



Analysis of Trends in Search and Retrieval of Intellectual Property-Related Information

23 September 2005

Clunies Ross St
Black Mountain ACT
www.cambia.org

Analysis of Trends in Search and Retrieval of Intellectual Property-Related Information

Marie Connett Porceddu *PhD*

Douglas Ashton

Neil Bacon

Nicholas dos Remedios *PhD*

Carol Nottenburg *PhD JD*

Shoko Okada *PhD*

Gregory Quinn

Wei Yang *PhD*

Richard Jefferson *PhD*

CONFIDENTIAL

Copyright 2005 by CAMBIA

CAMBIA, Patent Lens, Dekko and BIOS are marks used by CAMBIA. Names of other companies, search services and websites may be registered trademarks of other entities. No license to use these trademarks outside this confidential report is implied or expressly granted.

Report Overview.....	5
Chapter 1: Overview of current IP information systems.....	8
<i>Scope of the study.....</i>	<i>8</i>
Search strategy / queries	9
Features for comparison	10
Results and observations	12
Strengths and weaknesses of search sites.....	13
Selected National Patent Offices.....	13
No-charge providers.....	18
Commercial Providers	19
Thomson-Derwent providers	19
Independent fee-based or subscription-based providers	20
Non-patent literature	22
Traditional knowledge	25
“Value-added” information summary.....	26
Chapter 2: Discussion of Emerging Trends in Information Retrieval.....	28
<i>Information retrieval.....</i>	<i>28</i>
Inverted Index	30
Stemming.....	30
<i>Information Retrieval Models</i>	<i>31</i>
Boolean Search	31
Term Frequency - Inverse Document Frequency (TF-IDF).....	31
Proximity	32
Vector Space Models.....	32
Latent Semantic Analysis (LSA)	32
Cluster Analysis	33
<i>Inference Models.....</i>	<i>33</i>
<i>Dekko</i>	<i>33</i>
<i>Optical character recognition (OCR).....</i>	<i>36</i>
<i>Natural Language Processing (NLP) applied to searching Chinese, Japanese, and Korean (CJK) language patents</i>	<i>36</i>
Chinese Language.....	37
Japanese language.....	37
Cross language Information Retrieval	38
<i>Patent Claims</i>	<i>38</i>
Claims parsing.....	39
Types of claims	39
Quality of data for claims analysis	40
Chemical Searching	40
Searching Patent Data for Biological Sequences read on by claims	43
Patent Sequence Databases	44
Chapter 3. Emerging and potential new uses for IP information	47
<i>Searching of IP information for trade leverage.....</i>	<i>47</i>
<i>Searching of IP information to foster Australian technology development.....</i>	<i>48</i>
<i>Searching of IP information for putting Australia in a leadership position</i>	<i>49</i>
Chapter 4. Options for delivering improved IP information: Conclusions	51
<i>Recommendations from Chapter 1.....</i>	<i>51</i>

Important patent search features	51
Special Recommendations for IP Australia	51
<i>Recommendations from Chapter 2</i>	53
Special Recommendations for IP Australia	53
<i>Recommendations from Chapter 3</i>	55
Special Recommendations for IP Australia	55
<i>Final conclusions</i>	55

Report Overview

IP Australia commissioned this report to develop an informed view of the current and future directions in search and retrieval of IP related information as used by IP professionals within IP Australia, other IP Offices among the IP Australia customer base, and other informed customers of the global IP system.

Australian patent data is currently largely unavailable to the public in a unified, full text searchable form. Trademark data and plant variety rights data are set up in systems for ready searching, so the report concentrates on patent data. This report will examine some of the mechanisms that are available to searchers of IP data with an eye to identifying the best features that should be incorporated into the eventual service that IP Australia could provide for its stakeholders, both internal (examiners) and external (the public, particularly innovators and implementers of technology).

The report will also look at pathways to this future and interim solutions for the support of these stakeholders.

It is timely that IP Australia has identified a need to consider IP informatics, because of some significant changes in the landscape in recent years. Some of these changes relate primarily to information technology and the sociological change that advances in this area are bringing about:

- The volume of IP filings has greatly increased, and ever more exponentially the volume of data, including data accessible in Australia from overseas jurisdictions via the Internet, that can be accessed for consideration of prior art.
- Changed expectations about what constitutes acceptable search functionality and performance have arisen from growth of the Internet and popularity of search engines such as Google, which are simple to use but not very transparent or flexible for user control.
- The capacity of IT systems to store, search and process data has greatly increased.

The latter provides a new window on possibilities that neither national offices nor “value-added” commercial providers have yet realised, to add true value to IP data and metadata.¹

Why is this useful and timely? National offices such as IP Australia need to consider not only how to handle the volumes of data for proper prior art searches, but how to foster healthy domestic innovation for economic growth.

National offices worldwide could formerly view their main function as relatively autonomous provision of a receiving office for filings regardless of origin, but are increasingly subject to international patent harmonisation efforts and development agenda pressures. There have been many other changes in the landscape in recent years, related to trade, that in combination with IT and IP factors result in both disadvantages and opportunities for Australian innovation and the Australian economy:

¹ For example, in CAMBIA’s Canberra-based Patent Lens IT team, we are working on developments such as links to status information, tutorials, and tools for annotation, for exemplifying regarding licensing, that may be able to lead users of patent data to more robust conclusions on where the dominant rights are held.

- With the implementation of international trade agreements and growth of multinational business, patent filings from overseas constraining use of inventions in Australia are rising.
- Intellectual property rights in most technology areas no longer cover primarily simple inventions, but form an interlocking thicket of competing and overlapping rights to exclude. For multiple technologies this pattern has given rise to a monopsony, a market situation in which the product or service of several sellers is sought by only one buyer, which if unable to obtain all the necessary licenses is locked out of delivery mechanisms.
- Freedom to innovate and investment in innovation can be increasingly stifled by Fear, Uncertainty and Doubt (FUD) in the age of information, wherever the information is perceived to be potentially incomplete or ambiguous.
- A fee-based or subscription-based patent informatics industry preys upon these fears, and increasingly provides access for the privileged only to “value-added” information. Consolidation in the industry suggests a drive away from performance enhancements, while prices are likely to rise.

We suggest that the provision of a search service, offering not only full text searching of Australian patents but the harmonised datasets of other important jurisdictions, will support Australian economic development.

Much of the cost of developing technology *de novo* or extending the use of technology in Australia can be economised by uncovering enabling descriptions of technology and shedding light on where this technology can be freely used.

The intent of the enabling description requirement was to provide information on the technology so that it could be used outside the monopoly grant and readily improved. Thus, there is a latent potential in the patent system for prudent legal use of technology that has been described but is not subject to valid claims in a particular jurisdiction.

Where technology has been claimed, good patent informatics can support discovery of invalidating information such as prior art, possible invent-around and invent-beyonds. It is also reasonable to use informatics to encourage wide licensing and delivery of patented technology.

This report is structured in four components:

1. An overview of currently available IP information sites and approaches, with appendices providing commentary on and analysis of their strengths and weaknesses.
2. A discussion of emerging trends in searching and accessing IP information of special relevance to Australian stakeholders, together with likely future directions in this field. These trends include:
 - Information retrieval and ranking algorithms
 - International database searches and the languages that they contain
 - Claims parsing
 - Searching some of the types of matter claimed in many patent applications in Australia, such as biological sequences and chemical structures

In these two chapters, much of the information from other sites was inferred through a number of iterative searches in several art areas, including mechanical, chemical, biotechnology and nanotechnology. Due to our detailed knowledge of how our systems work, we have been able to provide exact information on the mechanisms that are used in the CAMBIA BIOS Patent Lens, and the subset of challenges presented with life

sciences as exemplary of the underlying information management and technology implications and issues at stake.

3. A discussion of emerging and potential new uses for IP information for IP Australia, its customers and other stakeholders.
4. Recommendations on options for delivering improved IP information and access services for the support of Australian innovation and the Australian public good.

This chapter concludes that all three preceding chapters support a recommendation and provide practical advice in a single major direction: that the most urgent and important step is implementation of a Boolean accessible, scalable search engine capable of doing full-text searches of Australian patent data, with unified searching as soon as possible of Australian data together with the largest patent data sets, namely the US, EPO and PCT data.

Status data on Australian patent applications and those in the same patent family in other jurisdictions should be incorporated as soon as possible in the portions or versions of the Australian database available to the public, so as not to disadvantage Australian inventors and investors.

Support for searches that assist in understanding breadth and interpretation of claims can and should be developed, even in the challenging areas of chemistry and biological sequence data. While the PCT data already requires a search engine and OCR that can handle Japanese language, eventual incorporation of at least the Chinese and Korean datasets has similar requirements and should be planned.

Chapter 1: Overview of current IP information systems

The primary need of IP Australia for intellectual property informatics analysis is related to patents searching and use of patents search information by its internal and external stakeholders. Examiners need ready access to the Australian patent data set, and they need to be able to do prior art searching that will support granting only claims with a reasonable presumption of validity. The public also needs ready access to the Australian patent data set, and the patent data sets of other jurisdictions, to do technology searching and freedom to operate analysis for Australian innovations.

Good informatics should support each of these needs, and IP Australia has already developed good informatics systems for trademark and plant variety rights cataloguing that support both the searching needs of examiners and the needs of the public users of these systems. Therefore, the report concentrates on informatics of patent and non-patent data required to support the perceived requirements of the innovation system reflected in patents.

Over recent years the volume of filings and related data has greatly increased, as has the capacity of IT systems to store, search, and process this material. At the same time the growth of the Internet has changed public expectations and indeed the needs of examiners and the public, particularly with respect to prior art searching.

National IP offices, fee-based and subscription-based data providers and public good providers have responded to these challenges in different ways. Clearly, there are a variety of potential measures of what constitutes acceptable search functionality and performance; by some metrics, very few providers meet the high standards of public requirements.

Scope of the study

This chapter presents an overview of the strengths and weaknesses of these different approaches based on the perspectives of two different broad classes of users:

1. The user skilled in patent terminology and searching, representing primarily examiners but also patent attorneys able to make ready use of interfaces that demand or offer, for example, Boolean queries, patent classification systems such as IPC and ECLA, and searches of patent-specific data fields such as “inventor” and “claims”.

The assumption was made that the primary requirement of such a user is to find information related to validity of patent claims. In particular, this user should be able to find, for any patent application, the entirety of prior art that should require modification of claims such that no claim reads on any available prior art.

The needs of this user are very important to the credibility and quality of the Australian patent system. There is an opportunity for Australian patents to be respected more widely if the metes and bounds of claims are clear and non-overlapping, and the presumption of validity following examination is well supported.

2. The innovator or user of technology, who may not be an expert in patent terminology (though possibly expert in the terminology of a particular field of art). In this document we refer to the “public” technology searcher.

This user has a strong economic interest in quality intellectual property informatics. The desire is to avoid wasteful investment, either in attempts to develop and patent technology that is already in the prior art, or in attempts to implement and market technology covered by exclusionary rights to which access may not be available.

A clear understanding of what is claimed must be supported by the availability of definitions and citations from the specifications (and any literature deemed relevant in

construing claims, such as contemporaneous dictionaries). For such a user, ability to search titles and abstracts only would be quite insufficient. Full text searching of specifications is a requirement.

Prior art is certainly of interest to this type of user, including both patent and non-patent data, but so is status information (where is this patent in force?) and applicant or assignee information (from which entities would practicing this technology require a license?).

The needs of this user are very germane to IP Australia's function within the innovation system upheld by Australian and international intellectual property law. If such users are facilitated in their attempts to invent new technology rather than re-inventing the wheel, and to commercialise technology in non-infringing ways, the economy benefits and social goods are multiplied.

In this report we provide user analysis of a number of representative public and commercial IP information providers. A set of queries was devised to allow direct comparison of the search engines, described below. Observations were made on a variety of parameters, services and features of special interest to each of the two types of users mentioned above. We also noted convenience metrics such as speed and whether patent documents can be downloaded in entirety as PDF files.

While analysis of all available providers would be well beyond the scope of this report, it was possible to choose the major providers, in that consolidation has been a key trend within the commercial sector, with multi-nationals such as the Thomson Corporation controlling a number of previously independent providers. A number of databases, such as the Derwent World Patent Index (DWPI), have become industry standards and are used by a number of otherwise competing providers. For this study we also looked at a number of additional providers, aiming to select primarily those that use somewhat different search approaches than the standard, and a number of national patent offices.

Despite the concentration of private search service ownership and the sharing of some core resources, the range and complexity of searching options available is such that conducting an effective search using existing tools is in no way straightforward for Australian innovators. Even these databases have gaps in data, sometimes caused by delays with access to raw data.² For Australian patent literature, we were unable to locate providers that would allow a comprehensive search of full specification texts.

Search strategy / queries

A number of test cases were designed for assessing the different databases. Eight US or EP utility patents or PCT patent applications in different art areas (chemistry / biotechnology; mechanical; electrical) and a design patent were chosen to represent technologies on which extensive searches were conducted, while up to five additional cases were used for more in-depth study of sites that claim to use divergent search technology. For non-US patents, each of the documents had an associated search report that could be utilised as one measure of whether the use of a particular database found the same prior art that a European examiner would have noted.

It is useful to remember that the claims in the original patent application may not bear resemblance to the claims in the issued patent. Rather than focus on claim structure, the approach of an examiner is to determine the inventive *concept* and base a search on that.

² <http://scientific.thomson.com/support/patents/coverage/latestupdates/closinggaps/>

In general, a patent examiner will approach finding prior art by a combination of IPC designation(s) (US classifications for examiners at the USPTO) or ECLA (EPO) classifications, keywords, and the inventor names. The latter are important because invalidating prior art is most often found in public disclosures by the inventors.

The public technology searcher is more likely to bypass the use of classifications, but may add the use of assignee/applicant names and may be able to use keywords in a way that reflects more in-depth field-specific knowledge.

Classification codes are assigned precisely on the basis of the technology field of the invention; this is where related art will be most likely found. Thus, for each of these patent documents, the initial search for examiners used the primary IPC. As expected, too many results were returned. Further refinement of the search used keywords. Given the amount of time examiners are generally allowed for a search, USPTO examiners tend to target obtaining 50-100 results.³ Titles, abstracts, and, at times, drawings are scanned to target the relevant documents.

Though searches on classification codes typically returned more documents than were relevant, in some fields of art they failed to find the most relevant documents. For example, inventions that are part of “nanotechnology” may contain elements of different art areas. To improve the ability to search and examine nanotechnology-related patents, a new cross-reference digest (to be eventually replaced by a classification schedule) has been established by the USPTO and other patent offices may follow suit, but no patent or application has yet been assigned to the new classification. Nano-scale objects themselves are not new, and for our main example case, US6689338 (“Bioconjugates of nanoparticles as radiopharmaceuticals”), searches using “nano*” in combination with the classification of the main inventive subject matter do find patent documents that appeared to be highly relevant in many of the databases.

Design patents were also included within the scope of this survey. A brief background on design patents is useful to know when examining the databases and results. The United States examines applications and issues patents for designs, but most countries do not. Instead, designs (often called industrial designs) are protected by registration, which may or may not entail an examination for novelty, or by copyright law, or sometimes both. WIPO administers a treaty, Hague Agreement Concerning the International Deposit of Industrial Designs, that provides a procedure for international registration. As of 26 April 2005, 42 countries are party to the Agreement. WIPO maintains a database of deposits under the Hague Agreement; the database is searchable, but it was not included as part of this study. The only design-related documents in any of the databases that were part of this study are US design patents.

Search strategy for design patents used classification codes as a primary term, which is the method most used by examiners, although it is less straightforward for the public user. The particular example chosen for this test was USD502885, a design for a watch dial face. Because it wasn't apparent that any database can take a range input for classifications (a short-coming of all the databases), keywords and a top-level class constitute the most efficient search.

Features for comparison

Searches involve multiple steps: the formulation of queries, input of the queries via an interface, translation of the queries into actions by the search engine(s), the interaction of the search engines with data, collation of results, presentation of the results in some ranked order, often followed by refinement of the query for the start of a new cycle. Thus, there are many points at which qualities of the various sites and services can differ, and many points at which faults in

³ Personal communication

transmission, collectively known as "bugs", can occur. At the same time, there are many points of intervention to make searches perform as the user intends. Features that are of interest to the needs of a patent examiner were chosen in four main categories: input, data, and output, and other user attributes:

Input criteria

- Ability to search by IPC and/or ECLA
- Complex Boolean structure
- Search by publication date range
- Wildcard options ('?', '*', etc) for character and string
- Ability to search terms "near" each other
- Assignee, inventor list to catch mis-spelling

Data sources

- Full text for searches
- English abstracts for JP
- INPADOC
- Design patents? Images?
- Non-patent databases at same site?
- Data coverage (countries)
- Other types of data available, such as chemical structures or biological sequences

Output criteria

- Results list can be ordered by file date?
- Results list can be ordered by publication date?
- Results list can display titles and abstracts
- Links are provided to full documents
- Can the relevant patent documents (and non-patent literature for available databases) be retrieved within the coverage provided by each database?

Other criteria

- Speed of returning results
- Ease of navigation

To represent not only the search needs of the examiner but also the "public" user viewpoint, we assessed the following additional criteria, with the assumption that the user would want to carry out a series of increasingly refined or increasingly broad searches related to technology investigation and freedom to operate (FTO), and then to find out considerably more information relating to the capability to use technology in particular documents.

Input criteria

- Ease of constructing a search query for the "non-expert" user
- Availability of combinations of search fields for narrowing searches without sole recourse to Boolean operators
- Availability of thesaurus or dictionary choices to assist with creating keyword sets (e.g. rice = ris = *Oryza*) and ability to search with keywords containing accents or in non-English languages
- Can search terms be saved/retrieved to refine a previous search?

Data sources relating to particular identified patent documents

- Accessibility of legal status information
- Ease of extracting definitions of terms in the claims from the specification
- Accessibility of cited references
- Accessibility of prosecution history
- Licensing information

Output criteria

- Option for relevance ranking of results
- Are documents other than English translated into English, and if so to what extent (abstract only? claims only?)

Results and observations

National patent offices (we examined EPO, WIPO, JPO, SIPO, KIPO, USPTO, and IP Australia) generally provide only their own patent data, but with quite varied interfaces.

By the other public sites and the commercial sites, the data coverage is generally within a minimal set: US, EP, PCT, sometimes JP and INPADOC. Other occasionally included countries are typically European countries - notably the United Kingdom, Germany, and France. Questel-Orbit appeared to have the largest set of countries covered, although coverage from some of these countries was admittedly inconsistent. The commonality of datasets is not surprising given that there are a limited number of organizations that produce the data; the providers here most likely procure their data from these few producers.

Factors that set apart these surveyed providers comprise input format, the results format and the ease of using the site. Although less important, convenience factors such as the speed of results being returned and navigation played a role.

As discussed below, most of the providers have an input interface that is at least usable by an examiner. Most of the sites use "fill-in the blanks" forms, in which the form contains a set of fields (e.g. assignee, title, abstract, inventor). As an alternative, there may be a blank box that accepts Boolean queries for which the user specifies the field(s) associated with each search term. Refining queries depended on the style of input. Depending on the provider, the user has to return to the form or can directly modify the query that is written out in standard Boolean format. As noted below, some sites retained a search history, making the refinement of previous searches more convenient. With regard to inputs, there was no apparent difference which technology area was being queried, with the following exceptions: design patents often required a different interface, and a user attempting to search chemical structures or biological sequences would be extremely limited in the ability to find prior art read on by the claims.

The biggest difference between providers, and usually the biggest deficiency area, is the output *i.e.* the results list. Capabilities to reformat and re-sort the list and to choose fields for display can greatly enhance finding the closest prior art where there are many search results. Unfortunately, the lack of transparency and inflexibility of Google is being mirrored by many intellectual property data providers, flexibility being sacrificed for seeming simplicity. The best providers in this area are Delphion and WIPS, and CAMBIA's Patent Lens.

A few of the commercial data providers also have available non-patent databases, some of which are also available to the public. Among the fee-based and subscription-based sites, DialogPro and STNeasy have a limited number of databases (about 100) and Dialog and Lexis have many hundreds each.

Whether these databases are sufficient for the examiner searching non-patent literature depends upon the technology area. For applications with claims to specific biological sequences, the searcher needs GenBank at minimum, which can be searched by the public cost-free through NCBI (to which CAMBIA provides a link) and optimally Derwent's gene sequence bank that also has sequences published in patents. For biotechnology (other than sequences), PubMed⁴ and BIOSIS are important non-patent databases. Several of the commercial providers charge fees for searching these databases, and the searches are usually separate from the patent data.

For chemical patent applications, a database where chemical structures are searchable is often necessary; STN, in connection with CAS, does have such a database (but not in STNeasy). CAMBIA will investigate the possibility of obtaining this data for public good use.

⁴ CAMBIA provides a link to PubMed and will investigate the development of APIs to this and other NCBI databases to make the searches more seamless.

The mechanical arts field typically relies mainly on patents for prior art. For computer and software technology, there is ongoing debate about where and how to do prior art searches of non-patent literature.

Most of the comments related to utility patent document searching also apply to design patents. Only one of the databases (PatentCafe) allowed the user to choose only design patents for the search. As design classifications are the primary search terms, it's unclear that this sub-division is much help because utility patents that may have a design element will be excluded from the results list, and vice versa. This feature may even be detrimental, as design patents can be valid prior art against utility patents and utility patents can be cited against design patents. Searching for design patent prior art in the non-patent literature is also gravely limited by reliance on classification for most design application examination.

Strengths and weaknesses of search sites

Selected National Patent Offices

IP Australia⁵

The site is difficult to navigate because it is in multiple sections with patents from different groupings available either in the web interface or via a downloaded software client:

- Patsearch⁶ (running on the Patent Administration and Management System – PAMS database): This database is accessible online, containing bibliographic and legal status information on innovation patent applications filed from 24/05/2001 and complete and provisional patent applications filed from 05/07/2002. Search fields for Patsearch include:
 - Patent number
 - Applicant/inventor name
 - Patent title
 - Option to exclude lapsed/withdrawn/ceased/expired documents
 - Option to exclude PCT applications and documents with non-Australian priority data
 - Option to restrict publication/filing date range
 - IPC
- Patent Administration System⁷ (PatAdmin, part of the Patents Mainframe Databases) information page: This database is accessible via a software client that must be downloaded and installed on a local machine from a different page on the IP Australia website.⁸ PatAdmin provides bibliographic and legal status information of patent documents (applications and granted patents) filed between 01/1979 to 04/07/2002. Search fields for PatAdmin (via the software client) include:
 - Application ID
 - Patent number
 - Provisional number
 - PCT number
 - WIPO number
 - Applicant/inventor name
 - Option to restrict date range

⁵ <http://www.ipaustralia.gov.au>

⁶ http://pericles.ipaustralia.gov.au/ols/searching/patsearch/search_page.jsp

⁷ http://www.ipaustralia.gov.au/patents/search_patadmin.shtml

⁸ http://www.ipaustralia.gov.au/patents/search_software.shtml

- Patent Indexing System⁹ (PatIndex, part of the Patents Mainframe Databases): This database is also accessible via a software client downloadable from the IP Australia website, and the title and number of patent documents (applications and granted patents) filed between 01/1979 to 04/07/2002, and seems to cater particularly for searchers to find prior art in these patent documents by using the IPC code. Patent titles and corresponding document numbers are searched by one or maximally two codes, connected by AND, OR or NOT. The resulting list can be narrowed down further by either conducting more IPC code searches, restricting the application year, excluding the lapsed/ceased patent documents, and using keywords in the bibliographic information (e.g. word in title or applicant).
 - AU published patent data (AAPS) searching main page¹⁰: This database is accessible online, containing full text images of published AU patent documents from 1975 and bibliographic information of AU patent documents between 1920 and 1974 (data incomplete). A 'quick' search for patent documents can be done by:
 - Application number
 - Patent number
 - Title

The 'Advanced' search provides searching for terms, dates, and numbers stated in the bibliographic information. Two terms/dates/numbers are available per search.

- Patent specifications main page¹¹: This database is accessible online, containing full text images of AU-A and AU-B patent documents from 17/12/1998 (these documents have navigation to the abstract, description, drawings and claims in the documents) and AU documents that did not enter via the two international routes (Paris and PCT) from 1975 (these do not have navigations to the separate sections in a patent document). Search for patent documents can be done by:
 - Application number
 - Patent number
- Application/serial number concordance main page¹²: This search finds corresponding numbers of patent documents that have been renumbered from the Patent Mainframe database system when they were moved to the PAMS database, using the application number (old or new system) or the patent serial number.

None of the patent search/retrieval engines offer a term search within particular fields of document text, i.e. claims, description, full text. Searching for information on the Patents mainframe can be very difficult for a public, non-examiner user because of the requirement for knowledge of the commands used on the telnet interface. Full understanding of the coverage of each database (Patents mainframe and PAMS) is needed for the searcher to be able to retrieve information on particular documents.

IP Australia's website acknowledges that there are issues with data completeness and quality and a need to consult several distinct databases to conduct basic searches.¹³ It is further necessary to go to INPADOC for much of the status information, and we found many issues with accuracy and ongoing updates in the INPADOC data.

⁹ http://www.ipaustralia.gov.au/patents/search_patindex.shtml

¹⁰ <http://apa.hpa.com.au:8080/ipapa/qsearch>

¹¹ http://pericles.ipaustralia.gov.au/aub/aub_pages_1.process_simple_search

¹² http://pericles.ipaustralia.gov.au/ols/searching/patsearch/search_con.jsp

¹³ <http://pericles.ipaustralia.gov.au/ols/searching/content/olsPatents.jsp>

However, the mere fact that data are distributed over several systems during periods of redevelopment need not in itself impact negatively on public use of that information. For example, the Australian Trade Mark Online Search System (ATMOSS) has enhanced public access, while making use of data input and is maintained on a legacy mainframe system as well as a more recent web-based on-file filing system.

From the perspective of Australian innovators, it would be desirable to see:

- A single integrated interface that allows "one stop" authoritative searching of Australian patent documents, as well as legal status, and file wrappers
- Full text searching of the claims and specifications sections of Australian patents

European Patent Office (EPO)¹⁴

The EPO site provides wide data coverage with legal status and family data (INPADOC) of patent documents for over 40 jurisdictions. INPADOC patent family documents are summarised by patent number and not presented separately by publication (e.g. EP123456-A1 and EP123456-A2 are summarised as one patent document with A1 and A2 information contained in a single file).

A major drawback of the EPO public search site is that it doesn't support full text searching. Searches can only be done for one database at a time and no term searches can be done in claims or full text. For non-English patent documents, searching for terms in the title or abstract cannot be done and no manual Boolean search options are available. Furthermore, 'description' and 'claims' fields in the individual document information page are sometimes replaced by a corresponding WO or EP document if the information from the actual jurisdiction is not available. This can be misleading if you want to conduct claims analysis on a granted patent for a particular jurisdiction.

The EPO site does allow bibliographic searching of patent documents from a number of jurisdictions as well as to patent family and legal status data through INPADOC. However, EPO states that the INPADOC legal status is incomplete or not up to date, and therefore the information provided on the document information page may not be accurate. This same limitation applies to all providers that use the INPADOC data purchased from the EPO.

The epoline® Online Public File Inspection service¹⁵ gives access to prosecution history and status.

In addition, the EPO has been active in the development of software, such as EPOQUE Net, for the internal use of national offices as a revenue generation mechanism set up in co-operation with the private sector.¹⁶

United States Patent and Trademark Office (USPTO)¹⁷

The USPTO was one of the first offices to improve their public search functionality through initiatives such as web based full text searching. The USPTO also allows subscriber based FTP access to a comprehensive range of timely, high quality machine readable IP data, enabling third parties to offer innovative services that add functionality beyond that offered by the USPTO public search site. For example, this is the mechanism by which CAMBIA obtains full US patent and US application data.

The USPTO site is best for quick viewing or capture of text, although only US data is available here. The search interface is simplistic, with sophisticated searches being difficult, if not

¹⁴ <http://www.european-patent-office.org/index.en.php>

¹⁵ <http://my.epoline.org/portal/public>

¹⁶ <http://www.empolis.com/en/20D6DCCC63D14F7F80C2163B990DBFA5.php>

¹⁷ <http://www.uspto.gov>

impossible. To view images requires a specialised viewer of TIFF images. The output is a long list of patent numbers and titles, with the corresponding links leading to full text.

The USPTO provides online access to status information through the Patent Application Information Retrieval (PAIR) system, and there is a database of assignees that is updated regularly.

Japanese Patent Office (JPO)¹⁸

The JPO has an effective public search interface for users of Japanese language, but provides only limited English language access support¹⁹. As an example, for design patents, the English interface only provides designs search by registration number or application number of rejected designs, whereas the Japanese interface provides additional search options, including the text search option that was introduced with the incorporation of the electronic publication of registered designs from January 2000.

The JPO has been very active in the area of machine translation (MT) with a view to facilitating access in Japanese to patent documents published in European languages. There is less accessibility in the other direction. Electronic documents (1993 onwards) have machine translations of Japanese patent documents into English, which are of incomplete quality and still under development but sufficient for a person skilled in the art to be able to understand the content.

The JPO provides term searches within claims. Search terms cannot be easily refined. For example returning to the search page by the 'back' button on the browser and modifying the search term does not allow one to do another search on that page. Instead, the page must be refreshed and all terms and fields redefined and/or re-entered.

Patent documents before 1993 only exist as scanned images, so the claims cannot be accessed without obtaining the entire document. Searching for terms in full text is not available (possibly because the pre-1993 documents have not been converted into electronic format). Manual Boolean expression search (to create a personalised Boolean expression) is not available.

We found that inventor names were missing from certain PCT applications (possibly all PCT applications published before 1993) that entered national phase in Japan, making it difficult to search for patents using inventor names (PDF files of PCT applications published after 1993 also do not seem to have the inventor name stated in the national phase application).

Brief legal status is provided for Patent Abstracts of Japan (PAJ) documents where available.

State Intellectual Property Office (SIPO) - China²⁰

Use of this site is free for downloading patent documents, with images available as TIFF files, and legal status can be obtained, as at the CNIPR site (see below), but users may need to download a software navigator to view Chinese patent specifications and the site is slow in retrieving data. SIPO also offers **China Intellectual Property Net (CNIPR)**, which offers a Boolean search option with capabilities to refine the search. It is also possible to search legal status by application number or publication date²¹.

CNIPR provides information about whether or not a patent is granted or lapsed, and information on whether or when a patent application has been requested for examination, is abandoned, or

¹⁸ <http://www.jpo.go.jp/>

¹⁹ CAMBIA employs a native speaker of Japanese as a patent analyst who carried out the JPO searches.

²⁰ http://www.sipo.gov.cn/sipo_English/default.htm

²¹ CAMBIA employs a native speaker of Chinese as a patent analyst who carried out the SIPO and CNIPR searches.

deemed as withdrawal, etc. A synonym search for possible words with same or similar meaning is also available.

However, the site is not free – a patent search at CNIPR site requires registration and the purchase of reading cards, despite the fact that searching (without downloading) for Chinese and international patents is free for cardholders. Reading cards can be purchased at three levels - cards valued at 500, 1000 and 3000 Chinese Yuan (approximately \$80, \$165, \$485 Australian dollars respectively). These cards allow users to download 500, 1250 and 6000 pages of Chinese patent specifications respectively, or 0, 5000 and 20,000 international patent documents respectively.

The searchable parts of a patent at both sites are limited to the bibliography page including:

- Publication Number
- Publication Date
- Title, Address
- Application Number
- Application Date
- Inventor Name
- Abstract
- International Classification
- Applicant(s) name
- Attorney/Agent.

Korean Intellectual Property Office (KIPO)²²

A search facility in English is offered via that Korean Intellectual Property Institute. At this site, only the interface for "Patent & Utility Model Search" is translated into English, while more is available for users of the Korean language to perform keyword searches. It is possible to retrieve documents by using English search terms only if the title or abstract contains English words.

There is also a Korean Patent Abstract (KPA) search with the interface in English, but only for Korean patents published with English abstracts, from 1979 for examined patent publications and 2000 for unexamined patent publications. The latter does, however, provide legal status information in English.

World Intellectual Property Office (WIPO)²³

WIPO's data coverage is small, with PCT documents available only from 1997, severely limiting its usefulness. One advantage though, is the links to other documents such as the examination report and priority application(s). However, international reports can only be accessed as the PDF files, less readily searchable.

WIPO recognises the limitations of its data as a resource for finding prior art and improving the quality of patents internationally, and is working to improve this. WIPO has added full text search functionality for recent PCT applications to its on-line site and further improvements are promised. For the subset of applications that have this search functionality, terms can be searched within full text and claims. Terms can also be searched in French.

WIPO employees have been cooperating with CAMBIA and recognise consonance with the goals of CAMBIA's BiOS Initiative.²⁴ On the WIPO site, as on CAMBIA's Patent Lens, it is possible to manually create a Boolean expression and it has the Boolean operator 'NEAR'. The

²² <http://eng.kipris.or.kr/Search/Search.html>

²³ <http://www.wipo.int/patentscope/en/>

²⁴ <http://www.bios.net/daisy/bios/48>

period range can be specified (by weekly increments only). Electronic descriptions are available and the claims can be readily accessed.

For status information, each individual document has a link to a chart describing the national phase entry deadline for PCT member States.

No-charge providers

Espacenet²⁵

While this site has improved considerably over the years, the extent of coverage has not increased significantly and substantial disadvantages limit its usefulness. Notably, the results list is disorganised; no obvious parameters dictate the order of display, nor is there any ability to impose some order on it. Thus, unless the search is limited to just the EP data, it is a formidable task to wade through the results. In addition, full PDFs are not yet available for download.

Surf IP²⁶

This site, which has both a free-of-charge and a “premium” fee-based component, is maintained by the Singapore government and advertises a large data set with access to at least nine national patent office databases and other technical, business and search engine sources. A few datasets are not found elsewhere (Singapore; Taipei and Thailand). It was also of particular interest for access to the Korean and Japan patent office databases, but no documents from those databases could be retrieved during our trials (repeated error messages on the results page). The information retrieval speed of this site is extremely slow (over a minute per search), and in some sessions every search attempted resulted in either a time-out or no results.

Although with every query we found that irrelevant patent documents were retrieved, our analysis of the sources of these problems may be instructive for IP Australia if consideration is being given to Boolean interface design. We found that many such errors were due to a failing of the Boolean operator bracket function, which resulted in difficulties specifying combinations of search terms. Also, as in Google’s default search, structured search terms do not seem to be connected by Boolean AND, but rather OR, which pulls out documents that do not have one or the other term, particularly in documents from the EPO and UK Patent Office. There was, unfortunately, also no wildcard option available.

Patent Lens (CAMBIA)²⁷

For the purposes of this report we sub-contracted an expert US patent to review the CAMBIA Patent Lens. The consultant’s opinion was that compared to other no-cost providers, this site is easier to use and better organised, consisting of a well laid out input page, with search capabilities of multiple fields, and a clean results page that can be sorted on a number of parameters.

The consultant remarked that the coverage of countries is good and the ability to search full text documents is excellent - although clearly it would be desirable to extend the coverage of countries. Notably, Australian data was only available for a short period, a shame because the site is one of very few that can provide full text searching of Australian data.

Currently the site is also limited (by IPC classification) to life science documents, but OCR is underway to provide patent documents in all fields within the next six months.

²⁵ <http://ep.espacenet.com>

²⁶ <http://www.surfip.gov.sg>

²⁷ <http://www.bios.net/patentlens/simple.cgi>

CAMBIA's Patent Lens uses a fast and powerful query language developed in house for the specific purpose of full text patent searching. The same query language is used in the modification that was made earlier this year to incorporate INPADOC data. Due to proprietary restrictions around many other search engines, we are able to give a high level of detail only about this Canberra-developed language, Dekko, in Chapter 2 of this report.

A significant advantage of the Patent Lens is the ability to view the full patent document, which can be seen at the user's option as either text with highlighted search terms or PDF image. Giving the user options is part of the site ethos, and the developers are contactable by users if bugs or suggestions are to be communicated. Relevance ranking is in development and this also will be user-configurable.

Commercial Providers

Thomson-Derwent providers

The Thomson Corporation, a Canada-based electronic publishing conglomerate, has consolidated its position as a key provider of patent information through the acquisition of its competitors. In 1984, it purchased all remaining shares in Derwent, a key provider of patent databases. In 2000, it purchased Dialog²⁸, in 2002 it purchased Delphion, and in 2004 it acquired Micropatent, one of its main competitors. There have also been numerous smaller acquisitions along the way. As well as consolidating its Intellectual Property services, Thomson has continued to expand into the sciences, health and legal publishing and online databases.

Although Thomson-Derwent has acquired several of these larger data providers in recent years, thus far they have not unified many of the sites. Each of the individual sites has a different format, different set of data, different interface and different pricing. We chose the major sites mentioned above for our analysis.

Delphion²⁹

Delphion's speed has dramatically improved but the search interface is still clunky, and searching for a document by its number is a lesson in all the ways one can make "mistakes" formatting. It is possible to limit a search by date range, use wildcards, and search terms near each other. The results list can be highly modified in fields that are shown and there is standardization of assignee (applicant) names. A limitation is that the results are truncated at 500 results. When a document is viewed, post-grant status is available. Some analytical tools of varying utility are available, though some at extra cost, for e.g. graphical citation tools.

Delphion does carry the human-edited Derwent titles and abstracts database, and features forward and backward citation linking. It also clusters results by subject, but it is not clear on what criteria this clustering is based.

DialogPro / DialogWeb³⁰

The "Web" version is actually more sophisticated; for example, complex Boolean search strategy can only be implemented on the web version. DialogWeb also has many more non-patent databases that can be accessed for a price. Data coverage in the Web version includes China (bibliographic data and abstracts in English), as well as post-grant status and litigation information. Several useful options for sorting the output are found in the Web version. Some may find its command driven interface challenging.

²⁸ <http://www.dialog.com>

²⁹ <http://www.delphion.com>

³⁰ <http://www.dialog.com>

Micropatent³¹

Only the Patent Web portion was assessed, because each area requires a separate subscription. The search interface allows for some rather sophisticated search strategies, such as using ECLA codes, complex Boolean structure, limiting searches by publication date range and using a wildcard for truncation. A command-driven interface that would allow increased flexibility is available at a higher subscription fee. Results are returned quickly and refining searches is easily done from the search history interface. The site annoyingly opens new windows for everything, making navigation difficult.

Independent fee-based or subscription-based providers

Get-the-patent³²

This site was set up mainly to access document images for a low price with input of a patent number, but it offers some searching capabilities as well. The interface allows complex Boolean structure but limits a search to a single year and the search engine query language seemed in many of our searches to come up with bugs. Searching can be done by IPC codes, but not in all databases. The datasets are fairly extensive, but only the US is full text. Its strength still remains bulk downloads of patent documents. The bugs take this site out of the race when it comes to serious searching, though it could still be useful for downloading full images, which are in a proprietary format 1/6th the size of PDFs. However, these images are provided in a proprietary format (not PDF) that requires its own viewer.

Lexis³³

This service has an exemplary interface for inputting Boolean searches. Search terms can be applied to just about every field of a patent document, though it is challenging even for an expert patent searcher to find the help for formatting of some fields like IPC.

Unfortunately the output of results was not exemplary, with no formatting of the results list possible, and access is expensive.

Minesoft PatBase³⁴

This engine was originally developed for a specific client and is now available to others on a fee basis. It is a powerful search site, though unfortunately navigation of the site is quite challenging and there are some glitches and slowness.

While it allows wildcards, complex search structure, etc., it has a very useful browse function on names of inventors and assignees. Using this function, it is possible to capture documents by a group or individual regardless of bad spelling and typos and inconsistencies of the name (e.g. IBM = International Business Machine; Du Pont = DuPont).³⁵

Like the step searches of WIPS, initially the results come back as the number of results. A quick view of titles reveals if the search is on the mark or not. The output is thorough, with displayed fields highly controllable.

PatentCafe³⁶

The site uses latent semantic indexing searches as an alternative to Boolean searches. In the latent semantic search, a query consists of a sentence or a paragraph, from which the concept

³¹ <http://www.micropatent.com>

³² <http://www.getthepatent.com>

³³ <http://www.lexis.com>

³⁴ <http://www.minesoft.com>

³⁵ This would be quite valuable for a public user interested in freedom to operate considerations and CAMBIA is looking into a similar function for the Patent Lens.

³⁶ <http://www.iamcafe.com/index.asp>

of the sentence or paragraph is extracted to search for relevant patent documents. When our patent searching experts used text from claims to search, we found that every search tended to return a very large number of documents, the relevance of which was sometimes dubious. It also missed some of the known relevant patents where these used a somewhat different claims vocabulary.

However, the public user can benefit by a feature of the latent semantic indexing search that the patent results list page provides a set of 'related terms' (synonyms) generated from the context around the initial query words for each patent document that has been retrieved. This can be useful to further exclude/include patent documents on the list. Unfortunately, it doesn't also return the original query for editing, so refining search queries is less straightforward than for typical Boolean expressions.

Another useful feature is that the individual document information provides 'patent references cited' and 'cited by' (this seems only available for US patent documents and does not guide properly to WO documents, probably because of the metadata marking).³⁷

The website has some glitches, for example optional fields that proved to be mandatory. It doesn't work well with all browsers and all operating systems, and there is no provision of INPADOC family or status information, or prosecution history.

Questel-Orbit³⁸

The search input on this French company's site is quite good; searches by IPC, ECLA and date ranges are easily set up, as well as complex Boolean structure searches. Search terms are readily limited to be near each other or ordered. A search history allows combining prior done searches using Boolean connectors. This site also has more extensive data coverage (countries) than the others, legal status data, and value-added patent family data.

The weakness of this site is the results output; it is fixed order by patent number, although the results can be downloaded in a variety of formats.

One very nice feature in the output comprises links that allow the searcher to move through the database forward or backward in time following patent citations, a feature that CAMBIA is considering for incorporation into the Patent Lens.

Software for Intellectual Property (SIP)³⁹

This is a German site that has progressed over time from specializing in patent family data to a search engine with several features (e.g. ability to search using IPC codes, "near" searches, and even "fuzzy" searches for similar sounding words).

The user interface offers Boolean and structured query facilities that simply do not work, so we weren't able to perform our test queries.

The output format is not malleable, but it integrates well the family and legal data with the patents.

Univentio (PatentWarehouse)⁴⁰

Univentio is a company with long history in the IP information industry. For example, Univentio produced the "WIPO/PCT patents fulltext database" used by Thomson Dialog. Univentio has recently come under the control of LexisNexis, which has not yet been able to provide new subscriptions to the Univentio patent search site, the PatentWarehouse, although this is

³⁷ As such marking is not greatly different from other features the Patent Lens already incorporates, this is a feature that CAMBIA is exploring.

³⁸ <http://www.qpat.com>

³⁹ <http://www.patentfamily.de/>

⁴⁰ <http://www.patentwarehouse.com/>

probably a transitory problem which will soon be addressed under the new ownership arrangements. Accordingly, the following information was obtained from documentation rather than actual use.

The PatentWarehouse database contains intellectual property information, including full text of patent applications and granted patents, bibliographic information of patents, utility models and designs. Full text of patent documents are covered for US, EP, GB, JP, FR, DE, AU, AT, BE, CA, DK, LU, MC, NL, NO, PT, ES, SE and CH, and bibliographic information is covered for 70 jurisdictions. Five search options are available, including searches by complex Boolean expressions. The results page provide a list of retrieved patent documents with an indication of legal status using 'Trafficlight', a unique feature provided by PatentWarehouse. Patents that are in force (red), withdrawn/lapsed (green), or possibly in force (yellow) can be identified for each patent document with the three different colour codings. The claimed coverage of full text Australian patents is noteworthy.

WIPS⁴¹

This newcomer is run by a Korean company and uses an interface and output reminiscent of Delphion but with improvements. The search input is flexible and allows wildcards, near searching, date ranges, and assignee name standardization. The output features a "step search", which outputs only the number of results; when a desired number is returned, the list of results is easily viewed. Data coverage is standard, with only US and EP documents provided as full text. Ability to modify the results lists is complemented by ability to download in txt, xls, or mdb file types. Clustering results by IPC, applicant, application date, patent date and keyword are helpful.

One downside to this site is that to view images, proprietary software is required, which works only with certain browsers, and some graphical presentations require a Java Virtual Machine to be installed in the browser. We also found a number of bugs in the search. For example, full text search for words within a claim appears not to have worked.

Non-patent literature

For any type of prior art search, whether it be an FTO search for a particular technology or a patentability search for an invention, databases that contain non-patent or scientific literature are important. This is because a vast amount of technology development is published in scientific journals and in many art areas there is a substantial amount of publication that predates patent literature.

The examiners we spoke to indicated considerable use of general search engines such as Google for this purpose, which can certainly deliver large amounts of information, but is not relevance ranked for patent searcher needs, and extremely relevant documents can be missed. Citation databases contain literature that has been selected for publication by relevant bodies, the selection having been performed at least ostensibly by "persons skilled in the art".

To represent the many available free and paid databases, we chose one of each. Here PubMed and Current Contents Connect are briefly explained and compared. However, we note that many other databases are indeed available, and our the comparison in this report concentrates on typical features.

Before entering into this comparison, we also mention WIPO's efforts to create a common dataset of journals that can be used as sources of prior art in specialised fields, the Journal of

⁴¹ <http://www.wipsglobal.com/>

Patent Associated Literature, JOPAL.⁴² Currently this supplies only bibliographic data such as titles, authors and dates, but is the closest non-paying alternative to Current Contents Connect and is set up specifically for patent searchers.

PubMed

PubMed is a free search engine maintained by the National Library of Medicine (NLM) at the National Institute of Health (NIH), and is accessible through the National Center for Biotechnology Information (NCBI) website.⁴³ The databases that are covered by PubMed are:

- MEDLINE – database with over 12 million citations from over 4800 journals in the biomedical field from the mid 1960s to current.
- OLDMEDLINE – database containing 2 million citations from biomedical journals around the world between 1950 and 1965.
- In Process Citations – bibliographic information and abstract of articles not yet added to MEDLINE (record updated daily between Tuesday and Saturday).
- Publisher Supplied Citations – citations received in electronic format from publishers.
- ‘Out-of-scope’ citations – from general science and chemistry journals that index life science citations to MEDLINE full text articles submitted to PubMed Central by additional journals in the life science field.
- MeSH: Medical Subject Headings, a database containing a vocabulary thesaurus that is used for indexing articles in PubMed (similar to the IPC, but without codes). The MeSH database link finds the MeSH term based on a search, which then can be used as ‘term [MeSH]’ to search for articles in that particular MeSH category. MeSH terms can be combined, as well as narrowed or broadened from the initial MeSH term that was found from a term search.
- Journals Database⁴⁴

A basic search for retrieving citations is conducted with a single search box. Words in the following fields can be searched, often without the need of defining the field:

- Key concepts
- Author names (if the name is also a subject word, a search tag [au] can restrict the search to within authors)
- Journal title

Searches can be restricted in the ‘Limits’ option by using criteria such as language, publication type, publication date, and MeSH terms. Other options that are available include:

- Search for phrases
- Wildcard ‘*’
- Combining searches in ‘History’
- ‘Single Citation Matcher’ – to identify a specific article using bibliographic information
- Index of terms in ‘Preview/Index’
- Boolean operators AND, OR, NOT (default is AND)
- ‘Related Articles link’ to retrieve a set of articles relevant to the area of the article of interest, listed in the order of relevance
- Preview/Index – displays a quick result of the number of hits from a particular search, or provides an index of terms within a field based on a term search.
- ‘History’ provides function to combine search results

⁴² <http://www.wipo.int/scit/en/jopal/jopal.htm>

⁴³ <http://www.ncbi.nlm.nih.gov/>

⁴⁴ http://www.ncbi.nlm.nih.gov/entrez/getids_help.html#JournalLists

- There are special search categories and queries designed for medical practitioners and health service researchers

Search results are displayed with the author name(s), article title, journal name, issue and page, and the PubMed ID. The author name(s) is linked to the full bibliographic information including the abstract for those available (1975 onwards). Each hit has a link to either the abstract, free full text from the journal or free full text in PubMed Central (PMC), 'Related Articles' link to retrieve articles relevant to the area of the article (listed in the order of relevance), and 'Links' to other various resources or NCBI databases. The 'Link out' button in the 'Links' section provides links outside of NCBI that are related to the article.

Isiknowledge⁴⁵

This paid citation search engine is part of ISI Web of Knowledge, provided by Thomson Scientific (a Thomson Corporation group). The database contains bibliographic information on articles from approximately 7600 journals and 2000 books published since 1998. Areas of research that are covered are agriculture, biology and environmental sciences, social and behavioral sciences, clinical medicine, life sciences, physical, chemical and earth sciences, engineering, computing and technology, arts and humanities, business collection, and electronics and telecommunications collection. The search interfaces are user-friendly with available tag keys provided for all search pages, a table of available field tags and Boolean operators on the Advanced search page. The results are compiled in a list with the number of hits for each search, and the information from the retrieved articles (titles, author, abstract, and full text where available) is also easily obtained.

ISI Web of Knowledge contains the following main search engines:

- Web of Science – search engine with database containing articles in life sciences (Science Citation Expanded®, 1945-), social sciences (Social Sciences Citation Index®, 1956-), arts and humanities (Arts and Humanities Citation Index®, 1975-), chemistry (Index Chemicus®, 1993-; Current Chemical Reactions®, 1986 and INPI (Institut National de la Propriete Industrielle) archives between 1840 and 1985), and the Century of Science initiative with files backdating to 1900.
- Current Contents Connect – search engine with database containing information on articles from 7600 journals and 2000 books published since 1998.
- ISI Proceedings – database containing conference proceedings in the science, social science and humanities fields.
- Derwent Innovations Index – database containing over 23 million patents from 1963 from 40 jurisdictions.

In Current Contents Connect, information on publications from the following areas of research are available⁴⁶ (upon subscription to each area) for searching from 1998 to present:

- Agriculture, biology and environmental sciences
- Social and behavioral sciences
- Clinical medicine
- Life sciences
- Physical, chemical and earth sciences
- Engineering, computing and technology
- Arts and humanities
- Business collection
- Electronics and telecommunications collection

Search options

⁴⁵ <http://portal9.isiknowledge.com/>

⁴⁶ http://ccc02.isiknowledge.com:80/help/h_database.html#abes

There are four search options:

1. Classic search

Search by terms or phrases. Boolean operators AND, OR, NOT, can be used to combine terms within a particular field (selected from a range of bibliographic information, or topic subject). Search results that are compiled as a list with the number of hits can be combined with AND or OR to further narrow/broaden the search.

2. General search

Search by terms or phrases in the following fields (an index is provided for each field for browsing):

- Topic/subject
- Author/editor
- Group author
- Source title (journal title)
- Address (author's institution)

Terms can be connected with Boolean operators AND, OR, NOT, SAME, and language and document type can be restricted.

This search option is the only one that the search history cannot be accessed and edited, although a search conducted with this option will be included in the history.

3. Advanced search

Search by creating a Boolean expression. Operators AND, OR, NOT, SAME can be used to connect terms, and field tags can be added to each term to define the field (list of tags provided on the search page). Terms can be bracketed to indicate priority searching.

4. Browse

Search for articles by browsing by journal titles (linked in alphabetical order) or by research area.

Results page

The results of searches is listed in 'history', from which retrieved articles of a particular search can be viewed by clicking on the number of hits on the search results list. Retrieved articles are presented with the author name/s, title, abbreviated journal name, issue, page and publication year. Each article title is linked to detailed bibliographic information of the article, including the abstract. Access to full text is indicated under each article. Bibliographic information of selected articles can be retrieved and transported to personal citation management softwares such as EndNote and Reference Manager (both by Thomson).

Comparison

In searches for a particular life sciences query set, both searches retrieved the same documents. The capability to find particular documents based on this search seems to be similar between the two databases. If the field of interest for a particular search is in medical biology, PubMed offers a wider range of periodicals compared to Current Contents. On the other hand, searching for literature in fields of interest other than medical sciences would retrieve more documents with Current Contents.

Traditional knowledge

Although there has been heated discussion in the intellectual property community of the importance of averting IP protection awards that cover genetic material and processes that have long been known to traditional communities, at this point there is little access to traditional knowledge databases that may contain substantial prior art. Efforts to overcome this deficiency are being made at SIPO, the USPTO and WIPO.

WIPO has recently reviewed a wide range of traditional knowledge-related periodicals⁴⁷ and agreed that thirteen of these should be added to the list of published items of non-patent literature forming part of the PCT minimum documentation under Rule 34. The Meeting also recognised the importance of identifying further traditional knowledge-related databases suitable for use in international searches and agreed that the issues involved should be considered as part of a comprehensive review. WIPO is developing a Search Guidance Intellectual Property Digital Library (SGIPDL) which when available could be of assistance to (it is currently being reviewed by a task force of representatives from the International Searching Authorities).

“Value-added” information summary

Various patent information providers claim to add value to patent information in the following ways:

- Some providers facilitate searching by inventors and assignees through normalisation of variations in a person's or company's name (e.g. CSIRO = C.S.I.R.O.) to a single standardised form.
- Several sites use a common source of revised English titles and abstracts for patent documents in English as well as many other languages, on sale from Derwent. The Derwent titles and abstracts can be more informative than the originals, since the original titles are often minimal and the abstracts often don't represent the claimed matter well. Skilled users of IP data should not rely on title and abstract searches, but given that much searching is still performed on titles and abstracts rather than full text, this service to make at least some of them somewhat more informative can be viewed as adding some value.
- ECLA and IPC classifications are technology-based and designed primarily for patent examiner needs, and services tend to use one or the other or both, although some services provide neither. Derwent provides an industry based classification system to complement them, intended to satisfy the needs of patent information users who deal extensively with industry classification, primarily sophisticated patent attorneys; it seems to be used less often by examiners. We found that none of the three classification systems would be used extensively by the public technology searcher because of the breadth of the categories; full-text searching with user-configurable queries was preferred.
- Some providers allow searching of patent and non-patent technical literature at the same site, though few do this in a very integrated manner because the non-patent literature does not have the same formalised metadata⁴⁸. As much non-patent literature is available via the Internet, some searchers use Google to find it, but a lack of user-configurable relevance ranking can lead to failure to find the most relevant prior art; use of specialist non-patent literature compilations such as Medline is more likely to find that which those skilled in the art would identify as prior art. For occasional searching it is not inconvenient to use separate specialist citation databases via internet links, but ideally APIs into specialist citation databases such as Medline can be developed.

⁴⁷ www.wipo.org/edocs/mdocs/pct/en/pct_ctc_21/pct_ctc_21_3.pdf

⁴⁸ One approach to do this, being explored by CAMBIA, would be to mark and link literature citations from within patent documents, which would be easiest for USPTO-sourced patent documents due to the standard field for such citations. Forward and backward navigation via citations can be very useful for freedom to operate identification of dominating patents, though not identified as crucial for examiners.

- A very few sites provide chemical and biological sequence searching (discussed in Ch. 2).
- At this point, there is little access to traditional knowledge databases that may contain substantial prior art, although efforts to overcome this deficiency are being made at SIPO, the USPTO and WIPO. Public good providers will be following this with the same degree of interest as pay providers, but may be encountered with less suspicion.
- Several providers use enhanced family data and legal status, based on INPADOC but extracting or collating other information for presentation. It is possible, for example to cluster results by family using INPADOC and present this visually in a variety of ways.⁴⁹

⁴⁹ For example, CAMBIA uses tables identifying the priority documents cited.

Chapter 2: Discussion of Emerging Trends in Information Retrieval

This chapter analyses emerging trends in searching and accessing IP information, together with likely future directions in this field. The chapter begins with how some of the general information retrieval and ranking algorithms are applied to patent data. While many of the search engines that perform these algorithms are proprietary, CAMBIA is in a unique position to provide detailed information on the full text patent search engine developed by its staff in Canberra, Dekko, and aspects of its design that affect performance with respect to the major patent data sets.

We then turn to needs for searching and accessing IP information of special relevance to Australian stakeholders:

- Optical Character Recognition (OCR) of image data
- searching in international databases and the languages that they contain
- the specialised art of searching with a basis in certain types of claims, to determine coverage of patents (this section has a special focus on searching some of the types of matter claimed in many life sciences patents, such as biological sequences and chemical structures, because of the prominence of these types of claims and these art areas in Australian patenting).

One of the key factors driving the development of IR technologies has been the explosion in the quantity and diversity of machine readable information published on-line. Another factor has been the increase in the power and storage capacity of computing hardware compared to its cost. For example, Google allows subsecond searching across a claimed eight billion web pages. This reliability and scalability has been achieved by harnessing the power of thousands of cheap Linux PCs rather than high-end servers or mainframes.

Intellectual property searching covers a wide range of activities. A scientist interested in examples of how to use latest innovations in the field will have needs very different to an examiner charged with determining patentability. A trademark examiner searching for evidence of use will face very different challenges from an examiner searching for deceptively similar marks. An IP officer from a biotech start-up doing a FTO search needs to have access to data from the jurisdictions of their key markets as well as jurisdictions in which their research and production takes place.

The way in which IP information is accessed by IP professions and the public has changed greatly over the years. Card indexes, microfiche, mainframe based systems and midrange systems with dedicated line or dial up access have largely been superseded by web based search interfaces. "Web services" such as the European Patent Office (EPO) Open Patent Service (OPS)⁵⁰, are increasingly allowing organizations and even individuals to design their own interfaces to remote databases.

Information retrieval

Information Retrieval (IR) is the art and science of searching for information in documents or for documents themselves. IR draws upon cognitive psychology, linguistics, semiotics, information science, computer science and librarianship.

⁵⁰ <http://ops.espacenet.com/>

While user interface design and functionality are important to the usability of systems, the underlying IR technology utilised is also important. Technologies that may be excellent for small data sets with a few concurrent users may be inflexible and unable to scale. With the rapid increase in the capacity, speed and affordability of random access memory (RAM) over recent years, techniques that allow full text indexes to be held in memory have become practical. A form of index, known as an **inverted index**, in which any word in the corpus to be searched can be used as key to find the documents in which it occurs, has proved to be extremely scalable.

Various IR techniques make use of:

- the content of the documents being searched;
- meta-data describing the documents and relationships between them, *e.g.* citations;
- information external to the corpus being searched, such as dictionaries, thesauri and other bodies of relevant information.

A **search engine** is software that performs IR using one or more IR models. For text searching, the most prominent of these are the Boolean and the Vector Space models and their variants.

It should be noted that for this study, visual image searching was also investigated, for use in design patent searching, normal trademarks, and shape marks, The software available has significant technical limitations, and the greatest technical progress has tended to focus on security-related images such as faces and fingerprints rather than IP industry uses.

It would be useful to check this area of software development periodically, however, because although IP Australia trademark and design patent examiners indicated it would be of limited use for searching IP data (because most searching is done using classifications), searching of the non-IP literature could eventually be facilitated by such tools.

The two main statistics used to describe the quality of retrieval are

- **Precision** - The proportion of relevant documents of all documents retrieved:
 $P = (\text{number of relevant documents retrieved}) / (\text{number of documents retrieved})$

- **Recall** - The proportion of retrieved documents of all relevant documents available:
 $R = (\text{number of relevant documents retrieved}) / (\text{number of relevant documents})$

Thus, a search engine that retrieves all documents relevant to a query is said to have high **recall**, whereas a search that returns only relevant documents is said to have high **precision**.

Although the ideal technology would have both high recall and precision, in practice high recall can sometimes come at the cost of irrelevant hits (low precision) and high precision can come at the cost of the omission of some relevant hits (low recall).

These parameters are important to understand and control in the choice of an algorithm for intellectual property searching, because an examiner doing a prior art search may tolerate high recall with some loss of precision. However, precision may be more important to a researcher doing a technology search.

Ideally, the expert searcher should be able to modulate the precision and recall directly, as well as through the choice of search terms and ranges such as date ranges within search fields. An interface that does not allow such control, *e.g.* the Google default interface, may not be ideally suited for intellectual property searching in general, although particular versions of such an interface may be suitable for particular types of searches.

The quantitative evaluation of recall and precision statistics requires relevance to be determined by some means other than the system being evaluated, *e.g.* human experts may be used to rate relevance with respect to an information need expressed in natural language. This is the

approach taken by the Text REtrieval Conference⁵¹ (TREC), which since 1992 provides a controlled environment for the evaluation of different IR techniques for specific tasks. The evaluation of the IR system includes the process of the human user translating the information need from natural language to the search engine's query language. Typically much information is lost in this translation, limiting both the maximum achievable precision and recall.

Recent advances in IR attempt to address this problem in various ways:

- providing rich query languages that allow expert users to more closely represent their information requirement (improves both precision and recall);
- broadening the search using search terms extracted from documents that match the initial query - either automatic query expansion or query expansion using some form of relevance feedback - (improves recall at cost to precision);
- broadening the search using Natural Language Processing (NLP) techniques including stemming and using synonyms from thesauri (improves recall);
- multiple iterations of human/computer interaction using NLP techniques to refine the search (improves both precision and recall);
- application of NLP techniques based on Artificial Intelligence (AI) in the determination of document relevance e.g. use of Bayesian inference networks (improves both precision and recall) and Word Sense Disambiguation (improves precision).

Inverted Index

An Inverted Index data structure is used to support most IR techniques. Each word in the corpus has an index entry, which specifies the documents that include that word.⁵² A search for documents that contain a word then becomes a simple lookup in the index.⁵³ The list of words may be reduced by using a 'stop word list'⁵⁴ or stemming. Index entries may be augmented with additional information such as the number of times the word appears in the document or the position of each occurrence of the word.

The inverted index is largely responsible for the impressive speed and scalability of current search engines. However, for large collections of documents the compilation of an inverted index can be a time consuming task, and the fast searching of an inverted index may require large amounts of computer memory. A large amount of research has been conducted on the problem of compressing inverted indexes to reduce these memory requirements.

Stemming

In most cases, morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of IR applications.⁵⁵

The key terms of a query or document may be represented by stems rather than by the original words. For IR purposes, it doesn't usually matter whether the stems generated are genuine words or not – thus, "computation" might be stemmed to "comput" – provided that different

⁵¹ <http://trec.nist.gov>

⁵² http://en.wikipedia.org/wiki/Full_inverted_index

⁵³ Paul E. Black, "inverted index", from Dictionary of Algorithms and Data Structures, Paul E. Black, ed., NIST. <http://www.nist.gov/dads/HTML/invertedIndex.html>

⁵⁴ Stop words are those words which are so common that they are useless to index or use in searches. In English some obvious stop words would be "a," "of," "the," "I," "it," "you," and "and". Hans Peter Luhn, one of the pioneers in information retrieval, is credited with coining the phrase and using the concept in his design and implementation of KWIC indexing programs (http://en.wikipedia.org/wiki/Stop_words)

⁵⁵ <http://en.wikipedia.org/wiki/Stemming>

words with the same 'base meaning' are reduced to the same form and words with distinct meanings are kept separate. Stemming provides several advantages:

- a query can find a document with different morphological variants of the search term (improved recall); and
- reduction in the number of distinct terms needed to represent the corpus reduces computer processing requirements.

Information Retrieval Models

Boolean Search

Most IR systems allow the use of Boolean expressions, explicitly or implicitly, to construct a query. The query may include Boolean operators such as AND, OR, and NOT and possibly a proximity operator such as NEAR. A list of documents that match the Boolean query will be returned by the IR system. If no Boolean operators occur in the query, the usual default is to interpolate an AND between terms: a query such as "intellectual property" will be interpreted as "intellectual AND property".

A skilled user of a Boolean search IR system will be able to refine searches by specifying Boolean expressions of arbitrary complexity.

However most users prefer their searches "pre-refined" by the application of one or more of the other models below.

Term Frequency - Inverse Document Frequency (TF-IDF)

TF-IDF is the classical IR technique for automatic determination of document relevance. The term frequency in the given document gives measure of the importance of the term within the particular document.⁵⁶ The inverse document frequency is a measure of the general importance of the term. A high weight in TF-IDF is reached by a high term frequency within in the given document and a low frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms.

There are many variations in the calculation of TF-IDF. Mostly the formulae used are *ad hoc* (determined by trial and error), but some recent variations are based on sound statistical methods based on natural language modelling. TF-IDF has proven to be efficient to implement and useful in a wide range of IR applications. It is the benchmark against which other techniques are compared.

The formulae used tend to be proprietary, but are known for open source software. For example, Apache Lucene uses the following formulae⁵⁷:

$$\begin{aligned} \text{TF} &= \text{sqrt}(\text{freq}) \\ \text{IDF} &= \log(\text{numDocs}/(\text{docFreq}+1)) + 1 \end{aligned}$$

CAMBIA's Dekko uses the following formulae to compute a normalised (ie in the range 0..1) TF-IDF score:

$$\begin{aligned} \text{TF} &= \text{freq}_{ij} / \text{max_occur}_i \\ \text{IDF} &= \log(N/n_i) / \log(N) \\ w_{ij} &= \text{TF} * \text{IDF} \end{aligned}$$

where

$$\begin{aligned} w_{ij} &= \text{TF-IDF score of term } i \text{ in document } j \\ \text{freq}_{ij} &= \text{frequency of term } i \text{ in document } j \end{aligned}$$

⁵⁶ <http://en.wikipedia.org/wiki/Tf-idf>

⁵⁷ <http://lucene.apache.org/java/docs/>

\max_occur_i = the maximum number of occurrences of term i in any document
 N = number of documents in the collection
 n_i = number of documents in which term i occurs at least once

The purpose of normalising the TF * IDF product is to allow relevance scores to be meaningfully compared across collections, allowing relevance results to be gathered from a "farm" of Dekko servers.

Proximity

When a query contains two or more terms, a natural measure of document relevance is apparent: within the document, what is the distance between each pair of terms? Intuitively, it seems clear that if a pair of words are adjacent (e.g. part of a phrase), the relation should be more significant than if they are separated by many other words. Elaborations of this scheme abound, some of them touching on aspects of so-called "natural language processing" - for example, applying different weightings depending on whether pairs of words appear in the same sentence or paragraph.

CAMBIA's Dekko uses a triangular matrix of minimum distances between word pairs, summed and (as for TF-IDF) normalised to give a relevance score in the range 0-1.

Vector Space Models

Vector Space Models use a vector in a high dimensional space to represent each document.⁵⁸ There are many variations, including Latent Semantic Analysis and Cluster Analysis (below), but generally the number of words in the corpus is reduced by using a stop word list, stemming and perhaps by replacing synonyms with the same word or even phrases with concepts.⁵⁹ Each remaining word in the corpus becomes a dimension in the vector space. A document is represented by a coefficient for each dimension, *i.e.* a vector in this space. TF-IDF is commonly used for the coefficients. Queries are mapped into the vector space in the same way as documents. The cosine of the angle between two vectors is most commonly used as a measure of similarity between documents or relevance of a document for a specific query.

Latent Semantic Analysis (LSA)

LSA is meant to solve two fundamental problems in natural language processing: synonymy and polysemy. In **synonymy**, different writers use different words to describe the same idea. Thus, a person issuing a query in a search engine may use a different word than appears in a document, and may not retrieve the document. In **polysemy**, the same word can have multiple meanings, so a searcher may retrieve unwanted documents with the alternate meanings."⁶⁰

LSA uses a Vector Space Model in which the number of dimensions is reduced by Principal Component Analysis, a well-known technique of linear algebra. LSA relies on vectors that are nearby in the low-rank space representing semantically similar concepts.⁶¹ The relation-significance of such nearness is often, but not always valid, so such criteria can lead to results which can be justified on the mathematical level, but have no interpretable meaning in natural language. The right number of dimensions appears to be important; the best values yield up to four times as accurate simulation of human judgments as ordinary co-occurrence measures.

⁵⁸ http://en.wikipedia.org/wiki/Vector_Space_Model

⁵⁹ <http://isp.imm.dtu.dk/thor/projects/multimedia/textmining/node5.html>

⁶⁰ http://en.wikipedia.org/wiki/Latent_semantic_analysis

⁶¹ <http://isp.imm.dtu.dk/thor/projects/multimedia/textmining/node10.html>

Cluster Analysis

Cluster analysis is a multivariate statistical technique that allows the identification of groups, or clusters, of similar objects in space.⁶² It is a technique that has a long history and wide application in many fields of science, resulting in a huge body of literature on the subject. The application of cluster analysis to IR has hence been studied since the early days of computing with the aim of improving the effectiveness of retrieval:

"The Cluster Hypothesis is fundamental to the issue of improved effectiveness; it states that relevant documents tend to be more similar to each other than to non-relevant documents, and therefore tend to appear in the same clusters".⁶³

As with LSA, the results of clustering methods can be seen to rely too heavily on purely mathematical techniques and empirical parameters but may be effective in conjunction with the other IR models mentioned here. Recent studies on so-called "post-retrieval clustering" hint at a way forward for Cluster Analysis techniques.

Inference Models

Information Retrieval can be regarded an inference or evidential reasoning process in which we estimate the probability that a user's information need, expressed as a query, is met by a document as "evidence".⁶⁴ The techniques required to support these kinds of inference are similar to those used in expert systems that must reason with uncertain information. Some of the inference models developed for expert systems can be adapted to the document retrieval task. The Bayesian inference network has been used to:

- support multiple document representation schemes. Research has shown that, even when retrieval against each individual representation has similar performance, documents that are retrieved using multiple representations have higher relevance;
- allow the results from different queries and query types to be combined. Given a single natural language description of an information need, different searchers will formulate different queries to represent that need. The same searcher may also formulate multiple queries for the same need, each based on a different strategy. Even when average performance is similar for each query, documents retrieved by multiple queries are more likely to be relevant; and
- facilitate flexible matching between the terms or concepts mentioned in queries and those assigned to documents. The poor match between vocabulary used to express queries and the vocabulary used to represent documents appears to be a major cause of poor recall. Recall can be improved by using domain knowledge to match query and representation concepts without significantly degrading precision.

Dekko

The Dekko search engine is the back end of the BiOS Patent Lens. Key features of Dekko are:

- full inverted text index, with position lists and word counts
- customisable XML parser for patent documents

⁶² <http://isp.imm.dtu.dk/thor/projects/multimedia/textmining/node11.html>

⁶³ Jardine, N. and van Rijsbergen, C. J. 1971. The use of hierarchical clustering in information retrieval. Information Storage and Retrieval, 7:217440.

⁶⁴ Turtle H and Croft W B, Evaluation of an inference network-based retrieval model, ACM Transactions on Information Systems. Volume 9 , Issue 3 (July 1991) Special issue on research and development in information retrieval Pages: 187 – 222 ISSN:1046-8188

- minimal perfect hash technique for fast word search
- Boolean, TF-IDF and Proximity Matrix information retrieval (IR) models
- written entirely in C; runs on any variety of Unix/Linux
- threaded dekho server application for fast response to multiple user queries

The Dekko system has three main components:

- the **dek indexer** scans a document collection and creates an inverted ASCII index
- the **dekbin compiler** takes an ASCII index produced by *dek *and makes a compressed binary index
- the **dekho server** uses a compressed binary index produced by **dekbin** to answer queries

The dek indexer

The first stage in most Information Retrieval systems is the creation of an "Inverted Text Index". This is just a list of all the words found in all the documents in a collection. For each word, a list is kept of all the files in which it occurs, and the positions at which it occurs in each of those files.

Key questions here include:

- what constitutes a 'word'?
- what words should be indexed?

The dekbin compiler

An ASCII inverted index can be a very bulky item - typically of the same order of size as the original document collection. Searching such an index could be a slow process. The **dekbin** compiler takes a number of steps to make this problem manageable:

- for efficient storage, the ASCII index is converted into a compressed binary form
- for efficient searching, words are stored in the binary index using a minimal perfect hash⁶⁵

The dekho server

This takes a binary index and perfect hash function as produced by **dekbin** and

- loads the essential parts of the binary index into memory
- maps the rest of the binary index into virtual memory (using mmap())
- opens a socket connection on a user specified port

Formatted queries are then sent to the port, and if the query makes sense a list of matching documents is returned. In theory, a user could telnet directly to the port and type in a query. In practice, only the most trivial of queries can be done this way - the query format is complex and tricky to type at the keyboard! Usually a client program, for example by generating a form, will initiate the connection, send a correctly formatted query, and collect the results⁶⁶.

Features of Dekko

As described above, the three major components of Dekko, the **dek** text indexer, the **dekbin** compiler, and the **dekho** server, together provide a solution to the task of indexing a collection of text documents (in this case, patent documents) and then searching the index for documents relevant to a query.

There are a number of unique features of patent data as sourced from the various Patent Offices (WIPO, EPO, USPTO, IP Australia)⁶⁷. For historical reasons, patent data comprise an

⁶⁵ provided by Bob Jenkins at <http://burtleburtle.net/bob/hash/perfect.html>

⁶⁶ Currently at CAMBIA we use a client written in Perl, but such a service can be provided with any other scripting language (Python, Ruby), Java, or C or C++. The fundamental requirement is the ability to connect to and communicate with a socket.

⁶⁷ Dekko has been purpose-designed at CAMBIA, specifically with the task of indexing these patent data.

interesting mixture of formats and encodings. We see data in 80-column card format, SGML/XML format, and text extracted from TIFF images using OCR, and variations in between. Much of the SGML/XML data is poorly formed and resists parsing by commonly available parsers. Embedded in the text we see what are now called “entities” in SGML and XML terminology. In older data, the formats of these entities do not conform to SGML standards, and entities are used to both encode data and to provide formatting and layout descriptions.

Similarly, we see various character encodings: 7-bit Asciii, ISO-8559-1 Latin1, 7 bit Ascii with some 8-bit characters derived from EBCDIC (mainframe) code pages.

The **dek** indexer is designed to be able to cope with all these variations of data format. It incorporates a purpose-built SGML/XML parser that is tolerant of badly formed documents and copes with unusual character sets. **dek** is highly configurable – new formats and styles can be accommodated by simply editing a configuration file.

The presence of OCR data (WIPO PCT documents, EPO pre-2000 documents) causes other potential problems (see below). The OCR process is never perfect, due to limitations in both the quality of the original images from which the text is extracted and the imperfect nature of OCR algorithms. This can cause the text contain large numbers of essentially random words. The Dekko system incorporates several filtering and compression techniques to reduce the size of the indexes. Without these filters, the PCT index would contain more than 40 million words.

Similar filters remove DNA and protein sequences from the searchable index. These sequences are no more than random strings of letters, in the case of DNA containing combinations of the letters A,C,G and T.

Having filtered the text, **dek** compiles an inverted index. This is stored on disk in ASCII format, and even with rigorous word filters the size of an ASCII index can be of the same order as the text from which it was derived. The **dekbin** component of the dekho system processes and analyses the ASCII index – it compresses the ASCII index by a factor of 3; it counts the frequency of words in documents (and sections of documents, eg abstract, title, claims). The word counts provide the basis for **relevance ranking** in the **dekko** server.

It is also in **dekbin** that a technique called **minimal perfect hashing** is used. For each word in the index, a single number is computed. Every word in the index gets a unique number; the numbers range from 1 to the total number of words in the index.

The eponymous **dekko** server is the program that searches a **dekbin** compiled **dek** index. The dekho user sends the server a query comprising a number of words; for each word the minimal perfect hash number can be computed rapidly (just a few machine instructions) and the details of the word – which documents it occurs in, where it occurs in each document, how often it occurs in the whole collection – can be retrieved from a known place in the index.

The **minimal perfect hashing** technique results in extremely fast word retrieval; this, combined with ability to store the bulk of the compressed index in memory makes **dekko** into a very fast Information Retrieval (IR) system.

It should be kept in mind that the essence of providing a good search capability for patent data is not necessarily the search algorithm itself – it is the ability of the engine to work together with all the data formats and structures that are present in collections of patent documents. That knowledge has been designed and configured into **Dekko**⁶⁸.

⁶⁸ The Dekko system has been successfully used in the CAMBIA BiOS Patent Lens for several years. It is currently used to index WIPO (PCT), EPO, and US patent data. For a time, Australian patent data was also being added, and although this service has lapsed for want of funding, **Dekko** can do it, has done it, and is in fact ideally suited to doing it.

Optical character recognition (OCR)

Over recent years the accuracy and speed of Optical Character Recognition (OCR) technology has greatly improved and it is now practical to perform OCR on very large collections of digital images cost effectively. CAMBIA has extensive experience in this area, having carried out OCR on millions of pages covering PCT, EPB and Australian patents (USPTO data is brought in electronically without the need for an OCR step).

Many patent filings in Australia enter national phase via the PCT. PCT applications should be considered prior art for Australian filings as soon as published, so another factor complicating searching of Australian patents and applications is that PCT applications may first appear in a language other than English, such as Japanese. There is discussion of adding Chinese as a language in which PCT applications are accepted. Optical character recognition of Chinese and Japanese characters may require special software.

We have recently conducted extensive trials of the latest OCR technology with a focus on the recognition of Chinese and Japanese as well as European languages. We found two OCR engines that with our applications achieved accuracy levels of over 99% on PCT patents published in Japanese and patent documents provided in Chinese.

IP Australia's web site⁶⁹ indicates that there may be some current data provision issues that could limit the usefulness of even a well designed search interface for full text or field based searching of bibliographic data, claims and specifications sections of Australian patents. To facilitate the future searching of specific sections of text, such as claims, it will be important to ensure that OCR data quality issues are anticipated so that they do not affect the facility with which data markup can be accomplished.

One way to address at least some of the data integrity issues is through the adoption of applicable international standards. For example, for the dissemination of Australian patents after OCR, the revised "wo-published-application" dtd format, outlined in the WIPO document C PCT 1037-76⁷⁰ may have some relevance. OCR quality of photocopied PCT applications being transmitted to Australia for national phase could be improved by instead obtaining the OCR that had been done for the PCT applications directly, and then the national phase information that is often stamped on hard copies could be added electronically. The recent decision by IP Australia to accept, from 18 July 2005, limited filing of international applications directly in electronic format in its capacity as PCT receiving office also offers an opportunity for a focus on data quality based on international standards.

Natural Language Processing (NLP) applied to searching Chinese, Japanese, and Korean (CJK) language patents

An emerging trend in IR relates to use of machine translation (MT) to allow searching across information in many languages. A number of popular search engines offer "on the fly" MT of web pages in number of languages into English, but there is a severe "lost in translation" problem that would greatly complicate reliance on this for technology searching. Nonetheless, cross language searching remains an area of intensive research and it is already proving valuable in some applications.

Patent data show a more formalised structure than natural language, however. This may reduce the challenge somewhat of using character-based text search algorithms. Our scoping of the Chinese patent data predicts that it is an accomplishable process to adapt in and invent search capabilities to integrate with the existing searchable patent data structures the handling

⁶⁹ see for example <http://www.ipaustralia.gov.au/pdfs/patents/fields.pdf>

⁷⁰ http://www.wipo.org/pct/edi/en/wo_publication_information/documents/pdf/circular_wo_changes_en.pdf, C. PCT 1037 - 76 of 8 July 8 2005

of non-ASCII characters that are much more complex. However, research needs to continue into the semantic models and utilities for differentiated thesauri.

Languages with large number of characters compared to the Latin alphabet or with an absence of white space between words pose particular IR challenges. Issues such as character encoding, word break analysis and specific processing for transliterated foreign words, abbreviations, and personal, organization and company names must be addressed for the effective searching of CJK language patents.

Fortunately, a large part of the informatics groundwork necessary for integration of Chinese information into databases can be extended into other language databases based on non-ASCII characters and other semantic structures. Thus further extension into patent information in the languages of other important trading partner jurisdictions such as India, the Middle East, Indonesia, Korea and Japan should be accomplishable as the program continues.

Chinese Language

Foreign proper nouns are usually transliterated into Chinese by using characters for their phonetic rather than semantic value. Each Chinese-speaking region may transliterate the same name differently: For instance, 'Novartis' could be translated into 诺瓦提斯 or 挪伐帝司, depending on the attorney who prepares the documents. In this example, the characters used for the translation are completely different. In Mainland China, the foreign names of individuals, companies and institutions that appear the bibliographic page of patents are normally transliterated into Chinese, which makes searching difficult. Fortunately, each region uses a relatively small and consistent set of characters when transliterating.

A further complication is the use of simplified characters in mainland China and traditional characters in Hong Kong and Taiwan. Although traditional Chinese and simplified Chinese share more than 70% of the characters, some characters (words) can be very different. For example, 'international' is '國際' in traditional Chinese and '国际' in simplified Chinese. Since the majority of Chinese language patent documents filed with WIPO or EPO are from Mainland China, simplified Chinese is more dominant. However, if the original documents were prepared in traditional Chinese, the search terms must also be entered in traditional Chinese to find the relevant documents.

Phrase and names can be abbreviated by taking a character from each part of the word or phrase. There is no clear rule as to whether these should be the first or subsequent characters. For example Beijing University, 北京大学, is usually abbreviated as 北大. Abbreviations are irregular and their use widespread, so constructing a comprehensive lexicon is a challenge.

Japanese language

The Japanese language uses a combination of four scripts with an extremely complex morphology and orthography. Natural Language Processing (NLP) techniques are essential for successful full text searching of Japanese text. The CJK Dictionary Institute (CJKI)⁷¹ specialises in CJK computational lexicography and has developed a lexical database with over two million Japanese and one million Chinese entries.

A major source of complexity in processing Japanese texts is the presence of an extremely large number of homophones. Many homophones are synonyms in some senses but not others and as a result it is hard to predict which an author will choose to use in a particular context. For example, the verb 'noboru' can be written using three different Chinese characters: 上 昇 can be

⁷¹ <http://www.cjk.org/cjk/index.htm>

used to mean "to move upwards", 登る can be used to meaning to "transport oneself to a high point" and 昇る can be used to depict the rise of an astronomical body. The point is that given that these words are all pronounced in the same way and that they have very similar meanings, there is ample scope for one to be used in place of another through error or expediency on the part of an author.

On the other hand, personal names (in most cases written in Chinese characters) have a variety of ways in which a single combination of characters can be pronounced (e.g. the first name 'Shoko' when written in Chinese '尚子' can also be read as 'Naoko' (most common) and 'Hisako'). This can create a problem when searching for an inventor's name if it has been transliterated into English, as all variations must be considered when searching.

One of the most diverse areas of the Japanese language in terms of orthography is words derived from English. The orthography of these "katakana" words is extremely variable and even native Japanese speakers wishing to search Japanese text may benefit from NLP techniques that allow English keywords to be entered to retrieve all katakana and Latin alphabet variants. These katakana characters are also frequently used for Japanese words on certain occasions, or for non-Japanese languages other than English. The former type of katakana characters is often seen in the cases of organisms that are used in scientific literature; e.g. rice is '籾 (ine)', which is written in katakana as 'イネ' for rice that is used as scientific material (not for agriculture). This creates a problem when using the word 'rice' as a search term, as (again) both katakana and Chinese writing must be considered.

Finally, the Japanese have several ways of writing numerals, similar to those explained in the section on Chinese above. In patent documents, the bibliographic information sheet uses Arabic numerals to identifying the different types of information (e.g. '(51) International patent classification', '(21) application number'), dates, references, and names of things like vectors, proteins, etc in the specification. All other numbers are mostly in unicode e.g. "200", Chinese characters are rarely used (except for old patent documents).

Cross language Information Retrieval

The Cross-Language Information Retrieval (CLIR) activities initiated within the Text Retrieval Conference (TREC) have stimulated much interest in Europe and Asia. The European Cross-Language Evaluation Forum (CLEF)⁷² supports TREC like workshops for CLIR using European languages, whilst the Japanese NTCIR workshops⁷³ support CLIR in Chinese, Japanese, Korean and English (with a particular interest in patents). A task at the most recent NTCIR workshop involved a patent examiner invalidating a claim in English by using English queries to identify sections within Japanese patents that would invalidate the claim.

Current CLIR techniques make use of parallel bodies of text available in both the query language and the corpus language (the language of the text being searched). For patents, professionally produced English abstracts may be used. This data is used in Machine Translation (MT) of either the query into the corpus language or documents into the query language. Both strategies provide good results, and combining both provides close to monolingual search quality.

Patent Claims

Because the body of a patent specification often contains considerable matter that is not claimed because it describes the prior art which may be in the public domain, but which is used to understand the definitions that form the bounds of what is claimed for exclusionary rights,

⁷² <http://clef.iei.pi.cnr.it>

⁷³ <http://research.nii.ac.jp/ntcir/outline/prop-en.html>

analysis of claims should focus on the parsing of the technical language in the claims but carried out in light of the specification is necessary for more sophisticated analysis of freedom to operate.

As outlined below, the formal technical style of claims language allows certain computational linguistic techniques to be employed that are not applicable to patent documents as a whole.

Claims parsing

In determining the extent of protection conferred by a patent, the claims are the most important section of the patent. Claims analysis is central to the examination process and in determining freedom to operate and patent infringement. Having access to the complete text of the claims section of a patent is a prerequisite to claims analysis, but this information alone is insufficient for the clear delineation of the metes and bounds of what is claimed. When interpreting claims scope, it is important to interpret the claims in the light of any definitions that may exist in the patent specification, including references to diagrams, figures and other non-textual information. Where the specification is unclear, the file wrapper may clarify matters. Should further clarification be required, external references such as dictionaries may need to be consulted.

Types of claims

Claims can be viewed as falling into two broad types, independent claims and dependent claims. Claims can be further divided into categories based on what is claimed, a method claim or device claim, etc. Differing national patent laws mean that techniques and strategies needed to determine the metes and bounds of claims may differ. In much of Europe, claims tend to be interpreted according to a central claiming system, in which claims identify the "centre" of the invention. In the United States and England a peripheral claims system, which identifies the limits and bounds of what is claimed, applies.⁷⁴

There are a number of special types of claims, which can be identified by reference to their structure and content. "Jepson claims" typically may include the phrase "wherein the improvement comprises". Markush claims, often used in chemistry but occasionally seen in all other fields of art, may include the phrase "consisting of" or "comprising" followed by a variable list of possible substituents. Product-by-process claims typically use phrases like "Product obtained by the process of claim...". Beauregard type claims often include the phrase a "computer-readable medium".

Computational linguistic techniques have been applied to patents claims with an number of aims in mind, including computer-aided drafting of claims and the machine translation of claims from one language to another, the conversion of claims into more readable forms through the reformatting of claims into a more linguistically natural structures, and the alignment of sections of the description with the relevant sections of the claims.⁷⁵

Patent claims have a peculiar structure that challenges natural language processing (NLP) techniques. Claims are an extreme example of very long sentences with an abundance of telescopically embedded clauses. There are, however, aspects of claims grammar such as use of predicates⁷⁶ and aspects of the linguistic specificity of claims drafting style that allow the use of NLP techniques such as Rhetorical Structure Analysis (RSA) using cue phrases⁷⁷, and combination of methods exploiting grammatical formalisms with predicate lexicons.⁷⁸ Although

⁷⁴ http://www.european-patent-office.org/tws/tsr_97/chapter3.htm

⁷⁵ http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/OPEN/OPENSUB_Akihiro_Shinmori.pdf

⁷⁶ http://en.wikipedia.org/wiki/Predicate_%28grammar%29

⁷⁷ <http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-PATENT-ShinmoriA.pdf>

⁷⁸ <http://acl.ldc.upenn.edu/W/W03/W03-2008.pdf>

many of these techniques have advanced in recent years, these technologies are yet to be widely adopted.

As the volume of patent applications increases and the need for the translation of patents and non patent literature grows, as a result of trade and patent co-operation treaties, computational linguistics will play an increasingly important role. In the areas of patentability, freedom to operate (FTO), invalidation and infringement searching across multiple languages, NLP techniques have the potential to give significant competitive advantage to those organizations that successfully deploy the technology.

Quality of data for claims analysis

While some patent offices, such as the USPTO and EPO, publish patents in a format that makes claims easy to programmatically identify and search, this is not universally true. In the case of Australia, where claims are published as part of an image facsimile of the whole patent document, optical character recognition (OCR) and text parsing techniques are required just to identify the claims section. If claims have been amended when a PCT application enters national phase, programmatic identification of claims is dependent on the conventions and procedures adopted by publishing authority of that nation. If all that is provided is a cover sheet with hand written amendments or sections of claims ruled out, programmatic access to and analysis of claims becomes problematic.

Publication of the claims sections of patents in machine readable form (e.g. Standard ST.36 Recommendation for the Processing of Patent Information using XML) can be seen a first step in facilitating innovative software to add functionality such as linking terms used in the claims to definitions in the specification, file wrapper, or external references.

WIPO has announced that as of January 2006 it will produce "Weekly Published International Applications" on DVD (formerly known as Rule 87 DVDs) in a new format that includes search quality OCR data with the claims clearly marked.⁷⁹ The "wo-published-application.dtd" could be seen as setting a minimum benchmark for the publication of patent documents.⁸⁰

Specialised Data Searching

Two areas of specialised data searching are presented here, chemical and sequence searching, because the art areas in which they occur are of special significance to Australian industry, and because they present particular challenges for claim construction.

Chemical Searching

Searching the patent literature for chemical substances poses a unique set of challenges.

Chemical names can have many equivalent synonyms, but in the literature chemical substances are often represented by chemical structures, which can also be depicted by a number of equivalent nomenclatures.

Claims over chemical structures are frequently presented in Markush⁸¹ form, in which a particular claim can prophetically cover a large variety of substances in which a named structure is represented. The metes and bounds of a given claim are often provided by various transitional phrases such as "consisting of" or "comprising". A Markush structure is a claim with

⁷⁹ http://www.wipo.int/pct/edi/en/wo_publication_information/

⁸⁰ http://www.wipo.int/pct/edi/en/wo_publication_information/samples/wo-published-application-v1-4.html

⁸¹ named after Dr. Eugene Markush, whose 1923 patent became a test case for the inclusion of multiple, independently varying functional groups in the description of a chemical invention (this is the source of the figure shown).

multiple "functionally equivalent" chemical entities allowed in one or more parts of the compound.

For an indication of the challenge, on the figure page that follows, the Markush structure shown covers more than 150,000 individual compounds, so even though relatively few actual words for substituents are used, a claim on this structure would read on 150,000 possible variations of the core molecule that would theoretically need to be checked in the prior art.

Searching for specific substances requires a database containing an index of **chemical fragment codes**. Searching Markush structures requires a specialised "**Markush**" database. Both types of searches require dedicated chemical structure tools (computer programs) that allow standard chemical structures to be converted to and from chemical fragment codes and/or Markush structures.

The following are the chemical structure databases available currently, both full fee-paying services with non-trivial costs.:

- **Derwent/INPI Merged Markush Service** (also called Markush **DARC**), comprises the Derwent Markush index merged with the INPI Pharmsearch, with data from 1987 onwards, available on *Questel Orbit*.
- Chemical Abstracts Service (CAS) **MARPAT**, available online on *STN Web* and available via the desktop PC programs *STN Express* and *SciFinder*, with data from 1988 onwards.

Both these databases have richer data indices for more recent patents than for earlier patents.

An encouraging trend in the area of chemical searching is the use of Bayesian search algorithms, exemplified by Reel Two⁸², which can use thesauri defined by the user to translate chemical names (CAS, IUPAC, common, SMILES) to structure. The structures selected by the user can then be utilised as a starting point for iterative searches in which additional synonyms and related terms found in the searched database are brought forward to the user as options. However, good knowledge of chemical terminology by the searcher is clearly required. Additionally, this is currently set up primarily for in-house dedicated and proprietary databases and is not available as a web service.

⁸² <http://www.reeltwo.com>

Searching Patent Data for Biological Sequences read on by claims

Biological sequences can be broadly categorised into two types: sequences of nucleic acids (DNA and RNA) identified by the coding letters A, U or T, C, and G, which identify nucleotides with particular base-pairing relationships; and polypeptide, peptide or protein sequences (linear polymers of amino acids either identified by a single letter alphabetic code or a three letter code, e.g. "S" or "Ser" for the amino acid serine)...

The challenge in claim construction is that biological function is usually not dependent on the exact sequence, and similar or homologous functions can be "encoded" by many variations that might be only weakly similar to the particular sequences in the specification. Thus, claims are seldom over a single exact sequence, and tend to be worded such that they read on many sequences that may be found in the literature.

For example, a typical claim wording is "An isolated polynucleotide having a sequence (to some named percentage of the sequence such as 65%, with higher homology being "preferred", and higher homology up to 95-99% "most preferred", "particularly preferred", or "especially preferred") **homologous** to SEQ ID No. 1, or a **portion / fragment** (of specified length) thereof" or "An isolated polynucleotide which **hybridizes** (usually under some specified conditions) to SEQ ID No.1, or a **portion / fragment** thereof"⁸³.

The language used in sequence claims is also very often similar to that used in chemical claims using the "Markush" notation. The metes and bounds of a given claim are provided by various transitional phrases such as "consisting of" or "comprising" multiple sequences.

There is no general rule for the required percentage agreement or the stringency conditions for chemical hybridization, because biological function of an enzyme and particularly the portion of the sequence encoding the active site may be very highly conserved, whereas function of a structural protein may be seen in variants that are less similar.

A further complication in prior art searching is that the literature may or may not indicate the biological function related to a particular sequence accurately, so for a particular sequence claim it may be challenging to find the many sequences in the prior art that could be related by sequence similarity, by potential hybridisation or by function.

Furthermore, amino acid sequences are encoded by "translated" DNA sequences, which occurs by a biological mechanism that reads sequences or "frames" of three nucleotides each, so a deletion or insertion error in the nucleotide sequence can result in a different "translation". Thus, for comparisons involving a relation to a predicted protein sequence, DNA sequences must be checked in three "frames" on each of the two base-pairing strands of each DNA molecule. For example, shown on the figure page above is a sample output in which a DNA sequence being searched has matched a sequence in a patent document, although there are short gaps that suggest the protein sequence would not match.

Biological sequences are usually given as separate listings as well as shown in figures or tables within the specification or claims section of a patent. Short sequences (less than 20 residues) may be quoted within the text of a sentence, and not all sequences in the specification or sequence listing will be covered by the claims; for example, sequences claimed by others or in the public domain may have been used in the examples of a patent specification to find or compare the sequences that are claimed.

The two most commonly used sequence search algorithms are variations of FASTA and Basic Local Alignment Search Tool (BLAST) as developed by the US National Center for Biotechnology Information (NCBI), or the similar, "local alignment" algorithm FASTA, both of

⁸³ Examples are taken from the Examination Guidelines for Patent Applications relating to Biotechnological Inventions in the UK Patent Office

which use heuristics to search DNA and protein sequences. For a comparison of sequence search algorithms for protein sequences, see Shpaer et al. (1996).⁸⁴ Most of the patent search service vendors provide these algorithms for searching patent data for sequences, though with varying flexibility of use. In general it is desirable, depending on the claims, to do:

- DNA sequence query against DNA sequences, both for similar sequences and complementary sequences that would hybridise to similar sequences
- Protein sequence query against protein sequences
- Protein sequence query against translated DNA sequences
- "DNA sequence translated into 6 possible reading frames" query against protein sequences

These algorithms have a tendency to trade off completeness for search speed, so many users do not realise how important user specification of parameters is to determine the precision of the output. The sequences being compared are aligned within a moving "search window", and controls for adjusting input parameters include:

- search matrix (e.g. blosm62)
- "word" length (the length of the moving search window in which matches are made)
- gap penalty (for adjustments due to small insertions and deletions in one sequence relative to another)
- cut-off threshold for expected frequency ("E score")
- cut-off threshold for percent identity over the sequences being compared
- options for displaying output, such as displaying alignments, etc.

Because different patents can claim different degrees of relatedness for sequences (e.g. from exact matches to as low as 40% identity), searches by examiners and searches for freedom to operate should be performed initially with small word lengths, low gap penalties and low cut-off thresholds with the intention of finding all related sequences, including distantly related ones. Then depending on the claims language of the patent documents of interest, each patent document found to contain a sequence that a claim may read on can be checked to determine whether it should be discarded or included.

Patent Sequence Databases

Four of the major national patent offices provide public-accessible sequence databases:

- USPTO provides sequence data to the National Center for Biotechnology Information (NCBI) database
- European Patent Office provide sequence data to the European Bioinformatics Institute (EBI) database
- Japan Patent Office provides sequence data to the DNA Database of Japan (DDBJ)
- World Intellectual Property Organisation (WIPO) provides sequence data (PCTGEN) from PCT application via their web site

⁸⁴ Shpaer E et al (1996) "Sensitivity and Selectivity in Protein Similarity Searches: A Comparison of Smith – Waterman in Hardware to BLAST and FASTA. Perkin – Elmer, Genomics 38, 179 - 191

There are two commercial and one non-commercial online search services that allow combined searches against some or all the above sequence databases. These are:

- Thomson/Derwent "GENESEQ" (for in-house searching)
- Chemical Abstract Service "STN" (online searching of the GENESEQ database)
- Patent Informatics "PatGen" (online searching of publicly available databases listed above)

Thomson/Derwent provide the largest single database of sequences extracted from patents, GENESEQ, and an "early availability" database called "GENESEQ FASTAlert" that provides timely access to newly published patent sequences, before those sequences are fully annotated and added to the full GENESEQ database. For data from 1981 onward, GENESEQ includes all nucleic acids (10 or more bases in length), amino acids (4 or more residues in length) and all PCR primers and probes contained within patents derived from the claims, examples and general disclosure. This covers basic patents from 41 issuing-authorities including US government and PCT patent applications.

GENESEQ is available on the web, via STNweb and in flat file (EMBL) format (enabling integration into in-house bioinformatics systems). Many IP offices prefer the latter for confidential searching of pre-publication sequences.

STN provides access to the Thomson/Derwent GENESEQ database (called DGENE in STN) as well as access to the PCTGENE database.. Polypeptide and nucleic acid sequences in DGENE and PCTGEN are searchable using three search tools:

- BLAST® sequence similarity searching from the National Center for Biotechnology Information (NCBI)
- GETSIM FASTA based sequence similarity searching from FIZ Karlsruhe GmbH
- GETSEQ sequence code match searching from FIZ Karlsruhe GmbH. Useful for short and/or highly conserved sequence queries.

Input sequences can be manually entered (up to 200 residues), read from a file uploaded to the server, or may be recalled from a previously saved search.

PatentInformatics⁸⁵ provides an online service for searching sequences extracted from patents, but also advertises an in-house database solution that can be installed for a particular user's database with weekly data updates. The sequence data is extracted from the publicly available databases listed above (NCBI, EBI, DDBJ, PCTGEN).

PatentInformatics maintains the following online services:

- PatGen DB Lib - This is a basic keyword search tool that enables the searching of bibliographic data. The user can search for the title, abstract, inventor, applicant and date.
- PatGen DB Tax - This tool has been set up to enable the searching of patent genetic sequences based on the known sequence taxonomy. This is useful to make queries such as, for example, how many patents disclose Fowlpox virus sequences.
- PatGen DB Blast - Basic Local Alignment Search Tool (BLAST) is used to perform sequence searching. In this case patents can be searched through the matching of related genetic sequences.

⁸⁵ <http://patentinformatics.fdns.net/>

PatentInformatics, like Thomson/Derwent GENESEQ, also provides an in-house search that can be installed for confidential searches of pre-publication sequences, for example by an examiner of an Australian patent that did not enter via the PCT. Unlike GENESEQ, PatLAMP uses a "LAMP" based (Linux, Apache, MySQL, PHP/Perl) server, which includes:

- Linux - Suse 9.2
- Webservices: SOAP
- XML processing: DOM and XSLT
- Patentinformatics - PatIndex, PatBLAST, PatServ
- Bioinformatics - BioPerl, BioPerl DB, BLAST, CLUSTALW

While this service is very imperfect (for example, the search output is limited to ten results, no matter how broad the claim around the sequence being searched), PatGen's presence in the market suggests that provision of such a search tool is something that a national office such as IP Australia could also envision for public good, and the tools are available for an open source solution.

Chapter 3. Emerging and potential new uses for IP information

This chapter considers emerging and potential new uses for IP information for IP Australia, its customers and other stakeholders. We consider that these customers and stakeholders comprise both internal (examiners) and external (the public, not only innovators and implementers of technology). In the increasingly globalised world of trade, we recognise that the Australian innovation system must produce innovations that can be delivered to markets, most of which are outside of Australia.

In the context of the worldwide trend for patent reform, it is now possible to conceive of wise use of search trends to put Australia in a leadership position rather than in catch-up mode, to show total transparency and allow the leverage of many rights (currently held by Australians or not), into Australian industry, environment and social prosperity.

Searching of IP information for trade leverage

The last twenty years has seen an unprecedented increase in worldwide patenting. Over 99% of patents originate from inventions outside of Australia. Increased compliance with TRIPs amongst our Asian and other trading partners is seeing their national patent systems inundated with patent filings from major overseas corporations, effecting new barriers but possibly also new opportunities for timely Australian trade access for new primary industry products, advanced technologies, and services in key markets.

IP barriers may be as significant or more than tariffs and other trade barriers in affecting market pull. Though an Australian entity may develop and own a technology, injunctions can force cessation of production and sales and delay or stop imports in countries where blocking patents are in force, even if such patents are vulnerable to later legal challenge

Thus, we see a necessary extension of IP informatics and search considerations to emerging markets of interest. Indeed, given Australia's relatively close geographic location to Asia, its strong trade history, and the close political alliances with many of the relevant nations, it would be contrary to the interest of Australia for IP Australia to miss the opportunity to consider how the intellectual property information can in effect be leveraged for the continuing industrial, cultural and financial development of our national export markets.

Searchable and timely information on what technology is being constrained in overseas patents is important to a very wide community of stakeholders in IP Australia:

- Inventors, examiners and legal professionals have an interest in not wasting money with developing and prosecuting technology that has already been invented elsewhere.
- The financial community and public good funding both aim to make wise investments in research and development. To base investment decisions on the commercial deliverability of the resulting innovations will require continually updating information on the opportunities and barriers in existing technology and target markets.
- For commercialisers and production industries interested in commissioning, delivery and uptake of the inventor community work product, ownership patterns of important technologies and attendant rights to practice are a major issue for planning acquisitions, partnering, plant locations, and other decisions with wide economic implications.

A number of Australia's key markets are emerging economies with recently developed IP systems with an understanding that provision of effective public access to their IP systems is necessary to foster local innovation and economic growth. Clearly Australian researchers, business people and other stakeholders in the innovation process must also have

straightforward means of finding, combining and utilizing the IP information they need from internal and overseas jurisdictions to make good decisions about technologies, potential markets, and partners.

To ensure that Australian innovators, investors and industries are not disadvantaged, IP Australia must consider long-term investment as well as rapidly placed interim solutions that begin with the right search and retrieval technology and that are scalable to continue expanding while long term solutions are set in place; the latter is not only vital but urgent.

- Basic functionality such as full text searching of Australian patents, including claims and specifications, is a minimum requirement.
- Effective access to IP data from key markets is also central to ensuring that Australian research and innovation can be commercialised.

Searching of IP information to foster Australian technology development

The intent of the enabling description requirement was to provide information on the technology so that it could be used outside the monopoly grant and readily improved. Thus, there is a significant potential in the patent system for wise and legal use of technology that has been described but is not subject to valid claims in a particular jurisdiction.

Much of the cost of developing technology *de novo* or extending the use of technology in Australia can be saved by uncovering enabling descriptions of technology and shedding light on where this technology can be freely used. CAMBIA's mission focus has been development of IP information as a reservoir for understanding of the constraints related to existing technology and where windows are open for development or improvement. Our view and experience is that this can be highly successful. This view is also shared by the EPO:⁸⁶ "Because of a lack of information, existing inventions are re-invented, problems that have already be solved are solved again, and products that already are on the market are re-developed. Duplication of efforts in this way costs European industry US \$ 20 000 000 000 every year - simply because of the lack of information... 80% of technical information is published in patent documentation - and often nowhere else. "

Key jurisdictions in which a lot of innovation is occurring, including China and Korea, and for software and drug production processes, India, and Brazil, are increasingly focusing on development, use and enforcement of their patent systems. Rendering these databases accessible, interlinked and thus much more informative could create a major opportunity for technology discovery. The opportunity is greater if it involves mechanisms to ensure that patent status information (e.g. INPADOC) is incorporated in a relevant manner⁸⁷. Status information from the USPTO and other mature patenting systems is currently available via their patent offices, but not in a friendly interface that allows comparison with other jurisdictions. With status information users could more readily locate matter in patent documents that has passed into the public domain in the jurisdiction(s) of interest or that may not be patented in particular jurisdictions as the result of exhaustion of patenting rights. Leads obtained through this searching can then be verified and further developed to allow the leverage of technologies (whether the rights are currently held by Australians or not), into Australian industry, environment and social prosperity.

⁸⁶ <http://www.european-patent-office.org/patinfopro/index.shtml>

⁸⁷ Currently INPADOC information is available with paying database services and via Espacenet and CAMBIA, but CAMBIA appears to be the only provider that is improving the interface actively in the past year to make it more user-friendly for searching by the general public (including business, science, policy and other specializations) as well as by IP professionals. CAMBIA is also looking into ways of making the in-force status a field that can be queried in searches.

Searching of IP information for putting Australia in a leadership position

In the context of the worldwide trend for patent reform, it is now possible to conceive of wise use of search informatics to put Australia in a leadership position rather than in catch-up mode by showing increased transparency and integration.

The cost and time involved in obtaining data lengthens the patent drafting process and the examination process. However, unless data searches actually uncover the relevant prior art, they do not contribute to increased rigour of patent grants. Instead, they contribute to opportunity costs in innovation where investment has been misdirected into technology that already exists elsewhere.

Patent grants unrelated to deliverability, furthermore, increase rather than decrease the fear, uncertainty and doubt (FUD, a technical term used in the software industry) which is a disincentive to investors and innovators.

Better data-handling will support and is supported by a commitment, consistent with the obligations of TRIPS and other agreements but also with the public good of Australia, to examine the innovation system for where incentives are occurring and for which behaviour. The concept of a monopoly grant is, with varying emphasis in different jurisdictions, an economic exchange, as an incentive for enabling disclosure of inventions in the arts useful to the public and bringing leverage for capital recruited into the downstream portions of the innovation chain. Blows to investor confidence upon the discovery of invalidating prior art that should have been made known before the patent grant do not contribute productively to the economy.

- Conduct highly effective prior art searches of the patent data of Australia and other jurisdictions, non-patent literature and traditional knowledge, so that all prior art that would be read on by the claims of patent applications or invalidate them is found. This is visible action on a commitment to the rigour of the patent grant, so that patent monopolies are granted for truly novel and inventive work and not for innovation that is obvious in view of the prior art.
- So that the presumption of validity is commensurate with the actual rigour that went into the allowance of past patents, enhance the accessibility and decrease the cost of the opposition process. Similarly, develop and encourage any other comment processes that will harness the ability of the wider public to uncover prior art that might otherwise have been missed.

Service to IP Australia in this section of the report requires discussion of the revenue implications. IP Australia may consider that fees must be raised to cover increased search costs in order to gain such increased respect. The following recommendations offer concomitant opportunities to increase respect for the patent system in ways that may have a smaller effect on patent examination costs, or to mitigate effects on revenue streams while supporting the domestic economy in a wide-ranging way.

- Provide and support other means of disclosure for innovations, by disseminating awareness of incentives to innovation that may be less costly, in both real and opportunity cost and time, different from the process of obtaining a monopoly grant.

In software development, the availability of the “open source” alternative for generation of largely pre-competitive innovation in a “protected commons” has resulted in thousands of projects with over a million contributors worldwide that are bringing billions of dollars to the profitability of small, medium and large enterprise⁸⁸. This model can be followed in other industries.⁸⁹ Patent applications are still being filed in this area for truly inventive technologies for which a competitive advantage is desired by the applicant. However, the known alternative of the open source protected commons has allowed many innovations to come to light that might not have been disclosed in patent applications. Examples include innovations too small to justify the cost of patenting, still comprising useful adjustments in implementation or a productive basis for improvement. Some of the most active inventors in this area are Australia-based. “Examination” is provided by the peer review of the user community, and incentive by known capability for such technology to be used in deliverables.

- Provide incentives to license patented technology widely so that it can be used and improved. For example, the Brazilian patent office has undertaken measures such that the charges for annuities are much higher for entities that cannot show that the technology is being licensed in-country. In a similar vein, the patent grant may be subject to earlier termination if the patentee cannot show evidence of implementation in-country. Similar measures in Australia could provide disincentives for an overseas patentee to maintain an Australian patent that largely serves the purpose only of excluding Australians from using the technology.

⁸⁸ <http://sourceforge.net/index.php> (the content at this site is continually updated and therefore cannot be guaranteed)

⁸⁹ 'Sharing Your Innovations is Potentially Profitable', Wall Street Journal, Europe, March 24, 2005 (<http://www.bios.net/daisy/bios/532>)

Chapter 4. Options for delivering improved IP information: Conclusions

Recommendations from Chapter 1

Important patent search features

The following search engine features have been identified as being of varying availability from the multiple search service sites investigated, but selected as important or useful for users of Australian patent data such as examiners and public technology searchers:

- full text search with support for searching within fields, e.g. entire document, title, abstract, inventors, assignees, claims;
- an extensive Boolean query language (including proximity operators) for expert users;
- a field-based search interface for less expert users, with the ability to convert the current search into the Boolean query language for finer control;
- wildcard characters in search terms (the basic function is to match any number of trailing characters, but some systems allow wildcards at the start or middle of a search term and some provide different wildcards for a) 0-1 characters, b) 1 character and c) any number of characters);
- ability to restrict search by ECLA and/or IPC classifications can be important for examiners, but should allow the use of multiple classification codes simultaneously;
- ability to restrict search by date ranges.

The biggest difference between providers, and usually the biggest deficiency area, is the output, *i.e.* the results list. Capabilities to reformat and re-sort the list and to choose fields for display can greatly enhance finding the closest prior art where there are many search results. Ranking of results by various means, usually by date or patent number, is useful (date ranking particularly for priority date-based searches), ideally supplemented criteria such as IPC classification or nearness of search terms to each other in the text. Surprisingly few sites allow user-configurability of relevance ranking⁹⁰. Unfortunately, the lack of transparency and inflexibility of Google is being increasingly mirrored by many intellectual property data providers, flexibility being sacrificed for seeming simplicity. The providers that show the best balance in this area are Delphion, WIPS, and CAMBIA's Patent Lens.

Special Recommendations for IP Australia

IP Australia's website acknowledges that there are issues with data completeness and quality and a need to consult several distinct databases to conduct basic searches.⁹¹ It is further necessary to go to INPADOC for much of the status information, and we found many issues with accuracy and ongoing updates in the INPADOC data.

⁹⁰ CAMBIA and WIPO have incorporated nearness of search terms as an optional tool for the expert user, which can provide some of the advantage and less of the disadvantage of latent semantic indexing (see Ch.2 recommendations); CAMBIA is beta-testing additional user-configurable relevance ranking options for release before the end of the year

⁹¹ <http://pericles.ipaustralia.gov.au/ols/searching/content/olsPatents.jsp>

However, the mere fact that data are distributed over several systems during periods of redevelopment need not in itself impact negatively on public use of that information. For example, the Australian Trade Mark Online Search System (ATMOSS) has enhanced public access, while making use of data input, and is maintained on a legacy mainframe system as well as a more recent web-based on-file filing system.

From the perspective of Australian innovators, the most immediate needs are:

- A single integrated interface that allows "one stop" authoritative searching of Australian patent documents, as well as legal status and file wrappers
- Full text searching of Australian patents⁹²
 - full text search with support for searching within fields e.g. entire document, title, abstract, inventors, assignees, claims;
 - an extensive Boolean query language (including proximity operators) for expert users;
 - A field based search interface for less expert users, with the ability to convert the current search into the Boolean query language for finer control.

It would be of further interest to users of Australian patent data to be able to search the databases of other important jurisdictions simultaneously, e.g. for prior art, ideally simultaneously with searching the Australian data. CAMBIA currently obtains and marks up the USPTO, EPO and PCT data so that it can be served up and searched in a common format, and provides INPADOC data to expand the search results. Additionally of interest would be to include databases of important trading partners in which extensive patenting is occurring, such as China and Japan.

The following facilities would be further desirable for IP Australia stakeholders, and can be developed within the above framework:

- Accommodate new structured query language approaches to support user-designed queries that integrate content searches with jurisdiction-specific status information such as in-force duration and maintenance.
- Thesaurus and dictionary assistance in search formulations, to associate professional phrase and word synonyms, ranked for their closeness.
- APIs into specialist citation databases such as Medline, and into databases for chemical and biological sequence searching.

The following types of searches that could be relevant to IP Australia needs are not currently provided by any of the patent information providers, although the rate of development by which addition of these search types could be envisioned is rapid:

- Allowing searches to be restricted by in-force status (issued, expired, abandoned, etc.) is not currently supported by any patent information provider we tested, but could be provided through appropriate informatic use combining INPADOC, metadata marking, and application of business rules.
- Incorporation of traditional knowledge databases, to help avert IP protection errors similar to those experienced in some jurisdictions where applications have

⁹² CAMBIA is currently able to provide all the above using its proprietary Dekko full text search engine, and indeed does already provide this for a particular date range of Australian patent data denoted with life sciences IPC codes that had been available for OCR. With the ongoing expansion in scope of all CAMBIA's databases (USPTO, EPO, PCT) from life sciences IPC codes only to all fields, if digital (e.g. TIFF) images of sufficient quality for OCR were provided to CAMBIA the Australian data could rapidly be integrated into CAMBIA's full text search services or provided separately in full.

been granted for traditional plant cultivars, medical treatment methods and the like.

Recommendations from Chapter 2

This section draws conclusions concerning patent searching from our research into developments in Information Retrieval (IR) and our tests of various patent information providers.

The inverted index is the basic Information Retrieval (IR) technique used to provide high performance in searching large data sets. This is linearly scalable technology, meaning that doubling the number of computers available for searching doubles the number of users that can be handled with the same response time.

Latent Semantic Analysis (e.g. PatentCafé's "concept" search) has not proven to be as effective in search precision as conventional IR techniques as yet. It provides unsatisfactory precision for the types of searches that a patent examiner typically carries out, but as the various algorithms become improved it could be useful to the public technology searcher.

Any IR technology chosen for patents should be capable of handling world languages. This allows coverage of PCT applications in, for instance, Japanese. Multilingual capability includes the use of Unicode and support for Chinese, Japanese and Korean (CJK) searching.

The patent IR systems surveyed do not currently use Natural Language Processing (NLP), although NLP is clearly an emerging trend promising significant improvements in search quality i.e. achieving a closer match between the user's intent and what the search engine finds. NLP will be important in English searching and even more so for CJK and Cross Language Information Retrieval (CLIR). CLIR has already been shown to be useful in discovering prior art in foreign languages.

Visual image searching was also investigated, for use in design patent searching, normal trademarks, and shape marks, The software available has significant technical limitations, and the greatest technical progress has tended to focus on security-related images such as faces and fingerprints rather than IP industry uses. It would be useful to check this area of software development periodically, however, because although IP Australia trademark and design patent examiners indicated it would be of limited use for searching IP data (because most searching is done using classifications), searching of the non-IP literature could eventually be facilitated by such tools.

Special Recommendations for IP Australia

Our analysis of the currently available IP information systems and emerging trends has highlighted a number of key areas we recommend be addressed to increase the accessibility and quality of information to support innovation.

- Use of a search engine based on an inverted index, such as Dekko or Lucene. The inverted index is the basic Information Retrieval (IR) technique used to provide high performance in searching large data sets. This is linearly scalable technology, meaning that doubling the number of computers available for searching doubles the number of users that can be handled with the same response time.
- Use of a search engine that supports world languages (Unicode and Chinese, Japanese, Korean support are particularly important). This will allow immediate coverage of PCT applications in, for instance, Japanese, as well as expansion for the future.
- Monitoring use of Natural Language Processing (NLP) in the future. The patent IR systems surveyed do not currently use Natural Language Processing (NLP), although NLP is clearly an emerging trend promising significant improvements in

search quality, *i.e.* achieving a closer match between the user's intent and what the search engine finds. NLP will be important in English searching and even more so for Cross Language Information Retrieval (CLIR). Bayesian approaches may, however, already be integrable via a selective interface with inverted index full text search engines.⁹³

One way to address data integrity issues is through the adoption of applicable international standards. For example for the dissemination of Australian patents that have been OCR'd to address point 3 above, the revised "wo-published-application" dtd format, outlined in WIPO document C. PCT 1037 -76⁹⁴ may have some relevance. The recent decision by IP Australia, to accept, from 18 July 2005, limited filing of international applications in electronic format in its capacity as a PCT receiving office also offers an opportunity for a focus on data quality based on international standards.

Given the percentage of Australian applications lodged through the PCT system, opportunities exist to make use of original PCT scans, supplemented with electronic capture of any amendments made upon entry to national phase, rather than simply rescanning amended printouts of the PCT originals.

It should be kept in mind that the essence of providing a good search capability for patent data is not necessarily the search engine itself – it is the ability to understand the data formats and structures that are present in collections of patent documents. That knowledge has been designed and configured into **Dekko**⁹⁵.

The presence of OCR data (WIPO PCT documents, EPO pre-2000 documents) causes other potential problems. The OCR process is never perfect, due to limitations in both the quality of the original images from which the text is extracted and the imperfect nature of OCR algorithms⁹⁶.

The Dekko system incorporates several filtering and compression techniques for this before compiling an inverted index. The **minimal perfect hashing** technique results in extremely fast word retrieval; this, combined with ability to store the bulk of the compressed index in memory makes **dekko** into a very fast Information Retrieval (IR) system.

Searching the prior art for chemical structures and DNA sequences, as detailed in Chapter 2, presents unique challenges. Because of the way that such compositions of matter tend to be claimed, even more so than with word searches there are problems with ambiguous search terms, which often lead to search results that are not relevant or omission of relevant documents from the search results list.

Tools are becoming available and better developed for Bayesian approaches to literature searching, which though not practical for full-text searching, may help overcome some of the challenges to efficient searches using DNA and amino acid sequences and chemical names by associating the query term with function, species, genes, proteins or chemical diagrams.

⁹³ This approach is being explored at CAMBIA for particular types of information (see Ch. 2 recommendations)

⁹⁴ http://www.wipo.org/pct/edi/en/wo_publication_information/documents/pdf/circular_wo_changes_en.pdf, 8 July 2005

⁹⁵ The Dekko system has been successfully used in the CAMBIA BiOS Patent Lens for several years. We currently index WIPO (PCT), EPO, and US patent data. For a time, we provided full text search capabilities for Australian life sciences patent data as it arose, and although availability of Australian data to CAMBIA has lapsed, **Dekko** is in fact ideally suited to working with it.

⁹⁶ CAMBIA has tested OCR engines in a separate project outside the scope of this study. Despite the imperfections seen in all OCR engines, it was possible to identify two OCR engines that can handle typical patent data in English and related languages and in CJK (Chinese, Japanese and Korean) characters with 99% accuracy.

While currently such searches are primarily done using subscription or fee-based databases, in-house search tools can be installed for confidential searches of pre-publication chemical structures and sequences, for example by an examiner of an Australian patent that did not enter via the PCT. The best solution for IP Australia may be the use of open source software modules, including Bayesian approaches and BLAST tools, that can be integrated with a search engine such as CAMBIA's or one that IP Australia could develop for both these areas of searching, which are important in the Australian patenting landscape.

Recommendations from Chapter 3

The key market jurisdictions mentioned in Chapter 3, which include Japan, Korea and China, and soon including Indonesia, India, countries of West Asia, and Latin America, are increasingly focusing on development, use and enforcement of their patent systems, and these need to be transparent to all who wish to do business in these nations.

Much technology that already exists can be found in the patent literature and is available to exploit where no monopoly grant is in force. Thus, provision of the patent literature of other jurisdictions, rapid publication of patent applications, and availability of status information can be very valuable as a stimulus to innovation.

Increased rigour of examinations and other measures can bring about increased respect for the patent system and more likelihood that it serves the purpose of providing an incentive for enabling disclosures of technology useful to the public.

Special Recommendations for IP Australia

We suggest that the provision of a search service offering not only Australian patents but the harmonised datasets of other important jurisdictions will support Australian technology development and deliverability. A major opportunity for Australian innovators would be created by providing support to render the databases of Japan, Korea, and China (and eventually other emerging economies) accessible and interlinked for concomitant searching with Australian data.

Similar interests would be served by incorporating APIs to collections of other potential sources of prior art, databases of chemical formulae and other kinds, which can perhaps be carried out via APIs in view of the recommendations of Chapters 1 and 2.

There is an opportunity for Australian innovations to be accepted in these nations with increased reciprocal appreciation for the quality of the Australian patent information system. IP Australia should act on a commitment to increased rigour of the patent grant, so that patent monopolies are granted for truly novel and inventive work but not for innovation that is obvious in view of the prior art. The most important step in this is to conduct highly effective prior art searches of the patent data of Australia and other jurisdictions, non-patent literature and traditional knowledge, so that all prior art that would be read on by the claims of patent applications or invalidate them is found. Blows to investor confidence upon the discovery of invalidating prior art that should have been made known before the patent grant do not contribute productively to the economy.

The overall goal of the recommendations of Chapter 3 is to see the steps forward in the Australian innovation system that will provide best support to the domestic economy and standing in the region. In the context of global patent reform, such recommendations are well placed to put IP Australia in a position of increased leadership.

Final conclusions

All three chapters support a recommendation and provide practical advice in a single major direction: that the most urgent and important step is implementation of a Boolean accessible,

scalable search engine capable of doing full-text searches of Australian patent data. Ideally this would be coupled with unified searching as soon as possible of Australian data together with the largest patent data sets, namely the US, EPO and PCT data.

Should IP Australia require strict confidentiality of search parameters for pre-publication prior art, two discrete search interfaces and servers could be used using largely identical search engine technology and software. These could be maintained in house, or through an out-sourced facility.

Status data should be incorporated as soon as possible in the portions or counterparts of the Australian database available to the public, so as not to disadvantage Australian inventors and investors. Support for searches that assist in understanding breadth and interpretation of claims can and should be developed, even in the challenging areas of chemistry and biological sequence data. While the PCT data already requires a search engine and OCR that can handle Japanese language, eventual incorporation of at least the Chinese and Korean datasets has similar requirements and should be planned.