R-Tools: Mediawiki extension for full-scale statistical computing

Juha Villman
National Institute for Health and Welfare
Department of Environmental Health
Nelaniementie 4, 70701 Kuopio, Finland
+358 29 524 6442
juha.villman@thl.fi

Einari Happonen
National Institute for Health and Welfare
Department of Environmental Health
Nelaniementie 4, 70701 Kuopio, Finland
+358 29 524 6846
Einari.happonen@thl.fi

ABSTRACT

Wikisystems are proven to be good for producing text and knowledge in collaborative manner but they are not designed to handle large amounts of numerical data. We needed a system that is capable for producing text and run calculations from datasets. For this purpose we created Opasnet which is a Mediawiki with integrated statistical computing extension and an external database for data. In our demonstration we will show how R (statistical software) can be integrated into Mediawiki as an extension (R-Tools) and how it can be used directly from wiki pages. This extension enables users to write R-code, run it and see the results of the calculation on the wiki page. R-tools can use data from external databases and this functionality is also demonstrated. First R-Tools demonstration was held at Wikisym 2012 in Linz. Now we will focus on its new features developed within this year.

Categories and Subject Descriptors

H.2.8 Database Applications: Scientific databases, Statistical databases; H.5.3 Group and Organization Interfaces: Collaborative computing, Web-based interaction

General Terms

Design, Experimentation

Keywords

Mediawiki; Open knowledge; Open data; R; Opasnet; MySQL

WikiSym '13, Aug 05-07 2013, Hong Kong, China

ACM 978-1-4503-1852-5/13/08.

http://dx.doi.org/10.1145/2491055.2491100

1. INTRODUCTION TO OPASNET

Opasnet is a website (running on Mediawiki) and open workspace for a mass collaboration project that aims to improve societal decision-making. Opasnet is hosted by THL (National Institute for Health and Welfare, Finland), The Department of Environmental Health. It has been used for almost 5 years as a main workspace in our modelling and assessment unit. The original motivation was to improve environmental health assessments and thus decisions related to environment and health. However, as the methods have developed and the project has grown, the scope has been widened to policy-making in any field.

Opasnet is based on open participation by anyone interested, free distribution of information, and strict application of the scientific method. Main idea behind Opasnet is that assessments should no longer be done in closed expert groups that produce static reports that may or may not answer the questions a decision-maker actually has, and that are only as credible as the expert group is. Instead, two improvements are needed. First, an assessment should be built on an explicit information need that is defined by an open deliberation between experts, decision-makers, and stakeholders. Second, everything in the assessment - including premises, data sources, modelling, and conclusions - is open to scientific criticism.

2. R-TOOLS: INTEGRATION FOR STATISTICAL COMPUTING IN MEDIAWIKI

Mediawiki is designed mainly for encyclopedia use so it is good for displaying text and images. Assessments in Opasnet require also text and images but often we need massive amounts of numerical data and software to run statistical analyses from that data. For that purpose we have developed R-Tools Mediawiki extension and Opasnet Base database.

R is a programming language and software environment for statistical computing and graphics that is widely used for statistical and data analysis. It is free software that compiles and runs on a wide variety of UNIX platforms, Windows and OS X. Off-the-shelf (and through using selected "packages") it is possible to use the software for Monte Carlo simulation, production of complicated graphics (over which the user has complete control) and a host of other applications.

R-Tools extension makes it possible to write and run R-code directly within wiki page. Using R-Tools is possible to do a vast variety of statiscal analyses as well as display data on Google Maps. Starting a R-code run using R-Tools generates a process in special R-server and output of the calculation is generated into a special wiki page or within the same page. R-tools can also generate graphics and diagrams if needed.

For users who don't want to learn R-code we have developed a special user interface. Using this interface it is possible to easily change parameters used in R-code without changing the code itself and run models with custom parameters. This all can be done directly from a wiki page.

R-Tools uses R platform that has been installed on a separate server. Extension is being distributed freely so it can be installed on every Mediawiki project where it is needed. It is also possible to use our R-server from other Wiki installations.

3. OPASNET BASE – DATABASE FOR ALL THE DATA IN THE WORLD

Calculations made using R-tools often require huge amounts of numerical data. It is quite obvious that this data cannot be soundly stored in wiki pages. For this purpose we have developed Opasnet Base.

Opasnet Base is technically a migration of MySQL and Mongo databases. It is designed to be flexible enough to store information in almost any format: probability distributions or deterministic point estimates; spatially or temporally distributed data; or data with multiple dimensions. It can be used as a direct source of model input data, thus making it possible to use shared input information sources such as population data, climate scenarios, or dose-responses of pollutants. Opasnet Base is integrated into Mediawiki as an extension which has its own special page for browsing and uploading the data.

We have developed special libraries to R making it possible for R-Tools to easily use store and access data from Opasnet Base and use this data within wiki pages.

4. DEMONSTRATION

On our demo we will focus on R-tools extension and its new features. Especially the focus will be on new and graphical features that haven't been introduced earlier. Only very basic of Opasnet concept is being explained. Demo will show to the audience how easy it is to alter and generate R-code and run your code directly from the Mediawiki page. We will also demonstrate a special user interface which enables users to alter variables used in R-code. This makes it possible for users without any skills in R to run models with their own set of parameters.

In addition to R-Tools demo also some details about Opasnet Base basic usage is being demonstrated. We will show how you can upload your own data into database and how to use that data in R models.

Hopefully the audience will benefit from the demonstration in many ways: they will learn that R is a very capable software for any scientific calculation needs and it can be considered as a substitute for commercial statistical software. Audience will also see that developing models and calculations in an open platform with crowd sourcing could be much more efficient compared to working with your models on your own.

R-Tools is under continuous development so we hope to get good and critical feedback from the audience so we can enhance the software even more further. So far we have worked with relatively small group of people so it is essential to know what larger crowd think how we should change and developed R-tools features and usability..

5. TECHNICAL REQUIREMENTS

Demonstration does not require any special technical requirements. Our system should work on any modern computer and browser with internet access. Opasnet can be accessed from: http://en.opasnet.org.