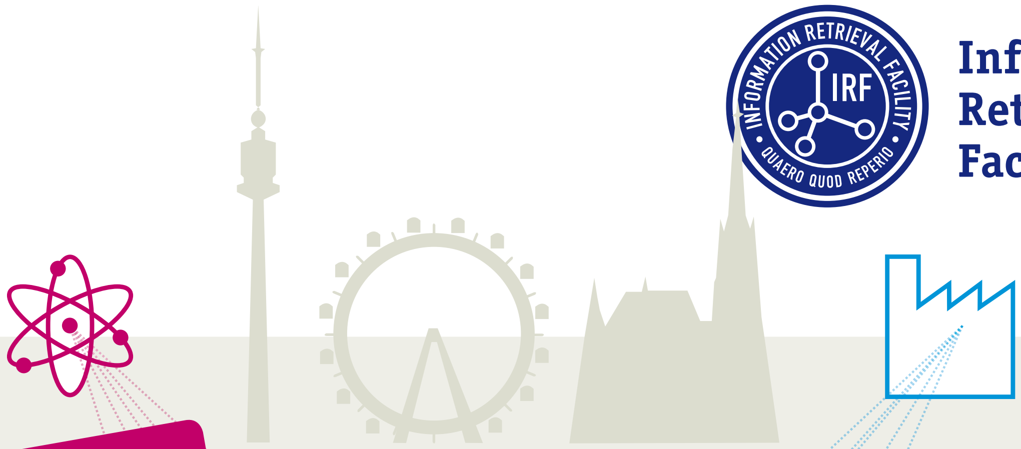
Information Retrieval Facility



Next IRFC / IRFS:
6 - 9 June 2011 (tbc)

MAY 31 – JUNE 4, 2010

1st IRF CONFERENCE
STRATEGIC SEMINAR
3rd IRF SYMPOSIUM

VIENNA /// AUSTRIA

REVIEW

1st IRF SCIENTIFIC CONFERENCE

The first IRF Scientific Conference was held on 31 May 2010 in Vienna with the participation of about 70 academics and information professionals, fulfilling the vision of a multidisciplinary scientific forum. The presentations dealt with information retrieval from different perspectives, such as natural language processing, evaluation measures and alternative models. Many of the papers considered applications in the challenging area of patent retrieval with use of diverse approaches including logic-based retrieval, conditional random fields and probabilistic retrieval models. All accepted papers and posters are published by Springer in the Proceedings of IRFC 2010 as part of the Lecture Notes in Computer Science series.

Prof. Mark Sanderson (University of Sheffield) opened the conference with a brief outline of the history of search evaluation before tackling the problem of designing test collections. He described some pioneering but relatively overlooked research pointing out that the key problem for researchers isn't the question of how to measure searching systems accurately but how to accurately measure people.

M-Dyaa Albakour from the University of Essex talked about attachment prediction - the task of automatically identifying email messages that should contain an attachment. His team proposes an analysis based on the identification of individual sentences within an email that refer to an attachment. This finer-grained approach outperforms previously reported document-level attachment prediction in similar evaluation settings.

Neil Newbold from the University of Surrey presented a new approach to ranking that considers the reading ability (and motivation) of the user. His team investigated using readability to re-rank web pages. Results to date suggest that considering a view of readability for each reader may increase the probability of relevance to a particular user.

Erik Graf from the University of Glasgow explored the benefits of integrating knowledge representations in prior art patent retrieval. Key to the introduced approach is the utilization of human judgment available in the form of classifications assigned to patent documents. In general the proposed knowledge expansion techniques are particularly beneficial to recall and result in significant precision gains. As a low-cost resource that is up-to-date, Wikipedia recently gained attention as a means to provide cross-language bridging for information retrieval.

Benjamin Roth from the Saarland University showed that standard Latent Dirichlet Allocation (LDA) can extract cross-language information valuable for IR by simply normalizing the training data. Furthermore, his team showed that the combination of LDA and Explicit Semantic Analysis (ESA) yield significant improvements.

Jay Urbain from the Milwaukee School of Engineering explored the development of probabilistic retrieval models for integrating term statistics with entity search using multiple levels of document context to improve the performance of chemical patent search. His team reports better results than those achieved at the 2009 TREC Chemistry track.

A typical evaluation of a retrieval system involves computing an effectiveness metric, e.g. average precision for each topic of a test collection and then using the average of the metric, e.g. mean average precision to express the overall effectiveness. However, averages do not capture all the important aspects of effectiveness. **Mehdi Hosseini** from the University College London explored how the variance of a metric can be used as a measure of variability.

David Hawking, Chief Scientist at Funnelback Internet and Enterprise Search in Australia, closed the conference with a practical and commercial perspective on a half century of electronic information retrieval. He concluded that tools must be created to help users to better cope with the limitations of systems, for example spelling suggestion tools and query expansion tools.

Hamish Cunningham from the University of Sheffield and **Stefan Rueger** from The Open University, respectively General Chair and Programme Chair, were extremely satisfied with the outcome of this first IRF Scientific Conference, both in terms of participants and quality of papers (11 accepted out of 20 submissions). The planning of the 2nd IRF Scientific Conference can begin!

All presentations, photos and videos are available at:
www.ir-facility.org/events/irf-conference

IRF STRATEGIC SEMINAR: INTELLECTUAL PROPERTY AS A KEY TO BUSINESS GROWTH?

An exceptional panel of international experts met in parallel to the 3rd IRF Symposium to discuss Intellectual Property as a key to business growth and value creation. They presented clear indications of best practice in this area, exploring both policy issues and technological decision-making breakthroughs. Participants were on the one hand senior operating executives from research-driven companies like the Austrian voestalpine and the US pharmaceutical group Pfizer, on the other hand innovation management and intellectual property experts. The lively discussions delivered concrete cost-effective IP optimizations on the best way of managing intellectual property and the performance of an R&D department.

Desai Narasimhalu, director of the Institute for Innovation and Entrepreneurship of the Singapore Management University, stated that successful innovation management is the only competitive advantage any enterprise can have. Innovation management begins with the creation of a culture of innovation within a company and continues with the introduction of several innovation management processes. The IP Audit which gives a recommendation for the commercialization strategy of IP assets is an important component of successful innovation management

Setting up an IP department requires careful analysis. The IP strategy must be discussed and agreed with all decision-makers within the company. The structure of internal work processes and the extent of involvement of external partners like patent attorneys depend largely on the answers to questions like ”Do we want to use IP to refinance our R&D spending? Do we covet access to new technologies? Do we want to hinder competitors?“ ”There is no silver bullet: Your IP strategy and your corporate structure will determine how you set up your IP department“, concluded **Franck Cuypers** from PricewaterhouseCoopers.

Mehdi Hosseini, University College London

IRF Strategic Seminar

According to **Gerald Landl**, Head of IPR at voestalpine Stahl, patents are an important information source for competitive intelligence and technical know-how. That is the reason why patent search and analysis is integrated in the research and development process of voestalpine, which set up a R&D project in order to improve the efficiency and quality of the search for relevant documents and in order to simplify the analysis. Together with the Knowcenter Graz and m2n Intelligence Management they developed an easy-to-use visualization tool that gathers the know-how of all researchers within the company and enables the early identification of high potential technical areas.

The chemical and pharmaceutical industries are probably the most patent driven industry domains. ″Billions of dollars are spent on chemical research. And companies rightly believe that effective R&D is the lifeblood of the business, or certainly for those intending to stick around in chemicals", was quoted from ICIS Chemical Business, October 2009. But, although the nature of R&D has changed significantly in the past decade, the methods for analyzing the performance of R&D organizations have largely remained the same.

Nils Omland presented at the IRF Seminar an analysis tool developed by the Otto Beisheim School of Management that takes these changes into account: The new Patent Asset Index measures patent quality according to technical and market aspects and is more accurate than existing methods. The pharmaceutical industry is looking for solutions to cope with the exponentially growing quantity of potentially relevant patents which must be analyzed accurately, sometimes down to a single molecule. **David Walsh**, information specialist for patents and chemistry, presented the approach of Pfizer: The US company has implemented new semantic technologies like text mining to create an in-house patent database with Web 2.0 functionalities – Pfizerpedia.

The topic of patent evaluation was addressed from 2 complementary perspectives. **Pekka Sääskilahti** from Nokia Corporation impressed with a demonstration of how the game theoretic model can provide an answer to the growing need for patent evaluation. **Anthony Trippe**, director IP Analytics of 3LP Advisors presented a rather unconventional approach: It does not matter how much patents are objectively worth, nobody will want to acquire them if they are not valuable to potential buyers. He showed in his presentation several possibilities to increase the desirability of patents or a patent portfolio.

Manuel Desantes, former vice president of the European Patent Office, focused on the political aspects of innovation and intellectual property, stating that the last twenty years have shown exponential changes in the environment surrounding an intellectual property system which has changed very little. To meet the challenges of the information society, we need a patent system serving innovation, competitiveness and development, was his conclusion.

More information about the speakers as well as pictures of the Strategic Seminar are available online:
www.ir-facility.org/events/strategic-seminar

3rd IRF SYMPOSIUM – BENCHMARKING RELEVANCE

The Imperial Riding School Vienna was home of the 3rd IRF Symposium from 1 to 4 June 2010. About 150 delegates from academia and industry met to discuss the advances in information retrieval science impacting patent information retrieval, and to test new systems and prototypes exposed by solution providers and research groups. More than 40 international speakers addressed the topics of evaluation, interfacing, semantic annotation, image retrieval, chemical patent retrieval and multilingualism in plenary sessions and workshops. While the IR scientists have now a better understanding of the specific needs of patent searchers and have begun to integrate them in their research, there is a need to involve the technology implementers to develop test applications.

Claudia Tapia, Research in Motion

Yike Guo, Imperial College London

James Boyle, co-founder of Creative Commons and Science Commons and one of the most influential creative thinkers and expert of Intellectual Property in the information society, was invited as keynote speaker at the 3rd IRF Symposium. He addressed the audience of leading academics, information retrieval and intellectual property experts, asking: "What if the Web really worked for science?" and inviting to "Reimagining data policy and intellectual property". Although the World Wide Web was originally invented to help spread science, it is in fact easier to use the Web to buy products than to use it for scientific research. To find and get access to academic texts is a cumbersome task due to the publication policies of information providers and the different types of data needed. Boyle concluded that we need to rethink our approach to research policy, intellectual property, and the competitive norms of science if we want the Web to really work for science.

Evaluation Methodologies: Towards a quality standard

The first session, chaired by **Mark Sanderson** from the University of Sheffield and **Teresa Loughbrough** from Unilever, addressed a topic that was ubiquitous at the 3rd IRF Symposium: Evaluation Methodologies. In the plenary session **Wim Vanderbauwhede** from the University of Glasgow presented the results of a survey commissioned by the IRF with 81 patent analysts from 13 different countries. The results highlight the gap between the patent analysts' requirements and the available functionalities of Open Source Information Retrieval toolkits (e.g. Lemur/Indri, Lucene and Terrier). Wim concluded that many of the essential features for patent search professionals (e.g. exporting search queries history, keyword navigation, trend analysis etc.) are lacking in these toolkits, and either these features have to be incorporated or a specific patent search toolkit has to be developed.

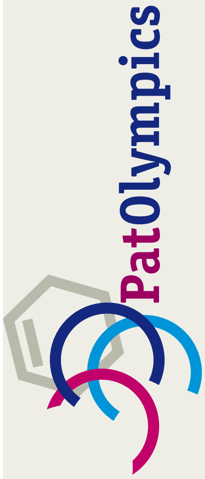
Linus Wretblad and **Joni Sayeler** from the Swedish Uppdragshuset presented a new quality model for patent data based on the assumption that there should be a standard assuring a minimum quality of the material. Wretblad and Sayeler introduced a quality model assessing different degrees of qualities: Gold level (ground truth – absolutely correct), Silver level (good for machine translation), Bronze level (allows searchability), and the Zero level (dead information). The Zero level document was humorously referred to as "vampire document". The higher the number of vampire documents in a collection, the lesser the performance of the machine translation and other natural language processing applications will be. Finally, **Michael Dittenbach**, an independent information access solutions engineer, presented a benchmark evaluation of different information retrieval toolkits and ranking models in the context of patents. The result showed that the ranking models only slightly change the outcome, but the merge of the result set of different indices generates overall better results. Dittenbach's conclusion that there is still a gap between what academics consider important and what professional searchers actually need explains the results of Vanderbauwhede's survey presented earlier.

Evaluation was also the focus of a workshop, where **Giovanna Roda**, Information Retrieval Specialist, presented the outcome of the 1st Cross Language Evaluation Forum Intellectual Property (CLEF-IP 09) track. Roda concluded that the Cranfield model to evaluate IR system reflects poorly the day-to-day research tasks of a patent searcher and that future tracks should take this into account. **Mihai Lupu** from the IRF presented the learnings from the Text Retrieval Conference Chemistry Track (TREC-CHEM 09), which aims at providing professional users with the means to choose a tool best fitted to their needs. But, as he underlined, an evaluation track is fundamental research, meaning that results are expected in a 3-year timeframe which reduces the attractiveness of such projects for the industry. **Jason Baron** from the US National Archives and

Cynthia Barcelon-Yang, Bristol-Myers Squibb

James Boyle, keynote speaker

Desai Narasimhalu, Keith van Rijsbergen, Yves Chiaramella



PatOlympics Prototype Evaluation Class

Two PatSports (ChemAthlon and CrossLingual Retrieving) were the focus of the new interactive prototype evaluation campaign called PatOlympics. 5 referees (3 for ChemAthlon and 2 for CrossLingual Retrieving) worked with 3 participants and their systems. The objective: A new way to directly interact with 'the other side' and to concentrate discussions away from powerpoints and into real systems under development.

In the end, three medals were awarded:

- ChemAthlon: BiTeM Group (University of Applied Sciences, Geneva, Switzerland)
- CrossLingual Retrieving: Spinque (Netherlands)
- Jury's favourite: BiTeM Group (University of Applied Sciences, Geneva, Switzerland)

The event took place in the afternoon of the second day of the IRFS, in the exhibition hall. This provided a very open atmosphere, where participants, referees and observers mingled and enjoyed an interesting exchange of ideas. In rounds of just under 30 minutes each, the referees sat down and worked with the teams and their systems to solve a task they had prepared in advance. The systems' scores were then computed based on the number of relevant documents retrieved.

The referees were also asked to rate each system on a "user friendliness" scale between 1 and 5. In this sense, the winning system, from the BiTeM group, had a "lovely interface" according to Teresa Loughbrough, while Tony Trippe qualified it as an "interesting system, but not optimized for chemistry".

The logistics of the event were provided by the IRF and included a web interface for both participants and referees to submit relevant documents, as well as an API specification to allow participants to submit documents directly from their systems. The scoreboard, showing the final scores, is still available online at **http://patolympics.ir-facility.org/**

The organizers wish to thank the Referees: Teresa Loughbrough, Tony Trippe, Henk Tomas, Monika Hanelt and Pierre Buffet

Records Administration contributed via video in presenting the results of the TREC Legal Track 2009, which evaluated competing search methods used in the context of US civil litigation (E-Discovery). He showed how the legal profession may approach the question of "standard-setting" in this domain. To conclude, **Anthony Trippe** showed the difficulty to apply a similar approach to patent information retrieval, the complexity of the tasks making a standardization very difficult.

Interfacing: Breaking the cycle of command line use?

The second session about Interfacing was chaired by **Joemon Jose** from the University of Glasgow and **Anthony Trippe**, 3LP Advisors. **Kalervo Järvelin** from the University of Tampere presented an approach to analyzing task-based information accesses and query formulation within the field of Molecular Medicine (MM). Järvelin concluded that search tasks in information intensive domains, such as MM, share common features with patent search. For instance, in both fields a searcher may conduct multiple concurrent query sessions in various information channels. The next speaker **Gerhard Fischer** from Syngenta also concluded that professional search interfaces need to support flexibility, e.g. support complex search terms construction combined with classification and codes etc. He finds that command line interfaces support the flexibility required by professional

3^{rd} IRF Symposium

Gerhard Fischer Syngenta,
Kalervo Järvelin University of Tampere

Leonardo Workshop

The purpose of this workshop was to discuss how visual workflows can contribute to more efficient information search and analysis in making the process simpler to conduct and easier to share with the broader community.

One important question raised was: What shall the next generation of search interface look like? Some of the participants from the intellectual property field maintained that their day-to-day search requires command lines interface due to the complexity of the search query. But Boolean queries are hardly reproducible, countered the proponents of visual pipelining. Another highly relevant topic was the presentation of the search results: There was a general agreement among participants that a traditional list of retrieved relevant documents is not good enough.

There is a need for tools that create different type of reports such as patent landscaping, technology landscaping etc. Furthermore, the workflow and the technologies used need to be transparent in order to insure quality of the conducted search. The need to integrate different data resources and libraries and, by doing so, reduce the transfer time between different information tools and resources is paramount – this is where visual workflows could bring groundbreaking improvements.

Leonardo is an application framework with a complete and convenient tool chain focused on the IR domain. It has been designed on the workflow paradigm, which allows IR problems to be solved similarly in a scalable and extensible environment. It can be described as a sophisticated search interface builder.

searchers better than graphical interfaces, which are too "rigid". In the third and forth presentations of the interfacing session the focus was more on workflows. Prof. **Yike Guo** from Imperial College London presented how to build an IR application with workflow technology based on a cloud computing platform. Meanwhile **Karen Abraham**, informatician at Unilever, reported how the workflow approach saved time for the Unilever scientists, since they were able to conduct searches both on external and internal literature through the same information interface, and the retrieved information was also presented in a user friendly format. The following panel discussion gathered arguments in favour of and against breaking the cycle of command line use by using new approaches like visual pipelining. This is a question that strongly polarizes the IP participants: While some of them would never want to part from their Boolean-based tools, others long for new approaches towards more sophisticated user interfaces.

Cross-lingual patent retrieval and translation: Dreams and reality

Karim Benzineb from SimpleShift opened the session with the results of a small survey that illustrates the needs in multilingual patent search and the satisfaction with automated patent translation tools. Not surprisingly, the main target languages come from the Asian region. The concentration on one input language (English) could help raising the quality of translation tools. While the increase of output quality remains the main objective, the costs of translation tools are often highlighted as a barrier. **Jong-Hyeok Lee** from the Pohang University of Science and Technology in South Korea introduced R&D activities for patent translation of Asian languages and explained why patent machine translation is difficult (e.g. extremely long and complex sentences, high rate of out-of-vocabulary words). While statistical machine translation is the most popular technique, he sees a high potential in hybrid, multi-engine approaches for machine translation.

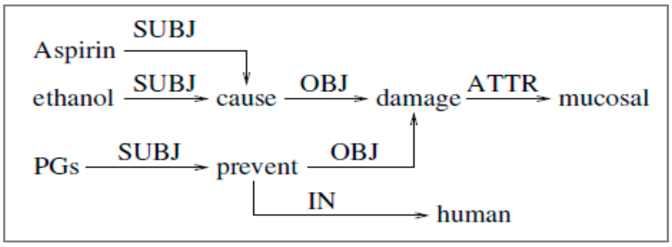
Jong-Hyeok Lee, Karim Benzineb, Jian-Yun Nie

Geetha Basappa, GE India Technology Center

Dominique Maret, Consultant in Information Retrieval and Machine Translation, focused on the issues in improving and measuring machine translation quality. He showed how quality evaluation could be customized to specific problems or user needs, in a way to obtain meaningful quality metrics. Finally, **Emmanuel Jelsch** from Katzarov Patent & Trademark Attorneys reflected on needs in a multilingual patent search from a practitioner's point of view. Jelsch concluded that the quality of machine translated patent has to improve, especially for the Asian languages. **Jian-Yun Nie** from the University of Montreal opened then a discussion with the audience which revealed the gap between the expectations of patent searchers and the current possibilities of machine translation.

Semantic Annotation: Nearer to market?

Borislav Popov from Ontotext presented state-of-the-art annotation tools for measurements. The GATE software was used to create the semantic annotation. Borislav presented systematically the different phases of the implementation from measurement identification, normalization, and the final knowledge representation repository. The talk concluded with a demonstration of the retrieval tool enriched with semantic measurement annotations. The second talk presented semantic annotation techniques used in application for the aircraft maintenance. **Farid Cerbah** from Dassault Aviation argued that semantic annotation will provide end-user with advance search and navigation capabilities exceeding the bag-of-word systems. Then, **Kees Koster** from the University of Nijmegen presented a linguistic approach to conduct patent search with the PHASAR system. The PHASAR system uses linguistic information and display linguistic knowledge to the searcher, i.e. the searchers are allowed to specify different semi-syntactic relations. The system aims to capture dependency relationships between words via the triples, as shown in the figure below. In other words, the dependency triples are compositional structures of a sentence.



"In human PGs prevent the mucosal damage caused by aspirin and ethanol."

By using dependency triples the system aims to expand the bag-of-words to a bag-of-dependency-triples representation of a document. The dependency triples do not try to grasp logical reasoning and inference of each sentence in a document, they rather try to capture the 'aboutness' of a document. However, Koster concluded that the use of dependency triples still requires improvements in the state-of-the-art of syntax analysis in order to cope with heterogeneity of the textual nature of patent. **Michael Granitzer** from Know Center Graz and **Doris Ipsmiller** from m2n - consulting and development gmbh addressed the broad range of tasks associated with patent retrieval such as competitor analysis, trend analysis, prior art search, technology landscaping etc. They presented an application framework which combined statistical and visual methods with ontology driven applications framework. Also here a systematic presentation technique was deployed to give an insight to and transparency of the semantic technology used within IR field: From harmonizing exiting metadata, populating ontologies (e.g. entity and word sense disambiguation) to generating a search result analysis. Workflow techniques were deployed and become more flexible and user driven oriented. In the last talk, **Angus Roberts** from the University of Sheffield showed the application of semantic annotation to biomedical entities using the GATE software. Roberts concluded that life science semantic annotation is more than just annotating genes

Angus Roberts, Michael Granitzer, Doris Ipsmiller, Farid Cerbah

Stephen Adams, Magister Ltd.

and proteins. Linked data can be used to derive statistical models for knowledge discovery. Furthermore, semantic annotation tools and processes need to support a vast range of annotation styles to cope with the textual complexity of scientific literature and patent. The subsequent panel discussion was led by Hamish Cunningham from the University of Sheffield and Gerald Landl, voestalpine Stahl.

Image Retrieval: Light at the end of the tunnel?

Stefanos Vrochidis from the Informatics and Telematics Institute opened the session with a live demonstration of Patmedia, an innovative search engine for patent images that was presented at the IRF Symposium in 2008. **Gerard Ypma** from ASML tested the PatMedia application and presented his analysis: He was impressed by the processing speed of the engine (370 hits within minutes) and appreciated that the image search could be performed independently from the language. Gerard recommended to improve the presentation of results. **Massimo Ruffolo** from the Italian National Research Council presented VIEW, a general purpose approach to information extraction and wrapping from PDF documents that provides strong table extraction facilities. He stated that extracting information from tables and storing it in structured machine-readable form is of paramount importance in patent content analysis as in many other application fields. **Henning Müller** from HES-SO reported on a first experiment to adapt an image retrieval tool developed for the medical field to the patent field. The session, chaired by **Monika Hanelt**, Agfa-Graphics, ended with an open panel debate of patent image retrieval tools, the needs and requirements. There was also an exchange about the difference between what human consider conceptual similar and what is defined similar according to similarity algorithms.

Chemical patents:
Towards genuinely effective searching?

This session was chaired by **Peter Willett** from the University of Sheffield and **Stephen Adams**, Magister ltd. Stating that the most important information in a chemical patent is often the chemical structure disclosed or claimed, **John Barnard** from Digital Chemistry discussed then the prospects for the development of systems and databases to improve the searching of structural information in chemical patents. **Tim Miller** from Thomson Reuters addressed the challenges of indexing specific and Markush entities from patents and showed how the development of new technologies has impacted chemical patent searching over time. **Patrick Ruch** from the University and Hospitals of Geneva represented the text-based retrieval approach in this session: he presented the results of his group at the TREC Chemistry Track 2009. By applying section-specific weighting functions and meta-data filters on top of a state-of-the-art engine, featuring both Boolean and vector-space models and combining this with an original chemical named-entity normalizer, they showed that multi-modal approaches are very promising. **Nicko Goncharoff** from SureChem discussed the role of automated annotation in chemical search and presented ways of making automated annotation sufficiently precise and reliable for the needs of searchers.

The third IRF symposium was wrapped up, from an IR point of view, by **David Hawking** from Funnelback. Hawking concluded that the IP and IR communities now have more of an insight to the problems and methodologies used in the different fields. The next step is to join forces in order to create the future of patent retrieval systems. **Stephen Adams** from Magister Ltd agreed in the IP wrap-up with Hawking. But he also accentuated that these future patent retrieval systems deploying newer IR technologies need to be brought to the IP community as prototypes very soon.

All presentations, pictures and videos of the 3rd IRF Symposium are available online:
www.ir-facility.org/events/irf-symposium/2010/

Next IRF Scientific Conference & Symposium: 6 - 9 June 2011 in Vienna (tbc)