<table>
<thead>
<tr>
<th>Title</th>
<th>Eye of the Tiger</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Nováek, Vít; Groza, Tudor; Handschuh, Siegfried; Decker, Stefan</td>
</tr>
<tr>
<td>Publication Date</td>
<td>2009</td>
</tr>
<tr>
<td>Item record</td>
<td><a href="http://hdl.handle.net/10379/1315">http://hdl.handle.net/10379/1315</a></td>
</tr>
</tbody>
</table>

Some rights reserved. For more information, please see the item record link above.
EYE OF THE TIGER
Irish science has begun to establish itself as world-class in many progressive research areas, such as life sciences. The Digital Enterprise Research Institute (DERI) at NUI Galway has contributed to this trend by our recent achievement in the Elsevier Grand Challenge, a major international competition organised by the leading scientific publishing company.

As any other big player in the now knowledge-based economy, the science and health publisher Elsevier is seeking better ways of dealing with information overload. Conventional technologies can barely handle the vast amounts of data we have to cope with and this is especially true for life sciences, where knowledge comes from many different sources in immense quantities. Therefore, Elsevier invited members of the global scientific community to invent novel tools for interpretation and identification of meaning in the Elsevier’s life sciences publications.

Our team, composed of three DERI researchers (two PhD students and one senior supervisor), has made it to the Elsevier Grand Challenge final with CORAAL, a tool for intelligent search in oncology literature. Almost 70 teams from research institutes world-wide were by-passed, including high-profile competitors like Stanford University and the US National Library of Medicine. That surely proves the strong position of Irish research in the international environment. Read on to know more about this success story!

THE CHALLENGE The overall aim of the challenge was to select the most viable visions and prototypes from academia and bring them to industrial maturity, with support from Elsevier.

The competition had three stages: selection of semi-final proposals, semi-final round and the final. At all stages, the participants have been judged by a panel composed of reputable computer scientists, journal editors, Elsevier representatives and biomedical researchers.

The first stage ran through summer 2008 when 10 semi-finalists were selected out of more than 70 participants based on their initial project proposals. The semi-finalists were then provided with sample Elsevier publication data worked on their prototypes before presenting them to the judges in December 2008. The semi-final was organised as a one day workshop at MIT’s Stata Centre, where entrants presented talks and demos before the judges – a very exciting day, not only because of the remarkable venue, but above all because of the various innovative achievements being presented.

Only four could make it to the final, though, and we were very glad to see recognition of our hard work among the other well-established competitors from four continents. The final round will be organised at a major biomedical conference in April 2009. We, and the other finalist teams, will get worldwide media coverage in selected Elsevier journals and will have an opportunity to advance the Challenge prototypes with Elsevier’s support. In addition, the winner and the runner-up will receive $35,000 and $15,000 in prize money, respectively.

THE IDEA How did we get this far? First, we created quite a daring vision of how we would instruct the machines to better teach us life sciences better, employing our AI expertise in automated knowledge acquisition and integration.

Ireland’s DERI in Galway impresses the semantics world with their solution to make more sense of global life-science knowledge

By Vít Nováček, Tudor Groza, Siegfried Handschuh and Stefan Decker
Apparently, our proposal was not only visionary, but also promising and tangible – and with such enthusiastic reviews, we were among the 10 chosen by the judges. We then had about four months to prove that we could fulfil our visionary promises!

Several hundred thousand Elsevier articles were made available to us. Of those, we selected more than ten thousand that had something to do with cancer research or treatment, since this has been the primary application domain for our research. And then we started to play…

Essentially, we have combined two major research threads being pursued by our team. The first is a novel technology that allows for purely automatic extraction and exploitation of knowledge from arbitrary texts.

By ‘knowledge’, we mean machine-readable representations of concepts (e.g.: ‘cat’ or ‘animal’), their names (e.g: ‘cat’ or ‘felis catus’ associated with the cat concept), and relations between concepts (e.g: ‘cat’ is a ‘type’ of ‘animal’). From the real data processed within the Elsevier Grand Challenge, we were able to automatically find out, for example, that Acute Granulocytic Leukaemia is alternatively called Acute Myelogenous Leukaemia and that it is a different type of disease T-cell Leukaemia. Once our tool extracts and/or infers these facts, it is possible to easily search or browse them (details on how to do that are available at the tool website referenced below).

Our second research interest is architecture for mutually inter-linked publication repositories. The links are given not only by the explicit references, but also by more specific relationships between particular ideas and statements present in the scientific texts.

We combined this architecture with our framework for automated knowledge acquisition and made the incorporated Elsevier content accessible via an intelligent publication search interface called CORAAL (see Fig 1).

Using the interface, users can easily search for knowledge or terms in publications, and browse concepts or articles associated with the current search results. Both knowledge and publication perspectives of the search are mutually connected – one can easily find, for example, authors who write about certain genes that play a role in particular disease proliferation. This is not easy using conventional state-of-the-art tools. They do not expose relationships between concepts, only their names. Therefore, it is necessary to tediously go through many articles containing such names in order to find out those that deal with them in the particular context.

Our challenge prototype can be accessed at http://coraal.deri.ie:8080/coraal/

Typical interaction with CORAAL is outlined in Fig 1 containing three screenshots of the tool in action. The Scr1 screenshot shows the query ‘NOT acute granulocytic leukaemia : is a : T-cell leukaemia’. Using the CORAAL search syntax (see the Quick-Start link at http://coraal.deri.ie:8080/coraal/ for an explanation), the query is supposed to search for resources describing the respective terms as different concepts. In the Scr2 screenshot, we can see a sample of the query result – the negative certainty means negative-type relationship (i.e. difference). There is only one source article for the statement we were looking for. Details of the publication can be displayed by clicking on the respective link, as shown in the Scr3 screenshot.

FEEDBACK After applying the initial CORAAL prototype to the Elsevier data set, we brought it to our potential customers – bioinformatics experts and medical oncologists. They tried the tool and confirmed our hypothesis – they really performed better using CORAAL than with the traditional tools for biomedical literature search.

However, they also made lots of remarks on the actual usability of the interface and the content we expose. We managed to incorporate the most crucial feedback before the semi-final and scooped a success when presenting the result to the judges at MIT. We were named among the four challenge finalists, together with the teams from EMBL, a German bioscience research institute; Carnegie-Melon University (USA); and the Australian IT research institute CSIRO.

For the Elsevier Grand Challenge final, we decided to do a significant face-lift of CORAAL to increase its intuitiveness and applicability. With our typical users in mind – people who are not interested in the arcane processes that are going on ‘behind the scenes’, but who care about what information they can get and how easy it will be to acquire. While we have already worked with sample users, the Challenge judges recommended even deeper cooperation with prospective customers – the life scientists and medical practitioners.

For such an agile improvement of the CORAAL prototype, we organised two workshops with junior and senior medical oncologists. One, already done, involved presentation of the initial revamped version and served to get feedback we may incorporate in the ongoing re-development. The second workshop will involve a larger…
‘As any other big player in the now knowledge-based economy, the publisher Elsevier is seeking better ways of dealing with information overload.'