Number 89, April 2012

ERCIM NEWS

www.ercim.eu

Special theme:

Big Data

Also in this issue:

water current models from SOAP messages, can also be integrated. The system includes a database in which all SAR units at disposal of a Coast Guard Station are stored with details of their salient features. A friendly graphic user interface has been designed, where several graphic layers of information can be overlaid or hidden in visualization.

Other significant and innovative extensions of the IAMSAR procedures have been carried out during the development of the ICT-E3 project, and are now subject to a patent application. We wish to acknowledge the collaboration with the personnel of the Mazara del Vallo Coast Guard Station, their invaluable help and stimulating suggestions have been fundamental to the success of the activity.

Link: http://ceur-ws.org/Vol-621/paper21.pdf

Please contact:
Massimo Cossentino, Carmelo Lodato, Salvatore Lopes, Umberto Maniscalco
ICAR-CNR, Italy
E-mail: cossentino@pa.icar.cnr.it, c.lodato@pa.icar.cnr.it, s.lopes@pa.icar.cnr.it, maniscalco@pa.icar.cnr.it

Salvatore Aronica, IAMC-CNR, Italy
E-mail: salvatore.aronica@iamc.cnr.it

Wikipedia as Text

by Máté Pataki, Miklós Vajna and Attila Csaba Marosi

When seeking information on the Web, Wikipedia is an essential source: its English version features nearly four million articles. Studies show that it is the most frequently plagiarized information source, so when KOPI, a new translational plagiarism checker was created, it was necessary to find a way to add this vast source of information to the database. As it is impossible to download the whole database in an easy-to-handle format, like HTML or plain text, and all the available Mediawiki converters have some flaws, a Mediawiki XML dump to plain text converter has been written, which runs every time a new database dump appears on the site with the text version being published for everybody to use.

The KOPI Plagiarism Search Portal was developed by the Department of Distributed Systems (DSD) of MTA SZTAKI in 2004 as a plagiarism checker tool for carbon copy plagiarism cases. In 2010, the system improved by adding a unique feature to the search engine, the capability to find translated plagiarism. For this function it was necessary to include the whole English Wikipedia, as the number one source of potential plagiarism, into the database.

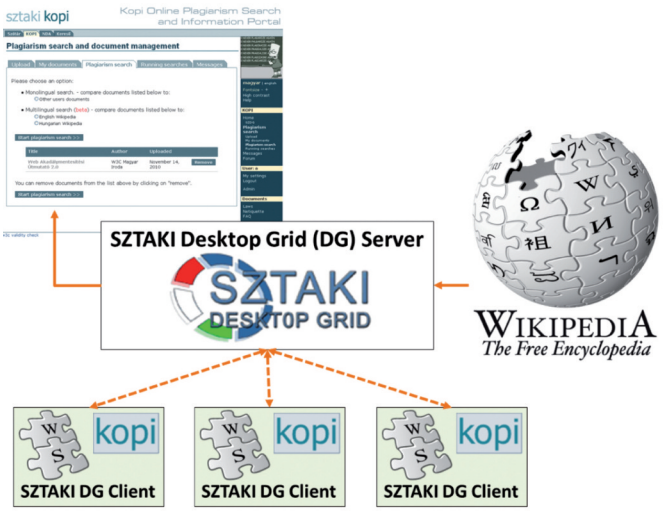
When the Wikipedia dumps were first downloaded from the server, all possibilities of converting them easily and quickly to plain text were examined, as plain text is easy to manipulate. Consequently, the whole content had to be run through a series of language processing steps. In most cases the available converters were not suitable for handling a larger chunk of XML dump, or the output was error prone and in many cases – when using, for example, a Mediawiki instance as a converter – the conversion was slow.

Consequently a new converter needed to be written based on the following features:

- article boundaries have to be kept
- only the textual information is necessary
- infoboxes – as they are duplicated information – are filtered out
- comments, templates and math tags are dismissed
- other pieces of information, like tables, are converted to text.

Wikipedia dumps are published regularly and, as the aim is to be up-to-date, the system needed an algorithm which is able to process the whole English Wikipedia in a fast and reliable way. As the text is also subject to a couple of language processing steps to facilitate plagiarism search, these steps were included and the whole processing moved to the SZTAKI Desktop Grid service (operated by the Laboratory of Parallel and Distributed Systems), where more than 40,000 users have donated their free computational resources to scientific and social issues. Desktop grids are usually suited for parameter study or "bag-of-tasks" type of applications and have other minor requirements for the applications in exchange for the large amount of "free" computing resources made available through them. SZTAKI Desktop Grid, established in 2005, utilizes mostly volatile and non-dedicated resources (usually the donated computation time of desktop computers) to solve compute intensive tasks from different scientific domains like mathematics and physics. Donors run a lightweight client in the background which downloads and executes tasks. The client also makes sure that only the excess resources are utilized so there is no slowdown for the computer and the donor is not affected in any other way.

The Mediawiki converter was written in PHP to support easy development and compatibility with the existing codebase of the KOPI Portal. The main functionality could be implemented with less than 400 lines of code. The result was adapted to the requirements of the desktop grid with the help of GenWrapper, a framework specially created for porting existing scientific applications to desktop grids.



SZTAKI Desktop Grid service

GenWrapper's primary goal is lowering the porting effort required for 'legacy applications', which either have no source code available or making changes to them is infeasible. In case of KOPI it allowed the development of the converter independently from the desktop grid and the result could be effortlessly deployed.

This new arrangement allows Wikipedia in any language to be converted and preprocessed in a couple of days. As these text versions can be used for several other purposes as well, they are shared and made available to everybody. Currently one can download the English (5.7 GB), Hungarian (300 MB), German (2.2 GB) and French (1.5 GB) versions (sizes are gz compressed size). Based on the project work and plans, other languages will follow shortly.

Links:
Wikipedia as text download link:
http://kopiwiki.dsd.sztaki.hu
KOPI portal: http://kopi.sztaki.hu
SZTAKI Desktop Grid: http://szdg.lpds.sztaki.hu
GenWrapper: http://genwrapper.sourceforge.net

Please contact:
Máté Pataki? MTA SZTAKI, Budapest, Hungary
Tel: +36 1 279 6269
E-mail: mate.pataki@sztaki.hu

Genset: Gender Equality for Science Innovation and Excellence

by Stella Melina Vasilaki, FORTH/IACM

GenSET was an innovative project aiming to improve the excellence of European science through inclusion of the gender dimension in research and science knowledge making. It functioned as a forum for sustainable dialogue between European science leaders, science stakeholder institutions, gender experts, and science strategy decision-makers, to help implement effective overall gender strategies. The goal was to develop practical ways in which gender knowledge and gender mainstreaming expertise can be incorporated within European science institutions in order to improve individual and collective capacity for action to increase women's participation in science.

Between March and June 2010, three genSET Consensus Seminars brought together 14 European science leaders to share knowledge and experience and arrive at a consensus view on the gender dimension in science and on the priorities for gender action in scientific institutions. The Science Leaders Consensus Panel represents extensive knowledge of different scientific fields and sectors, with over 500 years of scientific and leadership experience; involvement in appointing over 4000 researchers; direction of over 300 major research programmes and research funding of over €500 million; executive decision making through over 100 Executive Board positions; and research publication record exceeding 1000 peer reviewed research papers. They collaborated with a group of equally high-ranking gender experts, who provided expertise through lectures and research evidence during the Consensus Seminars.

The consensus recommendations call for action in four priority areas of the gender dimension in science: science knowledge making, deployment of human capital, institutional practices and processes, and regulation and compliance with gender□related processes and practices. All of these recommendations are meant to be included within an overall institutional science strategy. The work of the Science Leaders Panel has highlighted only the beginning of an important dialogue between gender experts and leaders of scientific institutions.

Here below there is a summary of the consensus recommendations:

- *Recommendation 1:*
  Leaders must be convinced that there is a need to incorporate methods of sex and gender analysis into basic and applied research; they must "buy into" the importance of the gender-dimension within knowledge making.
- *Recommendation 2:*
  Scientists should be trained in using methods of sex and gender analysis. Both managerial levels and researchers should be educated in such sex and gender analysis. Training in methods in sex and gender analysis should be integrated into all subjects across all basic and applied science curricula
- *Recommendation 3:*
  In all assessments – paper selection for journals, appointments and promotions of individuals, grant reviews, etc. – the use and knowledge of methods for sex and gender analysis in research must be an explicit topic for consideration. Granting agencies, journal editors, policy makers at all levels, leaders of scientific institutions, and agencies responsible for curricula accreditation, should be among those responsible for incorporating these methods into their assessment procedures.
- *Recommendation 4:*
  Research teams should be gender diverse. Institutions should promote gender diversity of research teams through a variety of incentives (eg quality recognition and allocation of resources) and through transparency in hiring.
- *Recommendation 5:*
  Gender balancing efforts should be made in all committees, with priority given to key decision-making committees. Panels for selection of grants and applicants must be gender diverse. This must be the goal for management teams as well.
- *Recommendation 6:*
  Institutions should seek to improve the quality of their leadership by creating awareness, understanding, and appreciation of different management styles. This can be achieved through training, self-reflection, and various feedback mechanisms. Diversity training, specifically, is essential in this process.
- *Recommendation 7:*
  Women already within scientific institutions must be made more visible. All public relations activities from scientific