



**Information
Retrieval
Facility**

IRF Academic Cooperation Opportunities

An Introduction

Reaching out to the global community of IR experts

The IRF promotes and facilitates cutting-edge industry-related research in Information Retrieval by identifying and addressing challenging topics through dialogue between academia and industry, communicating these topics to its members, and providing infrastructure for large scale experimentation.

IRF Mission Statement

The Information Retrieval Facility (IRF)

- Founded in **2007** in **Vienna, Austria**
- **Independent non-profit** research institute
- Member-based organization with over **200 scientific members from around the globe**
- **Managed by a Scientific Board** composed of leading IR experts
- Chairman of this Board is **Keith van Rijsbergen**, one of the founders of modern Information Retrieval

OBJECTIVES:

- Promotes and facilitates research in **large scale information retrieval**
- Fosters **knowledge transfer** between academia and industry



**Information
Retrieval
Facility**

Founding Members



Carnegie Mellon



UNIVERSITÄT
D U I S B U R G
E S S E N

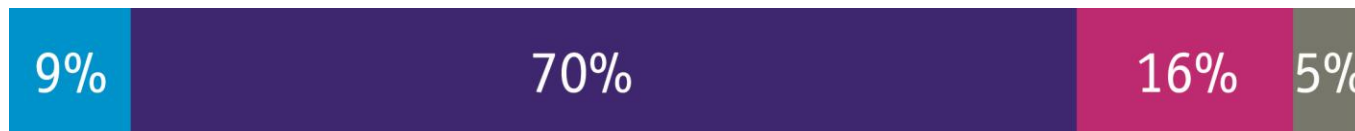
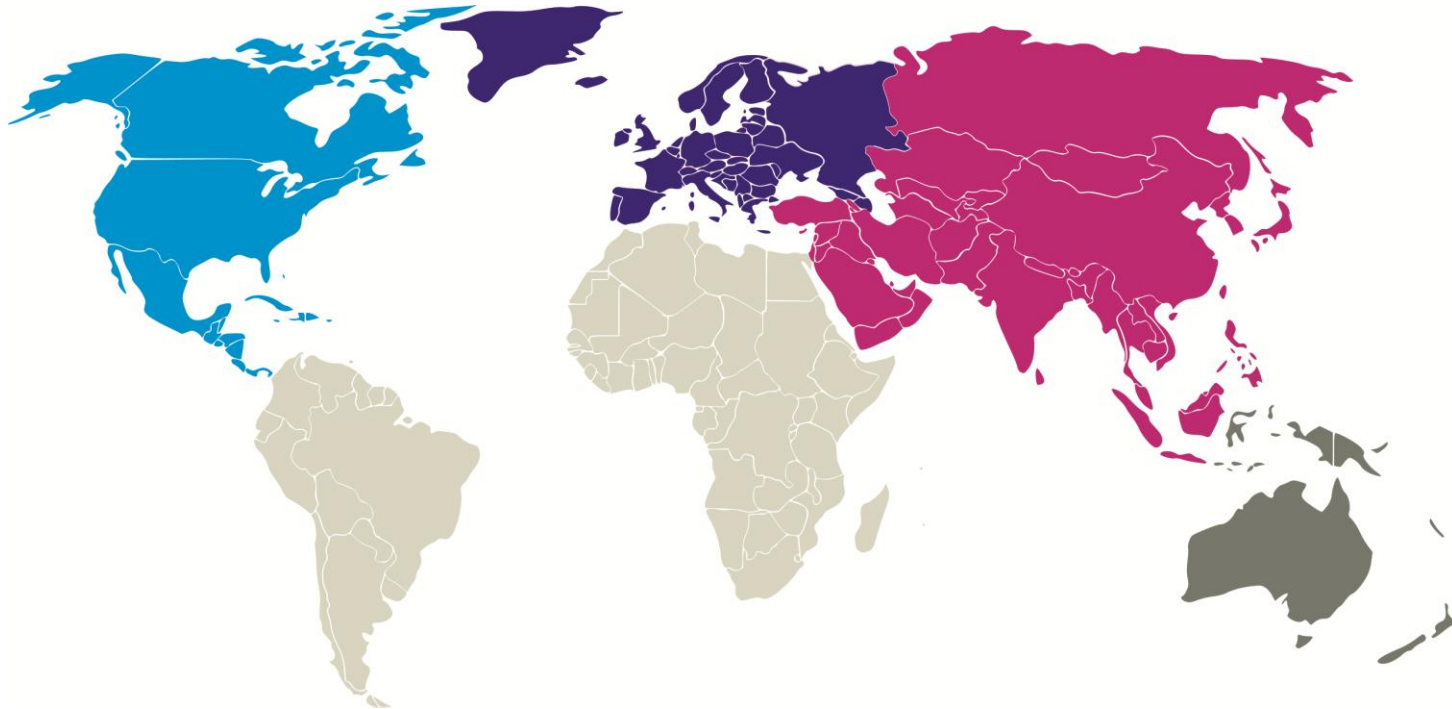
UNIVERSITÄT
D U I S B U R G
E S S E N



Geographical Distribution of IRF Members



Information
Retrieval
Facility



IRF Scientific Membership



Information
Retrieval
Facility

All applications are reviewed by the IRF Scientific Board

- Institutional Scientific Membership

For institutions that carry out Information Retrieval related research

- Individual Scientific Membership

For individual scientists who carry out non commercial information retrieval research

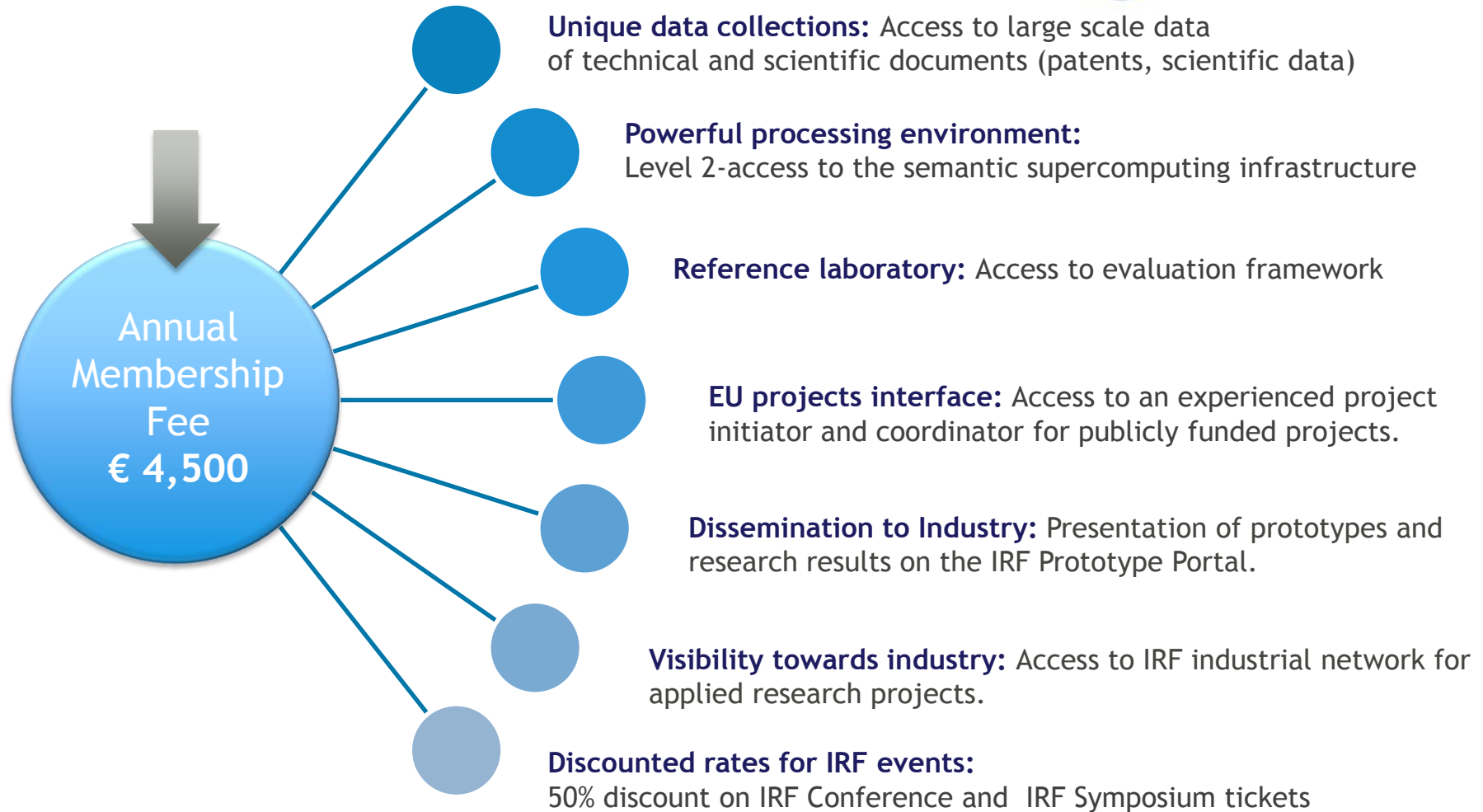
- Student Membership

For university level students in the context of non commercial information retrieval research and related subjects.

Institutional Scientific Membership



**Information
Retrieval
Facility**



Semantic Supercomputer

The IRF hardware infrastructure is one of the most powerful systems worldwide that deals with semantic processing of text.



**Information
Retrieval
Facility**

Large Data Collider (LDC)

- 320 Gbytes of main memory
- 80 Itanium CPUs running at 1.4 GHz
- 1 SGI Infinite Storage
- 1 mptsas SCSI controller
- 12 mptfc fibre channel controllers
- 4 SGI RASC RC100 FPGA
- 2 Broadcom BCM5704 Gigabit Ethernet interfaces

Medium Data Collider (MDC)

- 2 IBM x3950
- 32 Cores (4 quad core Intel Xeon@2.93GHz per node)
- 256 Gbytes of main memory
- 600 Gbytes storage on internal disks
- Production cluster for Java software and serial code.

Storage, SAN

Will be provided based on the needs of a given research project after approval by the Scientific Board.

Data Collections



**Information
Retrieval
Facility**

MAREC

Consisting of 19 million patent documents in different languages, that are normalized to a highly specific XML format, MAREC is the **first standardized patent data corpus for research purposes**.

MAREC serves as a global storage facility for high quality scientific, technical and business information.

MAREC at-a-glance

- 19 million XML documents
- From 4 patent organizations:
 - European Patent Office (EPO)
 - World Intellectual Property Organization (WIPO)
 - United States Patent and Trademark Office (USPTO)
 - Japan Patent Office (JPO)
- Unified fields, numbering scheme and citation format
- Comparable corpus

The MAREC collection is accessible to IRF Scientific Members to carry out non-commercial research.

Why are patents relevant for IR?



Information
Retrieval
Facility

- A patent is a **monopoly** granted by the state for a fixed period in return for **public disclosure** of an invention
- Patents are **complex legal** documents
 - Technical drawings, chemical formulae, OCR errors, multiple languages, classification, annotations...
- Patents are an important **economic factor**
 - Total worldwide economic value of patents: 500 billion USD p.a. [KIPO]
- Patent search is a well documented example of **professional search**
 - And a good example of collaborative search behaviors

Data Collections



**Information
Retrieval
Facility**

CLUEWEB

ClueWeb09 is a 25 terabyte dataset of about 1 billion web pages crawled in January and February, 2009.

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.

ClueWeb09 at-a-glance

- 25 terabyte dataset
- 1 billion web pages crawled in January and February, 2009
- Covers web content in English, Chinese, Spanish, Japanese, French, German, Arabic, Portuguese, Korean, and Italian.
- Is used by several TREC tracks

The ClueWeb collection is accessible to IRF Scientific Members to carry out non-commercial research.

Data Collections



**Information
Retrieval
Facility**

TREC-CHEM

The TREC-CHEM evaluation campaign aims at creating a reference collection for chemical information retrieval engines: It is run in 2010 for the 2nd time.

TREC-CHEM'10 at-a-glance

- 1.4 million patent documents from MAREC classified in category C or A61K (IPC)
- 170K scientific articles from
 - The Royal Society of Chemistry
 - Oxford University Press
 - And many other journals

The TREC-CHEM collection is accessible to IRF Scientific Members to carry out non-commercial research.

IRF Expert Pool



**Information
Retrieval
Facility**

The IRF manages a pool of experts in the fields of information retrieval, natural language processing and semantic technologies, to deliver specific consulting services to industry.

Information
Retrieval

Information
Extraction &
Visualization

Natural Language
Processing

Machine Translation

The IRF is looking for scientists with proven track records in these areas and interested in providing high quality consultancy for:

- **Training**
- **Evaluation & Benchmarking**
- **(Meta) data production**
- **Research**
- **Prototype building**

Experience levels: Postdoc, Professor and other faculty.

IRF Key Research Areas



Information
Retrieval
Facility

Multiple
Indexing

Automated
Text
Annotation

Automated
Information
Extraction

Automated
Document
Categorization

Complex
Professional
Search
Strategies

Image
Retrieval

Interfacing
Professional
Users

Statistical
Machine
Translation

Current Research Projects



**Information
Retrieval
Facility**

PLuTO: Patent Language Translations Online

- Funded by the EU (ICT Policy Support Programme). Duration: 2010-2013
- Consortium Partners: CNGL/Dublin University, ESTeam, Cross Language, WON
- Goal: Develop a collaborative online translation platform for patents

Key Technologies:
Machine Translation,
Translation Memory,
Information Retrieval

KHRESMOI: Medical Information Analysis and Retrieval

- Funded by the EU (ICT FP 7 Programme). Duration: 2010-2014
- Consortium Partners: HES-SO, Atos, Dublin University, Ontotext, University of Sheffield + 7 additional European partners
- Goal: Develop a multimodal, multilingual search and access system for biomedical information and documents

Key Technologies:
Automated Information
Extraction, Automated
Image Analysis and
Indexing,
Linking of Unstructured
Documents with
Structured Knowledge
Base, Cross-Language
Search, Adaptive User
Interfaces

Current Research Projects



**Information
Retrieval
Facility**

Service Detective

- Funded by the Austrian FFG (FIT-IT Programme). Duration: 2009-2011
- Consortium Partners: University of Innsbruck, Seekda
- Goal: Automatic discovery of web services

Key Technologies:
Web Crawling,
Semantic Annotation

LarKC: Large Knowledge Collider

- Funded by the EU (ICT FP 7 Programme). Duration: 2008-2011
- Consortium Partners: STI Innsbruck, AstraZeneca R&D, CEFRIEL, Max Planck Institute + 8 additional European partners
- Goal: Build an integrated platform for large scale semantic computing

Key Technologies:
Large Scale
Information Retrieval,
Semantic Computing

Other Research Activities



**Information
Retrieval
Facility**

- Phrase-based SMT technology (MOSES)
- Chinese to English automatic translation, customized for patents
- 4 million bilingual aligned sentences obtained from human translated patents

**Statistical
Machine
Translation**

Development of a technique to enable efficient image searches for a variety of technical drawings, such as flow charts, block diagrams, time charts and graph plots

**Image
Retrieval**

CLEF-IP (cross-lingual search of patents) and
TREC-CHEM (retrieval of chemical documents)

**Evaluation
Tracks**

- Development of a framework for professional search based on the workflow paradigm
- Goal: Translate extensive information search tasks in an easy-to-use graphical interface

**Interfacing
Professional
Users**

LEONARDO

A Framework for Information Retrieval



**Information
Retrieval
Facility**



leonardo

Leonardo is a premium toolkit framework for IR development:

- The workbench is built on Eclipse RCP and includes an open Software Development Kit (SDK).
- It allows to create custom tool nodes and widgets in an open source, drag n' drop modelling environment
- It allows to construct and deploy IR applications in less than a few hours

Leonardo sets a new standard for the human intelligence enterprise:

- It can plug into virtually any data source, enabling you to connect different kinds of information and integrate different tools.
- It is built on a IR Workflow model, which allows retrieval tasks to be solved similarly in a scalable and extensible environment.
- It allows to share workflows through publishing in a browser

Why join the IRF?



**Information
Retrieval
Facility**

- ➔ **Access to large scale experimentation environment**
The IRF manages a semantic supercomputing infrastructure, large scale data collections and evaluation frameworks
- ➔ **Access to publicly funded projects**
The IRF builds research consortia, writes proposals, manages projects
- ➔ **Access to industrial network**
The IRF connects to end-users willing to test prototypes, evaluate results or commission research projects
- ➔ **Join a network of leading IR scientists**
The IRF promotes academic knowledge transfer
- ➔ **Make publicity for your institution**
The IRF communicates its partnerships on marketing materials distributed at major IR conferences sponsored by the IRF and IRF events

IRF Scientific Conference // IRF Symposium



Information
Retrieval
Facility

Save the Date

6 - 9 JUNE 2011

IRF SYMPOSIUM

VIENNA, AUSTRIA



- The IRF Scientific Conference provides a multidisciplinary forum for young researchers and brings them into contact with the industry at an early stage.
- The IRF Symposium is a platform to discuss the specific challenges of patent searching and analysis and exchange ideas on potential solutions, with the objective of making the latest IR technologies available to the industry.
- The Exhibition at the IRF Symposium is *the* place to expose prototypes to end-users.

For more information please visit: www.irfs.at

Thank you!



**Information
Retrieval
Facility**

For more information about the academic cooperation possibilities
with the IRF, please contact:

**Prof. John Tait
Chief Scientific Officer**

Information Retrieval Facility
Tech Gate, Donau-City Strasse 1, A-1220 Vienna
Phone: +44 (7931) 793119
john.tait@ir-facility.org
www.ir-facility.org
<http://ir-facility.net>