World Patent Information 32 (2010) 120–128

Contents lists available at ScienceDirect

World Patent Information

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

Review

The text, the full text and nothing but the text: Part 2 – The main specification, searching challenges and survey of availability ☆

Stephen Adams

Magister Ltd., Crown House, 231 Kings Road, Reading, RG1 4LS, United Kingdom

ARTICLE INFO

Keywords:
Full text
Document segments
Description
Examples
Data values
Document quality
Proof-reading
Multi-lingual texts
Diacritics
Cross-language searching
Database survey
Review article

ABSTRACT

Effective searching of electronic full texts of patent documents requires both appropriate search engine technology and high quality source documents. This article reviews the impact of both the historic development of online searching and of document preparation upon the resulting databases. Many standards were developed at a time when patent documents were wholly paper-based, and may no longer be suitable as current guidelines for the preparation of full text electronic databases. Part 1 previously discussed the impact of applicant guidelines and patent office practice upon the usefulness of title, abstract and claim for retrieval. Part 2 reviews the main part of a patent specification, to understand the challenges of using this for various types of patent retrieval. A short survey of the main providers of full text patent information concludes the review.

1. Towards 'richer' text

In the first Part of this article [Adams Part 1 321 – in press], I discussed the underlying standards and guidelines currently in use by applicants and patent offices when creating and publishing patent specifications and periodical gazettes. Certain portions of the text material of these documents – such as the titles, abstracts and claims – form useful searchable elements in their own right, and were also reviewed in that paper.

This second Part of the article turns to consider the main body of the specification. Just as with titles and abstracts, the usefulness of the modern electronic full text reflects the means by which it was created, as well as being constrained by the available functionality of the search engine which is used to retrieve it. There is a general feeling amongst the patent search community that we have yet to exploit the full potential of full text sources, and I will explore some of the limitations and possibilities in this article.

1.1. Document segmentation

In the course of preparing the conference paper on which this article is based, I put out a call for comments on the IPI and PIUG discussion lists. The request was for searchers to suggest improvements to the existing forms of full text available. By far the largest single response was to request that patent full text should be made available with improved segmentation. Such segmentation can take two forms; either according to the form of information content (i.e. *where* in the document body the text appears, such as the description, examples or claims) or the data elements themselves (i.e. *what* the text represents, such as words, data values, units of measurement, captions and diagram labels).

At the present time, database producers are largely ignoring different content types. The most segmented files only offer the ability to distinguish between the body of the specification (often termed the 'description') or the claims. No finer detail is available. It is interesting to note that some publishing authorities, including the International Bureau of the PCT, recognise the usefulness of sub-dividing very large blocks of text into smaller segments. The PCT Administrative Instructions [1] allow for the possibility that the applicant can break their application text into parts such as Technical Field, Background Art, Disclosure of Invention, Brief Description of Drawings, Best Mode/Modes for Carrying Out the Invention, Industrial Applicability, Sequence Listing and Sequence Listing Free Text. Clearly not all segments will be applicable to all inventions, but it does provide a basic framework which could be used by database producers. The most common suggestions from users were for additional segmentation into one or more of the following: Prior Art, Description, Embodiment, Examples, Experimental and Drawing Captions. Whilst these two approaches to segmentation have items in common, it is clear that the PCT format is geared towards legally-significant distinctions (such as best mode) whilst the suggestions from the discussion lists were more suitable for information retrieval purposes.

☆ This article has been developed from a presentation by the author at the International Patent Information Conference, IPI-ConfEx, March 2007, Sorrento, Italy.

E-mail address: stephen.adams@magister.co.uk

0172-2190/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wpi.2009.06.002

In addition to segmentation according to 'part of document', there was some interest in improving retrieval by adding segmentation according to 'data type', such as words, tabulated values (including the ability to distinguish word-based or chemical structure-based tables, and for data capture to improve sufficiently such that tabular data are longer rendered as images), numerical properties, numerical ranges, sequence data and literature citations (patents and non-patent literature) from within the document (as opposed to examiner search reports which are already stored in a separate field, when present).

At first sight, such suggestions seem both perfectly achievable from the technical point of view, and useful for the searcher. However, there may be hidden hazards to adopting this approach. As soon as a database record is given a new field or sub-field, such as would be formed by one of the above segments, it is very tempting for searchers to assume that the field is populated both systematically (i.e. all records contain the field, even if it has a null value) and unambiguously (i.e. the field means the same thing in all records). Failure of either or both of these assumptions will hold dangerous consequences for the searcher.

If a field is not populated systematically, then running a search only in that field results in a *de facto* limitation to the sub-set of the database consisting only of those records which contain the field. If a field is not populated unambiguously, then it is quite possible that data which 'belongs' to this field may occur elsewhere in the document, and be missed in the search. Patent agents already use many techniques to obscure the true meaning within their clients' applications, and we cannot rule out the deliberate mis-use of document structure standards (such as putting 'examples' text into the 'disclosure' field) in order to hinder effective retrieval.

It is worth noting at this point that our databases for patentability searching are often based heavily (if not exclusively) upon unexamined applications. The content of these documents is the responsibility of the applicant, and although the content is subject to a formalities check, the patent offices have little scope for correction or re-arrangement of the text. We will see later in this paper some of the consequences of the patent offices' responsibility for the content of patents granted after substantive examination, and their adherence to explicit document standards.

There is already some concern in the searching community that there may be adverse quality implications as a result of increasing e-filing. Whilst these systems may eliminate re-keying errors, which is to be welcomed, their adoption means that there are fewer points in the processes between initial filing and publication at 18 months in which to capture and correct mistakes. The responsibility of the patent office at this stage is to publish the application *as filed*, which may contain any number of errors, with consequent implications for information retrieval. If systems for document creation and online filing are incompatible at the point of information capture, then documents of sub-optimal quality will be published and the databases containing them will be poorer quality. Consider one small example.[1] WIPO standard ST.9 [2] defines INID code 83 (information concerning the deposit of micro-organisms under the Budapest Treaty), which allows for the possibility of this data to appear on a front page. The WIPO standard ST.32 [3] defines corresponding SGML tags <B830>, <B831>, etc. for the same information. As a result of these standards, Questel has set up data structures based on the ST.32 tags, ready to accommodate the data into their file of full text and bibliographic European Patents. However, the EPO data input system EPASYS did not originally include a specific input field to enable the micro-organism deposit data to be captured, even if the applicant had tried to record the data on their application form. Furthermore, the instructions to applicants did not emphasise the need to distinguish this information. As a result, the content was either lost entirely or added to an obscure 'notes' field at the point of entry to EPASYS. Questel's data feed records show that the SGML <B830> tags are rarely populated. As a result, the mere existence of the field in the database structure becomes an unreliable indicator of the presence or absence of the data, and makes it effectively useless as a search tool. To an inexperienced searcher, the database summary sheets may suggest that comprehensive retrieval of micro-organism deposit information can be achieved by searching on the field, whereas much of the data are missing or placed elsewhere.

In case anyone is in doubt regarding the extent of errors in patent texts, consider a report by Intellevate, based on a survey conducted during January 2006 [4]. The survey identified proof-reading mistakes in 98% of a sample of US patents. Of these, 56% of errors were by the USPTO and 44% of errors by the applicant or the legal practice handling the filing. A comment is instructive:

> *"[Proof-reading] is also a predominantly lacklustre task that many law firms and legal departments would rather not have to take on."*

As a result of these observations, I would suggest that the introduction of further segments and search fields into full text databases should not be undertaken until better compliance with document creation standards is achieved. Applicants, patent offices and database producers need to invest in eliminating much of the incompatibility between theoretical input systems and practical everyday practice, and the consequent scope for error. Human elements are at work as well as technical ones, and both need to be addressed before we achieve better quality text applications, which could be segmented with confidence. Searching full text is hazardous enough, without introducing another "80% perfect" system of handing documents, and exposing ourselves to significant risk of missing relevant documents.

1.2. Diacritics and non-Latin character sets

This section will consider the challenges facing the publisher and searcher in relation to character sets. The current issues of understanding the Chinese patent literature are well known, but there are a range of issues much closer to 'home' for the Western searcher, which are equally in need of resolution before we can utilise patent full texts to their best advantage.

Amongst the member states of the EPO, the various national languages already include characters such as ð, þ (Icelandic), ø, å (Danish), ş, ı (Turkish), ł, ś (Polish) and д, л (Bulgarian), as well as the more familiar French, German, Spanish and Portuguese diacritics such as acute and grave accents, circumflex, cedilla and umlaut. These raise two distinct, but related, issues for the searcher: (a) how is the original text indexed? and (b) how can it be retrieved?

Unicode has been available as a method for indexing multi-lingual text for some years, and librarians and information specialists are well aware of its advantages and pitfalls, especially in relation to subject cataloguing [5]. However, patent searching is faced with a significant challenge due to the size and ongoing utility of its back-file. No amount of technical development will provide any retrieval improvement unless it is clearly understood how (or whether) the *existing* document corpus has been processed in order to allow systematic retrieval across time. Unfortunately, even major database producers have sometimes failed to archive accurate records of any changes in policy (such as the date when the Derwent World Patent Index began to accept American-English spelling forms as well as British-English). If a searcher cannot be sure whether (for example) a German ü (u-umlaut) has been indexed as 'u', 'ue' or 'ü' consistently across the 30+ years of a full text database, they will need to spend time identifying alternative retrieval

[1] Linda Williams, Questel, personal communication.