, the cleaning and standardization of several fields, the computation of normalization coefficients to correct for truncation or differences across fields in citation practice, the addition of new data (i.e. detailed information on technological classes, priority information, multiple assignees, inventor names and location, multiple inventors, applicants vs. examiners citations, attorney and patent agent names, re-examination requests and outcomes), and the link to complementary data (e.g. Patstat, data on litigations, geographical variables).

Construction of Japanese Patent Database and Preliminary Findings on Patenting Activities in Japan

Akira Goto and Kazuyuki Motohashi, University of Tokyo

The presentation by Goto and Motohashi illustrated the basic features of the Japanese Patent Databases, some comparisons between Japanese, EU and US citations data, and some descriptive statistics on patent and citation data in Japan.

The Japanese Patent Database has been developed by the Institute of Intellectual Property (IIP). It is currently available in Japanese but will be also available in English. The original source of data is the JPO Seiri hyojunka Data, containing the information generated through the acceptance of application to the examination process by the Japanese Patent Office.

The original data have been used to create the IIP Japanese Patent Database by using the NBER database as a benchmark for the selection of indicators to be included in the dataset. The Japanese

Patent Database includes information on 9,027,486 applications, 4,427,840 requests for examinations, 2,594,044 grants from 1964 through 2003, and many variables also present in the NBER database, including citation data and citation-based indicators like the generality and originality indexes. The information on patent right termination is also available.

Further developments in patent statistics and databases in Japan will link the IIP Database to the JPO's Survey on Intellectual Property Activities. They will also carry out activities going on in US and Europe like cleaning of application names, match with firm level data, and addition of other variables such as inventor information or post grant oppositions.

Cleaning Names (Applicants, Inventors), Matching Patent Data with Other Datasets

Rachel Griffith and Rupert Harrison, IFS

This presentation illustrated the aims, methods, difficulties and results of matching EPO patent data with Amadeus firm level data. This matching makes it possible to use accounting data with information from patents. However, it requires standardisation of company names to identify the patents owned by the same firm. The EPO-Amadeus matching is carried out in a way similar to the NBER patent database but there are differences depending on the specificities of Amadeus and EPO data, and of multiple European countries.

The EPO-Amadeus matching is currently carried out for 15 European countries. It uses unconsolidated account data for subsidiaries and consolidated data at the parent level. The matching is done country by country. The first step is automated followed by a manual check.

As a result of the matching process 47% applicants have been matched for the UK. The percentage of matched applicants is lower for other countries (ranges from 18% for France to 33% for Italy). Larger applicants are easier to be matched, as shown by the bigger share of matches weighted for the number of patents owned by the matched applicant (77% for UK, 61% for Germany, 42% for France, 52% for Italy, 43% for Spain, 54% for Sweden). Moreover, the share of matched applicants increases when searching only for companies that filed a patent in more recent years (1998-2002).

However, the matching process may produce some errors. The Type I error is the failure to match a company. This can be solved through manual matching and the search of additional information in other sources of data. The Type II error is the match to a wrong firm. Checks require complementary information supporting or rejecting the validity of the match. Finally, applicants could be matched to multiple firms. Again, additional information like the address or the ultimate owner can help identify the exact match.

The final output of the matching process is a set of tables of matching EPO applicant name-IDs with Amadeus firm name-IDs, and links to other EPO and Amadeus files.

General Discussion – 1st session

The general discussion on the first four presentations aimed at exchanging experiences and discussing the methods used for cleaning names and for matching applicant names to other datasets. It also highlighted the difficulties emerging from the various procedures. It was finally decided to coordinate the various on-going efforts. This included in particular sharing the procedures and a tool for matching names with the support of a website.

European Patent Citations – How to Count and How to Interpret Them

Dietmar Harhoff, Karin Hoisl, and Colin Webb, LMU

This presentation a) illustrated the specificities of the European search process of patent references and of the European Patent citation database, b) discussed how to solve some critical problems in citation analysis, and c) presented descriptive statistics and econometric analysis providing insights about the interpretation of citation indicators. The use of patent citations in economic analysis has been largely encouraged by the availability of the NBER dataset on US patents and citations. However, European patent citations differ considerably from the US citations, and they require specific treatment and corrections. Some of these issues should also be considered in the future version of the NBER database.

The database on European Patent citations uses the following raw data: OECD/EPO data (described in OECD Discussion Paper Webb/Dernis/Harhoff/Hoisl), EPOLINE references (12/2004) updated and checked with REFI (07/2005), EPOLINE data on procedural aspects (search dates) and OPS/ESPACE data on other than WO/EP documents.

This presentation showed how to solve some subtle issues on citations analysis. First, it described the specificities of the EPO Search Process and the classification of references used by the examiners, and their consequences for citation analysis. In particular, it emphasized the following aspects: the differences in the interpretation between examiner and applicant citations, the date used as a reference for the search, what type of documents are preferably referenced (early or later documents, language of application, access to relevant documents). It also provided detailed answers to some relevant questions like – how to deal with timing? Where to get the data on non-EP/WO documents? How to count references and citations? How to deal with the NPL references? How to get the date information?

Harhoff also showed some statistics on the calculation of citation lags, the treatment of references to the Non-Patent Literature (NPL), and the use of equivalent references from different patenting authorities. It also presented some results on the use and interpretation of citations indicators for assessing the quality of incoming applications, the patent characteristics by applicant type, and the impact of citations in value equations, in opposition likelihood equations, and in examination duration equations.

The citations data (updated to mid-2005) will soon be available on-line in an open environment allowing for comments by users.

The PatVal-EU Dataset

Paola Giuri, Sant'Anna, and Myriam Mariani, Bocconi

This presentation illustrated the PatVal-EU survey, some descriptive results of the survey and ideas for a future extension of the survey.

PatVal is a large-scale survey originally designed to be representative of the universe of patents in 6 European countries (France, Germany, Italy, the Netherlands, Spain and the United Kingdom) and subsequently extended to Denmark and Hungary. It covers all technological fields, deals with both for-profit and non-profit applicants, and collects information on small, medium and large business companies.

PatVal's main aim was to collect information about patents, inventors and the underlying innovation process on issues that had not previously been explored in depth because of lack of information in the patent documents. It also provides new proxies for variables like knowledge flows or patent value for which the present measures are subject to some limitations. Moreover, while some surveys on patenting activities have been already carried out, they have limited European coverage and are mostly biased towards large companies.

The presentation provided details about the design of the survey, the methodology, the response rate and some descriptive statistics on three areas: inventors, research collaborations, and use of the patents. In all of these areas, either the literature does not provide information on some relevant topic, or there is ambiguity in the existing measures, or the existing information is potentially incomplete.

For example there are important but under-studied research issues related to characteristics of inventors such as the inventors' life cycle, distribution of productivity across inventors, or the determinants of quantity vs. "quality" of their innovations. The main problem is that there is lack of data on inventors (not available in the patent document). PatVal contributes to fill this gap providing useful data on inventors' personal characteristics like age, gender, education, working experience and mobility, rewards, and other information specific to the patent-inventor (i.e. sources of knowledge that the inventor used to develop the surveyed patent, collaborations with other individuals, etc.). However, further data collection and cleaning work is needed on these issues.

Finally two extensions of the survey were presented: (i) the work already done and in progress for integrating the survey data with complementary data on patents, citations, companies, inventors, regional variables, technological classes; (ii) opportunities for an extension of the survey to other countries and subsequent periods, or for deepening some relevant issues with additional questions.

General Discussion – 2nd session

The final session aimed at discussing areas and financial opportunities for implementing new data collections. Three areas have been identified as being particularly important for new survey-based data collections: the use/non use/licensing of patents, the value of patents (particularly the patent premium as opposed to the value of an unpatented invention), and the inventors. In particular, the idea of an EPIP technical workshop focusing on the elaboration and design of a new PATVAL questionnaire, and the possible implementation of a PATVAL2 survey has been considered.

3. Program of the Workshop

EPIP WORKSHOP
"Patent Data for Economic Analysis"
24-25 February 2006
Venue: Room N03 at Velodromo Building – University Bocconi, Piazza Sraffa 13

Friday, February the 24th

1.15-1.30PM	Welcome and Introduction – Alfonso Gambardella, Univ. Bocconi, Jacques Mairesse, President of the Epip Association
1.30-3.00PM	**Basic Patent Data: Europe and the US**
	The PATSTAT Dataset – Dominique Guellec, OECD
	The New NBER US Patent Dataset – Bronwyn Hall, Berkeley
	Discussion
3.00-3.15PM	Break
3.15-4.00PM	**Basic Patent Data: Japan**
	Japanese Patent Data – Akira Goto and Kazuyuki Motohashi, University of Tokyo
	Discussion
4.00-4.45PM	**Cleaning Names (Applicants, Inventors), Matching Patent Data with Other Datasets**
	Matching EPO Data with Amadeus Firm Level Data – Rachel Griffith and Rupert Harrison, IFS
	Discussion
4.45-5.00PM	Break
5.00-6.30PM	**General Discussion**
	Summary of the Issues Raised in the Day
6.30PM	Adjourn
8.30PM	*Common Dinner at the Restaurant SCIMMIE - Via Ascanio Sforza 49 - MILANO - Tel 02 89 40 28 74, FAX 02 58 11 13 13*

Saturday, February the 25th

9.00-10.30PM	**European Patent Citations and PatVal**
	European Patent Citations – How to Count Them and How to Interpret Them – Dietmar Harhoff, Karin Hoisl, and Colin Webb, LMU
	The PatVal-EU Dataset – Paola Giuri, Sant'Anna, and Myriam Mariani, Bocconi
	Discussion
10.30-10.45PM	Break
10.45-1PM	**General Discussion**
	New data collections (e.g. PatVal-EU2)
	Summary of the Meeting
1PM	Adjourn

4. List of Participants

	NAME	AFFILIATION
1	Bordt Michael	Statistics Canada
2	*Colin Webb	OECD
3	Cremers Katrin	Centre for European Economic Research
4	Dambois Denis	European Commission - DG RTD
5	Danelutti Tea	EPFL - CDM - CEMI
6	Ebbensgaard Marie	CEBR
7	*Gambardella Alfonso	Univ. Bocconi
8	Gareth Macartney	Institute for Fiscal Studies
9	Gaulé Patrick	Ecole Polytechnique Fédérale de Lausanne
10	Giuri Paola	Sant'Anna
11	*Griffith Rachel	UNIVERSITY COLLEGE LONDON
12	Goto Akira	University of Tokyo
13	*Guellec Dominique	OECD
14	*Hall Bronwyn	UC Berkeley and U of Maastricht
15	Hamdan Intan	Ecole Polytechnique Fédérale de Lausanne
16	*Harhoff Dietmar	INNO-tec, Munich School of Management
17	*Kahin Brian	UNIVERSITY OF MICHIGAN
18	*Karin Hoisl	INNO-tec, Munich School of Management
19	Leone Maria Isabella	Luiss Roma
20	*Lotz Peter	COPENHAGEN BUSINESS SCHOOL
21	*Mairesse Jacques	IMRI – Université Paris Dauphine, F
22	Mariani Myriam	Univ. Bocconi
23	*Mohnen Pierre	MERIT, University of Maastricht
24	Montobbio Fabio	Univ. Bocconi
25	*Kazuyuki Motohashi	University of Tokyo
26	Nakagawa Takashi	National Graduate Institute for Policy Studies (GRIPS)
27	Oyamada Kazuhito	National Graduate Institute for Policy Studies (GRIPS)
28	Padula Giovanna	Univ. Bocconi
29	Romanelli Marzia	Univ. Pisa
30	Rullani Francesco	Univ. Bocconi
31	*Rupert Harrison	Institute for Fiscal Studies
32	Stephane Lhuillery	EPFL - CDM - CEMI
33	Sunami Atsushi	National Graduate Institute for Policy Studies (GRIPS)
34	Tateo Arimoto	National Graduate Institute for Policy Studies (GRIPS)
35	Thoma Grid	Univ. Bocconi
36	Valentini Giovanni	Univ. Bocconi
37	*Veugelers Reinhilde	K.U.Leuven
38	Vezzuli Andrea	Univ. Bocconi