Patent Images - a Glass-encased Tool

Opening the case

Mihai Lupu
Vienna University of Technology
Vienna, Austria
lupu@ifs.tuwien.ac.at

Roland Mörzinger
JOANNEUM RESEARCH Forschungsgesellschaft
Graz, Austria
roland.moerzinger@joanneum.at

Tobias Schleser
m2n consulting and development GmBH
Vienna, Austria
schleser@m2n.at

René Schuster
JOANNEUM RESEARCH Forschungsgesellschaft
Graz, Austria
rene.schuster@joanneum.at

Florina Piroi
Vienna University of Technology
Vienna, Austria
piroi@ifs.tuwien.ac.at

Allan Hanbury
Vienna University of Technology
Vienna, Austria
hanbury@ifs.tuwien.ac.at

ABSTRACT

The paper discusses the problem of patent image retrieval. It describes the issues faced when extracting semantic data of images in patents, as well as an integration framework between the data thus extracted and semantic information extracted from text. Combining the two sources of knowledge is on the wish list of many patent information users, as current systems either search only the textual data, or have extremely limited image processing functionality. In practice in the patent domain, depictions of the product or method are often vital to the understanding of the invention. Yet they are almost completely unsearchable. They are tools enclosed in a glass case, at which we can look, but of which we cannot really make use. The IMPEx Project (Image Mining for Patent Exploration) cracks open this case with a new focus on processing this particular type of images. This paper presents the motivations, status and aims of the project.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; J.1 [**Administrative Data Processing**]: [Business,Law]; I.4.9 [**Image Processing and Computer Vision**]: Applications

General Terms

Algorithms,Documentation,Design

Keywords

Patents, Patent images, Semantic data, Search

i-Know'12 September 05 - 07 2012, Graz, Austria

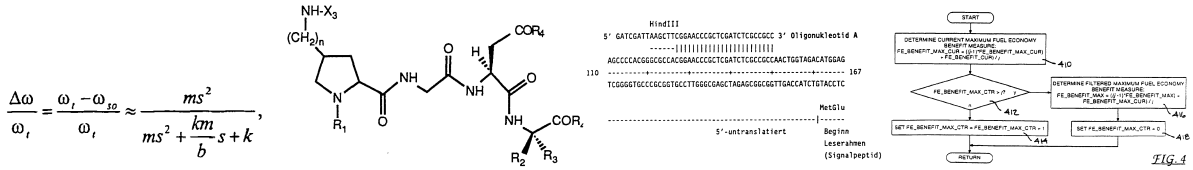
1. INTRODUCTION

The global patent collection is a trace of human technological development of the past 500 years. The patent system provides a common framework, in which inventors describe in great deal their innovations and make them public, in a national central repository. It is a source of not only market intelligence, but also of cultural heritage. Many of the patents have attached depictions of the inventions, most notably in the chemical, mechanical and electrical engineering fields. These depictions are to this day paramount to the understanding of the invention, and yet almost completely unsearchable. They are tools enclosed in a glass case, at which we can look, but of which we cannot really make use in an automated way.

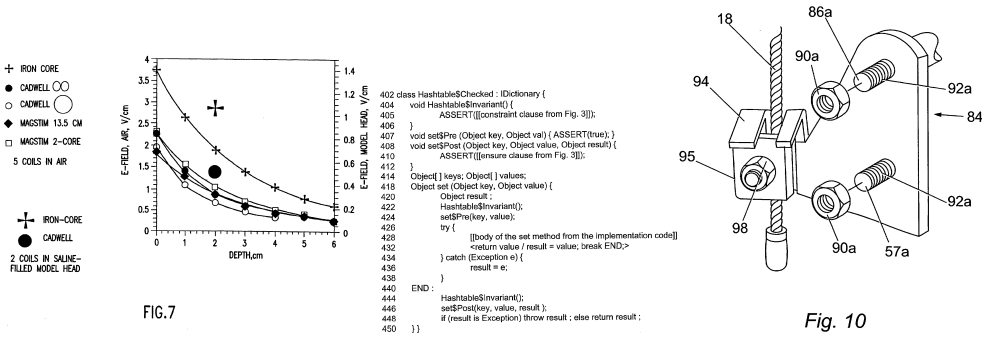
Currently, searchers relying on information in images have to manually go through hundreds or thousands of candidate images, selected only on meta-data of the patents to which they belong, or on text queries. This is prone to errors and a considerable burden on the searcher. To get an idea of the presence of images in patent documents, we looked at 16'002'652 patent documents. They represent the subset of the Alexandria Patent Data Warehouse[1] [5], which have either a description or claims. Among these 28% contained at least one image. The average number of images was 9.38 (with a standard deviation of 17.36), while the maximum number of images a document referenced was 8433.

In addition to the images mentioned above, explicitly identified as figures in a text document, we should also consider the fact that some national patent offices only provide patent documents as scanned PDFs. Such a file must be first correctly processed in order to extract the text and the images for further specific indexing. In fact, when we think of patent images we should not forget that a significant amount of textual information is captured in bitmap format. Important properties of a new method or entity are often described in tables, flowcharts, DNA sequences, or mathematical formulas. Even special characters are sometimes represented as an image in a patent. Figure 1 shows the 9 types of images most frequently available in a patent document.

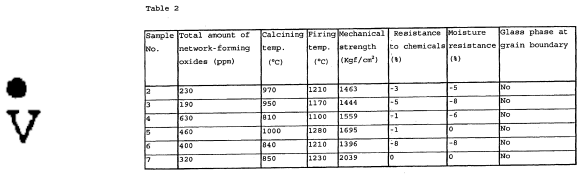
[1]kindly made available to us by Fairview Research, and currently available as IFI Claims®Global Patent Database `http://www.ificlaims.com`



(a) Mathematical (b) Chemical Formula (c) Genetic Sequence (d) Flow-chart



(e) Graph (f) Program Listing (g) Abstract Drawing



(h) Symbol (i) Table

Figure 1: Examples of types of figures in patents

1.1 Related Work

Very little work has been done on directly using information in images for patent retrieval. One of the first systems was proposed by Huet et al. [9]. PATSEEK [20] is an image-based search system for the patent database of the USPTO. More recently, PATMEDIA[2] [21, 18] was shown to perform well on patent image retrieval. Unfortunately, these papers all do evaluation on datasets of at most a few thousand images, not representative of the huge numbers of images found in patent collections. Even though publications on image retrieval in the patent domain are few, much work has been done on technical drawing retrieval that could have application in the patent domain, as pointed out in a recent survey of patent image retrieval [7].

1.2 Project Objectives

The Image Mining for Patent Exploration (IMPEx) Project[3] has as its core objective the extraction of semantic information from patent images. Such information may take different forms:

- **References** (e.g. "Fig.1") allow us to create links between the image and the text describing that image.
- **Sub-figures**: Patents contain several document pages. Making the distinction between pure-text pages, image-text-mixed and pure-image pages is necessary for further steps. Image pages often contains several figures, which need to be segmented.
- **Types of images**: As seen in Figure 1, there are 9 types of images, with considerably different feature spaces.
- **Sub-parts**: Particularly in mechanical engineering, figures depict complicated objects (e.g. Figure 1g). The inventive step of the patent may concern only one of these.

In addition to supporting the process of patent examination, the technology developed in IMPEx can also be applied, for instance, in the technical drawing domain in general, where searching for parts of drawings is becoming increasingly important. Modern CAD systems make production of technical drawings easier and quicker, and the number of technical drawings is rapidly increasing. Engineers may not always remember the components already designed, potentially by colleagues, and therefore require computer support in searching.

1.3 Motivation

The need for a project in this direction is obvious from Figure 1: images in patents have no colour information (not even grayscale levels), and yet they are semantically very rich. Their mining for the purpose of retrieval has to take into account specific features of each type. In order to do

[2] http://mklab-services.iti.gr/patmedia/
[3] www.joanneum.at/?id=3922

this, the processing is done in three stages, which we proceed to describe in the following three Sections:

1. Identify images in documents and classify their type (Section 2);
2. Mine the image for semantic information, specific to the identified type (Section 3);
3. Merge text and image in a semantic patent search system (Section 4).

In addition to this, Section 5 briefly describes current systematic evaluation efforts, and Section 6 summarizes the work and looks to the future.

2. PATENT IMAGE PROCESSING

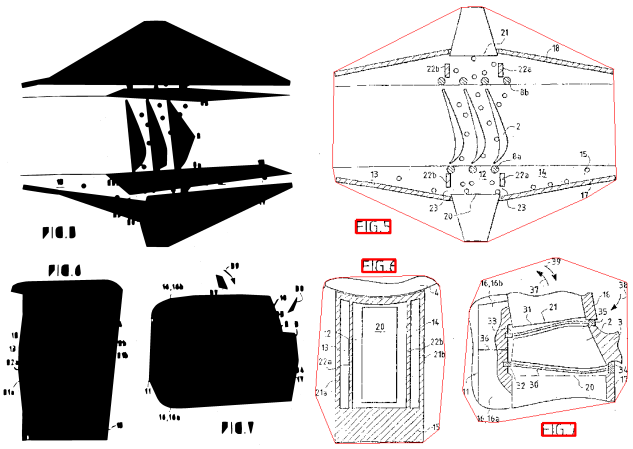
In patent pages with pure-image content, often multiple figures and their corresponding references are placed near each other, see Figure 2. These pages are usually separated from the textual descriptions. When a user is reading a section with a textual description referring to a specific figure, he has to go back to check the drawing section and manually search for the corresponding figure. For automatically linking descriptions with figures, we propose the following method.

First, pure-image document pages (as opposed to pure-text or image-text-mixed pages) are identified. For that purpose, connected component analysis [8] is applied on the black and white patent pages and all connected components whose area (the area of a rectangular bounding box area encompassing the connected components) is below the threshold of 20 pixels are removed. An image file is classified as pure-image if two criteria are met.

1. The largest bounding box of all connected component must cover less than 60% of the toal image. This excludes pages where tables and search reports cover almost the full page.
2. The ratio of the total area of all bounding boxes smaller than twice the median area to the total area of all bounding boxes must be less than 25%. This excludes pure-text and image-text-mixed content which features many small characters.

These criteria were empirically selected such that they produced a correct classification rate of 98% on 400 ground truth images of pure-image on the on the one side and pure-text and mixed content on the other side.

Second, the image is segmented into multiple meaningful parts, i.e. figures and their references. Similar to the work described in [11], we use optical character recognition technology and a regular expression to detect the location of one or more references, i.e. the region in the image containing the caption of a figure. After pre-processing for removing small information that is not useful (single isolated pixels), we apply morphological filtering (close operator for filling up holes) on the convex hull of the individual connected components. All components with overlapping convex hulls are merged. The distances between all references and the other connected components are computed. Based on minimum distances between them, the components are iteratively assigned to the references. An example for the individual connected components after cleaning and closing and the final segmentation result is shown in Figure 2.



(a) Closed Conn. Comp. (b) Segmentation Result

Figure 2: Example segmentation of document pages with multiple figures.

Third, for each of the segmented figures the type is automatically classified. SVM classifiers are applied on a variety of content-based features for black and white images [14], which range from texture descriptors over edge features to statistics obtained from optical character recognition (OCR). Results on a challenging dataset with 9 different image types show a classification accuracy of around 70%. These classification results can also used to validate text-based classification/annotation of the drawing type.

3. IMAGE INFORMATION MINING

Having obtained individual figures from the segmentation process, and knowing what kind of information they represent, we now move on to extracting specific features.

Mining images for semantic information is mostly realized by applying OCR technology and interpreting the recognized text. Typed and handwritten references are extracted in order to establish automatic interlinking between the patent's texts and drawing parts. Apart from this general technique to extract information from patent images, specific approaches focusing on certain figure types can be considered. For example, if you know that a figure is of type 'graph', the labels of the axis and the legend need to be determined. In sections with program code it might be interesting to distinguish between constants, variables and programming language constructs. The presentation of tables which are often rotated by 90° can be automatically fitted so that the text is easily readable from left to right. The following two sections go into more details on two such image types.

3.1 Flowchart Analysis

Among the different types of images present in a patent document, flowcharts are interesting for two reasons: First, they often represent the innovative step in the patent. A previously existing process is changed, or a totally new process is depicted, and this information is visible to the human examiner clearly in the flowchart. Second, they are approachable with improvements on today's techniques.

As seen in Figure 1(d), a flowchart is a graph. It consists

of a set of nodes and edges. To semantically process the data therein, the flowchart image analysis tool must, in a first step, be able to identify: 1. the number of nodes; 2. the type of each node (e.g. rectangle, diamond, oval, etc); 3. the text (if any) within each node, 4. the edges between the nodes and, 5. the type of the edge (e.g. continuous, dotted, etc.). Additionally, and specific to the patent domain, flowcharts images often contain annotations on nodes (e.g. 410, 412, 414 in Figure 1(d)). These should also be identified for creating the link between the image and the text in a subsequent system integration (see Section 4).

3.2 Chemical Search

Chemical search is not the focus of the IMPEx project, but it deserves mentioning here. Chemical image processing is special because it has received a lot of attention from the research community. It is extremely important for the patent domain, as the pharmaceutical industry is one that relies extensively on patents to protect its intellectual property. Table 1 shows the top 5 IPC[4] classes in the Alexandria collection at our disposal. As we can see two of the top 5 IPC classes rely on chemical formulations. The importance of chemical images and their particular nature are the reasons for which there has been a substantial amount of research in this area. The field of *chemoinformatics* is the one "concerned with the application of computational methods to tackle chemical problems, with particular emphasis on the manipulation of chemical structure information" [10]. Recent work has also been devoted to images found in chemical patents [17, 19, 22].

The main difference between the "standard" chemo informatics and the work related to patents is the frequent use in patent documents of so-called Markush structures [2]. Such structures are a form of wild-cards for chemical formulas, allowing the specification of very wide range of chemical formulas (a potentially infinite number) with one single depiction. However, unlike the regular expressions we are familiar with from text processing, the interpretation of Markush structures is still difficult to do fully automatically [3].

4. SEMANTIC SEARCH IN TEXT AND IMAGES

In IMPEx, semantic information and extracted meta-data of images is merged with meta-data and information gained from textual information of patents in the *m2n Knowledge Discovery Suite for Patent Exploration*, a system built on top of the *m2n Intelligence Management Framework for knowledge discovery* (m2n-kd). The system holds its data model as graph which is based on an ontology modelled in RDFS and OWL. This includes the domain data model (patents and images, see Section 4.1) as well as the application logic and user interface configuration.

The framework provides flexible and ready to use modules for knowledge discovery. Modules bundled for patent exploration include knowledge discovery methods which serve to extract information from unstructured data (text), enhance semantic data models and provide tools for linguistic text analysis, text based search, continuous surveillance of data sources using machine learning methods, filtering facets and topic clouds, document and meta-data preview, graphical analysis, workflow support and collaboration additions

[4]International Patent Classification scheme

Table 1: Top 5 most frequently used IPC classification codes in the Alexandria collection

	IPC Class	Title
1	H01	Electricity: Basic electric elements
2	A61	Human necessities: Medical or veterinary Science; Hygiene
3	H04	Electricity: Electric communication technique
4	G01	Physiscs: Measuring, Testing
5	C07	Chemistry;Metallurgy: Organic Chemistry

(sharing of search context, citations, boilerplates, etc.) to accomplish the wide range of search scenarios of the patent information user [1].

In IMPEx, this functionality is enhanced by using information extracted from figures for enhanced search and exploration features. Extracted meta-data is linked to the populated models representing patents in m2n-kd and made available to search and analysis services throughout the application. A brief description of the modelled ontology used as data model for patents and analyzed images is given in the following section.

4.1 IMPEx Ontology

The IMPEx ontology was built to represent data defining a patent (based on the MAREC[5] format) with additional focus on images in these patents. Figure 3 shows the image related part of the ontology, modelled in m2n-kd. The model includes the physical image ("Image File") including an Image which contains an arbitrary number of image parts ("Subpart"), defined by an image region. The image itself is subclassed into more specific image types ("Diagram", "Graph", "Drawing",...) and has several properties concerning the formal meta-data like resolution or dimension.

Images and subparts refer to annotations (extracted referring text in patent images like "Fig 1b") which constitute the link to references found in a patent's text.

Where appropriate and reasonable, the IMPEx schema makes use of other published schemata to include domain knowledge and enhance interoperability, for example:

- The schema for Units and Dimensions for Diagram Axis is taken from QUDT[6], a collection of ontologies which define base classes, properties, and instances for modelling physical quantities, units of measure, and their dimensions in various measurement systems with focus on science and engineering.
- The schema for image-file metadata is taken from the Nepomuk File Ontology[7] which provides a vocabulary for information extracted from files for example.

4.2 Data Exchange Using the Ontology

The above described ontology is modelled as RDF schema and defines a flexible and ready to use data exchange format.

[5]`http://www.ir-facility.org/prototypes/marec`
[6]QUDT - Quantities, Units, Dimensions and Data Types in OWL and XML, `http://qudt.org`
[7]NFO - Nepomuk File Ontology, `http://www.semanticdesktop.org/ontologies/nfo/`

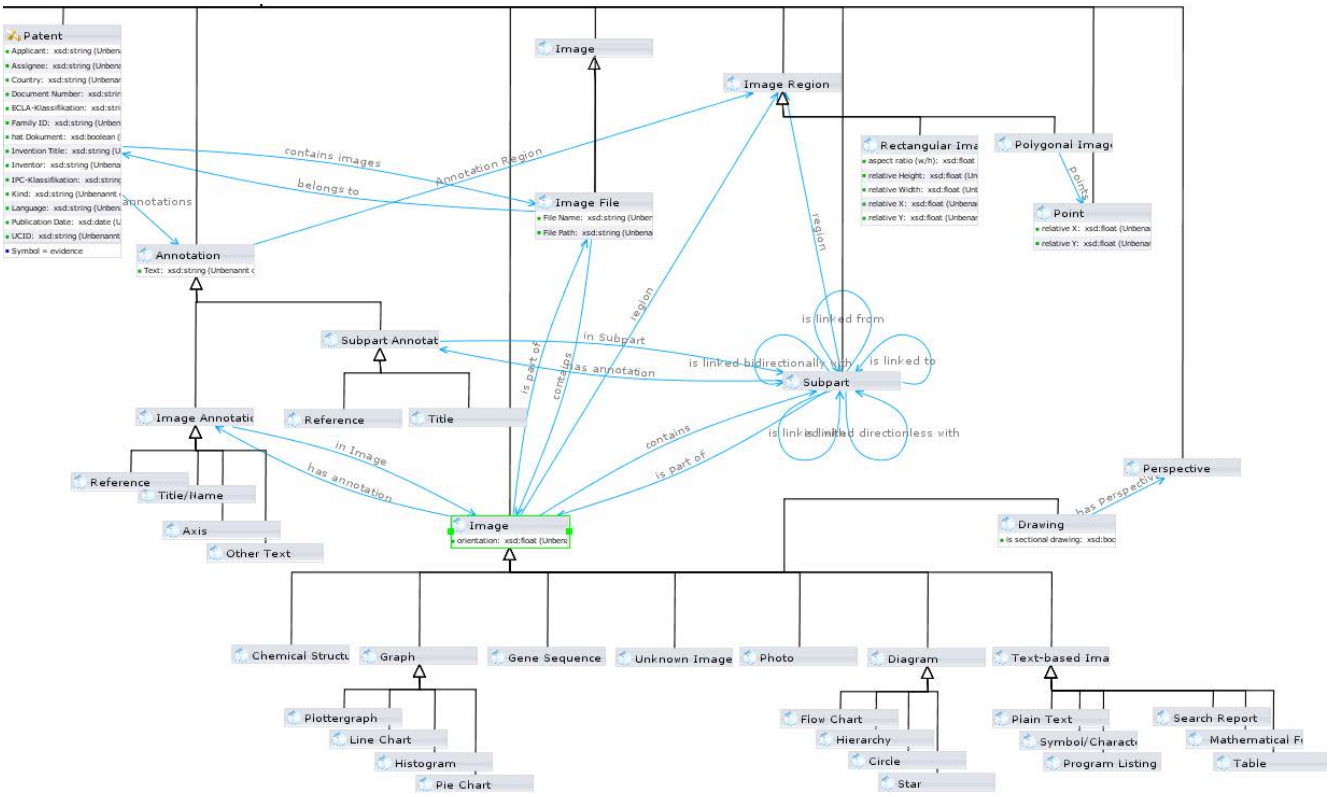


Figure 3: Part of the IMPEx ontology modelled in m2n-kd

For example, models representing patents are first populated by m2n-kd through textual knowledge discovery means and are then transferred to the image processing engine. They are returned with additional information extracted from image files, thus further populating the ontology.

The following listing shows the (incomplete) model of an example patent's image annotations in Notation3 (N3) syntax[8], based on the schema definition described in Section 4.1. The image is a hypothetical one, containing in its top half a drawing of an engine, in the lower left hand quarter a line chart with performance measurements for said engine (x-axis: time, y-axis: power) and in the lower right hand side a flow chart with some kind of process model. The listing shows only the "top half part" and the beginning of a subpart definition.

```
ex:ExamplePatent
  :patentAnnotation ex:I11DrawingTitle,
  ex:I11DrawingReference, ... .

ex:ImageFile1
  :image ex:Image_1_1, ex:Image_1_2, ex:Image_1_3 ;
  nfo:width "1500"^^xsd:int ;          #px
  nfo:height "6000"^^xsd:int ;         #px
  nfo:colorDepth "24"^^xsd:int ;        #bits/pixel
  nfo:horizontalResolution "300"^^xsd:int ; #dpi
  nfo:verticalResolution "300"^^xsd:int ;   #dpi
  nfo:aspectRatio "0.25"^^xsd:float .       #w/h
```

[8]Notation3 (also known as N3) is an assertion and logic language which is a superset of RDF, `http://www.w3.org/TeamSubmission/n3/`

```
ex:Image_1_1  #top half of ImageFile1
  a :Drawing ;
  :imageFile ex:ImageFile1 ;
  :imageRegion
 [
    a :RectangularImageRegion ;
    :rirX "0.0"^^xsd:float;
    :rirY "0.0"^^xsd:float;
    :rirWidth "1.0"^^xsd:float ;
    :rirHeight "0.5"^^xsd:float ;
    :rirAspectRatio "0.5"^^xsd:float ;
         # aspect ratio (absolute measurements)
  ] ;
  :drawingPerspective :FrontPerspective ;
  :drawingIsSectional "true"^^xsd:boolean ;
  :imageAnnotation ex:I11DrawingTitle,
                   ex:I11DrawingReference ;
  :contains ex:I11Piston, ex:I11Valve,
            ex:I11SparkPlug .

ex:I11DrawingTitle
  a :ImageTitle ;
  :annotatedImage ex:Image_1_1 ;
  :annotationText "Steam Engine"^^xsd:string .

ex:I11DrawingReference
  a :ImageReference ;
  :annotatedImage ex:Image_1_1 ;
  :annotationText "Fig. 1a"^^xsd:string .
```

```
ex:I11Piston
  a :Subpart ;
  ...
  ...
```

Note, that resource URIs are given meaningful names for readability in this example (e.g. `ex:11DrawingTitle`). However, this is not done nor needed in reality as semantics of resources are defined by the ontology and resources classified by their type ("a") property.

The ontology is populated in this way with image processing means and gets connected to the application graph in the m2n Knowledge Discovery Suite for Patent Exploration. Thus, the data is made available to search and analysis widgets seamlessly.

4.3 The Integrated System

We briefly present the recent achievements for the integrated system. Besides standard m2n-kd functionality as described above, emphasis is laid on references as finding objects (see Section 4.3.1) and the semantic patent viewer (see Section 4.3.2.

4.3.1 *References as Finding Objects*

The m2n-kd suite allows a flexible definition of so-called *indexing objects*. The structure is defined by selecting a subset of the schema (by cutting out a part of the graph) and gets applied to all resources in the model that match this subset. This allows to not only index on a document basis but to define arbitrary objects, which can later be searched for and used for data visualization.

In IMPEx, a particular class of indexing objects, *references*, is created, which represents references in patent text (for example *"Figure 3b"*, *"see tab. 2"* and so on). Indexing objects of this class describe not only the reference itself (in general, multiple annotations match a particular reference), but also the context of all reference mentions for a particular reference. For example, a particular reference may be mentioned in four different paragraphs and be referred to as "FIG 3b" once and "Figure 3B" in all other cases. The indexing object definition defines all paragraphs which describe the reference as its textual "content", along with additional meta-data such as a normalized label ("fig. 3b") and a link to the patent. Once indexed, the user can then find these reference objects by means of full-text or semantic search; the before mentioned "content" property serve as data basis.

As described in the previous section, reference objects are further enhanced in IMPEx. In general, these references refer to image BLOBs[9] (i.e. *"FIG 3b"* refers to an image which shows figure 3b). This information is added to reference objects and used to present the particular image to the user when exploring a reference result list or a patent.

Figure 4 shows a screenshot of m2n Knowledge Discovery Suite for Patent Exploration. The result list contains patents and reference objects as described before. Note that these references have normalized reference labels (for example, "fig. 1") and contain information about the referring patent, the image BLOB and the context of references. Actions on these reference objects (not shown in screenshot) include opening a structured view of the meta data of the object (including a larger view of image BLOBs), opening the referring patent in a semantic viewer, searching for similar references and bookmarking the item for later use.

[9]Binary Large Objects

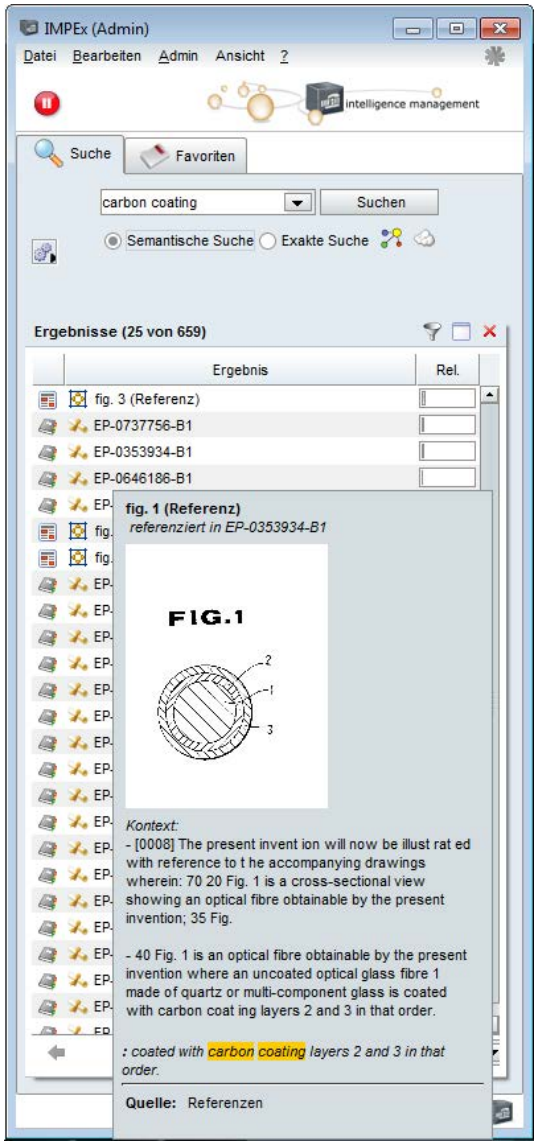


Figure 4: Result list showing patents, reference objects and a tooltip on reference object "fig. 1"

4.3.2 *Semantic Patent Viewer*

The Semantic Patent Viewer shown in Figure 5 brings everything together and is a core component of m2n Knowledge Discovery Suite for Patent Exploration. It shows PDF, Office and HTML Files and supports search hit highlighting and browsing. The key feature, however, is the listing and highlighting of extracted concepts, such as the before mentioned references.

Extracted references in patent text are listed in a concept tree on the left hand side, each including a list of occurrences in the document. On the right hand side, the document (patent) itself is shown, search hits are highlighted in orange and references in blue in this configuration. Reference objects can be clicked to show their properties in a tooltip preview, such as the image BLOB and further meta data.

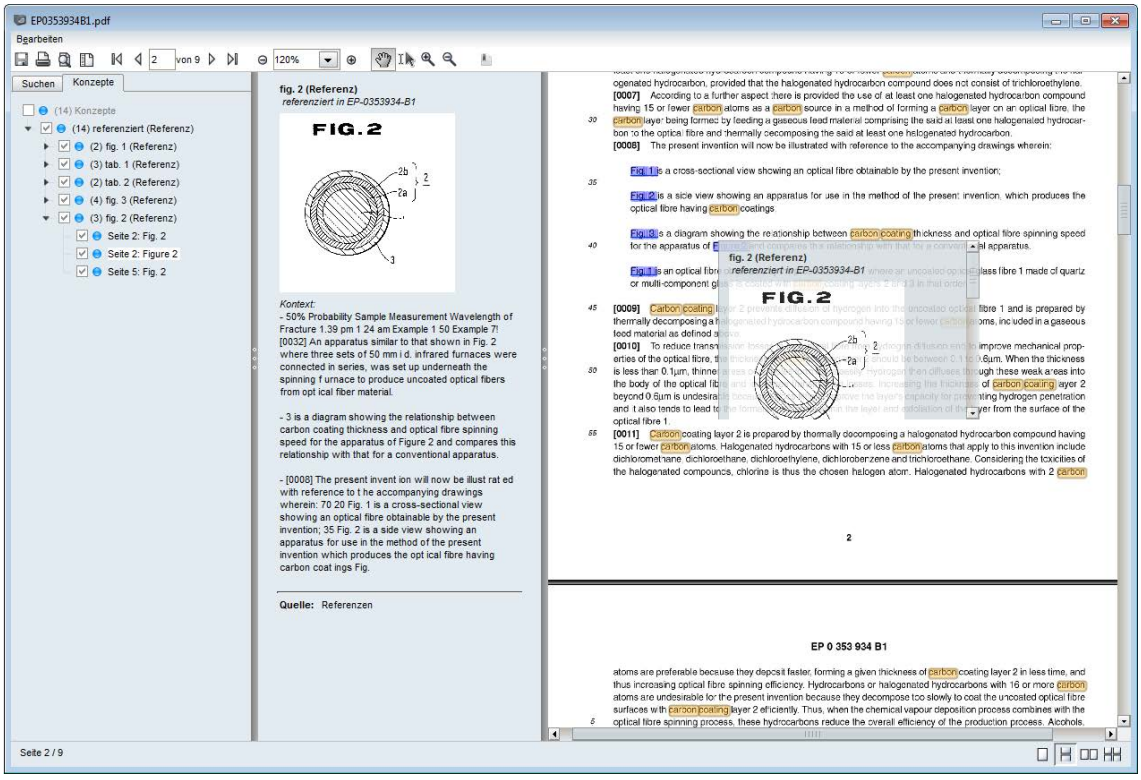


Figure 5: Semantic patent viewer with concept tree and highlighted search hits (orange) and references (blue) which can be clicked to browse reference properties such as the image BLOB

5. BENCHMARKING EFFORTS

The IMPEx project is committed to the use of publicly available benchmarks to test its image classification and processing components. This section describes the involvement of the project partners in standardized evaluation campaigns.

Benchmarks are commonly made available via standardized evaluation campaigns such as ImageCLEF[10] or TRECVID[11]. In general, benchmarks allow systems to be compared objectively and improvements to be monitored across different versions and parameter changes. The campaigns organized up to 2010 have however little utility to the task at hand. Practically all benchmarks consist of colour images, or at most grayscale images such as black-and-white art photos or radiology images. As documented before in this paper, this is not the case in the patent domain. The first benchmark with patent images for the general domain was at CLEF-IP 2011 [15]. The proposed task (the use of both images and text for patent retrieval) received however only one participation. A similar experience was observed in TREC-CHEM 2010 [12], where although images were made available to participants, none of them actually used the image data. We venture to assume that is due to the difficulty of the task. The benchmarks for patent image processing must be much more focused. It comes with the lack of information in the images themselves, which must be compensated by greater assumptions about what they contain. Consequently, a 2011 track on chemical image recognition received wide support from the community [13], and in 2012, we will organize a track on flowcharts recognition.

The experience of the past three years on creating benchmarks for patent image processing tools has thus lead us to the conclusion that not only must the benchmarks be technically approachable by the research community, but also that using a general purpose task (in this case "retrieval") lead the potential participants away. Instead, extracting semantic information from the images and evaluating the correctness of this information, allows us to make progress both on the side of the core research, as well as on the side of working towards the ultimate goal of patent image retrieval.

For each of the types of images listed in Figure 1, we must have a distinct processing method, and a distinct evaluation measure. For those that have already be done (i.e. chemical images), as well as for those that are in the pipeline (i.e. flowcharts) evaluation criteria have been developed. For chemical images, we take advantage of considerable research in the area and accepted standards [13]. For flowcharts, the evaluation measure is based primarily on graph similarity measures [16, 4], with points added for the correct recognition of the text within the cells of the flowchart.

Some of the other types of images are subjected primarily to OCR based techniques (genetic sequences, program listings, symbols, tables). Mathematical formulas fall somewhere in between, and attempts to create benchmarks have been already demonstrated [6]. The elephant in the room is the "Abstract Drawing" class of patent images. As the example in Figure 1 shows, they can be extremely complicated and difficult even for humans to understand. A benchmark for such images is therefore still on the future work list and

[10] http://www.imageclef.org/
[11] http://trecvid.nist.gov/

will presumably be so for the medium, if not the long term.

6. CONCLUSIONS

The IMPEx project is the first coherent attempt to make the semantic information within patent images compatible and searchable in conjunction with semantic data extracted from text. The approach taken involves image segmentation, extraction and classification, followed by a processing method specific to the type of image at hand. For some type of images there is already a considerable amount of work and know-how (e.g. chemical images). For others, it is extremely difficult to see how to overcome the semantic gap within a medium time-frame (e.g. abstract drawings). The lack of traditional image features in this context (e.g. colours, textures) makes it particularly difficult, and we have approached the problem in a pragmatic and user-minded way. Although only in its first half, IMPEx has generated a prototype based on the m2n Knowledge Discovery Framework which brings together text and images for the patent information user. At the same time, the project opens the door to other researchers interested in the area by organising evaluation campaigns and making data available. Future work includes the full processing of flowchart images and the incorporation in the m2n-kd of the semantic similarity search specific to this type of images.

7. ACKNOWLEDGMENTS

This work was supported by the Austrian Research Promotion Agency (FFG) FIT-IT project IMPEx (No. 825846). M. Lupu, F. Piroi and A. Hanbury are also partially supported by the EU Network of Excellence PROMISE (FP7-258191).

8. REFERENCES

[1] D. Alberts, C. B. Yang, D. Fobare-DePonio, K. Koubek, S. Robins, M. Rodgers, E. Simmons, and D. DeMarco. *Current Challenges in Patent Information Retrieval*, chapter 1 : Introduction to Patent Searching - Practical Experience and Requirements for Searching the Patent Space. Springer Verlag, 2011.

[2] J. M. Barnard and G. M. Downs. Use of markush structure techniques to avoid enumeration in diversity analysis of large combinatorial libraries. `http://www.daylight.com/meetings/mug97/Barnard/970227JB.html`, (visited 03/2012) 1997.

[3] J. M. Barnard and P. M. Wright. Towards in-house searching of Markush structures from patents. *World Patent Information*, 31(2), 2009.

[4] D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(4), 2004.

[5] Fairview Research. Alexandria patent data warehouse. http://www.intellogist.com/wiki/Alexandria, 2011.

[6] U. Garain and B. Chaudhuri. A corpus for ocr research on mathematical expressions. *Int. J. Doc. Anal. Recognit.*, 7(4):241–259, Sept. 2005.

[7] A. Hanbury, N. Bhatti, M. Lupu, and R. Mörzinger. Patent image retrieval: A survey. In *Proc. of PaIR*, 2011.

[8] R. M. Haralick and L. G. Shapiro. *Computer and Robot Vision*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition, 1992.

[9] B. Huet, G. Guarascio, N. J. Kern, and B. Mérialdo. Relational skeletons for retrieval in patent drawings. In *ICIP (2)*, pages 737–740, 2001.

[10] A. Leach and V. Gillet. *An Introduction to Chemoinformatics*. Springer, 2007.

[11] L. Li and C. L. Tan. Associating figures with descriptions for patent documents. In *Proc. of DAS*, 2010.

[12] M. Lupu, J. Huang, J. Zhu, and J. Tait. TREC Chemical Information Retrieval - An Evaluation Effort for Chemical IR Systems. *WPI Journal*, 2011.

[13] M. Lupu, Z. Jiashu, J. Huang, H. Gurulingappa, I. Filipov, and J. Tait. Overview of the trec 2011 chemical ir track. In *Proc. of TREC*, 2011.

[14] R. Mörzinger, A. Horti, G. Thallinger, N. Bhatti, and A. Hanbury. Classifying patent images. In *CLEF (Notebook Papers/Labs/Workshop)*, 2011.

[15] F. Piroi, M. Lupu, A. Hanbury, and V. Zenz. Clef-ip 2011: Retrieval in the intellectual property domain. In *CLEF (Notebook Papers/Labs/Workshop)*, 2011.

[16] K. Riesen, X. Jiang, and H. Bunke. Exact and Inexact Graph Matching: Methodology and Applications. In C. C. Aggarwal and H. Wang, editors, *Managing and Mining Graph Data*, volume 40 of *Advances in Database Systems*. Springer, 2010.

[17] N. M. Sadawi, A. P. Sexton, and V. Sorge. Performance of MolRec at TREC 2011 Overview and Analysis of Results. In *Proc. of TREC*, 2011.

[18] P. Sidiropoulos, S. Vrochidis, and I. Kompatsiaris. Content-based binary image retrieval using the adaptive hierarchical density histogram. *Pattern Recognition*, 44(4):739 – 750, 2011.

[19] V. Smolov, F. Zentsev, and M. Rybalkin. Imago: open-source toolkit for 2D chemical structure image recognition. In *Proc. of TREC*, 2011.

[20] A. Tiwari and V. Bansal. Patseek: Content based image retrieval system for patent database. In *ICEB*, pages 1167–1171, 2004.

[21] S. Vrochidis, S. Papadopoulos, A. Moumtzidou, P. Sidiropoulos, E. Pianta, and I. Kompatsiaris. Towards content-based patent image retrieval: A framework perspective. *World Patent Information*, 32(2):94–106, 2010.

[22] M. Zimmermann. Chemical structure reconstruction with chemocr. In *Proc. of TREC*, 2011.