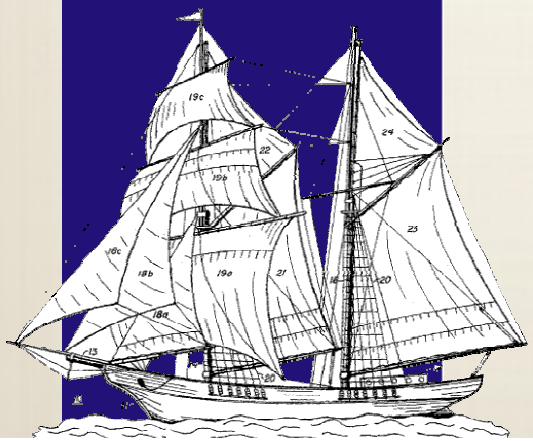
John Tait, Chief Scientific Officer, IRF
Francisco Webber, CEO Matrixware & IRF

The Information Retrieval Facility and its role in Professional Information Research

The IRF Mission

- To bring the latest information retrieval technology to the community of patent professionals and other professional searchers.
- To bridge the gap between the information specialist and patent data.
- To maintain a facility that enables large scale information retrieval and in depth patent and other complex data processing.



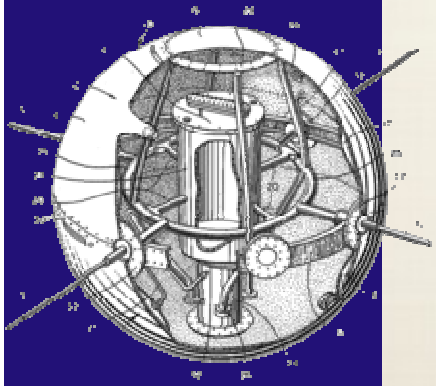
Patents - General

Intellectual Property (IP):

Across the world there are about **60 million patents**

Patent documents formed the most important shared information pool:

- Knowledge and research
- Innovative capacity and commercial strength
- Legal information



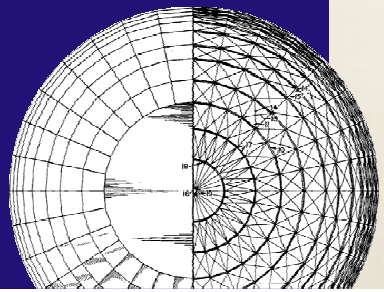
80% of world technical-scientific knowledge can be found in patent documents – in some branches of industry the number is significantly higher still

Patents – Commercial importance

Intangible Assets:

Innovation improves competitivity, creates jobs, promotes growth and secures prosperity.

- **The only valid and binding instrument** to protect innovation
- **An important commercial asset** – a monopoly on the use of an invention
- The issue of licences has become a **significant revenue source** for many companies



Patent Retrieval

Problems/challenges:

With the increasing further development of data-processing, the **volume of digitally available, unstructured information** is also growing.

Patent experts these days are still using **10-15 year old technology.** Patent searching and assessment is often laborious and time-consuming.

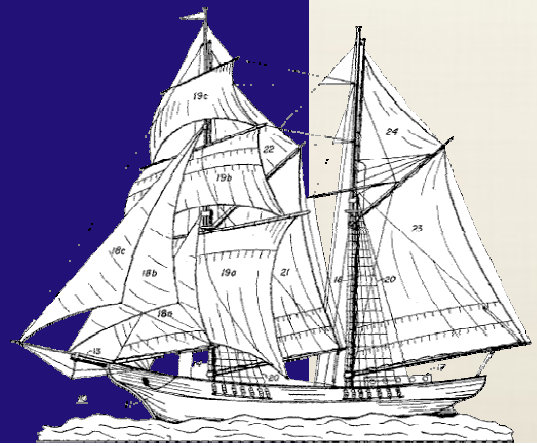
Consequences:

- Each year €60 billion (inside the EU) spend on **double inventions**

UK Patent No 1

From 1617
Engraving and Printing
Maps Plans etc. 5 pages

Adobe Acrobat
Document



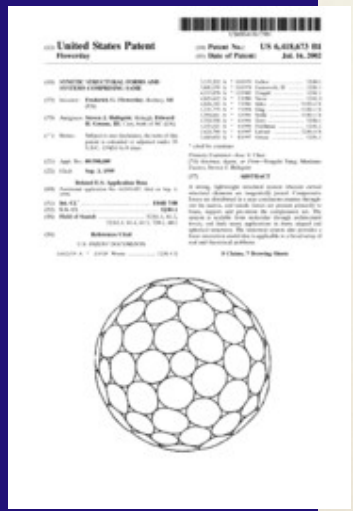
US Patent 7 089 111 (2005)

Vehicle Navigation System and Route Guidance Method – 13 pages

Adobe Acrobat Document

Distinctive Patent Search Characteristics

- **High Recall:** a single missed document can invalidate a patent
- **Session based:** single searchers may involve days of cycles of results review and query reformulation
- **Defendable:** Process and results may need to be defended in court



Matrixware

- Established in 2005
- Headquarters in Vienna
- Has over 50 employees, an expert team of software developers, technicians, mathematicians, language experts and other specialists

Field of activity: Information Retrieval in the segment of Intellectual Property

Products: innovative solutions for searching and categorising patent data

Methodology: semantic and statistical

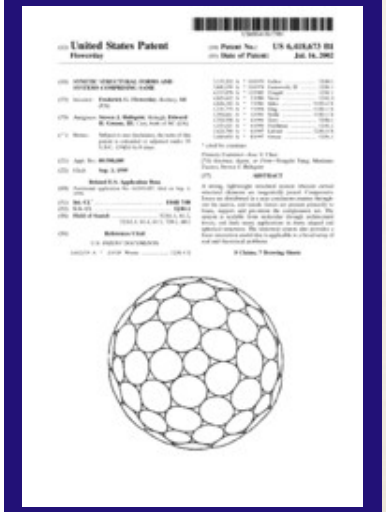
Matrixware

Objective: using the latest technologies, to automate data analysis, processing and use of existing information

Means adopted: Matrixware is constructing and expanding an extensive corpus of international patent literature

Patent data corpus: The basis is a complete world patent register, which is filled in by Matrixware with the fulltext documents

- With very highly controlled, transparent quality
- Enriched with meta-information and secondary literature

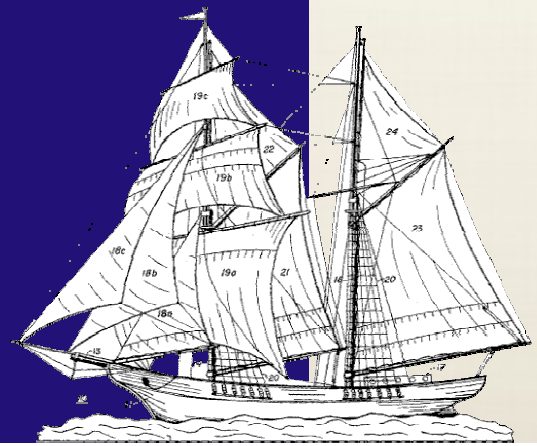


Information Retrieval Facility (IRF)

A platform initiated by Matrixware which:

- improves the global transfer of knowledge between the areas IP and IR and
- promotes collaboration between experts on the development of new research methodologies for international patent data

The IRF provides researchers with one of the **largest invention databases in Europe**.



IRF - History

November 2006
IRF Foundation Conference in Vienna

December 2006
Silicon Graphics
becomes a sponsor of the project

April 2007
IP Expert committee appointed

November 2007
1st IRF Symposium
in Vienna

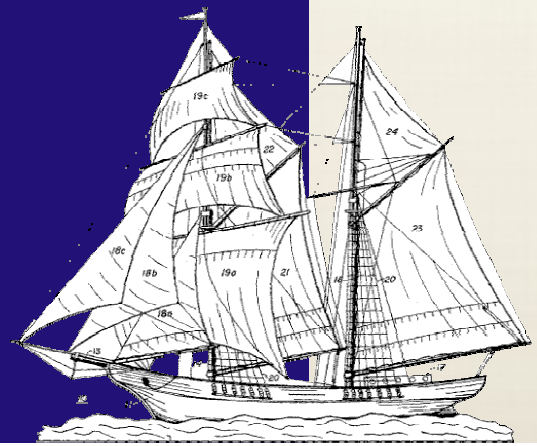
July 2008
2nd IRF Conference

November 2008
2nd IRF Symposium
in Vienna

- 2006: **Matrixware** approaches a group of world leading information scientists with the idea of establishing a text **research institute**
- November 2006 nine professors from leading universities agree to form founding scientific board
- Dec 2006: **Silicon Graphics** is acquired as a supporter and sponsor
- April 2007: **IP Expert Committee** is appointed
- September 2007 CSO appointed
- November 2007 first IRFS

IRF – Areas of activity

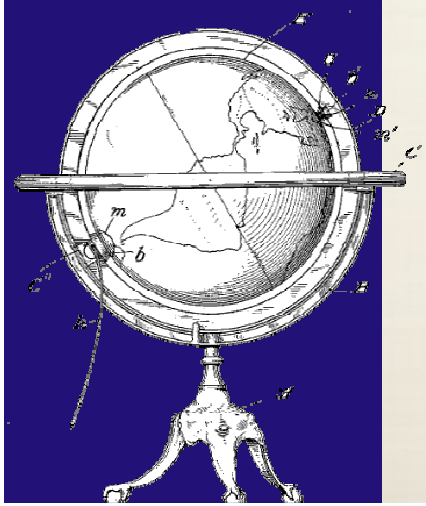
- Information Retrieval **Experiments** on a large scale
- Ongoing **publication** of scientific results
- **Scientific consultancy** for industrial companies and government organisations
- **Continuous and further training** of young IR-researchers



IRF – Target group

The IRF is aimed at the following groups:

- **Researchers** in the area of IR and related fields
- **Students** of these subjects
- Information management experts in **industry**
- **Patent offices** and **government organisations**

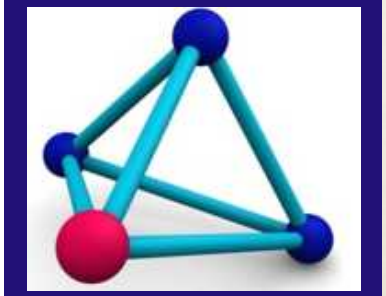


IRF – Players

Partner-Universities

Current Projects Include

- Accessability of Information (Glasgow)
- Quantum IR Models (Glasgow)
- Semantic Analysis of Patent Data (Sheffield and Nijmegen)
- CEA List
- Umass Amherst
  - Language Modelling for Patent Retrieval
  - OCR for patents



IRF – Importance for the location

The headquarters of the IRF: Vienna

From here the IRF is continually expanding its reputation as an **international "Centre of Excellence"** in the area of Information Retrieval.

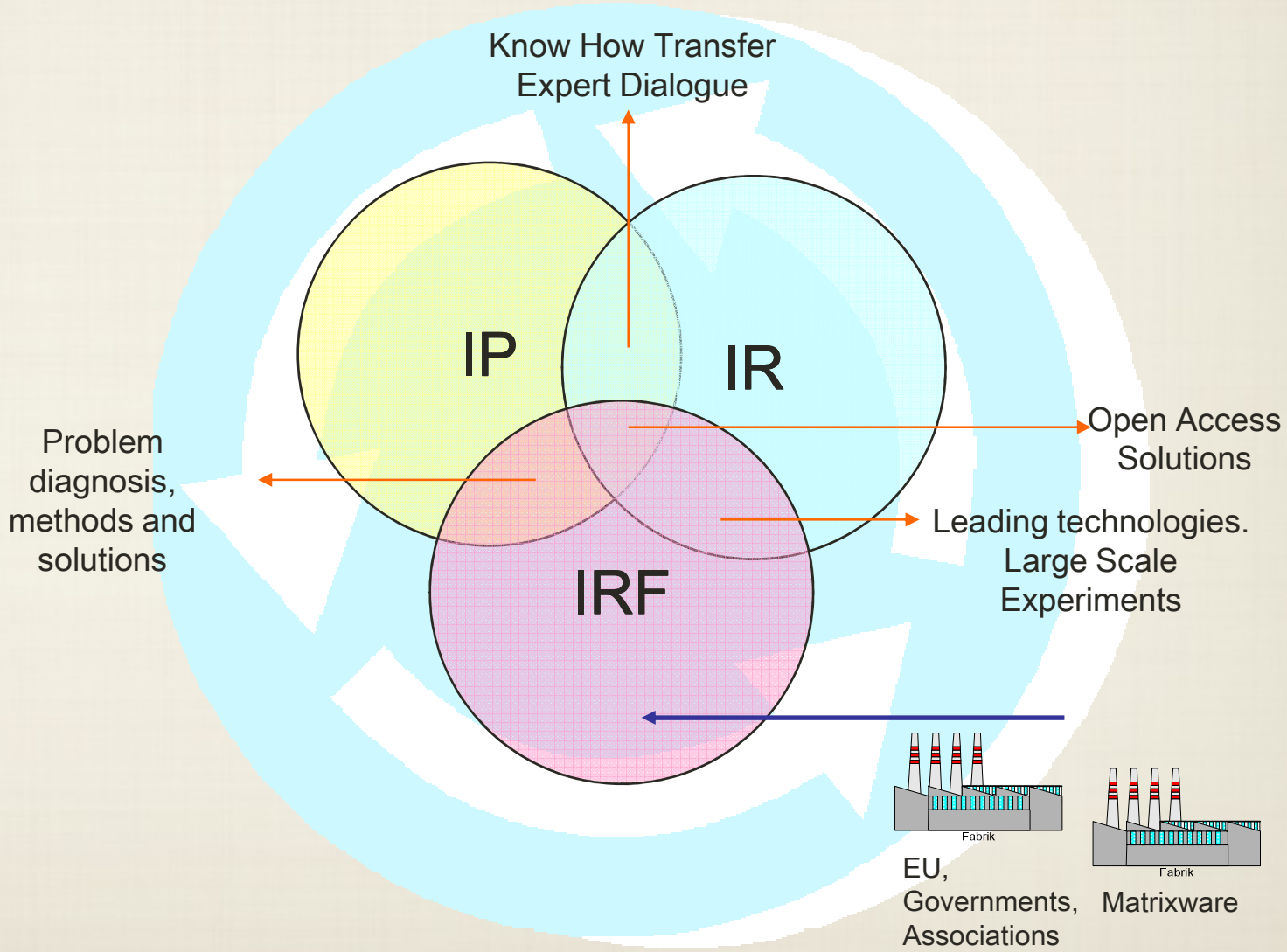
The IRF supports the international IP and IR community as an **open reference laboratory**.

IRF – Innovation cycle

One of the tasks of the IRF is to develop models, methods and standards in order to create **a bridge between science and industry.**

→ The interaction between theory and practice creates a **sustained innovation cycle.**

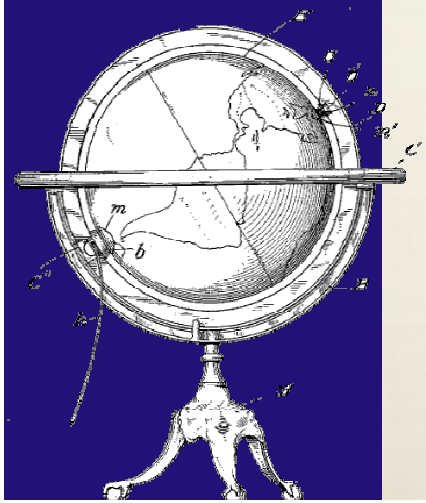
The innovation cycle



IRF – Semantic Supercomputing

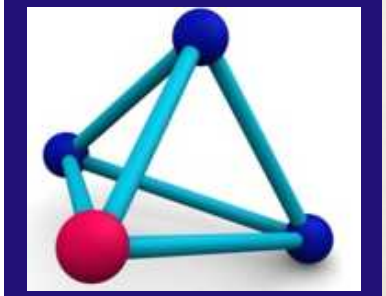
For the processing of multi-terabyte corpora the IRF maintains a **high-performance computer architecture.**

- High-performance supercomputer
- The latest supercomputing software
- The latest “Configurable Computing”
- The latest database technology



IRF – facility details

- SGI Altix 4700
  - 80 cores –> 40 CPUs (Itanium - IA-64) 1,4 GHz
  - ~ 300GB Memory (307)
  - 4FPGAs (Type: RC100)
- High performance XFS file-systems
  - ~40 TB Storage
- Software:
  - Linux
  - Lemur/Indri, Terrier, diverse JavaSDKs (Sun, BEA)
  - SGI NUMA Tools
  - Caché (object oriented database)
- Data
  - Patent Corpora (only text, no pictures and drawings) :
  - USPTO ~ 103 GB (with XML tags, ~68 GB without tags) EPO ~ 134 GB
  - About 2.6 million documents



IRFS – November

- 5 areas of emphasis
  - Data quality
  - Language barriers
  - Corpus enrichment
  - Tools for IP professionals
  - Tools for management & research
- 12 subject topics
- 30 experts giving presentations
- Over 100 participants

For the first time: Interesting expert discussions on the principal problems of information search in patent documents.

SCIENCE MEETS INDUSTRY
IRFS 2007 Vienna
Information Retrieval Facility
SYMPOSIUM
8-9 November 2007, Marriott Hotel, Vienna, Austria

IRFS2008

6th & 7th November
in Vienna

Thank you for your attention
Any questions ?

www.ir-facility.org
www.matrixware.com