World Patent Information 32 (2010) 279–286

Contents lists available at ScienceDirect

World Patent Information

journal homepage: www.elsevier.com/locate/worpatin

Quality assurance in the EPO Patent Information Resource

Miguel A. Albrecht *, Rex Bosma, Trudy van Dinter, Jean-Luc Ernst, Koen van Ginkel, Fenny Versloot-Spoelstra

Data Resources, European Patent Office, Patentlaan 2, 2288EE, Rijswijk, The Netherlands

ARTICLE INFO

Keywords:
Quality assurance
EPO
Patent Information Resource
PIR
Master databases
Data standardisation
DOCDB
REFI
FTM
MCD
ECLA
NPL
OPS
Classification data
Citation data
Examiner workflow
esp@cenet®
INPADOC
Patent families
RID
Reference identifier

ABSTRACT

In increasingly competitive global markets, access to consistent and dependable information on innovation is becoming more and more indispensable. This article describes the efforts performed at the EPO to create and maintain reliable master databases of prior-art information. Data standardisation and patent family rules applied at the EPO are described in detail and how master databases support the workflow of examiners. The authors discuss the benefits that the EPO, its partners and the community at large obtain from these activities. Finally, a look at what lies ahead reveals exciting potential developments for the near and mid-term future.

1. Introduction

The globalisation of economic markets have lead to a globalisation of innovation. Industry, patent offices and patent information professionals struggle to keep both the volume and consistency of global patent data under control. Being able to trace inventions throughout regional markets in a reliable manner has become an essential part of economic competition and an important element for patent offices to master their workload. The key to solving this dilemma lies in reliable and consistent patent information.

Quality in patent information is a much-debated topic [2–6]. Many definitions of quality have been put forward and, while they are all valid, the question remains: what does quality mean to you? Are there tangible benefits to be claimed from investing in quality in patent data?

This article will put the case for a strongly affirmative answer to that last question.[1] In a nutshell, data quality translates at the EPO into a data mining and refinement process which helps us to obtain sufficiently clear patent data and puts us in a position to carry out classification and search tasks in a time-efficient and resource-friendly manner. We will describe the efforts that the EPO is making to first turn patent data into information and then transform it into knowledge. We will estimate the quantifiable benefits and explain how the pursuit of quality has changed the EPO's view of documentation with respect to master databases and business rules.

Section 2 of the article summarises the work that has been and is being done at the EPO to create and maintain master databases for prior-art information. Section 3 describes the standardisation rules that are applied to data and how patent families and citations can become powerful sources of information for examiners.

* Corresponding author. Tel.: +31 70 340 4720; fax: +31 70 340 3320.
E-mail address: malbrecht@epo.org (M.A. Albrecht).

[1] Editor's note: In a parallel article from the EPO, Scott describes a wide range of other efforts throughout the EPO to maintain and enhance the quality of patent information available to the Office's examiners and the public alike [22].

0172-2190/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wpi.2009.09.001

Section 4 outlines how master data is used not only to create searchable databases but also to support examiners' day-to-day workflow. Section 5 details the benefits of master data, while Section 6 takes a look at what lies ahead.

2. The EPO Patent Information Resource (PIR) – some facts and figures

It has always been part of the EPO's mission to deliver high-quality searches with a view to granting strong patents. As a consequence, the office has worked continuously to develop and enhance its patent data collection strategies in order to ensure optimum quality in its data stock.

At the end of the eighties, the EPO's automation strategy was to maintain the independence of its internal search tool, known as EPOQUE [10], and to enable searching in a maximum of available data sources. As a consequence of this strategy, it started to build a comprehensive repository of patent data.

Over time this repository has become an invaluable resource, much appreciated not only by our own examiners but also by the IP community at large. The EPODOC search file in EPOQUE is used on a daily basis by examiners at patent offices worldwide. EPO data products delivered on a regular basis find their way into a wealth of commercial offerings. Many statistical studies rely heavily on the PATSTAT [17] database.

With the introduction of relational database technologies (RDBMS) in 2004, the EPO adopted Master Data Management methodologies, enabling it to make a quantum leap in its focus on data quality. Where the old database infrastructure allowed for a maximum of one million records per year to be added or updated, the new system has made such limits obsolete.

By way of illustration of the scale of this effort, 18.9 million documents were added to the collection in the period 2004–2008, and 5.4 million corrections performed on them (Fig. 1).

The high peaks in 2004 and 2006 reflect the treatment of backfile data that the EPO had not been able to load before. This considerable increase in volume was only made possible by a major system migration – from a conventional to a relational database – that was completed at the end of 2003.

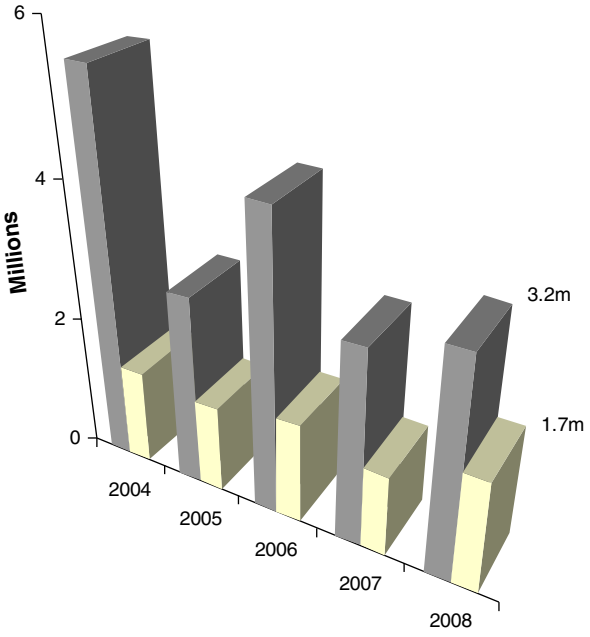

Fig. 1. Volume of patent documents loaded into the EPO PIR (in grey) and the number of corrections applied to them (in yellow). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The EPO Patent Information Resource (PIR) collects together all the EPO documentation master databases:

- DOCDB: master database for patent bibliographic data

  DOCDB plays a central role in the PIR. Originating from the old INPADOC-PFS [19] file, DOCDB has evolved substantially both in volume and content. It is the connection between all the other patent master databases and contains well over 69 million patent documents.

  In what is a quite unique process, the data is procured from 92 national and regional patent offices and then collected and validated, corrected and standardised to form the core of the EPO Patent Information Resource.
- REFI: master database of citation data

  REFI contains a wealth of information on cited patent and non-patent literature relating to international, regional and national applications and publications. These references are key sources of information for EPO examiners and other users of patent information. They enable the quick retrieval of prior art documents or, failing that, help to prepare a more targeted search strategy. As such, attention to coverage and quality are of the utmost importance to the key users.

  The EPO citations database currently contains more than 90 million references (to both patent and non-patent literature) relating to over 13.1 million applications/publications.

  In the period 2004–2008, 28.9 million citations were added to the database, over 1.8 million corrections were performed by human input and an equivalent number automatically via software.

  Fig. 2 shows the number of citing and cited documents available in the EPO PIR per year of publication. The continuous growth of the number of citations per search report gives a clear indication of the growing complexity of patent applications in the last decades – a topic discussed in Refs. [7,8,13] among others.
- FTM: full-text master database

  The FTM database contains all full text and corresponding embedded images of the publications and applications available at the EPO. Since 2006, all incoming patent applications have been systematically OCRed and loaded into FTM, offering examiners direct access to viewing full text, comparing versions and supporting the draft of examiner communications with applicants. When the backfile load operation is completed at the end of 2009, this database will house over 23 million patent documents.
- MCD: master classification database

  The MCD database [9,11] contains all IPC (International Patent Classification) symbols worldwide, with around 250 million allocated symbols for 61 million documents. More than 95% of DOCDB documents have an IPC symbol. The MCD plays a central role in managing the re-classification of documents under the IPC8 procedures.
- ECLA master database

  The ECLA master database supports the distribution for classification of incoming documents, as well as the allocation and maintenance of ECLA (European CLAssification [15]) symbols. More than 38 million ECLA symbols have been allocated to patent families.
- NPL: master database of relevant non-patent literature

  The master database for non-patent literature contains the bibliographic data of all NPL documents cited by EPO examiners. A complex matching procedure brings together NPL documents with identical technical content. The results, i.e. families of NPL documents, are stored in the database as described at the end of Section 3.