SAB Newsletter

Added by Steven Vale, last edited by Steven Vale on 19 Oct, 2009 ()

MSIS Sharing Advisory Board

Software Sharing Newsletter, Issue 1: October 2009

In This Issue:

- Introducing the SAB
- Collaborating on Software Projects
- GIS: Statistics and Open Source Software
- What is R?
- RELAIS: An Open Toolkit for Record Linkage

Editorial

(Marton Vucsan, SAB Chairman)

Welcome to the first newsletter of the MSIS Sharing Advisory Board (SAB for friends)!

The SAB can trace its roots back to the last century, when a United Nations Economic Commission of Europe (UNECE) initiative called the Statistical Computing Project marked the beginning of international collaboration to try to create computer programs for the production of official statistics. In those days the political situation was very different. Travel in large parts of Europe was restricted, some countries were rich, others were poor or were starting on a different road. It was felt that collaboration between nations in the UNECE region could be very beneficial to emerging economies, whilst also giving benefits to richer countries. This collaboration has continued into the current century albeit in different forms, including the annual meetings on the Management of Statistical Information Systems (MSIS), in which IT directors and specialists from national and international statistical organizations exchange ideas and information about their statistical computing practices.

MSIS has established the SAB to coordinate and encourage collaboration between organizations with a goal very close to the original one: To unlock for one another the knowledge and the programs to create statistics with the emphasis on the principle that emerging economies are entitled to benefit from the work of others. Our role in this process will be that of facilitator, creating tools such as a web repository. We will guide and support the creation of a common reference architecture to ensure that the different components we are sharing will fit together efficiently.

For this to succeed, the span of attention needed is much longer than that of a single project. It is a strategic activity which has to survive individual projects and has to accommodate the fact that individual projects will only partly deliver on these goals and primarily serve other more direct goals.

So here is our first newsletter. It includes content from many sources, including Eurostat, who also gave us the initial template of the newsletter (thanks!). We hope you will enjoy this newsletter and encourage the people in your organisation to read it and to help us to achieve what we have set out to do. We also encourage you to submit articles on relevant topics for future editions.

Introducing the SAB

The MSIS Sharing Advisory Board (SAB) celebrated its official launch at the Eurostat IT Directors Group meeting on 21/22 October 2009.
The mandate for this work comes from the Conference of European Statisticians, who are closely following our progress.

Highlights of the current work programme include setting up a brokerage facility for those interested in collaborating on development projects and maintaining an inventory of existing software and initiatives that are open for sharing. The SAB will also study models for collaborative development projects, including open source and other approaches. We will develop guidelines based on experience and good practice. The full work programme is available here.

The members of the SAB are introduced here, however, everyone can contribute via the MSIS wiki. For a user account, please contact us via sab.stat@unece.org.

	Marton Vucsan, Statistics Netherlands (Chairman) Marton has worked in statistical computing for 31 years and is currently senior project manager responsible for things like process integration and enterprise architecture of the statistical processing domain.
	Antonio Consoli, Eurostat Antonio has a background in IT and he is currently project manager for systems architecture in Eurostat's unit managing information systems for statistical production.
	Karen Doherty, Statistics Canada Karen has 30 years experience in IT in a variety of roles and is currently the Director General, Informatics, at Statistics Canada.
	Trevor Fletcher, OECD Trevor is currently Head of the Statistics Information Management & Support division at the OECD Statistics directorate. Previously led the Analytical/Statistical system development group.

Alistair Hamilton, Australian Bureau of Statistics
Al has a background in information architecture (data and metadata), and currently heads the data management section, working on a roadmap for statistical information management in the 21st Century.

Carlo Vaccari, ISTAT
Carlo is head of the software development division in ISTAT. He has 30 years of experience in public and private companies on software, databases and IT architectures. Father (3), blogger and biker.

Rune Gloersen, Statistics Norway
Rune is currently Director of the IT and Statistical Methods Department in Statistics Norway.

Valentin Todorov, UNIDO
Valentin has a background in mathematical statistics and software engineering as well as industrial and academic experience, and is Management Information Officer in the Research and Statistics Branch of UNIDO.

Steven Vale, UNECE
Steven is currently head of the User Services Section of UNECE, responsible for statistical information systems, dissemination, metadata, quality and other cross-cutting methodological issues.

Collaborating on Software Projects

(OECD)

Since the development and implementation of the OECD Statistical Information System (SIS) for the management and dissemination of statistical data and metadata, we have received many enquiries from other organisations with similar

statistical data management requirements to see if the software could meet their needs. Several of these bodies have since adopted the SIS for their own use or are currently evaluating the software. SIS consists of four major components:

- *StatWorks* - manages the import and validation of raw data and the calculations of output series
- *Metastore* - manages all metadata content
- *OECD.Stat* - the central repository for all statistical data and metadata dissemination
- *Pubstat* - publication and dissemination of statistical outputs.

There are several reasons to share statistical or other software. The most obvious is the savings in time and money in re-using or adapting an existing technical solution to meet processing requirements. Any new features subsequently developed by one of the collaborating parties can then be shared and re-used by other partner(s) at minimal cost. Secondly, the use of common shared systems can promote the use of statistical standards, for example, SIS components widely use the Statistical Data and Metadata eXchange (SDMX) standard for inputs and outputs. Using the same software platform for statistical data management also facilitates the production of joint data collection or dissemination exercises between organisations.

GIS: Statistics and Open Source Software

(Eurostat, edited by P. Cavallini, http://www.gfoss.it/)

Statistics is a wide reaching and multidisciplinary field applied to a variety of different sectors, from finance to market, from forecasting to geo-statistics to understand demographic changes. This last area has seen notable successes in recent years thanks to open source software.

There are many types of software that manage geographical information, but most of them only collect coordinates to represent points on maps. However, some collect more points of interest for citizens and are referred to as GIS (Geographical Information System) software. GIS software is not only interesting because it applies algorithms, e.g. to find the minimum distance between two points, but also because it can associate user needs with maps and perform a market strategy. Collecting the right data is an important process that stands before all possible strategies.

During recent years, the GIS sector has become an interesting field for both open source developers and public administrations. Open source developers can produce customized software dealing with heterogeneous data while public administrations take advantage of open source software for dynamic maps, and can save money. A lot of communities exist around open source GIS software and could be merged. They have the same interests but have developed their software in different ways.

Fundamentally there are two types of GIS software, standalone applications and web applications. The number of standalone applications is lower than the number of web applications, though the former are more widespread. Some standalone applications are already quite mature, but all of them deserve attention. Most of them share the same libraries to represent and transform data and can run on a variety of systems. GRASS (Geographic Resources Analysis Support System) is a remarkable example of an open source application that has been continuously improved. It can handle raster, topological vector, image processing, and graphic data. It was developed and maintained by the U.S. Army from 1982 to 1995. Since then, GRASS was picked up by academics and the project now has its headquarters in Trento, Italy. It can be used from the command line, or through several graphical front ends, most notably QuantumGIS.

MapServer is a web application originally developed by the University of Minnesota in cooperation with NASA, which needed a way to distribute its satellite images. MapServer is oriented to represent spatial data (maps, images) for the web. Another good example is OSGeo. A large community has been created under the name "Open Source Geospatial Foundation", a non-profit organization, where users and developers exchange ideas while contributing to project development. The OSGeo works as incubator for new projects. It has been sponsored by different companies such as Autodesk, the company which produces the dominant CAD software.

The proliferation of these open source projects raises two problems:

1. Common standards to exchange data between GIS applications. This is a huge problem since the use of GIS software started with each country developing their own geographical interfaces.

2. Getting data and creating up-to-date, accurate databases of geospatial information is expensive. In this case there is real need to share data.

The Open Geospatial Consortium (OGC) is leading the development of standards for geospatial and location based services. Many countries have already joined. Open source software is perfect to act as a glue between multiple proprietary systems ensuring interoperability in the long-term. An example of adoption took place in Spain at the end of 2002, when the Council of Infrastructure and Transportation began a global migration towards open systems. One example of new software was a GIS tool, gvSIG, developed from scratch to fulfill the needs of the Council. GvSIG conforms to OCG standards and European Union directives.

At European level, the IDABC programme finances the Open Source Observatory and Repository (OSOR) project. The repository is a first approach to push public administrations to use open source software (GIS included) from all over Europe, sharing data and possible solutions.

In conclusion, the GIS sector is becoming more and more interested in open source projects. OSGeo is a first and very important step to provide a common platform where it would be possible to exchange ideas and work of the GIS community.

References:

- State of open source GIS software: http://2009.foss4g.org/presentations/
- Open source Tools for GIS professional: http://www.gisdevelopment.net/magazine/
- GRASS GIS: http://grass.osgeo.org/
- QuantumGIS: http://qgis.org/
- MapServer: http://mapserver.gis.umn.edu/
- OSGeo: http://www.osgeo.org/
- OSOR: http://ec.europa.eu/idabc/en/document/6728/

What is R?

(UNIDO)

The world of commercial statistical software is dominated by a few very well known names - SAS, SPSS, Stata, etc. The situation with free software is similar. Although there are hundreds of free tools for solving specific statistical problems, if we talk about a comprehensive statistical environment which could compete with the main commercial packages, the choice is very limited and the main option is R.

According to the R-core development team, R is "*a system for statistical computation and graphics. It provides, among other things, a programming language, high-level graphics, interfaces to other languages and debugging facilities*". R is an open source implementation of the S computer language (developed by John Chambers and others in the 1980's). The development of R is managed by the R-core team. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.

R provides a wide variety of statistical and graphical techniques - linear and non-linear modeling, classical statistical tests, time-series analysis, classification, clustering, robust methods, etc. Hundreds of specialized statistical procedures for a variety of applications are available from the Comprehensive R Archive Network (CRAN) in the form of contributed R packages, which can be downloaded in source form or installed directly from the R console. A few examples relevant for national and international statistical organizations are: survey analysis (*survey, pps, sampling, sampfling*), handling of missing data (*VIM, mice, mi, mvnmle, mitools, EMV, mix, pan),* time series analysis, robust

statistics *(robustbase, rrcov, robust).*

More information on R can be found at the CRAN web site. You can also read a brief overview of R in this paper.

RELAIS: An Open Toolkit for Record Linkage

(ISTAT)

Record linkage is a process that aims to quickly and accurately identify if two (or more) records represent the same real world entity. Record linkage can be performed for different purposes making it a powerful instrument to support decisions in business and government institutions.

RELAIS has been designed and developed to decompose the record linkage process into its constituent phases and to address each of these phases with the technique most appropriate to the nature of the data. The constituent phases are:

- Pre-processing of the input files
- Choice of the identifying attributes (matching variables)
- Choice of the comparison function
- Creation of the search space of link candidate pairs
- Choice of the decision model
- Selection of unique links
- Record linkage evaluation

Several record linkage systems and tools have been proposed, in both the academic and private sectors. Some provide the user a certain degree of flexibility, e.g. allowing a choice of comparison functions. However, none of them allow multiple choices for each record linkage phase, or dynamically build a record linkage workflow by combining the most appropriate technique at each phase.

The RELAIS toolkit comprises a collection of techniques for each record linkage phase, which can be dynamically combined to build the best record linkage workflow, given a set of application constraints and data features. The current version (2.0) includes the following functionalities:

- Creation of the search space of candidate pairs by means of the cross product of the input files;
- Search space reduction using a "blocking" method or a "sorted neighbourhood" method;
- Choice of the matching variables;
- Choice of a distance function for approximate string comparisons;
- Deterministic decision models (Equality or Rule based);
- Probabilistic decision model (Fellegi and Sunter);
- Reduction from N:M to 1:1 matching solution;
- A data profiling phase that helps the user to choose the best blocking or matching variables.

RELAIS is an open source project. There are at least two reasons for this choice. First, there are many possible techniques that can be implemented for each of the record linkage phases: relying on a community of developers such a set can be increased and maintained very rapidly. Second, in recent years there have been several independent efforts to develop record linkage projects for specific purposes, which have not led to the best generic solution. An open source approach gives the possibility to gather together work already done and make it available to the community for the most appropriate usage.

RELAIS is mainly implemented in Java, due to the well-known features of strongly typing and platform independence; some phases are implemented in R, (see previous article). It has been implemented using a relational database architecture, based on a MySQL environment, which is also in line with the open source philosophy of the RELAIS project.

For RELAIS software and documentation, see: http://www.istat.it/strumenti/metodi/software/analisi_dati/relais. There

is also a [Relais page](#) on OSOR.

Labels

(None)

Powered by a free **Atlassian Confluence Community License** granted to UNECE. Evaluate Confluence today.

Powered by [Atlassian Confluence](#) 2.8.1, the [Enterprise Wiki](#).