MAREC

Information Retrieval Facility

www.ir-facility.org

The **MAtrixware REsearch Collection (MAREC)** is a subset of the Matrixware commercial grade Alexandria repository, a global storage facility for high quality scientific, technical and business information. MAREC is intended as basic dataset for research in areas such as information retrieval, natural language processing or machine translation, which require large numbers of complex documents.

MAREC consists of 19 million patent documents, nearly half of which include full text in English, French or German, normalised to a highly structured XML format developed by Matrixware for the IRF. The standardised fields include dates, countries, languages, references, person names, and companies as well as subject classifications such as IPC codes. MAREC is optimized for comparison purposes, as many documents are available in similar versions in other languages that refer to the same distinctly attributable invention.

The 19,386,697 XML files measure a total of 621 GB and are hosted by the supercomputing infrastructure of the Information Retrieval Facility. Scientific members of the IRF have full access to this collection, free of charge.

If you are interested in experimenting with the MAREC data, please contact **membership@ir-facility.org**

MAREC AT A GLANCE

- 19 million XML patent documents
- from 4 patent organisations:
  - European Patent Office (EP),
  - World Intellectual Property Organisation (WO),
  - United States Patent and Trademark Office (US)
  - Japan Patent Office (JP)
- unified fields, numbering scheme and citation format
- comparable corpus
- free access and support for IRF members

DOWNLOAD A SAMPLE

MAREC's 19 million patents take up more than 600 GB. You can download a sample of one week's data, i.e more than 20,000 patents at http://matrixware.net/prototypes/marec.

DATA COLLECTIONS

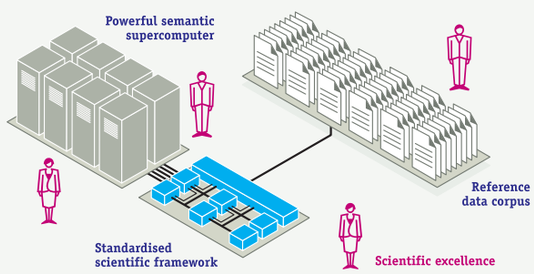
The IRF provides a number of test data collections that have either been developed by the IRF, by one of its members or by third parties. These data collections can be used freely for scientific experimentations. In addition to MAREC, the IRF provides access to the ClueWeb09 data. The ClueWeb09 dataset was created by the Language Technologies Institute at Carnegie Mellon University to support research on information retrieval and related human language technologies. It is a 25 terabyte dataset of about 1 billion web pages crawled in January and February, 2009. The crawl order was best-first search, using the OPIC metric. The crawl was started from about 28 million URLs that either

- had high OPIC values in a web graph produced from an earlier 200 million page crawl, or
- were ranked highly by a commercial search engine for one of 4,000 sample queries in one of 10 languages.

This dataset covers web content in English, Chinese, Spanish, Japanese, French, German, Arabic, Portuguese, Korean, and Italian. The dataset is used by several tracks of the TREC conference.

If you are interested in accessing the ClueWeb09 collection, which is available on the IRF infrastructure, please send an email to: membership@ir-facility.org

THE IRF: OPEN SCIENCE – OPEN INNOVATION



The Reference Laboratory for Information Retrieval

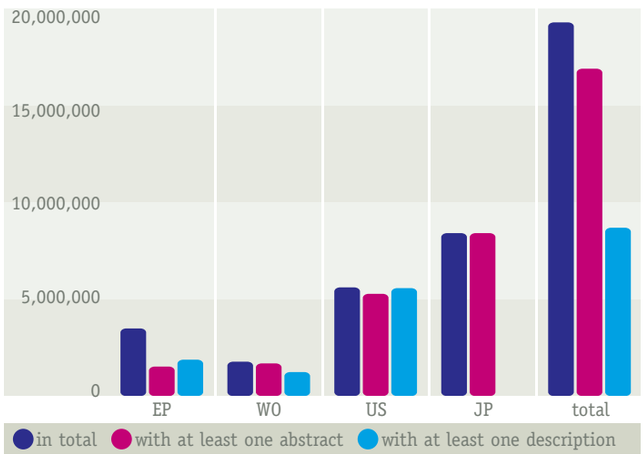
The Information Retrieval Facility (IRF) is an independent international research institute, based in Vienna, Austria, dedicated to the promotion and facilitation of research in the fields of information retrieval, data mining and machine learning on a large scale with a primary focus on patent retrieval. The IRF is controlled by a Scientific Board of leading experts in information retrieval, chaired by Keith van Rijsbergen, and committed to the concept of open science. The IRF provides a powerful semantic supercomputing infrastructure, and offers scientists unique free-of-charge access to test data collections. The IRF takes part in numerous research projects, e.g.: the development of semantic technologies, the processing of natural languages, automated image recognition, or the development of next generation search engines. By linking researchers with end-users from the industry the IRF nurtures a sustainable innovation cycle and ensures a high scientific standard and mutual acknowledgement of requirements and solutions.

MAREC
Statistics

Information Retrieval Facility
www.ir-facility.org

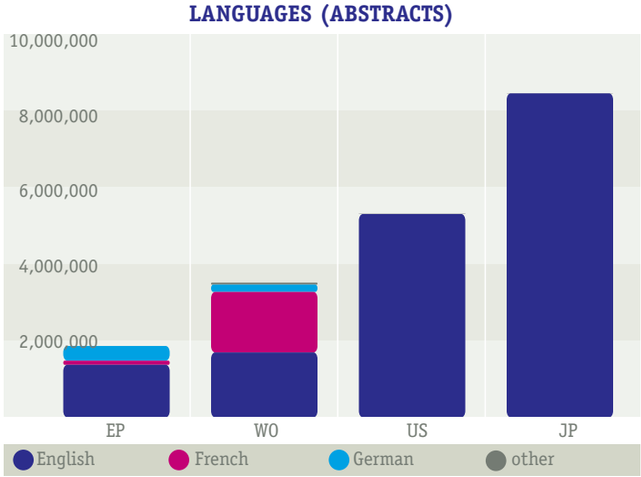
The textual elements of a patent are the abstract, description and claims. Each section has a different purpose: the abstract gives a short and general summary of the invention, the description gives background information for understanding the invention and the claims section is the legal scope of protection of the patent. In MAREC all textual elements are marked with section label and language label. Not all documents in MAREC have all sections – the following graph shows distribution among the MAREC documents on abstract and description.

NUMBER OF DOCUMENTS



The language distribution in the documents is shown in the following graph:

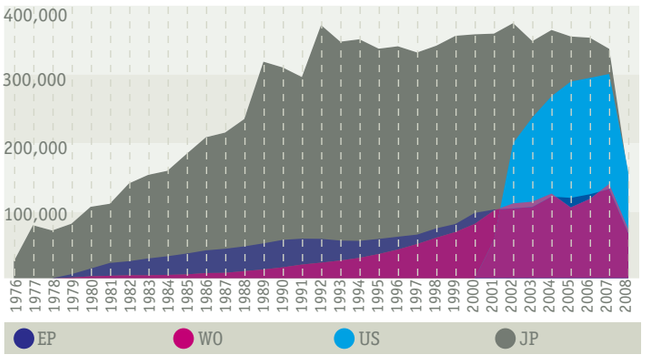
LANGUAGES (ABSTRACTS)



Patent documents change over time i.e. patents have different versions produced at different stages of the patent life cycle (the main stages are application and granted). For instance, in the UK it is accepted to have an initial title which does not need to be informative but after 18 months the title should have been modified so that it indicates the subject matter.

MAREC documents cover a time period from 1976 to 2008 for both patent applications and granted patents. The following graph shows the distribution over time of applications for patents:

DOCUMENTS PER YEAR (applications, 2008 until end of June)



MAREC documents are classified according to schemata such as the International Patent Classification (IPC) and European Classification (ECLA, based on IPC but more fine grained). The IPC is arranged in a hierarchical structure of 8 sections, divided into 120 classes, 600 subclasses and 70,000 groups.

SECTIONS:
A - Human Necessities
B - Performing Operations, Transporting
C - Chemistry, Metallurgy
D - Textiles, Paper
E - Fixed Constructions
F - Mechanical Engineering, Lighting, Heating, Weapons, Blasting
G - Physics
H - Electricity

CLASS: two digits following the section symbol (e.g. H 01 – BASIC ELECTRIC ELEMENTS). A patent can be classified into more than one class (multi-classification).

SUBCLASS: capital letter following the class symbols (e.g. H 01 J – ELECRIC DISCHARGE TUBES OR DISCHARGE LAMPS).

GROUP:
Main-group: integer followed by a slash and 00 (e.g. H 01 J 9/00 – APPARATUS OR PROCESSES ESPECIALLY ADAPTED TO THE MANUFACTURE).
Sub-group: integer counting from 01 at the right side of the slash, (e.g. H 01 J 9/38 – EXHAUSTING, DEGASSING, FILLING OR CLEANING VESSELS).
A third digit at the right side of the slash is interpreted as a further subdivision (e.g. H 01 J 9/385 – EXHAUSTING VESSELS).