Distr.
GENERAL

Working Paper No. 14
26 March 2008

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2008)
(Luxembourg, 7-9 April 2008)

Topic (iii): Exchange/sharing/re-use of components, common models among statistical offices

NSI/ISI Statistical software: Issues and a way forward to maximise re-use and minimise integration efforts

Invited Paper

Prepared by Andrea Toniolo Staggemeier, ONS, UK

Abstract

1. This paper will present short case studies from the Office for National Statistics (ONS) in the United Kingdom when integrating Statistical Software provided by other statistical agencies, as well commercial off the shelf packages from third parties. The paper will discuss the following main concerns:

(a) There is some great work being done within National Statistical Organisations on specialised statistical software. This is great software and works very well.

(b) The challenge is that it is hard to predict what the long term support will be, whether there will be updates for the software, and how additional functionality can be added to meet specific requirements.

(c) So the question to be resolved is - how do we turn very high quality 'unsupported' software into very high quality software with a real and guaranteed future that we would all be happy to invest in?

2. Examples will be given using a variety of statistical value chain services, from Data Collection, Data Preparation, Primary and Further Statistical Analysis, and Statistical Disclosure Control. The paper will propose areas where standardisation could be achieved amongst National and International Statistics Institutes (NSI/ISIs) to reduce the burden on using Statistical tools regardless of the IT architecture of choice.

I. INTRODUCTION

3. Historically NSI/ISI(s) have been sharing new methodologies and IT products or best practices for a long time. ONS encourages professional development and sharing knowledge as a matter of principle. However, when efforts to integrate NSI/ISIs software products into high profile projects start to outweigh the cost of building it with preferred technologies, concerns are often raised regarding the merits of these collaborations.

4. Back in 2003-2005, ONS undertook a major review of the statistical methodologies available within several statistical packages produced by a variety of organisations, i.e. national/international institutions, and third party commercial off the shelf statistical packages.

5. Extensive evaluations were carried out to establish the statistical quality of the different methodologies implemented within NSI/ISIs products, but little IT architecture investigations were undertaken, creating an unbalanced view.

6. The problem IT departments encounter now is centred on the fundaments of systems integration, serviceability, supportability, and is aggravated by a mixture of technologies used to build statistical products/services[1].

7. This surely is not only an ONS problem, as any other organisation wanting to take advantage of the knowledge communities would face the same dilemma should they take what is already available and try to blindly support some of the statistical products. The same is true if they take the methodology definitions and redevelop them to suit each individual IT standard and need.

8. The short case studies in this paper are not aimed to detail all the issues one may have with any of the software used at ONS, but to stimulate thought on how we could piece this information together in order to resolve the issues.

9. A Service Oriented Approach (SOA) was recently presented (Duoba, 2006) for the development of statistical packages using the Statistical Value Chain (SVC) as the NSI(s) business process model. After the presentation, ONS and EuroStat have received interest from other NSI/ISI(s) to collaborate using this approach. Expressions of interest came from Latvia, Estonia, OECD, Norway, Germany, Eurostat/UNECE, Ireland,and the Netherlands. Do the majority of the countries that manifested their interest have similar problems integrating and supporting statistical applications from other agencies as ONS does? If so, clearly the need to articulate our preferences for a sustainable model of production of Statistical systems needs to be reconsidered by the present main providers of statistical systems and services.

10. A possible reason why some of the NSI/ISI(s) which developed statistical systems did not come forward could be associated with the nature of how some of their systems have been developed, i.e. a research type approach to development, and therefore large and substantial investments made by the providers so that each tool could be used through business service layers of applications in a seamless way.

11. Therefore, is paramount that IT departments of NSI/ISI(s) communicate better with each other and in the same forum where statistical modelling of systems currently interacts.

12. For instance, should a clear statistical model metadata be available, this could be used as centre of common understanding. In principle, this may sound like a large job to may of the readers of this paper, but presumably, and most definitely, it is a matter of making the existing metadata for the existing statistical systems available to other IT/MD departments. This would allow new systems that integrate those components to be designed in the most appropriate manner for each of the business needs.

13. The next section will elaborate on the Statistical Value Chain (SVC) steps shared amongst other institutes.

[1] The term Statistical Service can be interchanged with Statistical Products and is used to highlight the potential mistreatment of statistical models.

II. CASE STUDIES

A Data Collection

14. Blaise is the ONS social survey data collections software of choice. In the last few years ONS has dedicated major efforts to harmonise and consolidate the majority of its social statistics collection system. This project is called Field Data Collection Modernisation (FDCM), and is in preparation for the new Integrated Household Survey (IHS).

15. The new Blaise 4.8 architecture and scheduled further enhancements have been driven by a strong Blaise international user group community, which ONS Social and Vital Statistics Division actively participates. Unfortunately, this community deals more intensively with methodological and functional requirements than architectural and non functional requirements of the systems, therefore the ability to influence its development architecture in future releases is left entirely to the provider's choice. This may also be influenced by the vendors (Statistics Netherlands) current IT strategy and skill set.

B Editing and Imputation

16. ONS has chosen CANCEIS (Canadian Census Editing and Imputation System) and Banff (business data imputation system also provided by Statistics Canada) as the preferred packages for each distinct data type, i.e. demography and social data using CANCEIS, and business data using Banff. These two products are developed using different architecture approaches, one based on C/Windows/Sybase technologies and the other using C/SAS/Windows/Sybase.

17. One would expect that since the two products are from the same statistical value chain step they would share some common metadata model, or architecture principles. Unfortunately, that is not the case. Moreover, new development plans from Statistics Canada for CANCEIS takes the two products even further apart, at present CANCEIS is planned to be redeveloped in C#.

18. Integration between the tools in the majority of the cases is through manipulation of many different file formats incurring expensive transformations to data before it is consumed and after is been through the engines.

C Time Series Analysis

19. X12-ARIMA and more recently X13-ARIMASEAT tools have been used at ONS for many years. The recent modernisation programme for the UK National Accounts systems has developed a wrapper which integrates the latest version of X12-Arima tool (v177) as a black box. A few weeks later, another release was made available, with more functional enhancements which were desirable by our statisticians. The inability to communicate releases plans with new enhancements and list of bug fixes is probably the last thing in the minds of many day to day users, but it can easily bring a project to its knees.

20. Changing a release version should be easier if the API invocations were preserved and the added features provided through parameter driven options where backward compatibility is preserved. That is not only a problem we have observed using time series and seasonal adjustments tools, but also in a variety of other products.

D Statistical Disclosure Control

21. ONS actively funded the development of Tau-Argus and Mu-Argus, statistical disclosure packages for tabular and record level data types. Actively engaged in the CASC/CENEX/ESS projects, ONS helped to identify and define new methodologies in this field. In close collaboration with Statistics Netherlands and two private companies, Dash Optimization and Space-Time Research, ONS has helped to enable Tau-Argus functionality to be accessible through a more stable and robust tabulation engine (see Staggemeier et al (2007)).

22. Similarly to Blaise, CANCEIS, Banff, and X12-Arima tools, Tau-Argus, and by consequence Mu-Argus, presents signs of all the issues already reported. A particular concern ONS has is the lack of a license

agreement. Tau-Argus does not support any standard license agreement and comes with no guarantee that it will work. It also does not have any means of attaching any kind of support agreement to it. ONS has, on a number of occasions, tried to help to resolve this matter but recently has been obliged to write a disclaimer explicitly exempting themselves of any responsible for publications produced using Tau-Argus or Mu-Argus outputs.

III. Proposal for consideration

23. Create an IT development community amongst NSI/ISI(s) interested in making available statistical services/products.

24. Establish a governance agreement which comprises a sustainable development and support model for any service made available to the community.

25. Community members should establish a common development standard.

26. Here are other principles to be taken into consideration by community members:

1. Any statistical service should include enough methods to encompass the needs of the parties of the cooperation
   1.1. Be extendable to add new methods (parties own methodologies)
   1.2. Be generalised to fulfil all significant needs of the parties
2. Any statistical service created and made available by a community member should also publish full API(s) of the software enabling better integration
   2.1. when new release developments are planned the systems should first consider a SOAP approach
3. Statistical Standards and guides from international agencies should be use and new requirements for national standards proposed should be made public to all participants of the development community
4. Common vocabulary, metadata models, and data definitions are coherent and consistent at all statistical value chain building blocks
5. Ensure integrity, confidentiality, and security of systems and data at all times
6. User access through consistent and easy to use interfaces and from any appropriate languages
7. Sustainable agreement on maintenance and cooperation of the developed statistical services
   7.1. Procedure for inclusion of needs of other parties of the cooperation.
   7.2. Assurance of maintenance of the system (time scope)
   7.3. High level support assuring continuity.

IV. Conclusion

27. It is possible to create a community source for statistical services as long as the members are committed to continue support for this initiative.

28. A “redevelop from scratch” approach will not capture the interest of current service providers.

29. A more steady approach might bring better long term understanding of the issues that users of the products face, but may result in an increase of development costs and time.

30. The alternative proposed here is a steady-state convergence from isolated technologies and IT principles to a more sharable and joint initiative. The potential benefits to community members are:

- Realisation of the technical aspects of the statistical systems modernisation by harnessing the development and methodological capabilities of a number of partnering organisations
- Substantially reduced cost – “Build one, get three free”!
  - Development cost of “one” would rise, but other collaborators would be contributing their components for common use
  - Individual NSI/ISIs modernisation looks as unaffordable – is this the only way to get there in a foreseeable timescale?

- We already use components from other NSIs, but without a common architectural approach, integration is very expensive and many of these components are not scalable

This remains a high risk project, but with very large potential returns to all community members involved.

References

Douba, V (2006). Towards a Service Oriented Architecture (SOA) for the Statistical Value Chain. Joint UNECE/Eurostat/OECD Seminar on the Management of Statistical Information Systems (MSIS) Sofia, Bulgaria.

Staggemeier, A T. Lowthian, P. and Lee G. (2007) Applying Tau-Argus to SuperCROSS tables: A practical example using the UK Business Register Unit data. Joint UNECE/Eurostat/OECD Seminar on Statistical Computing Aspects of Statistical Disclosure Control Manchester, UK.