

# CNGL Undergraduate Students as Researchers Programme PROJECT DESCRIPTION

<b>Institution/Track:</b>	DCM	
<b>Project Title:</b>	Implementing a language modelling approach to information retrieval for structured document search at INEX	
<b>Suitable for students who are studying in the following areas:</b>	Computing, computer science, engineering – particularly those with a strong interest in programming. Some knowledge of information retrieval is desirable, but not essential.	
<b>Skills needed:</b>	Knowledge of C or C++. Knowledge of a scripting language for text processing such as shell scripting or perl Preferably knowledge of XML and DOM/SAX parsers.	
<b>Project Description:</b>	<p>INEX is an international workshop series on XML information retrieval. Each edition of the workshop sets a number of research challenges for information retrieval. The CNGL information retrieval group at DCU plans to participate in INEX 2010. As part of our work for this we need to develop some new software.</p> <p>The project will specify the required extensions to an existing software system and then implement and test this. Specifically it will involve extending the SMART system, which is a well established software application for information retrieval research, to include code to use the language modelling method for information retrieval for the INEX 2010 workshop. The project would involve the following:</p> <ul style="list-style-type: none"> <li>i) Understanding the code of the SMART system.</li> <li>ii) A preliminary study of information retrieval with reference to the vector space model and language models of information retrieval and basic XML retrieval.</li> <li>iii) Enhancing our existing language model implementation in SMART to handle INEX queries.</li> </ul>	
<b>The Role of the student &amp; benefits gained from participation in this project:<sup>1</sup></b>	The student will work with members of the CNGL information retrieval group at DCU. They will gain experience in understanding information retrieval technologies and an introduction to relevant research methods. They will also gain practical software engineering experience in the specification, implementation and testing of an extension to the SMART. SMART is coded in C, and so they will also gain significant experience in working with C. If successfully completed the extended software will be used for CNGL's participation in the INEX 2010 information retrieval workshop leading to co-authorship of the paper describing details of our participation.	
<b>Who will be working with you?</b>	The student will be working closely with Debasis Ganguly, postgraduate research student and Dr Johannes Leveling, postdoctoral researcher in the CNGL information retrieval group at DCU.	
<b>Short description of the group:</b>	The information retrieval group has 6 members: Dr Gareth Jones (group leader), 1 postdoctoral researcher and 4 research students. All members have strong interests in information retrieval research methods and experimentation and system implementation	
<b>Recommended Reading Material:</b>	An Introduction to Information Retrieval by Manning, Schutze, and Raghavan. Papers/thesis on the Language Modelling for information retrieval such as those by Ponte & Croft and Hiemstra & Kruj. Overview papers describing the activities of previous INEX workshops. Also papers from the Focused Track at previous INEX proceedings workshops.	
<b>Other information:</b>		
<b>For further details on this project please contact:</b>	<b>Name:</b> <b>Phone:</b> <b>E-Mail:</b> <b>Website:</b>	<b>Dr Gareth Jones</b> <b>01 700 5559</b> <a href="mailto:Gareth.Jones@computing.dcu.ie">Gareth.Jones@computing.dcu.ie</a> <a href="http://www.computing.dcu.ie/~gijones">www.computing.dcu.ie/~gijones</a>

<sup>1</sup> *This is an initial description of the role of the student and it is liable to change following discussions between the investigators and the student.*