ISSN 1977-0375

eurostat
Methodologies and working papers

European Plan of Research in Official Statistics (EPROS)

Main conclusions from the activities in the 5th Framework Programme

2007 edition

eurostat
EUROPEAN COMMISSION

Luxembourg: Office for Official Publications of the European Communities, 2007

ISBN 978-92-79-04705-3
ISSN 1977-0375
Cat. No. KS-RA-07-003-EN-N

Theme: General and regional statistics
Collection: Methodologies and working papers

TABLE OF CONTENTS

EPROS Projects

General Acronyms

CAI	Computer assisted interviewing
CAPI	Computer assisted personal interviewing
CATI	Computer assisted telephone interviewing
CBM	Current best method
EDI	Electronic data interchange
ESS	European Statistical System
NSI	National Statistical Institute
ONS	Office for National Statistics
FP	Framework Programme
ECHP	European Community Household Panel
EPROS	European Plan for Research in Official Statistics
ESDA	Exploratory Spatial Data Analysis
DOSIS	Development of Statistical Information Systems
DOSES	Development of Statistical Expert Systems
FP	Framework Programme
GIS	Geographical Information Systems
ICT	Information & Communications Technology
IST	Information Society Technologies
JOS	Journal of Official Statistics
JRC	Joint Research Centre
KA	Key Action
LFS	Labour Force Survey
NACE	Nomenclature d'Activité dans la Communauté Européenne
NORIS	Nomenclature Onon Research in Official Statistics
NSI	National Statistical Institute
NTTS	New Techniques and Technologies for Statistics
NUTS	Nomenclature of Territorial Units for Statistics
OCR	Optical Character Recognition
OECD	Organization for Economic Cooperation & Development
OMR	Optical Mark Recognition
ONS	Office for National Statistics
QNA	Quarterly National Accounts
ROS	Research in Official Statistics
SDC	Statistical Disclosure Control
SPC	Statistics Programme Committee
SILC	Survey of Incomes and Living Conditions
SINE	Statistical Indicators of the New Economy
SME	Small and Medium Enterprises
SODECE	Software Demonstration Centre
TDE	Touchstone Data Entry
TTK	Transfer of Technology and Know-how
VR	Voice Recognition
WHO	World Health Organization

Country Acronyms

BE	Belgium
BG	Bulgaria
CZ	Czech Republic
DK	Denmark
DE	Germany
EE	Estonia
IE	Ireland
EL	Greece
ES	Spain
FR	France
IT	Italy
CY	Cyprus
LV	Latvia
LT	Lithuania
LU	Luxembourg
HU	Hungary
MT	Malta
NL	Netherlands
AT	Austria
PL	Poland
PT	Portugal
RO	Romania
SI	Slovenia
SK	Slovakia
FI	Finland
SE	Sweden
UK	United Kingdom
HR	Croatia
MK*	former Yugoslav Republic of Macedonia
TR	Turkey
IS	Iceland
NO	Norway
CH	Switzerland

** Provisional code which does not prejudge in any way the definitive nomenclature for this country, which will be agreed following the conclusion of negotiations currently taking place on this subject at the United Nations*

FOREWORD

Rapid social, economic and technological changes are taking place in our societies. Unless statisticians respond to these changes and the associated new emerging needs, official statistics could be quickly regarded as outdated and irrelevant.

A focussed programme of research and development in statistics is always needed in Europe in order to provide the tools and techniques that will permit keeping pace with those new and changing circumstances. The promotion of research in official statistics is an integral part of Eurostat's activities. Since 1989, in close collaboration with DG Research and DG Information Society, Eurostat has stimulated research activities for developing, adopting and using new statistical methods and tools in high priority domains and in new emerging areas like the New Economy.

Eurostat, in its role of coordinator of the European Statistical System (ESS), has always tried to reinforce the co-operation between the ESS and the scientific community, as well as to facilitate the transfer of knowledge and technology and to accelerate the adoption by the statistical community of the results of the research activities.

The EPROS (European Plan of Research in Official Statistics) aimed to meet the new needs of producers and users of statistics. It was launched under the 5th Framework Programme for Research and Development. These projects, funded by the European Community, are described in this document.

The objective of this document is present the major achievements and successes of the research activities in statistics followed by Eurostat in the 5th Framework Programme and to introduce future research directions that might be implemented into the 7th Framework Programme that started in 2007.

Pedro Díaz Muñoz
Director

ACKNOWLEDGEMENTS

This document would not have been possible without the knowledge and enthusiasm of the author, Mr Deo Ramprakash. Furthermore, the help of the coordinators of the EPROS projects in verifying and complementing the project descriptions has been very valuable and highly appreciated.

Several colleagues of Unit B-5 'Research and methodology' of Eurostat have contributed to this publication. Mr Kimmo Rossi and Mr Pascal Jacques have greatly helped in creating a clear and logical structure of the document. Mr Knut Utvik played an important role in the development and the multiple revisions of the manuscript in order to make it a complete and comprehensive volume and Mr Mikko Saarnio helped a lot with the editing and consistency checks.

EPROS - OVERVIEW

1

eurostat

1. INTRODUCTION

Every production system, whether it is the production of vehicles or banking services, aims to increase efficiency and lower costs through research and development. The statistical production system is no exception. Statistical research seeks the deployment of latest statistical methodologies and information technologies in order to achieve the following specific benefits:

- Improved information service to the decision-maker and the citizen. This benefit is realized through making more timely and reliable data and more relevant analyses available;
- Lower statistical production costs. This is possible by making more extensive use of leading-edge information technologies and by rationalizing statistical systems;
- Reduction of respondents' burdens. Again, the route to achieving this gain is through more automated data collection and the more intensive secondary exploitation of existing data;
- Enhanced quality of official statistics. There is no substitute for installing effective quality control systems and procedures embracing every stage of the statistical production process.

Conducting research at the European level pools knowledge and experiences gained at the national level and helps to identify gaps in the research agenda.

The detailed profiles of individual projects of the European Plan for Research in Official Statistics (EPROS) are presented in Section II of this report; Section I below chapter looks across the project profiles and tries to highlight some common threads. It first sets the scene with an indication of the contribution which EPROS has made to European statistical research in general and more specifically to the European Statistical System (ESS). The ESS consists of the statistical organisations (National Statistical Institutes (NSIs) and other providers of official statistics) of the EU Member States and the EFTA countries, coordinated by Eurostat, the statistical office of the European Commission. At the highest level of consultation of the ESS is the Statistics Programme Committee (SPC), consisting of the Directors General of the NSIs and of Eurostat. A second tier of consultation is the Working Groups or/and Task Forces in specific areas of statistics, for example the EPROS Working Group and the EPROS Task Force.

Furthermore, Section I outlines possible future scenarios of research in official statistics. The predecessors of EPROS are then sketched, followed by the EU Framework Programme (FP) 5 context. FP5 was the source of funding of EPROS and it significantly determined the technological focus of the research. There is a detailed analysis of the statistical coverage of EPROS, including the transfer of technology and know-how, networking, dissemination and exploitation.

2. THE CONTRIBUTION OF EPROS

At the most general level and on the basis of the individual project profiles, it can be confidently asserted that EPROS has advanced the frontiers of knowledge in a number of leading-edge statistical and technological issues. For example, through INSPECTOR, we know significantly more about integrating a generic, distributed and flexible data validation system within the data collection process itself; through MANTLE about the use of such innovative technologies as night-time satellite imagery to produce surrogates of spatial socio-economic indicators; through MISSION about how different organizations can publish data on the web within a harmonised framework; and through STILE about better measurement of the labour market in the 'New Economy'.

Similarly, EPROS did make substantial progress on indicators through the programme on the Statistical Indicators of the New Economy (SINE). It developed a conceptual framework for integrating indicators on the new economy, quality-appraised existing Commission indicator sets, proposed new indicators and new uses of existing indicators, researched domain or thematic indicators and contributed to the methodology for deriving composite indicators.

Of course, it is in the nature of the incrementalism of scientific progress that in many cases as many questions were raised as answered. But, on the whole, the main legacy of EPROS was an increase in the knowledge stock of the ESS between 1999 and 2003, the formal duration of EPROS, (though the last project was not completed until 30 April 2006).

The knowledge created through EPROS was encapsulated in major software in at least 19 projects. These project outcomes were lodged in the websites of the projects concerned and were disseminated through conferences and publications. In almost all cases, NSIs were either consulted or involved as formal partners, thereby ensuring that the fruits of the research would not be dissipated but either be incorporated in the statistical production process or serve as a platform for further research.

EPROS aimed at a more cohesive ESS, with a greater systematic sharing of knowledge across national frontiers, between different institutions and different disciplines, and between large and small NSIs. Specifically, while there has always been cooperation amongst NSIs through such fora as Eurostat's working groups and task forces, it is generally more likely now than before that a NSI in one country will contact a University in another country to work together to their mutual benefit. For example, Statistics Netherlands and the University of Bath have become closely acquainted with each other through the NESIS project. This kind of networking can in due course be of even bigger benefit to the new member states.

Users at both national and international levels were encouraged through active consultation in every project to ensure that their needs were taken into account in project design and execution as well as specifically in the testing of software prototypes. In almost all cases they were invited to participate, and did participate, in workshops and conferences held through the projects. EPROS users spanned national policy makers, Directorates of the European Commission, public administrations and national research bodies. The hope is that this involvement of users would universally result in a habit of user-producer consultation, which is a main prerequisite for the costly production process to culminate in relevant statistics. The further hope is that the seeds of such networking and clustering would germinate, expand and deepen in a sustainable way, possibly resulting in more mobility across the ESS.

EPROS was a learning experience for the ESS in terms of organizing and managing substantial research projects and of working together systematically in large multi-national, multi-disciplinary teams. As a result of EPROS, the ESS should be more aware of issues concerning the optimal size of consortia, the internal organization of research and the funding of research within the NSI, between the NSI and Ministries and between the NSI and EU sources.

As the distribution of projects across research areas indicates, three areas seem inadequately covered, viz. Methodological issues, Advanced technology for data collection and Transfer of Technology and Know-how (TTK). Annex 1 gives the detailed contents of these categories. In particular, given the dynamics and central importance of nomenclatures in any statistical system, it is perhaps surprising that there were only two projects, CLAMOUR and STILE, in this sub-category. Similarly, the issue of improved access to microdata, which is quasi-political, was tackled through one project only, CASC, though CASC was supported by some of the leading EU NSIs and other experts in this field. TTK issues were researched through the only project in this area, AMRADS, which has demonstrated the need for more systematic mechanisms for best practice identification and standard-setting. VL?CATS was the only project on e-training in statistics, another key issue for the ESS facing a shortage of trained statisticians. In contrast, out of the 12 projects within 'Multi-data sources, integration and systematization', the sub-category 'Integrated statistical processing via metadata' was well subscribed, with 9 projects.

DISTRIBUTION OF OF PROJECTS BY RESEARCH AREA

NORIS heading plus	No of projects **
Methodological issues	4
Advanced technology for data collection	5
Quality issues	9
Data analysis and statistical modelling	13
Multi-data sources, integration and systematisation	12
Dissemination, disclosure control	9
Indicators	10
Technology and know-how transfer (TTK)	1

** *Projects may be classified in multiple NORIS headings*

In conclusion, EPROS to a substantial degree achieved its ultimate objectives of leading to *"more and more information, of higher and higher quality, faster and faster, more and more comparable, at lower and lower cost"*. These benefits are not all immediately apparent; in some cases, the gestation period for full exploitation and real impact of EPROS can be extended over many years to come. The ESS has to remain vigilant and seek either to utilise the results of EPROS as far as possible in the official statistical production process or to take forward the research until exploitable results are produced.

3. Future Research Orientations

3.1 Introduction

Since the formal end of EPROS in 2003, there have been major developments under FP6 and FP7. Thus, one cannot jump from EPROS to the delineation of a future research scenario without considering these intermediate landmarks as well as the substantial amount of activities undertaken by the Eurostat EPROS Working Group and EPROS Task Force. FP7 is considered in some detail because it is the present and the future up to 2013. However, it is only a part, albeit an important part, of the future. A more comprehensive research policy independent of the characteristics of the funding source seems highly desirable. For that purpose, one needs to learn, inter alia, the lessons thrown up by EPROS.

3.2 Sixth Framework Programme (FP6) 2002-2006

Eurostat did not have a dedicated action line (or Strategic Objective, as they were called in FP6). However, there were many projects with substantial statistical content and interest. Eurostat has been following these projects closely by attending relevant project events and inviting representatives of these FP6 projects to meetings organized by Eurostat. Some of these projects are (this alphabetical list is not exhaustive):

- **DECOIN – Development and Comparison of Sustainability Indicators** [2]

  DECOIN has three main objectives: to find out the most promising ways to develop methodologies and data quality of best-needed sustainable development indicators; to evaluate existing methods and analytical frameworks in order to assess the progress towards sustainable development, to elaborate forecasts and scenarios, and to identify inter-relationships between unsustainable trends in the EU, and; to carry out a analysis on the inter-relationships between selected unsustainable trends. Existing indicator frameworks, large-scale one-dimensional frameworks and scenarios of unsustainable trends are evaluated from the perspective of the EU Sustainable Development Strategy. are also evaluated. A detailed analysis of the inter-relationships of the development processes related to the unsustainable trends, their synergies and trade-offs, is carried out by utilizing analytical frameworks such as Advanced Sustainability Analysis (ASA), Sustainability Multicriteria Multiscale Assessment (SUMMA) and Multi-Scale Integrated Analysis of Societal and Ecosystem Metabolism (MSIASEM). Based on the analyses an integrated tool will be constructed for assessment of interlinkages and for forecasting.

- **ESEC – European Socio-economic Classification** [3]

  The project was designed to produce a European socio-economic Classification (ESEC) for use in comparative social science research across the EU. Comparative analysis of many aspects of the quality of life and of social cohesion, for example health, living conditions and economic situation of Europe's population, seeking to understand variation between member states, is hampered by the lack of an agreed, harmonised and validated classification of socio-economic positions. This project was designed to improve the state of the art in this area and to make demonstrable progress towards comparative research in a wide variety of areas relevant to the knowledge based society. By providing an essential comparative European research tool, the ESEC project also hoped to greatly facilitate subsequent analyses of intergenerational social mobility and inheritance of inequalities.

[2] *http://www.decoin.eu/*
[3] *http://www.iser.essex.ac.uk/esec/*

- **ESS – European Social Survey** [4]

  The European Social Survey (the ESS) is an academically-driven social survey designed to chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs and behaviour patterns of its diverse populations. Now moving into its fourth round, the survey covers over 30 nations and employs the most rigorous methodologies. The fieldwork has been funded through the European Commission's 5th and 6th Framework Programmes, the European Science Foundation and national funding bodies in each country.

- **EU – KLEMS – Productivity in the European Union: A Comparative Industry Approach** [5]

  The project created a database on measures of economic growth, productivity, employment creation, capital formation and technological change at the industry level for all European Union member states. This work provides an important input to policy evaluation, in particular for the assessment of the goals concerning competitiveness and economic growth potential as established by the Lisbon and Barcelona summit goals. The database facilitates the sustainable production of high quality statistics using the methodologies of national accounts and input-output analysis. The input measures include various categories of capital, labour, energy, material and service inputs. Productivity measures were developed, in particular with growth accounting techniques. The database is used for analytical and policy-related purposes, in particular by studying the relationship between skill formation, technological progress and innovation on the one hand, and productivity, on the other.

- **I-CUE – Improving the Capacity and Usability of EUROMOD** [6]

  The main aims of this EUROMOD[7] -related project were to start the process of expanding EUROMOD to cover the 10 New Member States (NMS) of 2004 and to make EUROMOD easier to use, especially when dealing with 25 systems and datasets. The project ends with a final conference in spring 2008. The aim of I-CUE was to re-design and up-grade EUROMOD in the light of enlargement and of lessons learned from operating and using the first, prototype version. The project provide the basis for increasing EUROMOD's capacity to address a wide range of social science questions, incorporating the 10 New Member States of 2004, improving ease of use and accessibility, improving the quality of results by enhancing comparability across countries and reducing the resources necessary to maintain, update and develop EUROMOD in the future.

- **IECM – Integrated European Census Microdata (CIECM/HIECM/DIECM)** [8]

  The Integrated European Census Microdata (IECM) database, one of the world's largest integrated research infrastructures for the study of human populations, is under construction. The database will contain anonymised microdata samples encompassing as many as 50 censuses and totalling more than 70 million person records. The IECM subprojects form part of the International Census Microdata for Population Research (IPUMS) infrastructure. The Coordinating the Integrated European Census Microdata (CIECM) project organized an exploratory workshop to co-ordinate the European census microdata integration based on multiple rounds of population and housing censuses; the Harmonizing the Integrated European Census Microdata (HIECM) project contributes to the creation of the IECM research infrastructure by developing and implementing harmonized coding systems to permit comparisons across countries and time periods based on multiple rounds of population and housing, altering the existing coding system as necessary to accommodate the new samples; the Disseminating Integrated European Census Microdata (DIECM) project seeks to enhance and promote usage of the European census microdata through the implementation of an Internet based dissemination system to make the IECM infrastructure accessible to researchers and policy makers.

[4] *http://www.europeansocialsurvey.org*
[5] *http://www.euklems.net*
[6] *http://www.iser.essex.ac.uk/msu/emod/i-cue/*
[7] *EUROMOD is a multi-country Europe-wide tax-benefit model (http://www.iser.essex.ac.uk/msu/emod/)*
[8] *http://www.iecm-project.org*

- **INDI-LINK - Indicator-based Evaluation of Interlinkages between Sustainable Development Objectives** [9]

  One core objective of INDI-LINK is to contribute significantly to the further development of the EU SDI set, evaluating the existing SDI set and developing new concepts and methods for the calculation of "best-needed" indicators. The project reviews emerging policy areas and provide recommendations for next steps in SDI development. The second core objective of INDI-LINK is the assessment of interlinkages between different priorities of the EU SDS. It provides a quantitative analysis of past interlinkages using selected indicators with best-suited assessment methods and tools, putting a particular focus on sensitivity analysis of the results. The project tests different methods and extends existing simulation models for extending indicator time series through forecasting and provide estimations on future developments of interlinkages. The final objective of INDI-LINK is the presentation of conclusions for future SD policy making and for an effective implementation of the revised EU SDS. INDI-LINK; the project aims to identify most effective combinations of environmental, economic and social policy measures.

- **KEI – Knowledge Economy Indicators: Development of Innovative and Reliable Indicator Systems** [10]

  The project aimed to develop and improve indicators for the knowledge economy, including the analysis of aggregation issues and the use of composite indicators. KEI reviewed existing concepts and definitions of the knowledge-based economy and its key components; it developed main thematic areas in relation to the Lisbon and Barcelona objectives and used these themes to classify existing indicators and thoroughly explore data and indicator quality issues. Composite indicators were analysed in detail using both statistical and participatory approaches, including the use of multi-criteria methods, aggregation and weighting techniques, decomposition methods, and an evaluation of analytical and presentational techniques. Simulation methods were extensively employed to investigate the robustness of indicators and the conclusions based on them.

- **MICMAC – Bridging the Micro-Macro Gap in Population Forecasting** [11]

  In an ageing population, the demand for adequate health care services, pension systems and other social protection systems is paramount. The sustainability of high-quality health care and pension systems is influenced to a considerable extent by demographic change and by the way people live their lives (lifestyle and life course). Therefore, an adequate monitoring and forecasting of demographic change and of the lifestyle and life course of the population are a conditio sine qua non for the provision of health and social security to the people of Europe. This requires a methodology that moves beyond conventional projections of the population by age and sex. What is needed is a methodology that complements the demographic projections with projections of the way people live their lives. The objective of MicMac was to develop this methodology.

- **MEADOW – Measuring the Dynamics of Organisations and Work** [12]

  The coordination action MEADOW aims at setting out guidelines for collecting and interpreting harmonised European data on organisational change and work restructuring and their economic and social impacts, constituting a first step towards implementing a harmonised European survey instrument. The project integrates the perspectives of both producers and users by including research teams that have designed and implemented national survey instruments for measuring organisational change, innovation and work restructuring, as well as experienced users of such surveys. After a state of the art in surveys on organisational and work's changes, and concepts of the organisational change, priorities in measurement and basic definitions are fixed; then the measure of organisational change in employee surveys, and in employers surveys, in different countries are analysed and different statistical methods identified; a harmonised questionnaire developing core indicators is finally tested and revised.

[9] *http://www.indi-link.net/*
[10] *http://www.kei.publicstatistics.net*
[11] *http://www.nidi.knaw.nl/en/micmac/*
[12] *http://www.meadow.fr*

- **MOSUS – Modelling Opportunities and Limits for Restructuring Europe towards Sustainability** [13]

  The MOSUS project integrated within a macroeconomic, multi-sectoral framework three major themes of European policies: Sustainable development; Competitiveness and social cohesion in the knowledge-based society, and; Globalisation and international trade. Based on an existing economic model, the project developed and applied an integrated ecological-economic simulation model in order to quantify the interrelations between socio-economic driving forces and the state of the environment. The analysis was done within a multi-country, multi-sectoral macroeconomic framework, including trade flows within Europe as well as between Europe and all other economically relevant parts of the world. The model directly integrated comprehensive bio-physical data (material and energy flows as well as land use data) in European and global simulations up to the year 2020 and put them in relation to structural indicators of social and economic developments. Thus, this tool allows for formulating and evaluating scenarios of the economic and social/distributional impacts of key environmental policy measures and for presenting validated policy recommendations for responding to environmental changes.

- **PRIME – Policies for Research and Innovation in the Move towards the European Research Area** [14]

  PRIME was a network of excellence to develop long-term research and shared infrastructures on policies for research and innovation in the move towards the European Research Area (ERA); the overall objective of PRIME was to carry out the research and related structural actions needed to underpin policies for research and innovation in the move towards the European Research Area (ERA). The creation of the ERA involves the development of a fully-fledged new public 'actor'. Thus, in addition to national and regional levels of government, there is a third level of public policy-making in research and innovation. Such a three-level system is probably unique in the world; it will be difficult to develop the new policy approaches required for the goals of the ERA to be achieved with the current fragmentation of research capabilities (both geographical and intellectual) on issues which underpin the emergence of ERA policy.

- **SHARE – Survey of Health, Ageing and Retirement in Europe** [15]

  SHARE is a multidisciplinary and cross-national database of micro data on health, socio-economic status and social and family networks of individuals aged 50 or over. Data collected include health variables (e.g. self-reported health, physical functioning, cognitive functioning, health behaviour, use of health care facilities), psychological variables (e.g. psychological health, well-being, life satisfaction), economic variables (current work activity, job characteristics, opportunities to work past retirement age, sources and composition of current income, wealth and consumption, housing, education), and social support variables (e.g. assistance within families, transfers of income and assets, social networks, volunteer activities). In addition, the SHARE data base includes variables and indicators created by the AMANDA RTD-Project under the European Union's 5th Framework Programme. The data is freely available to the entire research community.

- **THESIM – Towards Harmonised European Statistics on International Migration** [16]

  The development, implementation and monitoring of a common European asylum and immigration policy requires a system of statistical information exchange on migration flows and asylum in the EU. However, these data are not always available and do not necessarily consider need for policy development. It is clear that more initiatives have to be developed in order to improve data collection and harmonisation of international migration statistics to provide reliable basic information and this is particularly important in the framework of European migration policy. The THESIM team organised 25 national meetings and through these meetings prepared 25 country reports on the national situation, as far as data collection on international migration, asylum, residence permits and citizenship. These meetings were organised with help of all statistical offices and the NCPs (National Contact Points) of the EMN (European Migration Network); all ministries and administrations involved in the field took part, those involved in legislation, in practical procedures or in data collection.

[13] *http://www.mosus.net*
[14] *http://www.prime-noe.org*
[15] *http://www.share-project.org*
[16] *http://www.uclouvain.be/en-7823.html*

- **WORKS – Work Organisation and Restructuring in the Knowledge Society** [17]

  WORKS was a pan-European research project which aimed at improving the understanding of the major changes in work in the knowledge-based society (KBS). Taking account of the global forces and of the regional diversity within Europe, the project investigated the evolving division of labour within and between organisations and the related changes at the workplace. The implications for the use of skills and knowledge, for flexibility and for the quality of work, as well as the impact on occupational identities, time use and learning of individuals were investigated in a comparative perspective. The role of the social dialogue and of the varieties of institutional shaping in Europe received particular attention.

3.3 Seventh Framework Programme (FP7) 2007-2013

During the consultations on FP7[18] , Eurostat raised the need for developing research in official statistics and secured important opportunities in the Programme. Two of the specific programmes – Cooperation and Capacities – specifically mention statistics as a tool for policy planning and impact analysis. Various Themes afford opportunities for statistical research. For example, in Theme 6 Environment, under the specific point 'Earth observation and assessment tools/assessment tools for sustainable development' there is the statement: *"Research will also seek to improve existing indicators and develop new ones to assess sustainable development policy priorities, and to analyze the linkages between them, taking into account the existing set of EU sustainable development indicators"*.

3.3.1 Cooperation: Socioeconomic Sciences and Humanities

The most significant current opportunity is in the 2007 work programme for Cooperation, Theme 8, Socioeconomic Sciences and Humanities (SSH), Area 8.6.3, Provision of underlying official statistics. The following proposal call text is reproduced from it:

> *"The overall objective of this research area is to contribute to improving the availability, quality and relevance of official statistics for use in policy. Research is needed on official statistics in order to guarantee that there are the best possible foundations on which to build indicators for use in all aspects of policy. This research may look at problems associated with existing official statistics as well as addressing needs for new official statistics.*
>
> *Priority will be given to research addressing issues related to key statistical policy areas such as: improvement of data quality, data integration, the statistical production process, data related to small areas and rare events and indicators and data delivery. Research improving the comparability of datasets and indicators through increased and appropriate harmonization and standardization. Linking sets of data from different sources and of accessing new sources together with assessing their impact. Increasing the availability of additional relevant and comparable disaggregations in order to solve the issue of "gaps" in statistics e.g. when gender breakdowns are not available. Methods to expand coverage to smaller regional units, sub-threshold and rare events. Streamlining of the statistical production and dissemination process and addressing the issues of quality in its widest sense including coverage, timeliness, comparability, confidentiality vs. usability, cost-effectiveness (including response burden), relevance, impartiality and reliability. Research should aim to build on the work of national statistical institutes or other organizations involved in producing official statistic. The work should be compatible with the European Statistical System and take into account work going on at the international level in order to improve comparability with third countries and linking with international organizations."*

It is important to note that Eurostat does not pro-actively initiate research actions but act as a facilitator.

In FP7, statistics are regarded as a support activity to the main EU policies rather than as a stand-alone operation, implying that opportunities for research should appear in all Thematic (and some horizontal) parts of FP7. The pro-active role in proposal initiation and preparation falls chiefly to NSIs acting in concert with Universities, research bodies, software houses and other organizations.

---

[17] *http://www.worksproject.be*

[18] *http://www.cordis.europa.eu/fp7*

3.3.2 Capacities: Remote Access

In FP7, support for the construction of new research infrastructures will be based on the roadmap for new research infrastructures prepared by the 'European Strategy Forum for Research Infrastructures' (ESFRI)[19] . Specifically, FP7 gives the possibility to launch 'Integrating Activities' under 'Research Infrastructures' with the objective of "promoting European wide access to microdata sets of officials statistics for research, leading to a European statistical system open to researchers." 'Research infrastructures' refers to facilities, resources and related services that are used by the scientific community to conduct top-level research in their respective fields. In the specific scope of wide access to microdata, it covers amongst others: knowledge-based resources such as collections, archives or structured scientific information; enabling ICT-based infrastructures such as computing, software and communications. The research infrastructures may be "single-sited" or "distributed" (a network of resources). The 'integrating activities' objectives are to provide a wider and more efficient access to and use of the existing infrastructures existing in the different Member States. They also aim to structure better and integrate, on a European scale, the way research infrastructures operate and to foster their joint development in terms of capacity and performance.

Such an infrastructure for remote access to official statistics microdata should mandatory follow the Integrated Infrastructures Initiatives (I3) model combining: Networking activities, Trans-national access and/or service activities and Joint research activities. Activities covered could include aspects such as: Definition of standards for accredited safe centers; Common framework and guidelines for output checking; Training of staff; Common procedures for approval of researchers; Secure connections between safe centers; On-line access to microdata sets in EU; Advanced SDC methods; Risk assessment; Input and output filters; Data Access Protection; etc.

3.4 Future Research Scenario for the ESS

The source of funding significantly shapes the statistical research programme. As already stated, funding through the Themes of FP 7 required statistical research to be regarded as a support activity of policies. In FP 5, in contrast, statistical research was treated as a transversal activity cutting across the main axes of the Specific Programme (SP) 2, Creating a user-friendly information society. The focus of this SP was the development and use of information society technologies. Thus, EPROS had to be opportunistically tailored to the nature and requirements of SP 2, including its administrative timescales. The EPROS projects were strongly oriented to the development of software and to the use of such technologies as the internet for the collection, analysis and dissemination of statistics.

Both the FP 5 and FP 7 orientations of statistical research were necessary. However, there is a case for having a more general "EPROS" that is not restricted in either scope or content by the nature of the funding source. Such a more comprehensive "EPROS" would start bottom-up from the full spectrum of needs and would roll forward from year to year to adapt to changing circumstances. It would be built up from:

- The pointers to further research contained in the EPROS project profiles themselves;
- The research needs of the ESS.

It is only after this more global research coverage that particular subsets can be selected and oriented to the requirements of particular funding sources.

Specific research needs tend to vary somewhat between NSIs but there is a broad consensus amongst them about the needs at the European level. In other words, even with their different domestic emphases, they share much in common. For example, it is highly unlikely that any NSI would deny the need for further research into confidentiality or nomenclatures or data integration or TTK.

Similarly, there would be a vast overlap between the research needs of Eurostat and those of NSIs. It is reasonable to assume that the ESS as a whole would speak with one voice on future statistical research. Thus, the fields of research suggested in the following list should command wide ESS support. They are expressed in broad terms to serve as hooks or pegs on which proposers can hang their specific research preferences. The list is neither exhaustive nor prescriptive.

[19] *http://www.cordis.lu/esfri/*

Methodological Issues

- Concept formation: rapid economic, social and technological changes require adaptation of corresponding concepts, definitions and variables. These concepts, definitions and variables underlie particularly classifications such as the "Nomenclature d'Activité dans la Communauté Européenne (NACE)" and International Standard Classification of Occupations (ISCO), which therefore require further research, for example into intelligent induction algorithms, to ensure their continuing relevance;
- International harmonization: this requires, for example, research into the cohesion between statistical subsystems across countries, such as between business data and the national accounts;
- Completing the universe: this concerns gaps that are so pervasive as to be almost endemic to all information systems, for example, gaps in estimates of international emigration.

Advanced Technology for Data Collection

- Research should take full advantage of the continuing improvements in information technologies to design, collect, capture and code data, with their metadata, in order to reduce respondents' burdens, lower costs and improve data quality. Such methods would include Automatic Coding by Text Recognition (ACTR), computer assisted interviewing (CAI), electronic questionnaires, electronic data interchange (EDI), remote sensing, satellite imagery and internet surveys, which are deceptively simple but require great care.

Quality Issues

- Reliability/timeliness trade-off: very often, there has to be a balance between quality and timeliness, which are competing requirements in any statistical system. More research is needed on how this balance varies between subject-matter domains, between countries and between international aggregates and individual countries;
- Variance estimation and resample methods in complex surveys: substantial further theoretical and empirical research is needed into variance estimation through linearization, replication and generalized variance functions, including longitudinal aspects;
- Incorporating data validation systems in the data collection process: the full pre-conditions for successful integration of generic, distributed and flexible data validation systems into processes of statistical data collection and consequential improvements in quality would seem to be a priority research issue;
- Assessment, control and reporting on the quality of administrative registers and other administrative sources: research should cover clarity, concepts, reference period, data up-to-dateness, coverage, errors, multidimensionality, record matching potential, confidentiality and comparability between datasets.

Data Analysis and Statistical Modelling

- Data analysis and knowledge extraction: this issue lies at the interface of statistics, database technology, pattern recognition and machine learning. The power of grid technology for mining very large databases needs to be exploited, (though the use of grid technology is not confined to data mining). The use of such techniques as neural networks/artificial intelligence in data mining and comparisons with classical statistical approaches need to be further researched;
- Micro-simulation models: there is a need for further research into the feasibility and use of such models, including longitudinal estimation, for policy impact analysis particularly in the enterprise sector;
- Timeseries analysis: there is a need for further research into seasonal adjustment, particularly such issues as finding the optimal solution to the seasonal adjustment problems of timeseries aggregation; and the use of ad hoc filters to separate the cycle from the underlying long-term trend; interpolation of high-frequency data; intervention analysis; the efficacy and added explanatory power of neural networks; and combination of sampling variability and uncertainty in the seasonal adjustment process. There is also a need for further investigation of the forecasting performance of dynamic factor models; and the optimal settings for the non-parametric tools involved in generalized dynamic factor models;

- Ontologies and semantics: methods and tools for annotation conversion; data archive design and analysis; cataloguing of statistical analysis tools; rich data representation, documentation of statistical analysis procedures applied to the datasets; and reduction of data fragmentation and computation fragmentation;
- Modern sampling methods: these may use additional known and perceived information in order to achieve better coverage of the population or more pinpointed targeting of variables than that possible through the fortuitousness of randomization. Thus research is needed on more modern approaches such as ranked-set sampling, composite sampling, adaptive sampling, capture-recapture methods, multiplicity sampling, snowballing and so on;
- Small area estimation/Geographical Information Systems (GIS)/Exploratory Spatial Data Analysis (ESDA), including methods for spatial smoothing: there is a case for member states to adopt model-based methods for estimating the characteristics of areas below NUTS (Nomenclature of Territorial Units for Statistics) 3 level. There could be more research into estimation methods that reproduced the underlying distribution of area means; and robust estimation of variance-covariance parameters to improve the coverage properties of confidence intervals and optimize the performance of composite estimators. There is still more research required into the development of integrated software for the capture, storage, analysis and visualization of spatial data, including digital maps.

Multi-Data Sources, Integration and Systematization

- Research is needed into: data integration at the micro level involving enterprise surveys with household surveys and involving administrative databases with survey databases; exact linkage amongst administrative sources themselves; the robustness of statistical matching; and related issues of the quality of the outputs from integration, warehousing, metadata, distributed technologies, data management systems and confidentiality;
- There is a need for research on the merging of data from real-time satellite remote sensing with statistical data from censuses, surveys and local administrative records. This process is linked to GIS to derive spatial profiling, maps and a parsimonious set of indicators that can be used to develop initiatives directed towards the environment and ecology, meteorology and climatology, and economic development.

Dissemination, Training, Disclosure Control

- Networking and collaboration: the ESS should galvanize itself to encourage and maintain the widest possible networks across countries, between different disciplines and between different institutions. Such networks, on an expanded scale, should be a lasting legacy of EPROS and the Framework Programmes. They would put the exploitation of research results on a more assured footing. Budgets should be found for a regular programme of user-producer conferences and seminars that would take stock of evolving research issues and of the exploitation of previous research outputs;
- Statistical literacy, including distance training and virtual classrooms: specific training and in particular interactive training solutions relevant to official statistics should be revisited with a view to reducing shortages of skilled statisticians, provide facilities for targeted human resources development and transfer of current best methods;
- Access to microdata: there is a continuing need for research into statistical disclosure control and the release of statistics according to flexible geographies; into the conflict between confidentiality and transparency, specifically the safe release of microdata and mesodata; and into cognitive aspects;
- New applications in IT: research is needed into innovative user-interface and visualization techniques, usability, user-friendliness, and interface issues (including ESDA)

Indicators

- Further research is required on all aspects of indicators, in particular on methods for achieving parsimony such as, for example, choosing a small number of resonant, (orthogonal) flagship indicators; increasing contextual details; deriving composite indicators; and using conventional statistical approaches such as exploratory factor analysis, clustering and latent variable methods. Research on indicators is the subject of calls for proposals under FP7, Cooperation, Theme 8, Socioeconomic Sciences and Humanities. Future research should include Sustainability Indicators and Indicators on the Millennium Development Goals.

Transfer of Technology and Know-how (TTK)

- The need for TTK is ever-important and further enhanced with the enlargement of the EU. The work of the EPROS project, AMRADS, should be further extended and deepened, e.g. by further supporting action projects, by conferences, seminars and workshops on current best methods, and through other information channels (e.g. reviving the idea of a Software Demonstration Centre (SODECE) for demonstration, best practice identification and standard-setting; support throughout by user-friendly documentation in important areas; web information and discussion fora; etc.)

3.5 POSSIBLE LINES OF DEVELOPMENT

Reflections on particularly the EPROS experience have thrown up ideas for possible future improvements mainly at a strategic level of research organization and management. These ideas are contained in the following lines of development:

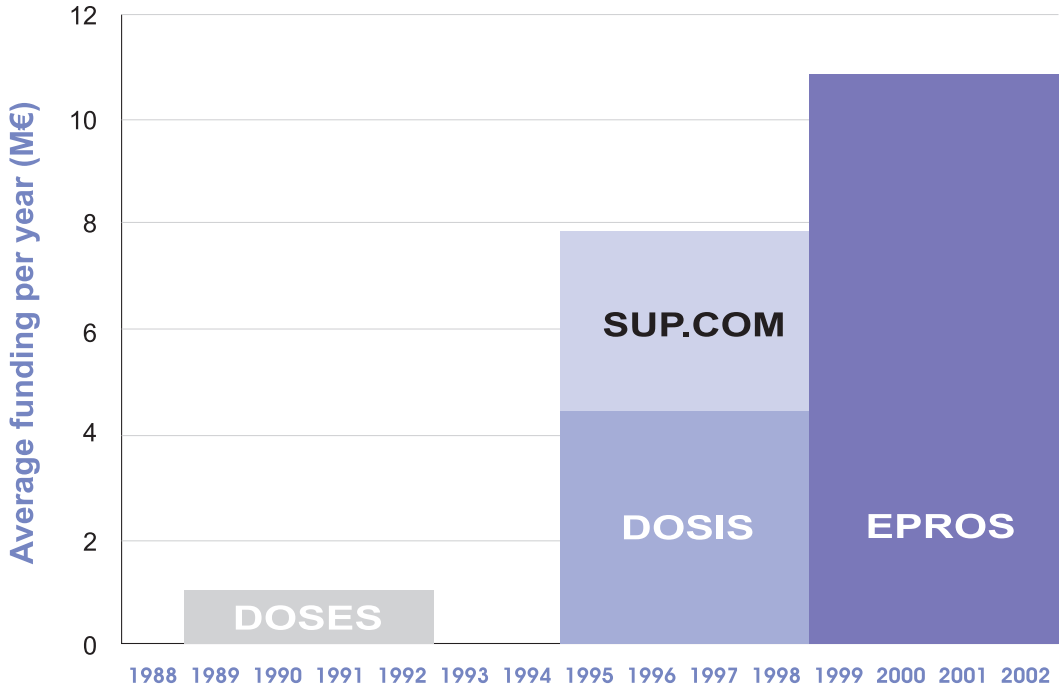
- There should be a comprehensive statistical research policy and a corresponding annually rolling "EPROS", with clear priorities;
- The funding of the larger "EPROS" should be sought not only from the Framework Programmes but collaboratively from the ESS;
- Such an "EPROS" should be constructed from the scenario indicated above, from the suggestions in the EPROS project profiles under FP5, from the experience gained in FP6 and to be gained incrementally in FP7, and from a bottom-up compilation of needs of NSIs and Eurostat;
- While this "EPROS" should explicitly be in support of policy and decision-making, with full involvement of users, it should not neglect longer-term infrastructure developments, such as integrated survey capacity building;
- This "EPROS" should seek to ensure the maximum exploitation, with the necessary safeguards, of leading-edge information technologies for the production, dissemination and use of official statistics;
- It should seek to promote mobility of trained statisticians within the ESS through fellowships and through any funding mechanisms available within the European Commission;
- The network built between stakeholders in EPROS and subsequent programmes should be pro-actively strengthened, extended, maintained and used;
- Some arrangements should be struck with research consortia in future that would allow project websites to be maintained;
- Dissemination and other mechanisms should be created to ensure, to a greater degree than at present, the exploitation of research results in the official statistical production process. Such mechanisms might require a more proactive role by Eurostat in feeding the research results to NSIs and monitoring the extent of their utilization in the production of statistics;
- Specific pro-active dissemination devices would include seminars and conferences intended to calibrate progress in research activity and to introduce fresh visions in a dynamic "EPROS";
- Mechanisms should be set up to expedite Transfer of Technology and Know-how (TTK) within the ESS, to facilitate demonstration and training, to assist with best-practice identification and standard-setting and to ensure user-friendly documentation; specifically, current best practices are not static but do improve with further research. One mechanism could be a software centre for demonstrating the results of research and imparting training in the process. Training could accelerate exploitation.

4. Historical Background to the European Plan for Research in Official Statistics (EPROS)

4.1 The predecessors of EPROS

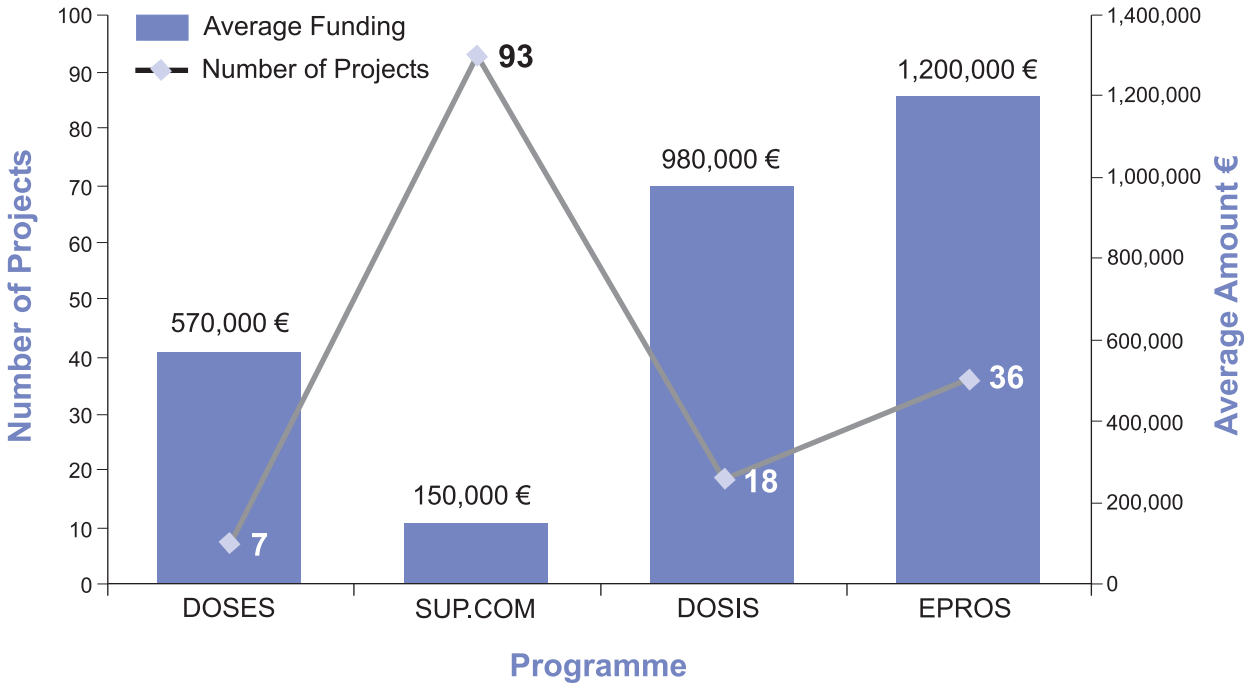
The historical background to EPROS, 1999-2003, should reveal both the evolution of statistical research and the total coverage of research topics over time at the European level. The earlier systematic statistical research efforts organized by Eurostat are as follows:

- DOSES (Development of Statistical Expert Systems; 1989-93) under the (FP2) 1987-91;
- DOSIS (Development of Statistical Information Systems; 1994-98) under the 4th Framework Programme (FP4) 1994-98;
- SUP.COM (1995-98) under the 4th Framework Programme (FP4) 1994-98.



The primary objective of DOSES was to improve the capacity for producing and analyzing statistical information by means of advanced processing techniques. More specifically, its aim was to stimulate and co-ordinate Community research in official statistics by developing artificial intelligence techniques in the form of expert systems. DOSES comprised twelve coordinated action projects and seven shared-cost projects over a period of about 4 years.

DOSIS was based on the same principles as DOSES, that is, it aimed at satisfying the needs of producers and users of official statistics. However, through statistically oriented projects in the domain of Information Technology and Telematics, it focused on the application of results for the improvement of the efficiency of the ESS. One aim was also to use the Framework Programme to encourage co-operation through the creation of multi-national consortia drawn from government, academia and the private sector.

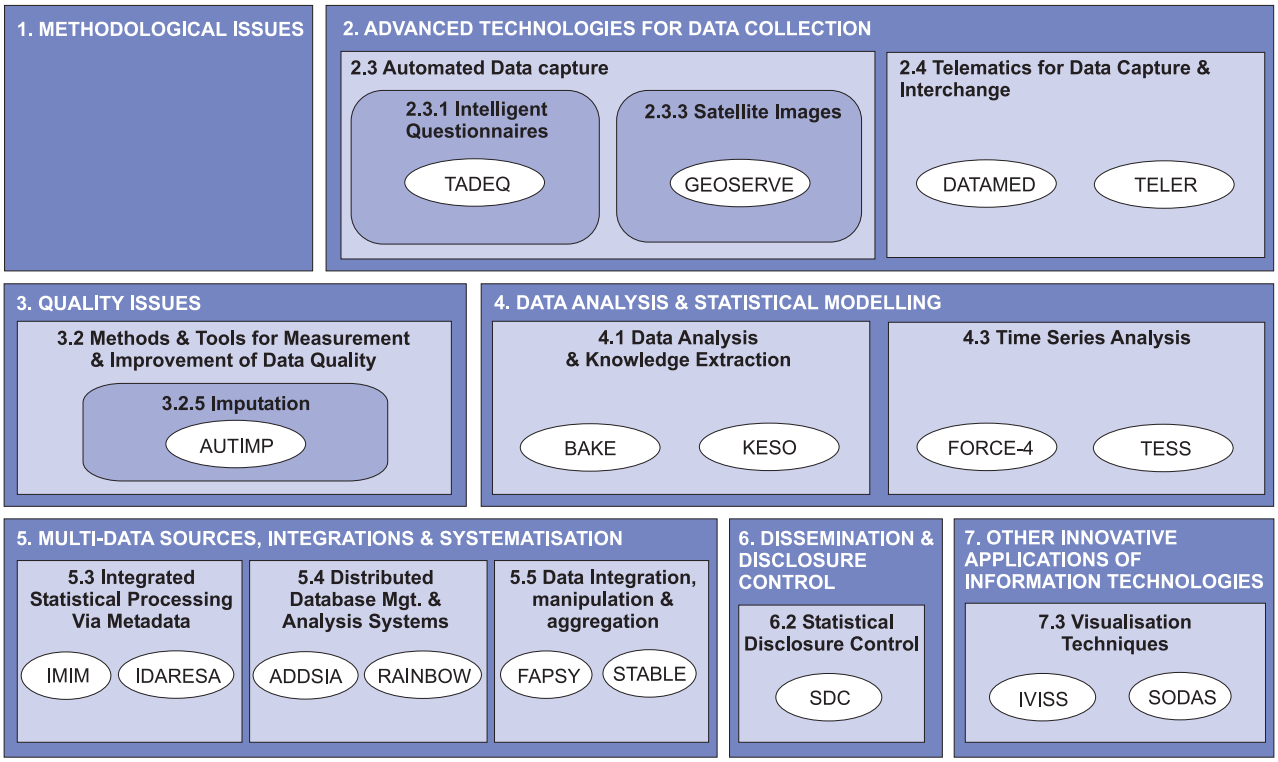


4.2 Outline of coverage of the DOSIS research

DOSIS generated 18 projects which are listed in the Annex. DOSIS research covered five very broad themes: data collection, data access, integrated statistical processing, data analysis and data quality. An overview of this research is as follows:

- There had been a trend towards more automated data collection because of the associated reduction in cost and in response burden and of the improvement in data quality. Thus, there was a growing use of computer-assisted interviewing (CAI), of electronic questionnaires and of electronic data interchange. However, CAI instruments became more complex. DOSIS research included the application of modern information technologies to make the documentation more readable and comprehensible and the use of electronic means for transmitting data, for example accounting data, from enterprises to public administrations;
- It is no use collecting data but denying the user adequate access to them. One obstacle has been the confidentiality both of microdata and of tables in which there is a risk of disclosure by inference. DOSIS research sought to optimize the balance between access and disclosure. Improved access also means providing the user with user-friendly, interactive software to help in the selection of tools and methods for the purpose in hand and to do so through the internet where appropriate. DOSIS research included the development of such software, including visualization, and the investigation of distributed data base management systems (DDBMS) for improved access to dispersed data, including specifically spatial data;
- The amount of data at one's disposal can be increased without new data collection. One way this objective can be achieved is by bringing together data from different sources either within the same country or between different countries. Reference has already been made to DDBMS. One crucial challenge in such data integration is in the development of infrastructures for the harmonization and management of the metadata that are attached to the data, such as the concepts, definitions, nomenclatures, descriptions of sampling schemas, even re-usable software modules and so on. DOSIS had a number of projects on the representation and management of metadata objects and their relationships;

- Data that are collected but not analyzed are a waste of resources. However, the analysis and modelling tools themselves can pose formidable challenges. One consequence of the exploitation of more sophisticated information technologies has been a marked increase in the size and complexity of databases, including timeseries and longitudinal data. DOSIS has researched data mining tools, including Bayesian Belief Networks, for the extraction of knowledge embedded in databases. DOSIS has also researched the seasonal adjustment of timeseries data and has investigated methods and tools for numerical symbolic data analysis techniques in NSIs and companies;
- There has been a tendency for the volume of data to expand without a commensurate increase in their quality. Statistics of poor quality can lead to serious policy mistakes. Research into the development and application of tools for quality assessments and corresponding corrections has always been a high ESS priority. However, the DOSIS research on quality was limited. It covered only automatic imputation, which involved the application of statistical methods for filling gaps in the data supplied by respondents.



SUP.COM statistical research projects were somewhat different from DOSES and DOSIS in the following ways:

- These were comparatively small projects with calls for tenders in each year 1995, 1996, 1997 and 1998;
- They were initiated for Eurostat (Support Commission) but executed externally through open tender;
- Totalled 95 different projects over the 4 years;
- Each project was of 12 months' duration;
- Covered topics right through the statistical production process, from data capture to dissemination.

SUP.COM could be counted as one precursor to EPROS because it included research aimed at the exploitation of new information technologies for the purpose of improving official statistics. Very briefly, the relevant projects covered such research as:

- Visualization of industrial statistics using high resolution PC graphics;
- Multimedia and new IT, including standardization activities;
- Computer graphics and Visualization/Study on the application of Parallel Coordinates to statistical data and training in this technique;
- Information highways and new technologies: access to, interchange and visualization of statistics-related information;
- Purchasing Power Parities, which sought to use bar-codes in place of visits to retail outlets;
- Integration of statistical software into statistical information systems;
- Network for Computer Assisted Training Systems (CATS);
- Virtual European Statistical Laboratory.

5. The Fifth Framework Programme (FP5) & EPROS

As already stated, EPROS was funded through FP5 and had to fit the research objectives of that Programme. The following sections elaborate the process by which EPROS was oriented to FP5. That is followed by a description of the principles that underlay EPROS and, given those principles, how EPROS was developed. This background should help in understanding the EPROS project profiles.

5.1 The FP5 context

In the late 1990s, Eurostat worked closely with the representatives of the Specific Programme (SP) 2, Creating a user-friendly information society, under FP 5 in order to produce the following niche for statistical research:

> *".. statistics are central to the information society, for which technologies offer new ways to attain the highest standards of quality and the widest and most rapid and accessible dissemination."*

> *" ..software technologies work will foster knowledge-based methods and tools which increase the usability as well as the capability of systems and the intelligence in the network; this includes the timely collection, production, dissemination and communication of high-quality information (including statistical and management information)".*

Compared with the view of statistics taken in FP 7, statistical research in FP 5 was considered a transversal discipline that cuts across the four Key Actions (KAs) of SP 2 and was hence labelled a Cross Programme Action (CPA). Thus, the calls for proposals for statistical and related research were realised within the following CPAs:

- CPA4 in 1999: New indicators and statistical methods;
- CPA7 in 2000: Socio-economic analysis for the information society;
- CPA8 in 2000: Statistical tools, methods indicators & applications for the Information Society.

CPA 7 came under the direct management of DG Information Society (DG 13). Thus, only CPA 4 and CPA 8 applied to EPROS.

The objective of CPA 4 was to develop and demonstrate new statistical tools and methods, to use statistical tools and methods in applications, and to develop indicators of the new economy. The objective of CPA8 was to develop new statistical tools, methods and indicators exploiting information society technologies; to demonstrate and disseminate their use in information society applications while serving the needs of official statistics within the ESS.

Projects were classified as follows:

(a) Research and development into generic statistical tools, techniques, methodologies and technologies;

(b) Statistical applications;

(c) Statistical indicators for the new economy (SINE);

(d) Transfer of technology and know-how (TTK);

(e) Supporting activities.

Implementation of SINE was, as far as possible, coordinated with the socio-economic research in SP 2 and with the SP on Human Potential and Socio-economic Knowledge Base. A SINE issues paper[20] was produced by the Commission on 15 July 1999, which served as a guide to potential proposers.

27 of the EPROS projects were formally categorized under (a) and (b) above, 6 under (c), 1 under (d) and 6 under (e). (Some projects are categorized under several categories; e.g. NESIS and STILE were classified as both SINE and Support Activities.) The last EPROS project ended on 30 April 2006. The total budget was €69.4 million, of which the EU contribution was €43.2 million.

5.2 The principles underlying EPROS

For EPROS projects to be successful, they were required to satisfy the following criteria:

- Have a clear and appreciable Research and Development (R&D) content. Research was defined according to the Frascati manual, as involving either the creation of new knowledge or new applications of existing knowledge;
- Fit the objectives and requirements of FP5 generally and SP 2 specifically;
- Therefore be both:
  - technology-driven, in the sense of exploiting advances in technology for statistical development, and
  - user-driven, that is involving users in the design and testing of prototype software;
- Have an actual or an identifiable potential application , with a clear dissemination and take-up plan;
- Rest on clear partnerships between national players and those at European level, with value-added, synergy, critical mass and network-building;
- Capitalize on past investments in statistical research with a view to their enhancement. For example, statistical disclosure control under FP 5 (project called CASC) built on the corresponding project, SDC , under FP4;
- Recognize that any IT-oriented research must be an ongoing activity in order to keep abreast of rapidly changing technologies.

5.3 The development of EPROS

EPROS evolved through a number of related activities:

- It was the culmination of systematic consultation over the two years 1997 and 1998 with DG Information Society (DG 13), with the DOSIS/EPROS Working Group of NSIs and with specially-convened think-tanks of academics at the leading-edge of statistical methodology and technology;

[20] *On the 15 July 1999, a 16-page seminal paper was produced by Messrs Chrissafis of DG Info Society and Deo Ramprakash entitled " Statistical Indicators for the New Economy (SINE), Towards 2000".*

- It was underpinned by bilateral visits undertaken by representatives of Eurostat to NSIs to ascertain their research needs;
- There were inputs into EPROS from various scientific conferences such as NTTS (New Techniques and Technologies for Statistics). The emphasis in the agenda of these conference was on current statistical research and on new ideas;
- It was endorsed by the Statistical Programme Committee;
- It was promulgated at a number of National and International Information Days convened by Eurostat over 2 years in the majority of Member States and it was therefore the springboard for the EPROS projects.

6 THE COVERAGE OF RESEARCH UNDER EPROS

This chapter describes the intended coverage of EPROS and its actual coverage in terms of the projects realised; the actual coverage is best organized and examined in terms of the results achieved under specific but broad headings, such as Methodology. These headings can be drawn from NORIS (Nomenclature on Research in Official Statistics) which was developed as an integral part of the evolution of EPROS in 1999, though its use remained internal to Eurostat. NORIS itself is spelt out in the Annex.

NORIS coverage was restricted to statistical tools and methods. Thus, by definition it excluded SINE, Networking, Dissemination and Exploitation.

6.1 THE INTENDED GIST OF EPROS

Given the principles enunciated above, there was substantial emphasis in EPROS on research designed:

- To adapt statistical concepts and nomenclatures to changing economic and social realities;
- To increase the use of automated data collection such as the internet, Computer Assisted Personal Interviewing (CAPI), Computer Assisted Telephone Interviewing (CATI) and satellite imaging in the context of GIS; to promote Electronic Data Interchange (EDI); to enhance data capture and interchange through electronic questionnaires;
- To improve data quality, including timeliness;
- To develop fresh theories of errors and error correction;
- To exploit new forms of data that arose as a by-product of technological applications in trade and commerce;
- To building prototypes of non-expert intelligent systems for knowledge extraction from large, complex and multidimensional databases;
- To timeseries analysis, particularly seasonal adjustment;
- To statistical integration through statistical matching and exact linkage, with more extensive use of administrative registers;
- To leading-edge or emerging information technologies in order to improve distributed access to microdata;
- To the maximization of intelligence, visualization, virtuality, interactivity and user-friendliness in the development of software interfaces;
- To the promotion of the transfer of statistical technology and know-how;
- To networking and partnerships between all stakeholders;
- To investment in human capital; and
- To the promotion of mobility, critical mass and institution-building.

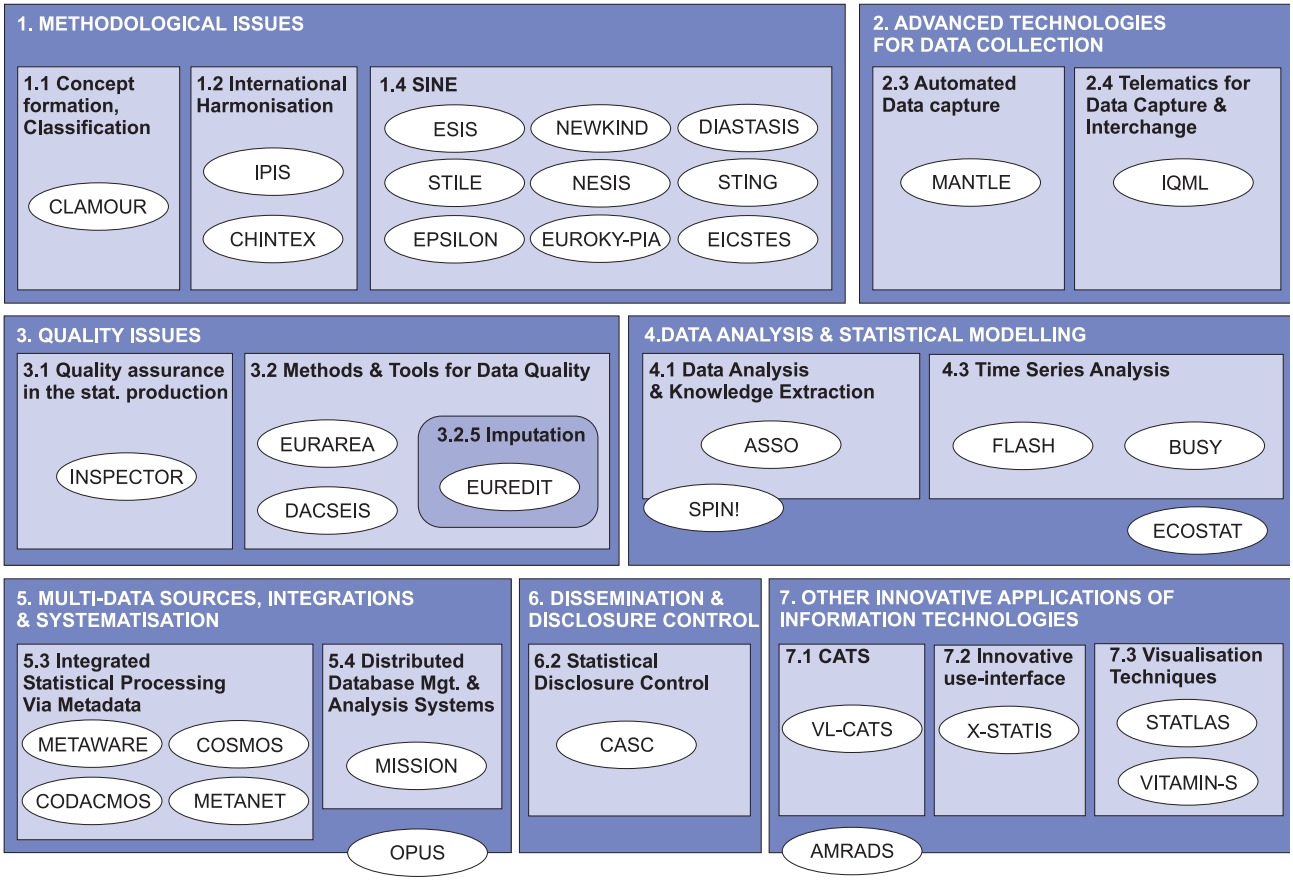
The ESS was first outlined in a Communication from the Commission to the Council in February 1992 and has been developed since then in full partnership between the EU member states and other members according to the principles of subsidiarity, partnership, transfer of competence and consultation. The aims of the ESS, which provides the institutional context of EPROS, are to:

- Use a system of standards, methods and organizational structures which are capable of producing comparable, reliable and relevant statistics throughout the Community;
- Provide the European institutions, in particular the Commission and the Governments of Member States, with the information they need to implement, monitor and evaluate Community policies;
- Disseminate statistical information to Europe, to businesses and to all concerned with economic and social matters;
- Seek to improve the statistical system in Member States and to support the advancement of the statistical system in developing countries and in countries that are in transition to a market economy.

At the highest level of consultation is the Statistics Programme Committee (SPC), consisting of the Directors General of NSIs of the Member States. It represents a delegation of authority from the Council of Ministers to act on the Council's behalf on statistical matters. A second tier of consultation is the Working Group of NSIs or/and Task Force in specific areas of statistics, for example the EPROS Working Group and the EPROS Task Force.

6.2 Actual coverage of EPROS

The actual coverage of EPROS is detailed in the following sections; where the centre of gravity of a project is clear-cut, it is classified only under one heading. Many projects contributed to more than one heading and are classified accordingly.



6.2.1 Methodological issues (NORIS 1)

6.2.1.1 Concept formation, classifications

Project	Comments
CLAMOUR	Aimed at improving the quality of classification systems, and the understanding of them, with the emphasis on economic classification systems such as NACE.
STILE	Explored the impact of the e-labour market on such nomenclatures as NACE (Rev 1) and ISCO 1988.

Economic restructuring requires a review of existing nomenclatures on economic activity (NACE), commodities, (CPA/CPS), trade (HS/CN) and occupations (ISCO). In the context of the continuing requirement to track the new economy, EPROS had called for greater explicit account to be taken of the technological characteristics and the knowledge content of the entities to be classified. Critical issues would include the relative merits of classification by product/commodities compared with classification by industry. Software' represents an increasing proportion of the ICT supply sector and the existing official statistics (based on NACE and ISIC, etc) do not provide adequate conceptual or operational categories to map and measure these important industrial and employment aspects of the expanding information economy/society.

As indicated in the table, both CLAMOUR and STILE have tried to grapple with a selection of these issues, with very valuable results. CLAMOUR was a new approach to units, economic classifications and the links between them. It had four complementary main themes: foundations; linguistics and automated coding; users' needs; and identifying and making tools for classifying modern industries. Out of a consortium of 7, 5 were NSIs, led by Office for National Statistics (ONS) in UK. However, there were a number of nomenclature issues in this project that were not addressed. For example, the industries for which actual model-based data were collected were necessarily limited. Therefore, further model-based data collection was recommended. Also, the model has to be adapted in such a way that it is possible to obtain the information needed by means of a Blaise (CAI) tool that can be used in practice. As individual projects can only focus on a limited range of issues, the ideal would have been to have more projects on nomenclatures.

STILE's activities aimed at providing innovative methodologies and content for the statistical monitoring of the European labour market in the eEconomy, including assessing the relevance of the NACE and ISCO classifications and the quality of coding of businesses and professions in the eEconomy. The project concluded that there should be further work to deepen insights into the various aspects that directly or indirectly influenced ISCO, such as changing inter- and intra-organisational divisions of labour (networking and outsourcing), organisational changes (production concepts), changes at the workplace (eWork), changes in labour market behaviour (mobility) and work-force composition (new businesses and new occupations). As STILE was an accompanying measure, it could not delve as deeply in these topics as might have been desirable, which implies a possible need for further research in this area.

6.2.1.2 International harmonization

Project	Comments
CHINTEX	Sought to clarify whether it was necessary to have centralised, standardised survey instruments to achieve comparability between national statistics or if this objective could also be achieved by ex-post harmonisation of existing sources. The sources involved included panel data.
COSMOS	Promoted comparability through metadata inter-operability

Harmonization at European level enable comparative economic and social research and the identification of "best" statistical and policy practices. EPROS did envisage research on both input and ex post harmonization and harmonization through the use of metadata. The advantage of ex post harmonization is that it allows NSIs some flexibility in meeting data needs at European level from national sources, designed primarily for domestic purposes.

Both CHINTEX and COSMOS conducted research in these areas. CHINTEX found that for the vast majority of variables there seemed to be appreciable differences between input and ex-post harmonised surveys. Moreover, ex-post harmonisation of the type pursued in the project was resource-intensive, requiring considerable human expertise and time. That conclusion was somewhat unexpected; further research, from different angles, might be required to verify whether this conclusion is generally valid or whether it is specific to this ex-post harmonisation framework.

6.2.2 Advanced technology for data collection (NORIS 2)

Project	Comments
CODACMOS	Identified ways, including electronic data interchange, in which the cost-effectiveness and quality of the official data collection process could be improved while lowering the burden on respondents.
EICSTES	Developed automatic web data collection procedures using agent technology.
INSPECTOR	Integrated data validation system into the processes of statistical data collection, in order to ensure and monitor data quality.
IQML	Assisted collection agencies of raw data to build collection instruments in a variety of forms (e.g. CATI, CAPI), using a common metadata model which could facilitate the development of, and access to, a common metadata repository.
MANTLE	Assessed the potential of using satellite data to produce maps of light emissions and urban night-time intensity levels in the EU. These urban light emissions then formed the basis for deriving indicators of population size/density; urban population numbers; total energy consumption by sector of activity and other local characteristics.
STILE	Devised labour market modules for piggy-backing on existing social surveys, thereby avoiding the need to undertake completely fresh, costly data collection and avoiding additional response burdens.

EPROS had called for research mainly in the automation of data collection through the extended use of:

- Automated data capture:
  - "Intelligent questionnaires"
  - Bar codes
  - Satellite images
  - Touchstone Data Entry (TDE), Voice Recognition (VR), Optical Mark Recognition (OMR), Optical Character Recognition (OCR)
- Telematics for data capture and interchange
- Coding metadata.

CODACMOS and IQML covered these areas. In addition, EICSTES researched the use of the internet for data collection (and dissemination). INSPECTOR was also innovative by incorporating validation procedures in the process of data collection. (This project appears also under Quality). MANTLE developed a methodology for producing policy-related indicators (population, GDP, etc.) from night-time light emission data.

None of the EPROS projects had embarked on large-scale data collection[21]. NESIS did collect data but that was more in the nature of case studies to fulfil its project objectives, specifically to explore indicator metrics.

[21] *An example of one-off, large-scale data collection was the SINE project, SIBIS, which fell outside Eurostat's management.*

6.2.3 Quality issues (NORIS 3)

Project	Comments
CHINTEX	In the process of pursuing its main objective about ex-post harmonisation, this project explored related issues of non-response, attrition effects, imputation, quality of income data and the estimation of statistical models.
CODACMOS	In the process of pursuing its main objective about ex-post harmonisation, this project explored related issues of non-response, attrition effects, imputation, quality of income data and the estimation of statistical models.
DACSEIS	Pursued standardization and harmonization of variance estimation methods used to calculate sampling errors. To some extent, longitudinal data were included.
EUREDIT	Investigated and evaluated methods for automatic editing and imputation.
FLASH	Sought to improve timeliness, with accuracy, of quarterly national accounts.
INSPECTOR	Developed a data validation system to be integrated in the processes of statistical data collection, in order to ensure and monitor data quality.
IQML	Ultimately improved the accuracy and timeliness of statistical data collection from enterprises and individuals whilst at the same time reducing the collection burden.
MANTLE	Using regression and discriminant analysis techniques, evaluated the potential to use light emissions as a surrogate for socio-economic indicators across the EU.
STING	Improved the quality and timeliness of the indicators produced on technological innovation.

Statistics underpin policies at all levels and increase transparency. But statistics can discharge those functions only if they are reliable, relevant, timely, comparable, accessible, user-friendly, credible to the citizen and not too burdensome or costly to collect. Reliable statistics have been needed to meet the challenges facing the various actors in the European Union, who want more information, of higher quality, faster, more comparable and at lower cost. That was why the following words appeared in Specific Programme 2 (SP2):

".. statistics are central to the information society, for which technologies offer new ways to attain the highest standards of quality ....."

EPROS had called for proposals for further research to ensure that the setting of standards is based on scientific principles and statistical methods in each of the following quality dimensions:

- relevance of the concepts and data to particular purposes,
- accuracy of estimates,
- timeliness, i.e. reduced time lag between the data release and the data collection reference period,
- accessibility to and clarity of the available information,
- comparability of the information between countries and over time,
- coherence between different sets of concepts and statistics, and
- completeness of the data in specific statistical exercises.

Thus, CHINTEX has, inter alia, investigated important hypotheses about the data quality of panel surveys. CODACMOS aimed to improve the cost-effectiveness of the data collection process and the quality of official statistics while lowering the burden on respondents. DACSEIS developed practical and usable methods for variance estimation in complex multi-purpose sample surveys. Based on selection from a prior list of criteria, these methods enabled the user reliably to estimate variances in national surveys. EUREDIT was a key project on Quality, investigating, as it did, current methods for editing and imputation to establish current best practice. FLASH is interesting in that it had to wrestle with the perennial problem of the trade-off between accuracy and timeliness in the production of one key economic indicator. Normally, editing and validation tended to be seen as operationally separate activities from data collection but, treating validation rules as metadata, INSPECTOR developed a validation system as an integral part of the data collection process in order to assure and monitor data quality. IQML aimed at providing a solution for the collection of timely statistics to support regional, national, and Community policy making. Finally, STING suggested new indicators on scientific and technological progress in Europe.

On the whole, the issues researched in these 8 projects represent a fair coverage of the Quality domain. This wide coverage was required given the paramount importance of Quality in statistics.

6.2.4 Data analysis & statistical modelling (NORIS 4)

6.2.4.1 Data analysis and knowledge extraction

Project	Comments
ECOSTAT	Developed data mining software tools and advanced specialized statistical modelling to identify latent relationships in environmental phenomena.
SPIN!	Developed a spatial data mining system.
VITAMIN S	Developed a system for statistical visualization with a data mining perspective.

The need for intelligent approaches has increased as rapid progress in information technologies brought in its wake an exponential growth in very large databases, very often embodying complex interdependencies, multidimensionality, non-linearities, missing data and noise. Data have increased but there has not been a commensurate increase in knowledge. Thus, research in data mining techniques and the actual mining of large data sets were proposed in EPROS. The 3 projects, ECOSTAT, SPIN! and VITAMIN S, have all concentrated, in varying degrees, on knowledge extraction. Also, the application areas were interesting: environment in ECOSTAT; seismic and volcano data, as well as census data, in SPIN!; and the development and use of several visualization methods in VITAMIN S to facilitate the data analysis process.

Despite the following suggestion in EPROS: “However, these (intelligent) techniques require further, application-independent development in their own right and the establishment of links with classical statistical inference”, there was no project comparing classical statistical methodologies with data mining through, for example, neural networks.

6.2.4.2 Timeseries analysis (including now- and forecasting)

Project	Comments
BUSY	Addressed the statistical analysis of business cycles in the European Union.
FLASH	Sought to produce 'flash' estimates of the main aggregates of the Quarterly National Accounts of the Euro Area and the EU with a target delay of 40-45 days after the end of the reference period.

Traditionally, timeseries analysis has been one of the priority topics in the research agenda of NSIs, though much of the theoretical advances were arguably made by academia. Thus, EPROS had listed 10 specific areas for future study. BUSY, the only project in the timeseries field, involved key stakeholders in Europe, including 3 large NSIs; the analysis of business cycles in this project represented a major stride forward in the accumulation of knowledge in this area. Nevertheless, given the strong emphasis in EPROS on this subject, it is somewhat surprising that more projects did not materialize.

6.2.4.3 Small area disaggregation/estimation/geographical information systems (GIS)

Project	Comments
EPSILON	Conducted sustainability clustering at NUTS-III level, including the use of GIS web based technologies.
EURAREA	Sought to improve small area estimation methods currently used by NSIs.
MANTLE	Used a GIS based light emission model to produce maps and indicators of, for example, population size/density; typology; and energy waste.
OPUS	Used its statistical framework for the estimation of indicators of urban/regional mobility.
SPIN!	Offered new possibilities for the analysis of geo-referenced data.
STATLAS	Developed an integrated electronic atlas of statistical information, covering the EU at country and regional levels.

This area could have been highlighted more in EPROS. Small area estimation, GIS and ESDA (Exploratory Statistical Data Analysis) were relatively old issues but have been emerging more prominently in recent years. EURAREA, coordinated by ONS, researched the effectiveness of standard small area estimation techniques and developed enhanced estimators; the enhancements themselves were tested on data relating to unemployment status, income and household composition. The inclusion of 6 NSIs in the consortium indicates the strength of ESS interest in this subject. In 2006 the EPROS Task Force earmarked it for further research.

EPSILON delivered GIS web-based technologies. An interesting innovation of SPIN! was the integration of a spatial data mining system and visualization methods into GIS.

6.2.4.4 OTHER

Project	Comments
ASSO	ASSO designed methodology and software tools for the analysis of multidimensional complex data from very large databases, redefined as 'symbolic objects'; symbolic analysis is an extension of statistical data analysis methods to these complex objects.
X-STATIS	Provided software that would allow non-experts, particularly SMEs too small to recruit theirown statisticians, to select and apply appropriate analysis techniques and to interpret the results.

ASSO (Analysis System of Symbolic Official Data) sought further to improve and develop the methodologies and software developed in the DOSIS project called SODAS, which had conducted research on the use of 'symbolic objects'. A large user community was established and training and support in the use of symbolic objects provided. At the end of the project, a non-profit association was established to follow up the results of ASSO and assure continued maintenance, development, support and training. The results are being actively used by project partners, both non-commercial (universities, NSIs) and commercial (Small and Medium-sized Enterprises).

6.2.5 MULTI-DATA SOURCES, INTEGRATION & SYSTEMATIZATION (NORIS 5)

Data integration was a priority topic in EPROS. Because of the cost of collecting data afresh every time a need arose, the alternative of bringing together disparate data sets that already existed should be explored. For example, to derive multidimensional indicators or to carry out policy impact assessments would require merging different databases each covering only a subset of the required dimensions. But such integration is fraught with problems, not least that of the confidentiality of the underlying microdata. The full range of research topics envisaged in EPROS was:

- The use of administrative data for statistics;
- Multi-source data environments;
- Integrated statistical processing via metadata;
- Distributed database management and analysis systems;
- IT infrastructures.

6.2.5.1 Integrated statistical processing via metadata

Project	Comments
CODACMOS	Developed a common metadata model.
COSMOS	Clustered five IST projects with a shared interest in statistical metadata and metadata repositories.
DACSEIS	As an incidental task, developed efficient methods for combining data from surveys and registers.
IPIS	Integrated data through, inter alia, the development of a statistical metadata model.
IQML	Used a common metadata model which would facilitate the development of, and access to, a common metadata repository.
METANET	As a thematic network, developed proposals for standards in the methodology used for describing statistical metadata and statistical information systems.
METAWARE	Developed a standard metadata repository model for data warehouses.
MISSION	Developed a metadata model to support the query processing.
OPUS	Used metadata to generalise OPUS's results, using Bayesian inference linked to a comprehensive metadata and process metadata characterization of the inference process.

Data integration through metadata had attracted a relatively substantial number of projects. Of the DOSIS projects, 4 were in this area; of the EPROS projects, 9 were on integration through metadata. This has also been an appreciable activity under SUP.COM. Thus research on this topic was well-subscribed. That is to some extent understandable because:

- Metadata provides a means to harmonization which enhances the value of relational database management systems;
- Metadata improves the prospects of richer knowledge discovery;
- The prevailing ethos is "no data without metadata".

One driving force behind the research into metadata was the University of Edinburgh, which was the coordinator of several DOSIS and EPROS projects in this field. This institution was also a partner in CODACMOS and AMRADS, which aimed to identify and transfer best practice in metadata systems.

Metadata issues were central to METANET, slightly less so to the other projects in the table, which had other primary objectives. The aim of METANET, which was a Thematic Network, was to come up with standards by reviewing and consolidating the work on metadata models that had been carried out in DOSIS, SUP.COM and other EPROS projects. The projects, FASTER[22] , (Flexible Access to Statistics, Tables and Electronic Resources) and DIECOFIS[23] , (Development of a System of Indicators on Competitiveness and Fiscal Impact on Enterprise Performance), were not EPROS projects, but were managed by DG Information Society. However, they both covered integration. Based on a web environment with client side functionality, and built around the careful specification of metadata content, FASTER was designed to create a system for access to statistical and other data in a distributed 'virtual' environment. It allowed users to create their own personal data environment, derive all relevant contextual and supporting information and hence make the most productive use of expensive data. DIECOFIS integrated data from a variety of sources on enterprises in Italy, with a view to assessing the impact of taxes on their growth performance.

[22] *Flexible Access to Statistics, Tables and Electronic Resources (http://www.data-archive.ac.uk/randd/faster.asp?print=1)*

[23] *Development of a System of Indicators on Competitiveness and Fiscal Impact on enterprise Performance (http://petra1.istat.it/diecofis)*

6.2.5.2 Other integration

Project	Comments
ASSO	Investigated the possibility of joining databases from independent surveys.
IPIS	Integrated statistical data produced by different statistical producers through Europe in order to enhance data accessibility.
OPUS	Developed a generic statistical framework for the optimal combination of complex spatial and temporal data from surveys, censuses and other sources.

Comments were already made above about ASSO and IPIS. OPUS came very close to the original aim of EPROS in the field of data integration, which was to combine diverse data sources in a particular domain and to examine the feasibility of extending the merger to other domains. One conclusion of this project is noteworthy: "There was widespread support for the pragmatic Bayesian approach adopted by OPUS but the use of these methods inevitably placed more emphasis on the characterization of sources of uncertainty (sampling and non-sampling) affecting existing data sources. That would, in many cases, require new levels of rigour in the treatment of survey non-response and instrument/measurement effects".

6.2.6 Dissemination, disclosure control (NORIS 6)

6.2.6.1 Computer-assisted techniques for training

Project	Comments
VL-CATS	This project centred on e-training, i.e. provision of interactive teaching and access to educational material over the internet, specifically addressing areas relevant to official statistics.

The VL-CATS project developed a system based on groupware technology for delivery of distance training in official statistics, using a dedicated internet site with reference material and multimedia enhanced training courses. These courses covered areas such as Official Statistics and the European Statistical System; Data Analysis; Time Series Analysis; Inference and Statistical Modelling; Statistical Disclosure Control; Sampling Theory; and Strategic Management. Due to changes in the framework for training for official statistics, the launch of these training offers could not be made as originally foreseen. However, such dedicated interactive training offers for official statisticians would complement other training offers (traditional courses, workshops, etc), thus offering an additional important channel for human resources development and transfer of technology and know-how.

6.2.6.2 Statistical disclosure control

Project	Comments
CASC	Sought to enhance existing tools and to develop new methods for disclosure limitation through major extensions of the ARGUS software.

Confidentiality of individual data is a continuous critical issue for official statistics. Ever-growing computing power implies that threats of disclosure of confidential information are increasing; further development of methods and tools for statistical disclosure control is therefore a never-ending research issue in the ESS. CASC built on the DOSIS project SDC; CASC research was aimed primarily at the development of new methods for releasing business microdata while new optimization methods were sought for the protection of hierarchical and linked tables. The ARGUS software for disclosure control, developed through SDC and CASC, is freely available to NSIs.

6.2.6.3 New applications in IT: visualization techniques

Project	Comments
ASSO	ASSO produced new tools for visualizing Symbolic Objects and/or the results of Symbolic Analysis.
EICSTES	Developed a tool to visualize and map its results.
EPSILON	Combined GIS tools selected in the project to deliver, apply and validate a visualization model of the input and output data.
SPIN!	Advanced GIS by the development of new methods for the visualization of spatial and temporal information and of data mining methods.
STATLAS	Focused on cartographic visualization environment for statistical data.
VITAMIN S	Developed visualization tools for analyzing large survey data and time series data in order to discover data patterns, trends, associations, clusters and outliers.

EPROS had explicitly envisaged research in the area of visualization. The objectives had been to go beyond conventional static graphics in order to condense and represent a large mass of essentially correlated statistical information. This objective was to be achieved through research into new computer-graphics, image-synthesis and non-reversible image-compression technologies; through providing context-sensitive querying and help; and through extensive linking to parallel displays applied to statistical information, while staying close to the real needs of users. These six projects responded well to the EPROS appeal.

EICSTES, SPIN! and STATLAS have in common that the visualization developed related to the spatial dimension. VITAMIN S was more centred on visualization as a basic step in the exploratory data analysis process.

6.2.6.4 Other dissemination

Project	Comments
MISSION	Produced a software suite that allows statistical data providers to publish data on the Web

Within research into methods of dissemination, MISSION produced a modular software suite aimed at enabling providers of official statistics to publish data in a unified framework, and allowing users to share methodologies for comparative analysis and harmonization. The modules are distributed over the web and communicated via agents.

6.2.7 STATISTICAL SOFTWARE DEVELOPED

Given that EPROS was funded from the IST programme, the natural expectation was that the majority of research projects would come up with prototype software outputs, though of course the software concerned had ultimately to contribute to the better performance of some aspect of the statistical process. The following projects have done so in a major way:

Project	Comments
ASSO	Designed application software tools for the analysis of multidimensional complex data coming from very large databases in NSIs and other administrations.
BUSY	Developed a software tool for the statistical analysis of business cycles in the European Union. Prototype interface software was produced.
CASC	Extended the ARGUS software to accommodate new methods of data protection for both microdata and tabular data.
DIASTASIS	Developed the DIASTASIS Statistical Collaboration platform (SCoP).
ECOSTAT	Developed prototype software for environmental modelling.
EPSILON	Developed a software tool for generating sustainability indicators.
ESIS	Produced software called SPAD ESIS for measuring quality performance of companies and organisations from the customer perspective.
EUREDIT	Integrated the individual edit and imputation methods into a prototype software system.
FLASH	Implemented and assessed the overall methodology for producing 'flash' estimates of the main quarterly national accounts aggregates in computer programmes compiled in C++.
INSPECTOR	Produced a declarative validation process, and a generic, distributed and flexible data validation software system.
IPIS	Developed a modular system to assist public administrations in efficiently using and exploiting distributed information systems, with interface between the databases and the application.
METAWARE	Developed a common interface prototype to facilitate the connection of the metadata repository to different kinds of statistics production systems.
MISSION	Produced modular suite to enable providers of official statistics to publish data in a unified framework, allowing users to share methodologies for comparative analysis and harmonisation.
OPUS	Developed software to enable the combination of complex spatial, temporal and real time data in a statistically coherent fashion
SPIN!	Delivered an integrated software environment for spatial data mining.
STATLAS	Developed an interactive statistical atlas with a number of cartographic and statistical functionalities for portraying, comparing and analysing statistical data in a spatial context.
STING	Developed a computer-assisted system for the analysis of patents data.
VITAMIN S	Developed software for data mining and graphical data analysis through innovative visualization techniques.
X-STATIS	Provided software that would allow non-experts to select and apply appropriate analysis techniques and to interpret the results.

7. STATISTICAL INDICATORS

The SINE seminal paper had envisaged activities on the following aspects of indicators:

- Indicators as a generic issue (resonance, relevance, parsimony, taxonomy);
- Statistical quality of indicators, using Eurostat's quality criteria;
- Data collection on indicators;
- Indicators for policy development, monitoring and evaluation in terms of impacts;
- Domain or thematic indicators such as on health and environment;
- As a special theme, indicators on the new economy, including readiness, intensity, impact and outcome indicators;
- Indicators on sustainability;
- Methodological work on composite indicators.

The 6 SINE projects in the table below delivered, to varying extents, on most of these activities. NESIS and STILE were key indicator projects within EPROS. STILE delivered indicators on various aspects of the ICT-related labour market, such as jobs created, quality of work, mobility of ICT personnel, ICT vocational training, etc. The project also developed add-on modules on such ICT-related labour aspects for addition to the Labour Force Survey and generally for different business surveys.

The fundamental rationale for the NESIS project was that policy interventions could and should be guided by appropriate statistical indicators. Statistical information incorporating high-quality indicators is a fundamental requirement of a knowledge-based society and of transparent governance. These indicators should be capable of being applied at national, European and international levels, for purposes of benchmarking and policy learning. Accordingly, NESIS not only derived benchmarking indicators in the domain of the new economy but studied indicators as a generic issue. It also applied Eurostat's seven quality criteria in statistically appraising indicator sets drawn from various data sources; and it also proposed a novel quality assessment regime for indicators. NESIS quality-appraised EU indicator sets such as eEurope 2005, the EU's innovation scoreboard and the EU's science and technology indicators

NESIS had established explicit links with the 18 SINE projects, in particular BEEP (Best e-Europe Practices), SIBIS (Statistical Indicators for Benchmarking the Information Society), BISER (Benchmarking the Information Society e-Europe Indicators for European Regions), STILE and NEWKIND. The first three of these were not included in EPROS, but in the context of indicators, it is relevant to sketch also these projects.

There were basic differences between NESIS and BEEP; First, BEEP was interested in identifying socio-economic best practice in various domains of e-Europe, whereas NESIS was primarily interested in identifying statistical best practice in some of those same domains. Secondly, whereas NESIS had a strong interest in the use of statistical indicators for purposes of 'benchmarking', BEEP was concerned more with 'bench learning'. SIBIS attempted its own stocktaking of available new economy indicators and it sought to generate a range of new indicators based on a large-scale data collection exercise. SIBIS also developed "compound" indicators, the methodology of which could be used to achieve a numerical reduction. The BISER project aimed to develop, define and pilot statistical indicators for measuring and benchmarking the impact of the knowledge economy on Europe's regions, based on a model of the factors which influence regional development and on a general population survey. As already stated, STILE was concerned principally with the ways in which ICTs had been transforming work and how these changes could be measured. NEWKIND focused on industrial and enterprise performance in the new economy. EICSTES, ESIS and STING derived indicators in certain specific domains.

The data collection approach of the 18 SINE projects varied: SIBIS was characterised by large-scale surveys to test the usefulness of indicators; NESIS undertook small-scale pilot surveys as a precursor to indicator development concerning social capital, firm dynamics and other issues; BISER piloted a set of statistical indicators with a view to their being taken up by official statistical institutes; SEAMATE (Socio-Economic Analysis and Macro-modelling of Adapting to Information Technologies in Europe) made quantitative estimates of the microeconomic impact of IST in industry; BEEP relied on existing data; STILE fine-tuned statistics to match the e-economy and sought to develop ICT-related labour survey modules that could add on to the Labour Force Survey (LFS) and business surveys.

7.1 SINE PROJECTS

Project	Comments
NESIS	A key SINE project, identifying, assessing and developing indicators covering all aspects of the new economy
NEWKIND	Identified key issues and indicators related to the accumulation, distribution, and use of knowledge at the macro level and for two sectors of the new economy.
STILE	Developed indicators on the impact of the e-economy on the concepts and measures of the e-labour market, particularly telework, occupations and mobility.
EICSTES	Derived 74 indicators about the European Science-Technology-Economy System in Internet.
ESIS	Developed a new indicator system to measure customer satisfaction and perceived service quality for companies and organisations in the new economy.
STING	Developed a methodology for measuring technological innovation, from which indicators would be produced on a regular basis.

7.2 OTHER PROJECTS THAT PRODUCED INDICATORS

The following 4 EPROS projects were not formally classified under SINE, but also developed indicators. They produced indicators on internet use, the environment and policy impacts.

Project	Comments
DIASTASIS	Investigated how new indicators of internet usage could be developed for supporting decisions on the new economy.
ECOSTAT	Elaborated statistical indicators measuring environmental variables, in particular on water quality.
EPSILON	Delivered and clustered environmental sustainability indicators by exploiting Internet based IST. Developed composite indicators.
EUROKY-PIA	Sought to develop decomposable indicators permitting assessment of systemic strength and weaknesses, and link unambiguously indicators with policy goals to assess policy response indicators against policy outcome and contextual indicators

A few possible lessons for the ESS from these indicator activities are as follows:

- The new economy is very difficult to define and measure uniquely, simply and in a way that was both relevant to policy goals and universally acceptable. However, certain areas were identified as important candidates for particular attention, such as productivity, intangible assets including knowledge, innovation, ICT investment, e-security, the regional dimension, e-com/e-business, e-labour market and the household economy;

- The need for ad hoc, customized data collections to meet the immediate and specific requirements of individual research projects will continue. Some of these collection methods resort to "quick and dirty" expedients to meet urgent political imperatives. The data collection concerned is normally undertaken outside the control of the ESS. The question that arises is the lessons that can be learnt from these external activities for the strengthening of the regular production process of the ESS.
- The advantages of such special collections, for example the large-scale, EU-wide household survey undertaken in SIBIS, are that they can achieve high comparability between countries and great consistency between concepts for deriving truly multivariate measurements. Generally speaking, the disadvantages are that they are mostly one-off, not generally subject to the kind of rigorous quality control that is the hallmark of official statistics, cannot generate timeseries and cannot be readily integrated into the mosaic of ESS data;
- The need for new indicators did not necessarily call for new data but innovative uses of existing data. Thus, attention should be given to integrating existing data from different sources and to greater secondary exploitation. In this process, data from private sources such as EITO (European Information Technology Observatory) and RIPE (Réseaux IP Européens) should not be ignored.

8. Transfer of technology and know-how (TTK)

Project	Comments
AMRADS	This was the only EPROS project devoted systematically to TTK in its own right. It illuminated the issues concerned with the identification of current best methods (CBM). CBMs were identified in 6 statistical domains, viz business statistics, confidentiality disclosure, electronic data collection, metadata, timeseries analysis and statistical quality. The actual transfer of these CBMs to interested NSIs was through hands-on training.

AMRADS was designed to create the conditions for facilitating technology and know-how transfer within the ESS. It promoted a networked take-up culture as well as actual exploitation of specific research products and prototypes.
AMRADS showed that:

- Neither NSIs nor users are homogeneous. They are at different levels of development and have different interests;
- TTK is a process involving small steps; what was impossible yesterday is thinkable today and do-able tomorrow;
- TTK is generic, which means that the key findings of AMRADS transcend the project itself;
- Current best method (CBM) is a relative and dynamic concept;
- The level of development of CBMs varied between themes; For example, there was no CBM in the theme on Metadata, while good practices were more firmly established in the theme on SDC (Statistical Disclosure Control);
- A CBM need not always be a tangible product; It could be an idea or a list of prerequisites for effective implementation strategies;
- Available CBM systems have to be tailored to the needs and capabilities of the beneficiary organization;
- TTK can involve not only NSIs but also private sector organizations;

The real value-added of AMRADS was that it provided a roadmap to users about the relative merits and defects of adopting different systems and methods. Very clear and concise back-up documentation was essential. That roadmap was all the more necessary to allow users to navigate through the labyrinth of technical papers in the project website. The AMRADS training programme did focus concretely on CBMs in order to impart hands-on skills.

The subject of TTK is of continuing concern; A distinction should be made between (a) dissemination, which spreads awareness and may or may not lead to exploitation or TTK, (b) exploitation, which pro-actively takes research results to the market and could be aided by prior dissemination, and (c) TTK, which is the export of established, mature best-practice products and know-how mainly from one NSI to another, where the recipient could have learnt of the available CBM through prior dissemination.

9. NETWORKING / SUPPORT ACTIVITIES

Action at the European level has a clear added-value in promoting coordination, comparability, standardization, harmonization, synergies, critical mass, "bench learning", the identification of best practice and the establishment of centres of excellence. This process needs all actors to communicate, meet and steer the research effort productively, i.e. to network.

Institutional and international diversity in the network is important for the strengthening of the European Research Area (ERA). Possible actors are:

- Official and academic statisticians with theoretical and domain expertise from national and international institutions;
- Experts from the private sector;
- IT and database specialists;
- Where relevant, other disciplines, e.g. economists, playing their part in multidisciplinary teamwork;
- Data providers, such as businesses;
- Professional trainers in statistics;
- Statistics users ranging from the national policy-maker, the Commission Directorates, Central Banks, to the non-specialist layman;
- Professional statistical associations, such as the International Statistical Institute.

A particular contribution of EPROS was, for example, establishing cooperation between NSIs in one country and academia in another country. Networks could also be promoted through the use of distributed new technologies to create, inter alia, virtual laboratories.

Illuminating examples of networking, and dissemination, in the EPROS project were BUSY and FLASH. Interest groups were set up involving a community of potential users such as NSIs, Eurostat, the Directorate General for Economic and Financial Affairs of the European Commission (DG ECFIN), the Central Banks of the member states, the European Central Bank (ECB), the Bank of International Settlements (BIS), the OECD, private research institutions such as the National Institute for Economic and Social Research (NIESR) and the Economic Research Cycle Institute (ECRI), universities and public research institutes. Direct support to the dissemination of the software, BUSY, was provided through a helpdesk on web, within the site "Time Series Analysis for Official Statisticians". Through installation and demonstration on the premises, a list of test-users, which had received priority support, was drawn up. There was also support in the post-project phase. Thus, the network was wide and operational.

The following table summarizes the institutional make-up of each EPROS project. The extreme right-hand column shows the percentage of NSIs in each project. For example, out of the 14 members of the CASC consortium, 43% were NSIs. There were an (arithmetic) average of 2.33 NSIs per project and, interestingly, a higher average, 2.53, academic institutions per project.

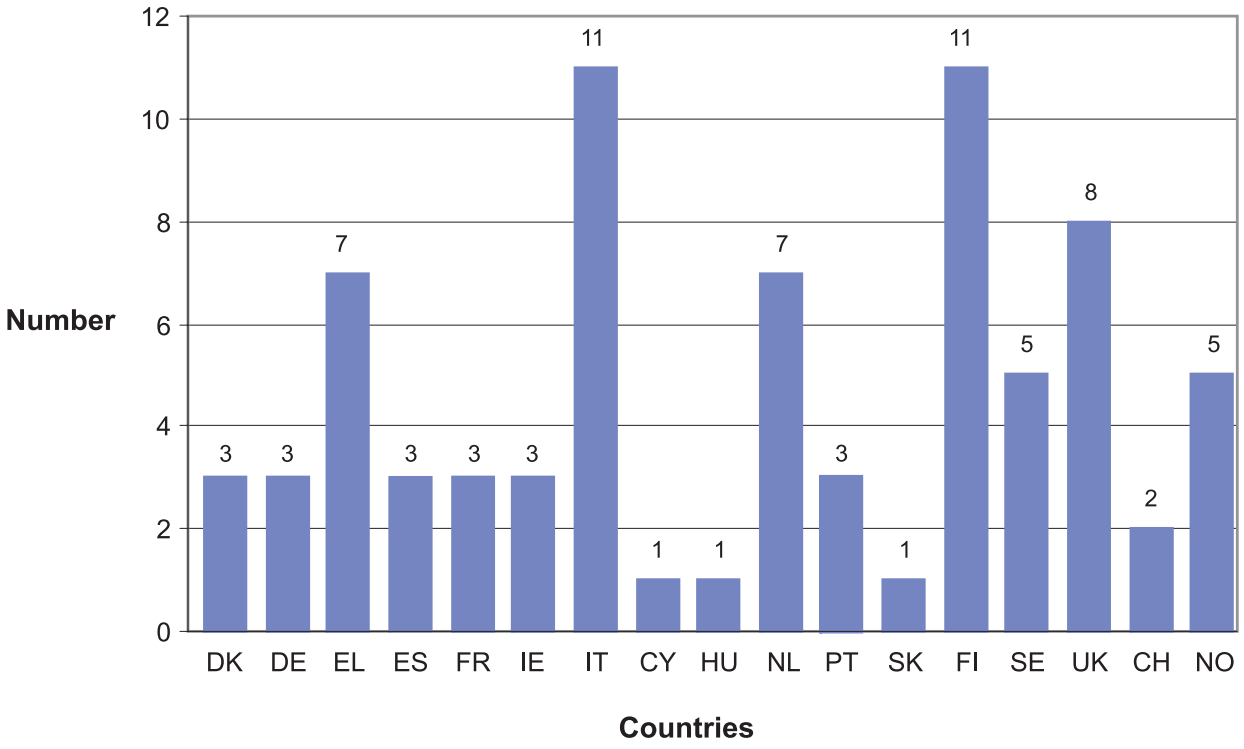
NETWORKING WITHIN EPROS PROJECTS

Projects	NSIs	Academia	Research bodies	Other public administration	Companies	Total	% NSI
AMRADS	4	2	1	0	2	9	44
ASSO	3	10	1	0	0	14	21
BUSY	3	1	2	0	1	7	43
CASC	6	7	1	0	0	14	43
CHINTEX	2	4	1	0	0	7	29
CLAMOUR	5	0	0	0	2	7	71
CODACMOS	5	3	0	3	2	13	38
COSMOS	1	4	0	0	3	8	13
DACSEIS	4	4	0	0	0	8	50
DIASTASIS	1	2	1	0	1	5	20
ECOSTAT	1	3	0	0	1	5	20
EICSTES	1	2	6	0	0	9	11
EPSILON	3	2	0	2	2	10	30
ESIS	0	2	0	0	6	8	0
EURAREA	6	4	0	0	0	10	60
EUREDIT	6	4	0	0	2	12	50
EUROKY-PIA	1	1	0	1	3	6	17
FLASH	1	0	3	1	0	5	20
INSPECTOR	2	2	0	0	2	6	33
IPIS	1	3	0	3	1	8	13
IQML	2	2	0	0	3	7	29
MANTLE	0	3	1	1	3	8	0
METANET	3	3	0	0	1	7	43
METAWARE	5	0	0	0	1	6	83
MISSION	3	3	0	0	1	7	43
NESIS	3	2	1	0	1	7	43
NEWKIND	2	0	0	1	0	3	67
OPUS	0	3	0	1	3	7	0
SPIN!	0	5	2	0	2	9	0
STATLAS	0	3	0	1	1	5	0
STAT-OBJECT	3	1	0	0	2	6	50
STILE	1	0	6	0	2	9	11
STING	2	1	1	1	1	6	33
VITAMIN S	2	1	0	0	1	4	50
VL-CATS	0	4	0	1	2	7	0
X-STATIS	2	0	1	1	2	6	33
Total	**84**	**91**	**28**	**16**	**54**	**275**	**31**
Average	**2,33**	**2,53**	**0,78**	**0,44**	**1,50**		

Note: *The World Health Organization was the only international organization which was formally a partner in EPROS (OPUS project); other organizations such as OECD, ECB, BIS and different DGs of the European Commission (ECFIN, ENTR, INFSO) participated in user groups, conferences, workshops, etc.*

Given the active role of the NSIs in the preparation of EPROS, it is interesting to illustrate the extent of their participation in the projects. The chart below displays the number of times (frequency) a country's NSI was a formal partner in an EPROS project. Many countries not indicated in this chart, as their NSI was not formally a partner in an EPROS project, were still involved; However, they had other roles, such as providing external expert advice, and participating in user groups and conferences, seminars and workshops arranged by the projects.

FREQUENCY WITH WHICH NSIS APPEARED IN EPROS PPROJECTS



10. DISSEMINATION

Nearly all of the EPROS projects provided a specific dissemination and exploitation plan. They also set up a project website and in many cases a help-desk to varying degrees of sophistication (including in certain cases visualization and multilingual support). Nearly all EPROS projects issued a project brochure; held workshops and international conferences; published scientific papers and books (including manuals where appropriate and the outputs in electronic forms); and undertook clustering with other EPROS projects to reap synergies. A number of projects arranged demonstration events and commercial fairs, and maintained their websites and provided consultancy services even after the end of the project.

Furthermore, presentation of progress and results to the Eurostat EPROS Working Group, consisting of researchers from all NSIs of the ESS, was compulsory. The projects were also strongly encouraged to submit papers and presentations of their research and results to key international conferences and seminars (e.g. the ISI conferences, the Q200x conferences, the NTTS conferences and ETK seminars, RSS conferences, etc) and to international publications in the research area(s) concerned.

The intricacy of the dissemination activities in EPROS can be conveyed by two examples, INSPECTOR and NESIS: The main objective of INSPECTOR's dissemination strategy included the promotion of the project's results and concepts to the main target groups, consisting of:

- Practising official statistics community, mainly NSIs, as well as other organizations which produced statistical information;
- The academic community involved in official statistics and IT;
- The broad user community.

The core dissemination instrument of INSPECTOR was the User Group, comprising NSIs, representatives of academia and other interests. This User Group acted through events organized in different European cities, including open workshops, demonstrations of the project results, press conferences, scientific publications and participation in information days; and through a distributed web site, publishing not only project-specific material but also material contributed by the network of participants, as well as a mailing list and an electronic newsletter.

First of all, INSPECTOR identified the target users and their requirements as a basis for the development of the project's classification scheme for data validation rules, the quality metrics, the data model and the system's architecture. Lists of users were established with direct mailing and encouragement to users to register and participate. In e-mail bulletins, the project disseminated a glossy leaflet, flyers and brochures, together with the URL of the project website. INSPECTOR's two workshops and conferences aired the project's results widely and obtained valuable feedback. INSPECTOR also disseminated its results, with synergy, through project clustering with other EPROS projects such as EUREDIT.

An interesting dissemination device was the NESIS Indicator Development Working Group (IDWG) and how it was used as a bridge into the NESIS final conference. The IDWG met periodically and comprised an almost standing body of eminent experts external to the project, including users, who scrutinized the intermediate outputs of NESIS and gave their opinions. The participants represented a wide variety of organizations such as the US Department of Commerce, OECD, DG Information Society, DG Enterprise, NSIs, academia and the business world. IDWG discussions fed directly into the final conference, which not only looked back at the challenges and achievements of NESIS over its 3 years but also tried to serve as a trampoline for launching a new programme of future work on the new economy. All the proceedings of the IDWG, the conference and the NESIS 8 workshops were disseminated through the NESIS web site.

11. Exploitation: using the research results

'Exploitation' is in this context meant as the use of the research results in the statistical production process of the ESS, though other uses are also important. Again, the IST Guide for Proposers emphasised, "The consortia should clearly identify the conditions required to maximise the exploitation of successful results". From the outset of the research programme, Eurostat ensured that gaps in exploitation arrangements were closed before the project contracts were signed. By the end of the EPROS programme, there was even greater awareness of the crucial importance of exploitation, as this had been underlined at every meeting of the Eurostat EPROS Working Group.

A few coordinators of EPROS projects, such as the coordinator of MISSION, considered exploitation in a more differentiated way; they considered that exploitation of the output could take two forms: direct and indirect. By indirect exploitation was meant the take-up of the ideas and approach of the project without directly using the product. This is the exploitation of the models and know-how that come out of the project. The participating NSIs would exploit the ideas and approach within their own organization. Direct exploitation meant the direct development of a marketable product. The coordinators concerned proceeded to caution that exploitation was not always easily achieved. In some cases, the software product could be highly specialized and, consequently, the market limited and the organizations forming this market heterogeneous. If the software was so specialized that it would not be 'off the shelf', this implied that installation and deployment would always generate a substantial amount of additional work and costs.

Exploitation or take-up depends ultimately on a number of factors such as:

- The extent of dissemination of the research outputs. Dissemination is a necessary but not sufficient condition for exploitation;
- The degree of specialization of the product and the extra work and costs required to adapt it into a more generic software;
- The extent to which users were formally involved in the consortium, or, outside the consortium, involved in development or testing of prototype software;
- The extent to which the research was concerned with the development of generic statistical tools, techniques and methodologies. Research in this category could be more technology-driven than user-driven, and even futuristic and visionary. There would in this case be a longer gestation period between research outputs and the market, even extending beyond the administrative time-frames of the Framework Programme(s);

- Projects under what are called Support Measures under FP5 would have more immediate prospect of exploitation because they would have been designed with that objective in mind. These would include projects which combined research with Demonstration, Take-up (access, assessment, best practice and trials), Concerted Actions and Thematic Networks.

There is exploitation of the results of many of the EPROS projects. For example, the results of the ASSO projects are currently being used by several project partners. These uses are both non-commercial (universities, NSIs) and commercial through Small and Medium Enterprises (SMEs). A no-profit association was set up to guarantee the continuous maintenance, training and support of the software (SODAS2) developed within the project. The projects results have been further developed both by individual project partners and amongst them acting in concert.

Similarly, the results of the METAWARE project are used both on a commercial basis and internally within the project. They were an input to the Neuchâtel Group and were presented at the meeting of the Eurostat "Metadata Working Group" held in Luxembourg on 7-8 June 2007. The software is continuously developed, particularly by a German company which was a sub-contractor in the project. At least five NSIs have in some way implemented parts of the project results in their metadata projects. In particular, the project outputs have been used by the Federal Statistical Office of Switzerland and also by Statistics Denmark and Statistics Sweden. There was also transfer of technology and know-how within the frame of the EU PHARE and TACIS programmes to East European countries and countries of the former Soviet Union.

EPROS PROJECTS

II

eurostat

AMRADS
Accompanying Measure to R&D in Statistics

Timetable	1/1/2001-31/12/2003 (36 months)

The Consortium

Member	Role	Institutional type	Country
Informer S.A	Coordinator	Company (IT)	Greece
Joint Research Centre (JRC, Ispra), European Commission	Partner	Public sector (research)	Italy
CS Italia	Partner	Company (IT)	Italy
Statistics Netherlands (CBS)	Partner	NSI	Netherlands
Office for National Statistics (ONS)	Partner	NSI	United Kingdom
Statistics Finland (StatFi)	Partner	NSI	Finland
Istituto Nazionale di Statistica (ISTAT)	Partner	NSI	Italy
University of Bocconi	Partner	Academia	Italy
University of Edinburgh	Partner	Academia	United Kingdom

Scope and Objectives

For many years, key stakeholders within the ESS were concerned about two interrelated issues: (a) the poor take-up of the results of R&D in statistics; and (b) very limited transfer of technology and know-how (TTK) amongst NSIs based on best practice and effective networking in order to achieve a general leveling-up of the ESS. Thus, AMRADS was seen as a measure that accompanied EPROS.

The specific objectives of this accompanying measure were:

- Through baseline studies, to ascertain users' needs for TTK in official statistics;
- To identify, to build awareness of, and to disseminate Current Best Methods (CBMs) in order to satisfy those needs;
- To transfer CBMs through a variety of means, mainly through exports of physical products, on-site rapid adoptionvisits and training;
- To create a networking environment and a culture that would be conducive to sustainability in TTK;
- To set up a common infrastructure that would facilitate TTK and the formation of in-depth thematic networks;
- To learn the generic lessons of attempting TTK in a way that could prove useful in the design of comparable actions in FP6.

Outline of methodology

The various activities would constitute a process, starting with the identification of users needs and ending with the transfer of CBMs to needy NSIs. The main axes of the methodology were:

- Each partner would form a network that would undertake studies on best practice identification and users' needs on specific themes;
- Based on these studies, reports would be produced for discussion at the kick-off conference of AMRADS, TTK1;
- As an outcome of TTK1, 6 themes would be finally selected for more intensive investigations through international Working Groups (WGs);
- These WGs would scrutinize the initial suggestions of CBMs and make proposals on dissemination strategies and training/roadshows;
- Workshops, (with wider participation than the WGs), would then be convened to examine these proposals and recommend actions. These actions would be taken mainly through training, through visits to selected NSIs on the feasibility of TTK and through, to a lesser extent, international information days; the emphasis was more on AMRADS acting as a facilitator to TTK rather than making universal commitments to the transfer of physical products;
- The outcome of all these activities would be discussed at a final conference, TTK2, with a view to learning the lessons for the design of support measures in future FPs.

One important medium of dissemination was the Training and Software Demonstration Centre (SODECE), a repository of technical documentation in various fields for demonstration with training, together with linked websites and helpdesks.

Main results achieved

The objectives were largely achieved, with some important exceptions that are outlined below. Needs were identified and so were the six themes in which AMRADS would operate. These themes were Business Registers, Quality, Timeseries, Metadata, Automated Data Collection (ADC) and Statistical Disclosure Control (SDC).

CBMs in each theme were selected and ratified by international expert meetings. The lessons about CBMs were:

- CBM was a relative and dynamic concept in terms of:
- Time: today's CBM is still imperfect, requiring further research to become a better CBM tomorrow;
- Space: there was no single universally ideal CBM, regardless of the level of development of the beneficiary NSI.
- CBMs involved interrelationships between themes. Thus a CBM in, for example, the ADC theme would contain elements from SDC and Quality;
- The level of development of CBMs varied between themes. At one extreme, there was no CBM in the theme on Metadata, which was still in its embryonic stages, while, at the other extreme, good practices were more firmly established in the SDC theme;
- A CBM need not always be a tangible product. It could be a list of prerequisites for and lessons from effective implementation strategies, for example, the pre-requisites for setting up good Business Registers and the lessons learnt from the exercise. In the field of metadata, there was no ideal, off-the-shelf metadata system. Available systems had to be tailored to the needs and capabilities of the importing organization;
- TTK could involve not only NSIs but also private sector organizations;
- Even after acknowledging these truths about CBMs, the identification of good practices in the thematic workshops was not as concrete and clear-cut as had been originally expected. It was not always easy to single out one practice and give it the imprimatur of best practice.

On the whole, the real value-added of AMRADS was to provide a road map to users about the relative merits and defects of adopting different systems and methods. That roadmap was all the more necessary to allow users to navigate through the flood of technical papers in the website. The training programme did focus concretely on identified CBMs in order to impart hands-on skills.

The premise underlying AMRADS activities was that TTK was feasible, that users' needs were known, that CBMs did exist and could be identified, that there were NSIs willing and able to export identified CBMs and that needy NSIs were willing to import and had the capacity to absorb the imports. One of the tasks in the AMRADS project was to verify these assumptions. That verification was conducted in two stages:

a) First, to send out a questionnaire to gather information on various aspects of TTK affecting the NSI;

b) Secondly, based on the data collected at (a), the AMRADS consortium selected a number of NSIs to visit in order to elaborate and complete the broad information provided in the questionnaires.

Given that the objective of the questionnaire was mainly to serve as an initial screening device for AMRADS to select the NSIs to be visited, the answers requested were somewhat superficial and subjective. More details and greater focus were achieved in the subsequent visits to NSIs. The main conclusions from these visits were positive. Examples of achievements were:

- On SDC, AMRADS created a sound basis for future exchanges of knowledge with the Romanian NSI. This NSI sent two members of the expert team that was visited to the training course organized by AMRADS on this subject in Ljubljana in March 2002. In general, this NSI was very willing to import software, to learn about best practices in the EU and had the capacity to acquire and adapt the underlying technologies;
- In each country visited, the basic Business Register system was quite sound in terms of coverage, basic concepts (enterprise, Local Kind of Activity Unit (LKAU)) and ordinary demographics. However, difficult questions such as enterprise splits and mergers, enterprise groups and historical files were still under development. Moreover, there was still substantial room for improvement in the analysis of the underlying microdata, including related questions of editing and imputation.

SODECE, which was important for AMRADS, never materialized. The development of this facility predated AMRADS and was being developed by Eurostat outside the project. However, AMRADS did largely compensate for the absence of this facility by holding training seminars in the selected six themes. Similarly, the International Information Days envisaged were not held because the FP 5 context in which AMRADS was conceived had changed with the passage of time. But, most surprisingly, the budget that was earmarked to fund visits to assist NSIs to establish products they had imported remained largely unused because of a lack of demand.

Possible impact on ESS and/or the end-user

AMRADS was inspired by the ESS for the ESS. The outcome of the exercises to identify CBMs was promulgated to the ESS, in the hope that those NSIs that needed the technologies and methodologies concerned could apply them to their own statistical production processes. Given that four out of the AMRADS consortium were NSIs, each responsible for one theme, the Eurostat working groups of which they were members were also informed of the AMRADS work. The NSIs that participated in the training should benefit. Ultimately, there should be a general leveling up of the ESS.

Dissemination and exploitation prospects

AMRADS was largely a dissemination project. There were the project website, brochure, help-desk, 6 international workshops, 6 training events that were widely attended, TTK feasibility visits and two major international conferences. AMRADS also acted as a channel of dissemination for a number of FP 5 projects
Suggested Further Work

In the interest of sustainability, the consortium felt that ESS stakeholders should make some commitment to the maintenance of the AMRADS website, with its rich deposit of material on best practice. TTK is a dynamic process that requires constant vigilance, commitment and awareness-building over a very long period of time in order to reap the full potential benefits of AMRADS.

Bibliography

- "SDC: from theory to practice" by Josep Domingo et al of the University Rovira i Virgili (Tarragona, Catalonia, Spain)
- "Transfer of Technology and Know-how": Kick-off conference proceedings report
- "Transfer of Technology and Know-how": Conference proceedings report

ASSO
Analysis System of Symbolic Official Data

Timetable	1/1/2001-31/12/2003 (36 months)
Website	http://www.info.fundp.ac.be/asso/

The Consortium

Member	Role	Institutional type	Country
Facultes Acultés Universitaires Notre-Dame De La Paix (FUNDP)	Coordinator	Academia	Belgium
Institut National de Recherche en Informatique et en Automatique (INRIA)	Partner	Public sector (research)	France
Instituto Nacional de Estatística (INE)	Partner	NSI	Portugal
Università degli Studi di Napoli Federico II - Dipertimento di Matematica e Statistica (DMS)	Partner	Academia	Italy
Centre International de Statistique et d'Informatique (CISIA)	Partner	Public sector (research)	France
Training of European Statisticians (TES) Institute A.S.B.L.	Partner	Training	Luxembourg
Université Paris-IX Dauphine (DAUPHINE)	Partner	Academia	France
Rheinisch-Westfaelische Technische Hochschule Aachen (RWTH)	Partner	Academia	Germany
Universidade Federal de Pernambuco - Centro di Informatica (UFPE)	Partner	Academia	Brazil
Euskal Estatistika Erakundea / Instituto Vasco de Estadística (EUSTAT)	Partner	Public sector	Spain
Statistics Finland (StatFi)	Partner	NSI	Finland
Faculdade de Economia do Porto (FEP)	Partner	Academia	Portugal
Università degli Studi di Bari - Dipartimento di Informatica (DIB)	Partner	Academia	Italy
University of Athens (UOA)	Partner	Academia	Greece

Scope and Objectives

The general objective of ASSO was to design methods, methodologies and software tools for the analysis of multidimensional complex data (numerical and non-numerical) coming from very large databases in NSIs and other administrations. Symbolic data analysis extends statistical data analysis methods to more complex objects. Individuals processed by these methods are complex in the sense that they represent groups of individuals, featuring variations amongst them. Within the context of the project, the complex objects are called symbolic objects. The ASSO project intends to improve the software built previously in order to render it more operational and attractive following users' requests, to add new innovative methods and to better meet the needs of NSIs.

Outline of methodology

The project started with an evaluation of the previous SODAS software in order to identify user requirements. In parallel, the scientific teams described their methods and wrote a scientific report. Both approaches were compared, and choices or fusions between methods were derived. A new document on "Meeting Users' Needs" was produced.

Its aim was to check if the users' needs were met and to show how the methods proposed by researchers would solve users' problems. More precisely, the ASSO project:

- Offered a tool to use, create, propagate and design statistical concepts by Symbolic Objects;
- Improved the building of Symbolic Data from a relational database, from a standard file (ASCII) with aggregated data or directly through editing;
- Introduced structured statistical metadata, thereby ensuring better quality of statistical results;
- Allowed the retrieval of new Symbolic Data and the propagation of the results in the databases;
- Added new measures of dissimilarity between Symbolic Objects;
- Added new Symbolic Analysis methods of unsupervised or supervised classification of Symbolic Data and Symbolic Objects;
- Gave new tools for the study of quality, stability and robustness of Symbolic Analysis;
- Gave new tools for visualising Symbolic Objects and/or the results of Symbolic Analysis.

Main results achieved

The main result was the software called SODAS 2, accompanied by a user manual and a help guide. The statistical methods used in the software are described in a book entitled Symbolic Data Analysis and the SODAS Software.

The SODAS 2 software is based on a modular architecture covering data management, treatment and visualization components, called modules, which were assembled in methods. Specifically, the ASSO project:

- Identified users' requirements, including NSIs', and evaluated the first version of the SODAS software. User requirements concentrated mostly on simplified data input and output functionalities, better help facilities but also support for complex analysis of sample survey data;
- Specified software kernel, mostly concerning data treatment, for example, the definition of the format for data;
- Highlighted scientific methods; partners had proposed extensions to the first SODAS software to include new symbolic objects and new methods to transfer from classical data analysis to symbolic data analysis;
- Prepared user benchmarks according to the definitions of user requirements and based on the general user interface with the SODAS software;
- Prepared data files for testing stand-alone modules, including treatments;
- Designed the metadata model. A new version that integrated benchmarks and developers' requirements was finalized. Visualization of metadata was integrated in the edit module;
- Implemented a stand-alone version of the modules for data management, methods and visualization;
- Specified workbench standards;
- Established a methodology for automatically chaining the modules in order to facilitate the use of the software;
- Integrated the modules in the workbench. Several versions of the integrated software were available, with improvements;
- Produced a Help Guide and User Manual;
- Built a User Club. Most of the members came from NSIs, with some others from universities and research centres. They were invited to participate in a validation phase.

Possible impact on ESS and/or the end-user

The ASSO project offers many benefits to end-users, including NSIs. It gives novel and robust approaches to statistical data and their analysis. It ensures data confidentiality by the use of aggregation. It allows the handling of data from complex questionnaires. It also offers the possibility of joining databases from independent surveys. These are all routine preoccupations of NSIs.

To elaborate, Symbolic Objects make the following contributions to data analysis:

- When applying exploratory data analysis to several databases, instead of merging them into one huge database, an alternative is to summarise each database by Symbolic Objects and then to apply Symbolic Data Analysis to the whole resulting set of Symbolic Objects;
- Symbolic Objects can easily be transformed into a query to a database and so they can be used in order to propagate concepts between databases;
- By characterising a concept, Symbolic Objects are able to join easily several properties based on different variables coming from different relations in a database and from different samples of a population;

Dissemination and exploitation prospects

A public website was created. A project presentation brochure was produced and widely circulated. Also, information on ASSO was disseminated to 200 private banks in Luxembourg. An electronic journal of Symbolic Data Analysis that would act as a dissemination platform for ASSO results was established. There was also a user group. The partners had attended different conferences and events where they presented papers on different aspects of the project's results. Specialized workshops were organised in such conferences as:

- "VIIIèmes rencontres de la Société Francophone de Classification", Pointe-à-Pitre, December 2001;
- GfKI Conference, Mannheim, July 2002;
- IFCS Conference, Cracovia, July 2002;
- ECML-PKDD Conference, Helsinki, August 2002.

Three training schools for symbolic analysis and use of the SODAS software were organised: Bilbao (EUSTAT) on September 2001, Helsinki (STATFI) on January 2002 and Porto (INE-FEP) on May 2002. The majority of participants were from NSIs.

Suggested Further Work

The work of the partners will continue. A lot of research is still under development. In particular, work is in progress in the following fields: symbolic objects described by variables that are continuous distributions (instead of histograms or intervals), lazy learning approach for classification of symbolic objects, assessing of cluster stability, linear regression model for interval-valued data, Markov model for symbolic data, dynamic clustering, use of metadata for interpretation of automatic clustering results, spatial pyramidal clustering based on a tessellation, etc.

The label SODAS software was registered and the consortium has continued to collaborate, to publish together and to participate in future conferences. The use of the SODAS software is free. It can be loaded at the Website www.assoproject.be .

BIBLIOGRAPHY

- A book on the project entitled: “Symbolic Data Analysis and the SODAS Software”, Diday, E., Noirhomme-Fraiture, M. (eds.), Wiley, (to appear, 2007).
- Bock, H.-H., Diday, E. (eds.) (2000): Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, Springer-Verlag, Heidelberg.
- Bock H.H. (2005): Optimization in symbolic data analysis: dissimilarities, class centers, and clustering. In: Baier, D., Decker, R., Schmidt-Thieme, L. (eds.): Data Analysis and Decision Support. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Heidelberg, pp. 3-10.
- Brito, P. (2002): Hierarchical and Pyramidal Clustering for Symbolic Data. Journal of the Japanese Society of Computational Statistics, Vol. 15, Number 2, pp. 231-244.
- De Carvalho, F.A.T., Cananee, I.C., Verde, R. (2001): Symbolic Classifier based on Modal Symbolic Descriptions. In: Book of Short Papers of the Meeting of the Classification and Data Analysis Group of the Italian Statistical Society, CLADAG-2001, Palermo (Italia).
- De Carvalho, F.A.T., De Souza, R., Chavent, M., Lechevallier, Y. (2006): Adaptive Hausdorff distances and dynamic clustering of symbolic interval data. Pattern Recognition Letters, 27 (3), pp. 167-179.
- Diday, E., Esposito, F. (2003): An introduction to Symbolic Data Analysis and the Sodas Software IDA. International Journal on Intelligent Data Analysis. Volume 7, issue 6. (December).
- Hardy, A., Lallemand, P. (2004): Clustering of symbolic objects described by multi-valued and modal variables. In Studies in Classification, Data Analysis, and Knowledge Organization, Proceedings of the IFCS’04 Conference, pp. 325-332.
- Malerba, D., Esposito, F., Monopoli, M. (2002): Comparing dissimilarity measures for probabilistic symbolic objects. In Zanasi, A.,Brebbia, C. A., Ebecken, N.F.F., Melli, P. (eds.) Data Mining III, Series Management Information Systems, Vol 6, pp. 31-40, WIT Press, Southampton, UK.
- Noirhomme-Fraiture, M. (2002): Visualization of Large Data Sets : the Zoom Star Solution. Journal of Symbolic Data Analysis, vol. 0, July 2002, electronic journal at http://www.jsda.unina2.it.
- Vardaki, M. (2004): Metadata for Symbolic Objects, Journal of Symbolic Data Analysis, vol. 2, November 2004, electronic journal at http://www.jsda.unina2.it.

BUSY
Tools and Methods for Business Cycle Analysis in the EU

Timetable	1/1/2000 - 30/6/2003 (42 months)
Website	http://eemc.jrc.ec.europa.eu/softwareBUSY.htm

The Consortium

Member	Role	Institutional type	Country
CS Italia	Administrative coordinator	Company (IT)	Italy
Joint Research Centre (JRC, Ispra), European Commission	Scientific coordinator	Public sector (research)	Italy
Centre National de la Recherche Scientifique (CNRS)	Partner	Public sector (research)	France
Institut National de la Statistique et des Etudes Economiques (INSEE)	Partner	NSI	France
Istituto Nazionale di Statistica (ISTAT)	Partner	NSI	Italy
Instituto Nacional de Estadística (INE)	Partner	NSI	Spain
GREQAM – University Aix-Marseilles II	Partner	Academia	France

Scope and Objectives

The objective of this project was to develop a software tool, namely BUSY, for the statistical analysis of business cycles in the European Union. The BUSY software was aimed at providing a guide to official statisticians through the main steps of a standard business cycle analysis so as to help in improving the knowledge of cycles in EU economies. Specific goals were to: (1) review and evaluate the main techniques available; (2) survey the practitioners needs and practices; (3) customize the most satisfying techniques in order to match the practitioners requirements; (4) organize the different steps of a typical analysis, making available a range of algorithms at all steps; (5) implement the overall scheme in a prototype software, and to (6) disseminate the software tool.

Outline of methodology

The project had the following components:

- Reviewing and evaluating statistical methods and procedures related to dating;
- Reviewing and evaluating statistical methods and procedures related to building composite indexes (leading/coincident/lagging);
- Reviewing and evaluating statistical methods and procedures related to forecasting the state of the economy;
- Reviewing methods in official statistics - the practitioners' side;
- Profiling and standardizing; production of prototype software, with user manual;
- Dissemination during and after the project, with the latter contributing to validation.

The BUSY program made available a selection of statistical techniques designed for conducting business cycle analysis both on small and large sets of time series. The program is downloadable at http://eemc.jrc.ec.europa.eu/software.htm.

Two broad types of statistical procedures for business cycle analysis were incorporated in BUSY. The first was the National Bureau of Economic Research (NBER) -type of empirical analysis that was based on descriptive statistics such as cross-correlations, coherences and phases of the cross-spectra and Bry and Boschan dating procedure. The second was model-based, the dynamic factor models, following the work by Forni et al. Both approaches were aimed at building composite indexes that were leading, coincident or lagging with respect to a reference series. These composite indexes were the main support of the business cycle analysis. The analysis proceeded only after the series had been transformed so as to be second-moment stationary. Several alternatives of stationary transformations were proposed to the users. Input data could be either in human readable or in Excel formats. All computational outputs were displayed in an HTML file that could be read either in BUSY or in Excel. The composite indexes produced were exportable either in Excel or in human readable file.

Main results achieved

The project outcome was a software for business cycle analysis called BUSY, together with a user manual. Besides a database manager and a graphic facility, the software comprised 3 modules:

- Module 1 was for data transformations. As business cycle analysis must be conducted on weakly stationary data and as most economic series have a persistent behaviour, a stationary transformation is necessary. The possibilities offered in BUSY are: linear detrending, first-order difference, yearly-difference, Hodrick-Prescott filter and Baxter-King filter.
- Module 2: NBER type of approach. This module was based on descriptive statistics following the NBER tradition. The tools available are cross-correlation analysis, coherence and phase analysis and the Bry and Boschan dating procedure
- Module 3: Dynamic Factor Models. This module interfaces the Generalised Dynamic Factor as proposed by Forni, Hallin, Lippi and Reichlin. Within this approach, all series are made up of a common component that loads the effect of common factors and of an idiosyncratic part. Composite indices are built on series cleaned of their idiosyncratic behaviour. The classification of series into pro or counter cyclical, leading, coincident or lagging with respect to the reference series is automatic. The dating of turns takes place on the common components.

The system also allows users to check the effect of different parameter settings or data vintages by offering the possibility of viewing simultaneously the output graphics of several work sessions.

Possible impact on ESS and/or the end-user

The objective of BUSY was to review, evaluate, and customize advanced statistical techniques for business cycle analysis in order to make them available for routine use in NSIs. The final aim was to implement a software tool able to guide official statisticians through the main steps of a standard business cycle analysis, so as to help in improving the knowledge of cycles in EU economies. As stated below, ISTAT had adopted it for analyzing the business cycle and also for helping in the computation of Industrial Production Indexes for the Italian economy. Thus BUSY should be of interest to all NSIs for their economic analysis, though it is not known how many of them are actually applying it.

Dissemination and exploitation prospects

BUSY'S dissemination was based on four lines of action:

1) The setting up of an interest group involving a community of potential users, as follows.
    - NSIs of the EU and Eurostat;
    - The Directorate for Economic and Financial Affairs of the Commission (DG-ECFIN);
    - The EU Central Banks; the European Central Bank (ECB),
    - The Bank of International Settlements (BIS);
    - The OECD;

- Private research institutions such as the National Institute for Economic and Social Research (NIESR) and the Economic Research Cycle Institute (ECRI);
- Universities and public research institutes,

Communication within this large community of statisticians was facilitated through:

- A Web-space located within the site "Time Series Analysis for Official Statisticians";
- Regular meetings: the consortium organised two meetings of that group in 2001 and 2002.

2) The second action was a direct support to the dissemination of the software BUSY, as follows:

- A Helpdesk on Web, within the site "Time Series Analysis for Official Statisticians",
- Installation and demonstration in the premises of a list of test-users, such as economic institutions, which had received prioritised support.

3) The third line of action concerned support in the post-project phase. The following had occurred:

- Maintenance of BUSY software by helpdesk support and processing of users' feedback;
- Continuation of interest-group with meetings;
- Training;
- Publications.

4) The fourth line of dissemination was a standard one for projects financed under FP5. It included:

- Presentation to conferences and special events, such as ETK 2001, the conference on Growth and Business Cycles in Theory and Practice', July 2000, at the University of Manchester and COMPSTAT 2004;
- EUROSTAT conference on business cycle analysis and future EPROS meetings;
- Articles and reports in the specialized literature;
- Advertisements in statistical literature;
- Use of web technologies, as the web-space and helpdesk already mentioned, and related support to users, maintenance and upgrades when necessary.

These dissemination activities had gone a long way to ensuring exploitation of BUSY. It was demonstrated to many institutions and it was tested at several places. Following the recommendation of the 2001 Panel Review, BUSY was presented to an interest group of about 40 delegates from the main EU Member States and Accession Countries economic institutions in Brussels on 15 January 2002. At a second meeting on 28 March 2003, BUSY was installed and demonstrated at DG ECFIN during February 2003 and, following the 28 March 2003 meeting, at the National Bank of Hungary and at the Agency for Economic Analysis and Forecasting, Sofia.

BUSY was also submitted for evaluation by students of the University of Parma in May 2003. It was tested at National Bank of Belgium, Bank of Italy, Bank of Luxembourg, OECD and much feedback was received from Pilar Bengoechea-Pere (DG ECFIN), Fabrizio Calabrese (Bank of Italy), Paolo Guarda (Bank of Luxembourg), Jan de Mulder (National Bank of Belgium) and Ronny Nilsson (OECD).

BUSY was presented as an invited lecture at a conference about Dynamic Factor Models organized in Rome on 21 June 2003 by University La Sapienza, where the emphasis was on practical experience with the use of dynamic factor models for business cycle analysis. Finally, BUSY was adopted by ISTAT for analyzing the business cycle and also for helping in the computation of Industrial Production Indexes for the Italian economy.

The final version of the BUSY program can be found on the BUSY web site. For further information or help using the BUSY program, please consult christophe.planas@jrc.it

SUGGESTIONS FOR FURTHER WORK

As the methodology implemented in BUSY is quite recent, - it has mainly been developed in the last ten years- a wide array of interesting topics is still open, including both empirical applications and theoretical investigations. Examples are:

- Application of dynamic factor models to the study of convergence and synchronization of cycles (a) in the Euro-area and (b) at the regional level;
- Further investigation of the forecasting performance of dynamic factor models;
- Optimal settings for the non-parametric tools involved in generalized dynamic factor models.

BIBLIOGRAPHY

- "Business cycles indicators: the sensitivity of generalised dynamic factor models to pre-filtering methods", by G.Fiorentini (University of Florence) and C.Planas (JRC), submitted to Empirical Economics.
- "Determining the Number of Factors in the Generalized Factor Model", by M.Hallin (Université Libre de Bruxelles) and Liska R. (JRC), forthcoming in Journal of the American Statistical Association.

CASC
Computational Aspects of Statistical Confidentiality

Timetable	1/1/2001-30/6/2004 (42 months)
Website	http://neon.vb.cbs.nl/casc/

The Consortium

Member	Role	Institutional type	Country
Statistics Netherlands (CBS)	Coordinator	NSI	Netherlands
Istituto Nazionale di Statistica (ISTAT)	Partner	NSI	Italy
University of Plymouth	Partner	Academia	United Kingdom
Office for National Statistics (ONS)	Partner	NSI	United Kingdom
University of Southampton (U. Soton)	Partner	Academia	United Kingdom
Victoria University of Manchester	Partner	Academia	United Kingdom
Statistisches Bundesamt (StBA)	Partner	NSI	Germany
University La Laguna	Partner	Academia	Spain
Institut d'Estadística de Catalunya (IDESCAT)	Partner	Public sector	Spain
Instituto Nacional de Estadística (INE)	Partner	NSI	Spain
Technische Universität Ilmenau	Partner	Academia	Germany
Consejo Superior de Investigaciones Científicas (CSIC)	Partner	Public sector (research)	Spain
Universitat Rovira i Virgili	Partner	Academia	Spain
Universitat Politècnica de Catalunya (UPC)	Partner	Academia	Spain

Scope and Objectives

This project aimed at the development of new methods for disclosure limitation and the enhancement of tools by implementing these new methods through major extensions of the ARGUS software. These extensions would aim at the production of safe business microdata and safe hierarchical and linked tables. Specifically, CASC aimed at:

- Expanding μ-ARGUS with several new techniques, such as micro-aggregation rank swapping, noise addition, PRAM (Post Randomisation) and model-based approaches. This would offer the opportunity for easily comparing amongst these methods and coming up with the most efficient selection in given circumstances.
- Expanding τ-ARGUS with options for hierarchical tables and also linked tables. The package should be able to offer the best suppression patterns attainable.

Outline of methodology

The methodology divided into two broad, technical streams. The first stream was devoted to the disclosure control of microdata and the second one to tabular data.

Microdata

- Several new techniques for disclosure protection were implemented. The need for these new techniques arose from the limitations of existing methods such as global recoding and local suppression. New techniques that were investigated were micro-aggregation, noise addition, PRAM (Post-randomisation) and masking techniques. The results were incorporated in μ-ARGUS. A special study into an alternative method for business data preserving the individual profile for each unit was also undertaken;
- Risk models that would help to assess the unsafety of microdata files were investigated. A study on record level measures yielded a research report on noise addition;
- A simulation of the intruder where attempts were made to undo the disclosure protection was undertaken. In addition, there was a study of the effects on the analytical power of the protected microdata file.

Tabular data

- The research included the disclosure protection of structured tables up to dimension 3 and higher, observing the dominance rule and the p% rule and different attacker scenarios.;
- There was further work on tables with hierarchical structures. This involved searching for the optimum suppression pattern in very complex models. The main approach was through research into new models and also the implementation and testing of these approaches. A second supporting approach was based on network flow algorithms. Besides these complex optimisation approaches, heuristic methods which aimed at a much quicker, near optimal solutions were also studied;
- One of the outcomes of these activities was a set of test-tables. These tables played the role of test-benches for the optimisation procedures.

CASC also conducted risk assessments. Research outcomes were not always predictable. Risk-reduction was achieved as far as possible through (a) the testing of the methodology and the software for their technical content and (b) testing of documentation for its user-friendliness and applicability to practical official statistical problems at NSIs.

Main results achieved

As the CASC project was designed to achieve practical results, the most visible outcome of this project was the twin ARGUS software packages: μ-ARGUS for microdata and τ-ARGUS for tabular data. Embedded in these packages were the detailed contributions of many CASC partners.

For microdata, a number of methods were investigated such as micro-aggregation, rank swapping, Sullivan's making method, qualitative micro-aggregation and Post-Randomisation. In addition, there was further development of the Franconi-Benedetti risk models to assess the disclosure risk per record and per household. All this plus the Dutch (simple) risk approach were incorporated in μ-ARGUS. The Sullivan's masking method was theoretically interesting but difficult to put into practice. Therefore, the CASC research switched focus to the development of record-linkage software that eventually could be incorporated as an extension of μ-ARGUS.

On the tabular side, CASC improved the German Hypercube method and devoted considerable attention to optimisation methods. A serious draw-back of the first version of τ-ARGUS was the lack of options for hierarchical tables. For secondary cell-suppression, this drawback greatly complicated solutions, and the resulting optimisation methods became more and more complex. Thus, in the current CASC research the classical optimisation models were extended, thereby enabling the protection of hierarchical tables. However, as these methods were computationally very onerous for large tables, a modular approximation was added.

Network based solutions also address secondary cell-suppression. These networks had been possible only for 2-dimensional tables with one hierarchy. CASC demonstrated that extensions were possible and so a very efficient variant for two-dimensional tables with one hierarchy was included in τ-ARGUS. Auditing software was developed to test the quality of suppression patterns.

Testing is always an integral part of software development. Thus, CASC assigned a special part of its efforts to several partners who were included in the project solely for testing.

Possible impact on ESS and/or the end-user

All NSIs have a statutory obligation to protect data while ensuring their maximum utilization. Thus, there is no surprise that the consortium included 6 NSIs. Walking the tightrope between data protection and maximum data access had required systematic and participative methodological research of the kind that was pursued in CASC and its predecessor project, SDC, funded from FP4. CASC attempted to solve practical problems faced by the ESS and its outputs were already being used by various NSIs and Eurostat. CASC software was well-documented with manuals and a context-sensitive help system. It is available free of charge and obtainable from the project web-site. Statistics Netherlands had also developed a training course.

Dissemination and exploitation prospects

During the project, participants had disseminated their research results at various statistical conferences. The AMRADS project also furnished a dissemination platform. Specific dissemination landmarks included:

- The first UN-ECE worksession on Statistical Confidentiality in Thessaloniki, 1999;
- UN-ECE worksession on Statistical Confidentiality in Skopje, 2001;
- Eurostat organised NTTS conference in Crete in June 2001;
- The Josep Domingo-Ferrer AMRADS workshop “SDC: From Theory to Practice" in Luxembourg in December 2001”, which led to the Springer publication “SDC: From Theory to Practice”;
- Joint UN/ECE and Eurostat worksession in Luxembourg in 2003. This led to a publication in the Eurostat series “Monographs in Official Statistics”, in which CASC-related papers were included;
- The ISI-session in Berlin in 2003 in which CASC results were presented;
- An Adieu-CASC conference in Barcelona in 2004, which resulted in a Springer Volume entitled “Privacy in Statistical databases 2004”.
- The AMRDAS training session in Slovenia

Suggested Further Work

Not all problems were solved by ARGUS and, moreover, there are always new challenges. SDC always tries to achieve a balance between the need for maximum information and the technical minimization of the possibilities for intrusions. With respect to the software, the consortium had decided that it could not make decisions on how the further development of ARGUS should be organized. It suggested, however, that if a commercial market for this software were to emerge and the future development of ARGUS could be guaranteed, the project team would assist as far as possible.

One area for further work is to accumulate experience in the use of the ARGUS software. The user-testing in the project had been designed to determine whether ARGUS was usable at the production stage by the European NSIs. The tests were carried out on sets of real data and the full results were contained in a project deliverable. In broad terms, these tests revealed that both software packages were well designed and their features corresponded to most of the needs of the NSIs. As NSIs tend to publish larger and more complex tables than other organizations, there is always a need to improve the capacity of τ-ARGUS. Also the protection of frequency tables is a topic for further research. Mu-ARGUS was found slightly more difficult to understand than Tau and, since practices for microdata protection varied greatly between NSIs, some testers pointed out that the package met only a part of their requirements.

As the production of safe microdata will always have its limitations, new directions have to be investigated to meet the needs of serious researchers. Access to microdata in safe centres and via safe remote access is becoming more and more important. But these options impose new challenges for disclosure protection.

Bibliography

- μ-ARGUS manual
- τ-ARGUS manual
- Inference Control in Statistical databases; Proceedings of the AMRADS workshop
- Privacy in Statistical Databases; Proceedings of the CASC adieu conference
- Various research paper, see our website
- Papers on masking methods for microdata by J Domingo-Ferrer, Ruth Brand et al
- Papers on Risk assessment by Luisa Franconi, Chris Skinner et al
- Papers in tabular data protection by Sarah Giessing
- Papers on Optimisation methods by JJ Salazar and Jordi Castro
- Papers on testing the methods proposed and the software by Mark Elliot, Luisa Franconi et al.

CHINTEX
The Change from Input Harmonization to Ex-post Harmonisation in National Samples of the European Community Household Panel

Timetable	1/1/2000 -30/6/2003 (42 months)
Website	http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Wissenschaftsforum/Chintex/Einfuehrung.psml

The Consortium

Member	Role	Institutional type	Country
Statistisches Bundesamt (StBA)	Coordinator	NSI	Germany
Johann Wolfgang Goethe-Universität, Frankfurt am Main	Partner	Academia	Germany
Deutsches Institut für Wirtschaftsforschung	Partner	Independent	Germany
University of Essex	Partner	Academia	United Kingdom
Centre d'Études de Populations, de Pauvreté et de Politiques Socio-Économiques (CEPS/INSTEAD)	Partner	Academia	Luxembourg
Statistics Finland (StatFi)	Partner	NSI	Finland
Åbo Akademi University, Turku	Partner	Academia	Finland

Scope and Objectives

The main objective of CHINTEX was to clarify whether it was necessary to have centralised, standardised survey instruments to achieve comparability between national statistics or if this objective could also be achieved by ex-post harmonisation of existing sources. Ex-post harmonisation means converting independent national sources to the common concepts, definitions, survey questions etc used in the ECHP. The independent national sources include long-running panels in Germany, UK and Luxembourg as well as Finnish administrative registers. In the process of pursuing his broad objective, CHINTEX investigates important hypotheses on the data quality of panel surveys (non-response reporting errors and panel effects) and of income.

Outline of methodology

The broad steps, which are to some extent inter-dependent, were:

1. The assessment of the level of harmonisation attained between the independent national panels and the ECHP with respect to variable conversion and the application of unique imputation and weighting schemas;
2. From the experience with conversion, the derivation of a taxonomy of conversion problems for non-trivial cases, with an indication of the statistical tools that should be used in each category;
3. Investigation of the extent to which the outcome of harmonisation is dependent on the use of different imputation and weighting procedures, particularly in relation to income;
4. An exploration of the impact on the harmonisation results of field-related factors that could not be harmonised ex-post. This step would cover non-response, attrition effects, quality of income data and the estimation of statistical models.

All investigations were based on the following data sources:

a) The input-harmonised ECHP for Germany, the United Kingdom and Luxembourg ran in parallel with national panels for the years 1994-1996, affording comparison between the different harmonisation strategies.

b) The Finnish input-harmonised ECHP data regarding income for the years 1996 and 2000 afforded comparisons with corresponding data from the Finnish register, particularly in respect of the crucial variable, income.

MAIN RESULTS ACHIEVED

The ECHP conversion projects in the United Kingdom, Germany and Luxembourg showed similar results. For only few variables there seemed to be no difference at all between input and ex-post harmonised survey. Specifically:

- There was clear evidence about problems unique to the ECHP conversion framework as distinct from issues common to all ex-post harmonisation of micro-data. Ex-post harmonisation of this type, that is the conversion of micro-data at a level very close to the ECHP questionnaire, was resource-intensive. It required considerable human resources and time. Staff experienced in both source and target formats was crucial for obtaining good quality results.
- Most of the conversion problems were rooted in differences between the concepts of the national source and the target survey. These concepts were heavily affected by national contexts. Solutions attempted were based mainly on ad-hoc decisions about withdrawing, combining or collapsing variables. Neither estimation techniques nor independent external data were use in the harmonisation process. The consequences were fewer numbers of successfully reproduced target variables and a lower degree of comparability for some converted variables. The conclusion is that the available harmonisation toolkit should be enlarged and systematised.
- In general, the transfer and transformation of information from internal or external sources were possible only in a qualified sense. Ideally, both the target survey and the independent source should cover the same population and comparable sampling designs and measurement functions.
- For BHPS (British Household Panel Study) and ECHP data, attrition was very high for the first wave of a panel but it decreased continuously thereafter. Therefore it was harder to transfer individual values to a previous wave within a fresh panel than within an older one in which the sample population did not change materially. The transfer of data was biased due to the variation of attrition across several strata of the population. There was a close correlation between attrition and citizenship because non-national respondents were more likely to leave the panel.
- Transfer bias was smaller with an old panel than with a new one. If conversion has to be applied, by utilisation of both internal and external data, the framework developed should ensure that harmonisation is considered with respect to all possible dimensions of the data and causes of bias.
- Unless attrition biases were very extreme, they were unlikely to affect the substantive conclusions of models. In particular, the longer period of cumulative attrition in older panels did not undermine the comparability of findings based on regression models.
- The research on data quality showed that changes in earned income tended to be underestimated by questionnaire information. The attrition bias showed a similar effect. For Finland, it was shown by means of register data that changes in household income and earned income, as well as changes in marital status, had an impact on attrition behaviour. There was also a bias towards an underestimation of mobility in household income and measures of inequality. All biases found were in accordance with the hypothesis that changes increased the risk of attrition. However, panel attrition did not alter the ranking of countries regarding their income distribution measures such inequality.
- A comparison between estimates based on Finnish ECHP and estimates based on Finnish registers for the same persons revealed substantial differences in the distribution of household equivalent income. These differences were stable over time. With respect to the poverty rate, there was a 50 percent divergence. Such large divergences would affect the ranking of European countries according to ECHP results. These distortions would be replicated in the EU-SILC.

- An observation from the work on non-response was that field work in the national ECHP sub-samples differed considerably. Specifically, the contact procedure in the UK ECHP was quite different from that in all other countries. In particular, in the UK ECHP the usual follow-up rules were not applied and consequently the contact rates were much lower, the workload per interviewer was much heavier and the average duration of the interviews was much shorter. Therefore, the impact of the data collection and the population characteristics on sample participation varied between different countries and between different types of surveys. Thus, there were significant differences in attrition and item non-response patterns in different surveys carried out in the same country, and between surveys conducted in different countries. The research also suggested that there were good survey practices which would minimise attrition and item non-response.
- On average, over the duration of a panel, there was a trend towards more precise levels of measurement. Therefore the quality of income data increased during the panel. The comparison between new panels and ongoing ones also showed that the cumulative losses in case numbers due to panel attrition in the ECHP were substantial, that is about 20 percent over 5 waves. But ongoing panels, such as the BHPS, and the German SOEP (Socio-Economic Panel), were more successful in preserving its sample size than was the new ECHP over the same period. The non-response bias at the initial wave of the Finnish ECHP was larger than the attrition bias over all other waves. Initial non-response bias and attrition bias might move in opposite directions and compensate each other over the long term. These various results supported panels with a long duration.
- For imputation, as much information as possible should be used in the algorithms concerned. With software like FVEware (Raghunathan, Solenberger and Van Hoewyk, 2002), MICE (Van Buuren and Oudshoorn, 2000), NORM (Schafer, 1999), PAN (Schafer, 2001) or corresponding procedures included in commercial software packages (e.g. SAS, SAS Institute Inc., Gary, NC, USA), this objective is rather easy to realise. However, if the data base is large, i.e. more waves are available and many variables are considered, then imputations may become computationally very laborious and it may become necessary to restrict the number of variables to be used in imputations. Data sets should flag cases with the imputed values.

Possible impact on ESS and/or the end-user

CHINTEX's research should be valuable for two main groups of users:

(1) Statistical Institutes

The results on harmonisation strategies are of interest to Eurostat and to all NSIs. The results should help NSIs to decide on strategies for providing comparable data requested by Eurostat. In the EU SILC context, the results should be directly exploited by those NSIs which are currently carrying out conversion of micro-data. They should gain insights into ways of improving their conversion process and to increase data quality. Furthermore, Eurostat should benefit directly from the quality analysis of ECHP sub-samples on improving imputation methods.

The use of research results is not limited to the panel work of the Statistical Institutes but also should cover the general context of adapting national statistical data to a harmonised European Statistical System.

(2) Academic Research

In addition, the outcomes should be of high relevance to academic researchers concerned with internationally comparable analysis in social science and economics, because they are often confronted with data of unknown quality with respect to comparability. To some extent, the investigation of panel effects and panel attrition have the character of basic research. The results should enlarge knowledge about panel methodology and should be of general use for researchers working on panel data.

Dissemination and exploitation prospects

It was the aim of the research consortium to perform its work in an open environment with permanent contact to colleagues interested in this field of research. CHINTEX itself benefited from the co-operation of institutes with manifold experience in the field of quality of surveys in social statistics. Statistical institutes as well as academic researchers were involved in the following ways:

- CHINTEX had co-operated with Eurostat as one of the main users of the research results to ensure that the knowledge which was necessary for the European decision-making process concerning the EU SILC was reflected precisely in the details of CHINTEX work plan;
- The co-ordinating institution had published the research output in a publication delivered to NSIs, to research institutes and universities world-wide. Besides, all project results were made available by publications in academic journals;
- The co-ordinating institution had integrated a CHINTEX web-page in their website with a general description of the project, of the research consortium and an abstract of every work-package. Several working papers were presented on the web-page.

There were the following events :

- Opening Workshop: 27-28 June 2000 in Wiesbaden
- Intermediate Workshop:29-30 November 2001 in Helsinki
- Final Conference: 26th to 27th May 2003 in Wiesbaden

The workshops enjoyed very good interaction. More than 70 participants from 18 countries attended the Final Conference thereby combining varied and excellent expertise around the themes of CHINTEX work packages and offered a forum to discuss project results.

- There were 10 presentations at international gatherings, 8 major publications and numerous working papers, all accessible on the project website.

Suggested Further Work

- There were important differences between refusals and other types of non-response (e.g. not known). This was an important area for further work;
- Further work could be undertaken with the national panels. However, it would also be helpful if organisers of cross-national surveys could pay more attention to this issue and make available better information (e.g. separate codes for different types of non-response and imputation flags at the level of the individual variable);
- The decision about the preferred duration of a household panel was also considered in CHINTEX in the context of panel conditioning, i.e. the respondent behaviour of participants. There was evidence of a trend towards less rounding, less use of income brackets and less under-reporting of incomes as the panel continued. This issue needs to be further explored;
- The use of scales with satisfaction scores on different subjects was also investigated. The scales were more informative the more calibrated they were. Thus, the SOEP scales ranging from 0 (totally unsatisfied) to 10 (totally satisfied) were more informative than the ECHP 5 point scale;
- The research comparing register with questionnaire information threw up valuable insights. These findings argue for more experimental studies in this area, especially for those participants in the EU-SILC where a direct comparison with register information is possible;
- Overall, cross-sectional surveys, and to a smaller extent short-term panels, suffered from a maximum of measurement error and of non-response bias. Both components were reduced to some extent by a longer operation of the panel. These results argue for panels with a long duration. But, again, there is scope for further research;
- Given the structure of the EU-SILC, there should be more methodological research on the quality of the longitudinal information. This survey offers excellent opportunities to study the effect of panel conditioning and panel attrition on a larger basis.

BIBLIOGRAPHY

- Minkel, H. (2003) Methodik zur Konvertierung von Panel-Daten unter Ausnutzung zusätzlicher Informationen aus anderen Erhebungen, Wirtschaft und Statistik
- Ehling, M., Linz, S., Minkel, H. (2004) Internationale Harmonisierung von Statistiken – Grundlagen und Beispiele aus dem Bereich der Haushaltsstatistiken, Wirtschaft und Statistik, Vol 1
- Behr, A., Bellgardt, E., Rendtel, U. (2005) Extent and Determinants of Panel Attrition in the European Community Household Panel, European Sociological Review, Vol 21
- Spiess, M., Goebel, J.(2005) On the effect of nonresponse on the estimation of a two-panel-waves wage equation, Allgemeines Statistisches Archiv, Vol 89,
- Rendtel, U. (2005) Plenary Meeting on Missing Data and Measurement Error, Allgemeines Statistisches Archiv, Vol 9.
- Nordberg, L. and Penttilä, I. (2001) A Comparative Study of the Quality of Income Statistics Based on Interview Surveys and/or Administrative Records, ISI 53rd Session, Seoul
- Hanisch, J., Rendtel, U. (2002) Quality of income data with respect to rounding, ICIS, Copenhagen.
- Neukirch, T., Rendtel, U. (2002) Empirical evidence for non-ignorable attrition in the Finnish sub-sample of the ECHP, ICIS, Copenhagen
- Rendtel, U. (2002)Attrition in Household Panels, Symposium on Non-response, Questionnaire Split and Multiple Imputation – Reports from Academic and Practice, Nuremberg
- Neukirch, T. and Rendtel, U. (2002)An Attrition Analysis of the Finnish Sub-sample of the ECHP Using Additional Register Information Statistische Woche, Constance
- Spiess, M. (2003)Compensation for Non-response: Weights and imputations, SOEP Anniversary Conference, Berlin
- Hanisch, J. (2003) Quality of income data in household panel surveys, ISI 54th Session, Berlin
- H. Minkel, H. (2003) Methodology of panel data conversion using additional panel and cross-survey information, ISI 54th Session, Berlin
- Neukirch, T. and Rendtel, U. (2003) Attrition biases and inverse probability weighting correction - A performance study using additional register information, ISI 54th Session, Berlin
- Nordberg, L. (2003)The Analysis of Income-Dynamics Using Interview- and Register-Based Income Data, ISI 54th Session, Berlin
- Rendtel, U. (2003) Attrition Effects in the European Community Household panel, ISI 54th Session, Berlin
- M. Spiess, M. and Goebel, J. (2003) On the effect of non-response on the estimation of a wage equation, Workshop on Item Non-response and Data Quality in Large Social Surveys, Basel.
- C. Kuchler, C. and M. Spiess, M. (2004) The Concept of Accuracy within the Imputation Process, European Conference on Quality and Methodology in Official Statistics, Mainz
- Minkel, H. (2004) Methodology for imputation of panel data in statistics on income and living conditions, Statistics Investment in the Future, Prague
- U. Rendtel, U. (2005) A Markov Model for the Decline of Initial Non-response Bias in a Panel Survey, ISI 55th Session, Sydney

CLAMOUR

Methodology, tools, users' needs and practical applications: Improving the quality of existing and future classification systems

Timetable	1/1/2000 - 31/3/2002 (27 months)
Website	http://www.statistics.gov.uk/methods_quality/Clamour.asp

The Consortium

Member	Role	Institutional type	Country
Office for National Statistics (ONS)	Coordinator	NSI	United Kingdom
Institut National de la Statistique et des Etudes Economiques (INSEE)	Partner	NSI	France
LexiQuest	Assistant partner	Company (IT)	France
Roland Rousseau	Assistant partner	Company	France
Statistics Netherlands (CBS)	Partner	NSI	Netherlands
Statistics Denmark	Partner	NSI	Denmark
Statistics Finland (StatFi)	Partner	NSI	Finland

Scope and Objectives

The role of the classifications, which try to represent the real world and structure it as clearly as possible, is more and more important in the information society. Most of the time classifications are described in a way which is not clear to end-users. Indeed, end-users may not even be aware of the underlying methodology. Although much has been achieved, the EU still did not have a coherent framework of classifications at its disposal that meets users' needs in all respects. CLAMOUR, CLAssifications MOdelling and Utilities Research, conducted research into methodology, users' needs and practical applications, with the aim of providing the means of improving the quality of existing and future classification systems, and understanding of them. The emphasis was on industrial economic classification systems, such as NACE.

Outline of methodology

There were four components:

- The linguistics subproject, in which the methods included linguistic engineering; recognition of full meaning techniques; recognition of compound words; analysis of contextual and co-textual disambiguation; analysis of complex sentences; construction of algorithms; and matching of corpuses;
- The foundations subproject, in which the methods fell in the following areas: (a) the construction of a model for the structure and activities of businesses; (b) applications for statistical units; and (c) applications for classifications;
- The users' needs subproject was mainly a large-scale survey in Denmark, Finland, the Netherlands and the UK into the needs of users of industrial classifications and how the existing statistical system (NACE) met those needs;
- A specially designed structured questionnaire (DPQ) for the collection of information from businesses, including organizations representing businesses, in order to construct a typology of activities of businesses and of the statistical units to which these activities apply; and the subsequent testing of the questionnaire in the Netherlands, UK and Finland.

MAIN RESULTS ACHIEVED

Overall, CLAMOUR has successfully advanced classifications theory across a number of themes and provided a sound basis for further developments in future. In particular, CLAMOUR results should be useful to those interested in developing automated classification systems; to any study of the fundamental building blocks of classification systems; to those who wish to have an up-to-date idea of the needs of EU classification systems users; and to anyone who wishes to develop technology for automated collection of survey data. The outputs from the individual project components follow.

Linguistics Subproject

The main objective of the linguistics work was to improve the quality of coding in classifications for six languages. In the event, it was possible to cover only English. The components of this study were disambiguation, compound words, highly structured descriptions, and semantic matching of classifications and lists (or corpuses).

From a linguistic point of view, the corpus study showed that the terms represented in these lists could be very complex, with coordination, negation and exclusion. These results are of interest to the statistical community, but not exclusively. Wherever there is an internal classification of a given concept and a need to provide an external reference, the method can be used to build a bridge between the given sets before manually finalizing the correspondence.

With regard to the distribution of terms in the lists, the study revealed that the frontiers between lists were not always easy to define. This was why, in CLAMOUR's proposal for automatic creation of lists, the focus was on making the distinction between objects and actions on the one hand, and services on the other.

The main goal of the analysis of NACE descriptions was to determine the kind of patterns (syntactic constructions) that were represented in the NACE corpus. The reference corpus was the English version of NACE. A program was developed automatically to classify the NACE wording, depending on the syntactic patterns to which they referred. 24 patterns were identified, covering 80 % of the descriptions.

CLAMOUR showed that Natural Language Processing techniques could be applied to nomenclatures for accessing, managing and comparing classifications. But the very specific domain and formulations of classification items implies that specific heuristics had to be implemented. These heuristics were described in the project deliverables.

Foundations Subproject

One product was a report describing the model. Its core was an extensive list of entities and attributes with precise definitions, together with relationships between entities. The model contained all the information needed for the definition of statistical units and activity classifications and was derived from users' needs as established by the users' needs survey and existing European regulations.

A second product was a report on the use of the model for statistical units and a prototype of a tool for their actual construction. The report included a list of definitions of statistical units in terms of the model, together with algorithms for their derivation. All types of statistical units defined in the Council regulation on statistical units and all types resulting from the users' needs study were included, with more than one definition and algorithm for cases where the Council regulation was ambiguous. The prototype was tested on the data set resulting from the tests of the DPQ.

The final output was reports on the use of the model for classifications and homogeneity measures, and a prototype of a tool for the actual coding of statistical units. The reports included the definition of a category of NACE in terms of the model. They also included a method for the model-based construction of an activity classification and an analysis of the possible use of the model for the construction of a system of multiple activity classifications. The prototype was tested on the data set resulting from the tests of the DPQ.

Users' Needs Subproject

The aim of the user needs research was to assess the suitability of current activity classifications, concentrating particularly on NACE Rev.1, for the purposes of a wide range of users. This work also fed into other parts of the CLAMOUR project dealing with the linguistic aspects of automatic coding, the modeling of business activities and structures, and the development of electronic data collection tools.

The results of this aspect of CLAMOUR can be summarized as follows:

- Users wanted more detailed classifications and better coverage of new activities;
- More responsive processes and tools were needed to deal with changing patterns of business activity over time;
- More flexible classifications and statistical units were required to better meet a wide range of needs. Multi-dimensional classification systems and a "building block" approach to determining structures and codes might help;
- The English language version of the questionnaire was not always as clear as it might have been, causing translation difficulties;
- The questionnaire seemed to have been more successful in personal interview approaches than in self-completion situations;
- The development of a multi-lingual web-based questionnaire tool would have facilitated data collection.
- The results of this exercise provided a very useful input into future revisions of particularly NACE.

Data Providers' Questionnaire Subproject

This aspect of the research concerned the progressive development and testing of the DPQ, culminating in DPQ-version 3. DPQ was not developed to cover all economic activities comprehensively and it would, in any event, require further future research. CLAMOUR tried to answer the basic question: can the information that is required for the simplified model be obtained with this questionnaire concerned? The extent to which this was the case was not proven. The tests carried out on a small scale were inconclusive in respect of companies of any size.

Possible impact on ESS and/or the end-user

A standard preoccupation in the ESS is the application of economic nomenclatures such as NACE and ISCO. Thus, all NSIs and particularly international organizations such as Eurostat and the UN Statistical Office should take a very close interest in the findings of CLAMOUR. These potential users are the counterpart of CLAMOUR's specific dissemination activities. Already the CBS is exploiting some of the project's findings. The CLAMOUR project has advanced a variety of approaches and provided a range of tools to enable classifications experts to improve existing and future classification systems.

Dissemination and exploitation prospects

The dissemination and use plan for CLAMOUR had involved four steps:

- Identification of the key products;
- Identification of the possible users of these products;
- Communication with the users;
- Exploring ways of ensuring that the products will be used by them.

The probable users are those in charge of developing or maintaining activity classifications, in particular, thebodies responsible for NACE/CPA, ISIC/CPC, NAICS and national activity classifications. Communication to them was through presentation of the deliverables concerned to a wide variety of bodies, such as Eurostat NACE/CPA Working Group (April 2001); the UN Classification subgroup meeting on classifications in New York (Spring 2001); and in future to meetings of the Eurostat NACE/CPA working. There was presentation also to national and international meetings concerned with business registers.

Suggested Further Work

Linguistics

CLAMOUR's version of the extraction rules already covered 80% of NACE descriptions. This extraction method was still in development and required further improvement concerning the precision of the items already extracted and concerning the extraction from descriptions not yet covered by the patterns.

Foundations

The industries for which actual model-based data were collected were necessarily limited. Therefore, further model-based data collection was recommended. Also, the model has to be adapted in such a way that it is possible to obtain the information needed by means of a Blaise tool that can be used in practice.

The quality of business statistics can be improved by applying the methodological basis for statistical units and classifications as developed in CLAMOUR's foundations subproject. This can be achieved through the higher precision of theoretical and operational definitions and through linkage with user needs.

Users' needs

- If it is to be used again, the questionnaire could be refined in the light of experience;
- The English language version of the questionnaire was not always clear, causing translation difficulties. A re-write to simplify and clarify the language should help to improve understanding and make translation easier;
- The questionnaire seemed to have been more successful in personal interview than in self-completion situations. The expense involved in personal interview argued for a simpler, clearer questionnaire more suitable for self-completion;
- The development of a multi-lingual web-based questionnaire tool should be considered for any future data collection exercise.

Data Providers' Questionnaire

An aspect that did become more important for future versions of the DPQ is how best to collect information from businesses. The tests carried out on a small scale in CLAMOUR were inconclusive. Therefore, the design of even DPQ-3, the final version within CLAMOUR, has considerable potential for further development into a very powerful and flexible tool that can be used to collect information on the activities and the organization of businesses.

CODACMOS
Cluster of Data Collection Integration and Metadata Systems for Official Statistics

Timetable	1/11/2002 - 31/10/2004 (24 months)
Website	http://www.codacmos.eu.org

The Consortium

Member	Role	Institutional type	Country
Istituto Nazionale di Statistica (ISTAT)	Coordinator	NSI	Italy
Statistics Finland (StatFi)	Partner	NSI	Finland
University of Edinburgh	Partner	Academia	United Kingdom
National Statistical Service of Greece (NSSG)	Partner	NSI	Greece
DESAN Research Solutions	Partner	Company	Netherlands
Statistical Division of Municipality of Milan	Partner	Public sector	Italy
Finnish Tax Administration	Partner	Public sector	Finland
University of Patras	Partner	Academia	Greece
Institute of Informatics and Statistics (INFOSTAT)	Partner	Public sector	Slovakia
University of Athens	Partner	Academia	Greece
National Social Security Institute	Partner	Public sector	Italy
Tietokarhu Ltd	Partner	Company	Finland
Statistics Norway (SSB)	Partner	NSI	Norway

Scope and Objectives

The CODACMOS cluster centred on four RTD projects viz. TELER, DATAMED, IQML and IPIS. However, clustering was only a starting point. CODACMOS aimed to go beyond the present and known situation on data collection. The main goal of the CODACMOS project was to identify ways in which the cost-effectiveness and quality of the official data collection process could be improved while lowering the statistical burden on respondents. The project would get as close as possible to the data source by bringing together data collection approaches with the tools, methods and strategy that aided the production of statistical data. Coordination at the national level was the best way of reducing administrative burdens by removing overlapping and sub-optimal data collection. Active interoperability between the current solutions would be investigated.

The specific objectives were:

- To review and to rationalize the state of art on the proposed solutions on electronic data collection and exchange of metadata;
- To identify two “experimental fields” and to implement demonstrations for the selected solutions on the integration of different data sources, such as existing archives /registers or other administrative data, used for statistical purposes based on close co-operation between the institutions concerned;
- On the basis of the demonstrations and through wide consultation, to specify EU-wide key issues for the standardization/harmonization of data collection models and methods and for the description of metadata standards.

Outline of methodology

To achieve its goals, CODACMOS would seek to improve the efficiency of data collection and to integrate data exchanges between NSIs, other Public Administrations and various data providers. CODACMOS would review the current state of the art in electronic data interchange and metadata by bringing together key researchers and actors from relevant projects and the representatives of international organizations, such as the SDMX Task Force. The project would add value to the concepts, models, tools and solutions for data collection by examining solutions, with associated metadata, in relation to enterprises and households.

There were four elements in the methodology:

1. The first stage would be to review the state of the art, then to identify the two main applications that would act as experimental fields in which the selected solutions and models could be demonstrated;
2. The demonstration phase of the project would involve all partners and would indicate the appropriateness of selected solutions in different contexts;
3. Workshops would then be organized in the areas of data collection, data integration and metadata. These workshops would bring together key researchers and other actors from the public and private sectors, including NSIs, National and Local Public Administrations, business associations, software developers, international organizations and academic institutions;
4. The main findings relating to the data collection solutions and their demonstrations would act as the framework for proposing a new integrated project for future FPs.

Main results achieved

In a nutshell, the main results were:

- The development of the common metadata model;
- Differentiation between different levels, (data level, collection process level, national level of co-ordination), in the optimization/integration of data collection and the statistical process as a whole;
- The achievement of close co-operation between NSIs, other data collectors and relevant public bodies;
- Production of the 'informative table model'/CoSSi;
- Determination of concepts and development of the tools for the various demonstrations;
- The analysis of benefits and deficiencies of innovative data collection tools with respect to specific situations;
- Formulation of recommendations for future research;
- Development of standards and interfacing tools for full questionnaire metadata.

The CODACMOS project:

- Provided a data collection strategy and tools for improving the quality of official statistics and for lowering the burden and costs for both data collectors and providers. This was done by improving the efficiency of data collection and by integrating data exchange between NSIs, Public Administrations and other data providers;
- Analysed and reviewed the current state of art in electronic data interchange and metadata by bringing together key researchers and actors from relevant projects and representatives of international organizations;
- Produced relevant scenarios to orient new projects for future research concerning data collection and systems of metadata. The most significant added-value was the new vision of CODACMOS integration process. The e-Government process could not be successfully realised without a unitary vision of the exchange, integration and dissemination of information. The process called for the construction of a data base containing integrated common knowledge of economic and social phenomena. Therefore, it was possible to construct a common portal dedicated not only to citizens but also to researchers and other interactive users;

- Demonstrated the feasibility of integrating primary and secondary data collection and deriving a common model and related standards of metadata for the exchange and collection of data between respondents and data collectors. The basic principle of CODACMOS cluster demonstrations was to manage and/or report an "optimised, new and integrated" electronic reply to a data collector by means of the technological solutions and common models selected as a part of an agreed strategy for data collection. CODACMOS added value to the concepts, models, tools and solutions for data collection by examining solutions affecting enterprises and households;
- Recognised that there were various projects dealing individually with metadata exchange, where the standardisation of metadata played a role in the selection and design of the IT solutions concerned. CODACMOS standardised the definitions of metadata, which rendered the task of electronic response easier and more efficient. The CODACMOS project took this approach some steps further by involving other administrations and institutions collecting data;
- Added value through its metadata model. This value-added was its contribution to the standardisation of processes and exchange, as well as the technologies used. Briefly, the value-added can be illustrated from the following properties of the model:
    - It included more than 150 metadata items which were selected in accordance with the needs expressed by OECD and IMF and it followed the definitions by UN, SDMX and METANET;
    - It incorporated the main stages of statistical data processing: data collection, processing, analysis and dissemination;
    - It took partial account of the harmonisation process concerning the correspondence between classifications and the conversion of measurement units;
    - It included information on quality issues;
    - It included seven operators (transformations) which kept the processing history of the manipulation of each dataset. It followed the Object Oriented paradigm.

Possible impact on ESS and/or the end-user

The project is highly relevant to the preoccupations of NSIs. Indeed their needs inspired CODACMOS. One key reason why a common model approach was needed was that Eurostat and other international organizations had recommended the adoption of a uniform methodology in this field. Common data structures were required to ensure interoperability amongst administrations at national and European levels. The use of XML/EDI-based solutions in Member States must be synchronized to prevent any country lagging behind.

Dissemination and exploitation prospects

CODACMOS had aimed to address the data collector and data exchange community as well as end users. A full list of interested organizations, top experts, producers, users and senior researchers was established. Most of the experts also participated as members of one or more CODACMOS Working Groups.
The channels of communication were the project website, project leaflet, posters, newsletters, publications in scientific journals, technical and demonstration workshops, conferences and journals and CODACMOS final publication (Deliverable D9.5), to be published shortly.

More specifically, CODACMOS carried out 15 demonstrations to show the value and feasibility of different data collection solutions that might improve the efficiency of the data collection process and/or improve the quality of official statistics. The consortium prepared the CODACMOS European Seminar on "The most effective ways to collect data and metadata: advanced strategies, tools and metadata systems", Bratislava, Slovakia, October, 2004. It prepared the following papers and participated in the corresponding conferences:

- H. Papageorgiou, M. Vardaki, "An Integrated Metadata Model for Statistical Data Collection and Processing", Sixteenth International Conference on Scientific and Statistical Database Management (SSDBM), Santorini, Greece, June 2004.
- Rutjes H., and Hartkamp, J., "Issues in "metadata" standards aiming at the reuse of questionnaires. Experiences from IQML, IT, and others", Royal Statistical Society / Association for Statistical Computing - Seminar on New Approaches to Software for Statistical Processing, January 2003.
- Hartkamp, J., Maas, R., "In Search of Generic Questionnaire Models", University of London, Institute of Education, London, UK, March 2004.
- Rouhuvirta, H.,"An alternative approach to metadata – CoSSI and modeling of metadata", CODACMOS European Seminar, Hotel Devin, Bratislava, Slovakia, October 2004.
- Myrylainen, J., "Summation of presentation over technologies given in CODACMOS Seminar", CODACMOS European Seminar, Hotel Devin, Bratislava, Slovakia, October 2004.
- Sorce, A., "E-government and interoperability of systems", CODACMOS European Seminar, Hotel Devin, Bratislava, Slovakia, October 2004.
- Hartkamp, J., "Data Collection Strategies to Reduce the Response Burden", CODACMOS European Seminar, Hotel Devin, Bratislava, Slovakia, October 2004.
- Hartkamp, J., "Future Research on Data Collection – Some General Notes", CODACMOS European Seminar, Hotel Devin, Bratislava, Slovakia, October 2004.
- Sorce, A., Dishnica, P., "Data collection strategy and integrated solutions", Conference of European Statisticians, Geneva, 13-15 June 2005.

SUGGESTED FURTHER WORK

Being mainly an organizational instead of technological challenge, CODACMOS faced the important question: "What is to be expected from decision-makers in the field of e-Government, from the NSIs and from Eurostat?" to facilitate an effective electronic Data Collection process that would minimize the burdens on respondents. To answer this question, the consortium proposed 10 areas in which further research would be necessary:

- A national portal for EDI (Electronic Data Interchange) for reporting by businesses and households to government agencies to be used by all governmental institutions collecting data from households, enterprises and other institutions;
- Primary EDI for establishing and operating administrative base registers for persons, workplace and dwelling (land/property/building/dwelling/ address);
- Central metadata register for reporting the obligations of businesses, where all government institutions must register their metadata before they are permitted to capture the associated data from enterprises, including the organization around the metadata register;
- Metadata model for data from individuals, which all government agencies must use when they collect data directly from individuals or request such data from other government agencies;
- Central metadata register for reporting on individuals, where all government agencies must register their metadata before they request associated data from individuals;
- A common EDI system for reporting enterprise accounts, internal procedures of the enterprise and direct reporting to government agencies such as tax, register of accounts for limited companies and the NSIs;

- An integrated EDI system for reporting from local government to central government;
- A coordinated system for reporting on employee jobs.
- Standard modules for operating longitudinal databases to derive statistics and analyse;
- Standard modules for automatic editing and imputation for administrative and statistical sources.

Furthermore, CODACMOS's recommendations were:

- The active participation of NSIs, to be stimulated by Eurostat, in e-Government at the European and National levels to promote statistics;
- Further support to extend the domain of the SDMX (Statistical Data and Metadata Exchange) initiative, including work on the Metadata Common Vocabulary;
- Support of open source principles for software distribution and maintenance of common statistical software;
- Active participation in the IDABC (Interchange of Data between Administrations, Businesses and Citizens) program to promote statistics.

BIBLIOGRAPHY

- Vardaki, M., “Statistical Metadata in Data Processing and Interchange”, Journal of Data Warehousing and Mining, (Ed.) John Wang, Montclair Univ. USA, IDEA Group Publishing.

COSMOS
Cluster Of Systems of Metadata for Official Statistics

Timetable	1/9/2001 -31/8/2003 (24 months)
Website	http://www.epros.ed.ac.uk/cosmos/index.html

The Consortium

Member	Role	Institutional type	Country
University of Edinburgh (UEDIN)	Coordinator	Academia	United Kingdom
University of Essex (UKDA)	Partner	Academia	United Kingdom
University of Athens (UoA)	Partner	Academia	Greece
DESAN Research Solutions	Partner	Company	Netherlands
Statistics Sweden (SCB)	Partner	NSI	Sweden
University of Ulster (UU)	Partner	Academia	United Kingdom
World Systems (Europe) Limited	Partner	Company	Luxembourg
Dimension EDI	Partner	Company	United Kingdom

Scope and Objectives

COSMOS is a Cluster which brought together five IST projects: MISSION, FASTER, IPIS, IQML and METAWARE, all of which had a shared interest in statistical metadata and metadata repositories. The objectives of the Cluster were fourfold: to build better metadata repositories by exchanging ideas and experiences in using metadata systems for the individual projects; to identify a common set of metadata objects, with agreed definitions, attributes and methods; to implement a demonstration subset of these objects to show interoperability of the developed systems; and to define a methodology for further developing this interoperability. The Cluster would link with other concertation activities in the area and international standards bodies such as the Object Management Group (OMG), UN/EDIFACT and ebXML. To achieve these objectives, COSMOS sought to:

- Maximise the interchange between the key developers of the metadata models in each project;
- Give an opportunity for all participating institutions to meet and exchange experiences;
- Demonstrate the interoperability of the project outputs, while respecting the exigencies of each project plan.

Outline of methodology

There were five elements of the Cluster's approach:

1. Key designers from each project would be brought together to work remotely to achieve a common understanding of the overall metadata framework and the object classes of the model;
2. All members of all five projects would meet to understand the implications of the common metamodel, to exchange experiences, and to input a user view. The main objective of the meeting would be to agree on a common core model identifying which of its elements could be implemented.
3. After the main conference, two activities would proceed in parallel. One would develop the APIs for the objects agreed at the main conference, and the other would define and refine the trial scenarios for the demonstration at the final project meeting;

4. At the final meeting, four representatives from each project, and 12 invited guests, would observe the demonstration of the trial scenario. The outcome would be an evaluation report compiled by all five projects of the trial itself and the impact of the Cluster's activities on the individual projects;
5. COSMOS would also interact with other networking and standardizing activities.

The major technical stages in the work were:

- Agreement of the work on model specification;
- Report on a common model: overview , metadata items selection and description, relationships and design methodology
- Report on the main conference: agreement on the basis for API specifications and trial scenarios; integration of trial scenario into the 5 projects;
- Acceptance of the initial inter-operability tests;
- Trial demonstration at the final meeting.

MAIN RESULTS ACHIEVED

The COSMOS project had the following broad outcomes:

1. The comparison of the cluster projects - in particular regarding their metadata terminology and models - and the production of the common model to enable the exchange the data;
2. Development of the additional software components for the cluster projects for the demos, including the COSMOS registry, RDF exporter, Metaware exporter, NESSTAR additions and IQML exporter;
3. Theoretical development of MISSION/IPIS statistical processing to enable these two projects to execute statistical operations together;
4. Organization of the final conference and performance of the demos;
5. Production of a strategy report based on the experience of the COSMOS project, and identifying key issues for further research.

To elaborate, a metadata repository was developed and the common model was defined. The IPIS metadata model was selected as the core COSMOS metadata model. Demos, practical and theoretical, showed the exchange of metadata between different projects through the repository and common APIs. The project also defined a methodology for developing interoperability for statistical research projects.

To those involved in larger integration or remote broker type projects, the concept of a simple core model with possible extensions was generally approved as well as the centralised registry concept and the technical architecture. However, for those who expected an ‘off-the-shelf’ statistical core model, the COSMOS model was not usable until further tested and enhanced by including information from other existing models, and the architecture would also probably also have to be enhanced in order to be useful.

COSMOS partners agreed that they had learnt about how to develop and integrate statistical information systems. The embedded nature of metadata models and systems with reference models versus a more contextual and functionally oriented approach had inspired some to think about the limits to more generic approaches, as well as the limits to the integration of existing systems, without these systems having to change or apply some common standard.

COSMOS demonstrated the possibilities for mapping between different standards, enabling metadata to be ‘exported’ from one system to another. This work also contributed to the development of the standards through identifying ‘gaps’ in the standards being used.

By defining the core model, the system architecture and the common API, a methodology for developing interoperability for the existing research projects was found. More specifically we focused on two sub-cases:

a) Importing a classification from METAWARE to MISSION

b) Importing Questionnaire data from IQML to FASTER.

c) IPIS/MISSION communication by:

   1. Export the IPIS dataset's meta-data in a format that can be accepted by MISSION.
   2. Import the meta-data into the MISSION system into a special frame.
   3. Create a new frame in the MISSION system that will contain the IPIS dataset's meta-data.
   4. Whenever a query on the particular meta-data is posed to MISSION, the system knows that it has to bypass the usual route:
   5. Acknowledge the fact that the query concerns IPIS data.

COSMOS brought together existing projects within the same field that focused on different aspects of the statistical process in an attempt to demonstrate the possibilities for integration, to develop common understanding and to share relevant knowledge within the community. It also identified future research directions for related projects and their interroperability, not covered into the COSMOS framework.

Possible impact on ESS and/or the end-user

In its final report, the COSMOS consortium had expressed quite candid views about the role of NSIs. It alleged that there was concern amongst the academic and commercial partners to projects involving the ESS that the ESS tended to work in isolation from developments within other organizations that dealt with statistical processes. Of course, statisticians with expertise in dealing with data collection and statistical processes were not confined to the ESS. There was a considerable amount of experience in a wide range of topics (survey management, data collection methods and tools, metadata management, data quality and comparability issues et cetera) outside official statistics that 'official statisticians' might benefit from, and vice versa.

Nevertheless, the ideas of common interfaces, shared standards and common portals should be of great benefit to all end users of statistical data, not only NSIs. The consortium expressed the view that one problem for users of statistical data was that each organisation presented its own data using different format, data exchange standards and interfaces. Search portals and common interfaces would lead to a much more flexible set of resources. The arrival of the Grid, and its related goals of sharing common resources, would supply the underlying infrastructure for sharing data, but the ability to participate would be dependent on domain-specific standards, and COSMOS should help to create these in a practical environment. Finally, since various metadata terminologies and models are used by each organisation, serving its individual aim, the adoption of a core metadata model, integrated into their systems will minimize data/metadata management and exchange mismatches.

Dissemination and exploitation prospects

There was a public website, a project brochure and a contact list of interested parties to discuss the project's ideas. The main conference was held in Greece in May 2002. All software developed during the project is available from Eurostat. In addition to developing specific services, it should be possible to develop a generic application on top of the ebXML registry that can provide generic services to a variety of user communities. This generic service can be marketed as such for use by value-added service providers and can be tailored to suit individual problem domains. This tailoring could be achieved by the development of services on top of the generic service, using the interfaces provided by the generic service.

Suggested Further Work

Further work was envisaged in the following areas:

- Interoperability, specifically, to establish means by which metadata can be collected at source and be re-used and complemented during the statistical process, ultimately making them available for public use;
- The development of tools for metadata collection for legacy material. Currently, the inclusion of legacy material into web-based delivery systems is selective. The development of generic tools and the adoption of recognised standards will alleviate this situation and enable more material to be placed in the public domain;
- Incorporation of developments relating to the production process for public statistics into the developments demonstrated by COSMOS. The production process was not heavily represented in the Cluster and all partners felt that this was an important omission. By including teams working on the production process, the gap between data collection and data distribution would be filled and, for example, issues of quality assurance and issues relating to the design, editing and construction of derived variables could be discussed and better supported;
- Data disclosure and confidentiality. This area will become increasingly important if the ESS is to capitalise on GRID developments which will enable data from a range of sources to be matched, combined and processed much more easily than at present;
- The COSMOS demonstrator created an embryonic system for managing data dissemination across distributed organisations using a 'google' type search engine based on off-the-shelf and royalty-free software known as a 'registry'. By investing in the further development of this work, the producers and disseminators of public statistics can lead the way in the provision of information on a wide-range of topics, tailored to the needs of particular user communities.

Bibliography

- Papageorgiou, H. and Vardaki, M. (May 2000) COSMOS Projects' profile, The COSMOS Final Conference, Athens, Greece
- Papageorgiou, H. and Vardaki, (May 2000) Definition of Indicators and Standards, The COSMOS Final Conference,, Athens, Greece
- Papageorgiou, H. and Vardaki, (May 2000) Quality/ Footnotes, The COSMOS Final Conference, May 2000, Athens, Greece

DACSEIS
Data Quality in Complex Surveys within the New European Information Society

Timetable	1/3/2001-31/5/2004 (39 months)
Website	http://www.dacseis.de

The Consortium

Member	Role	Institutional type	Country
University of Tübingen (UT)	Coordinator	Academia	Germany
Statistics Netherlands (CBS)	Partner	NSI	Netherlands
Swiss Federal Institute of Technology (EPFL)	Partner	Academia	Switzerland
Johannes Kepler University Linz (JKU/IFAS)	Partner	Academia	Austria
Statistics Finland (StatFi)	Partner	NSI	Finland
University of Southampton (U. Soton)	Partner	Academia	United Kingdom
Swiss Federal Statistical Office (SFSO)	Assistant Partner	NSI	Switzerland
Statistisches Bundesamt (StBA)	Assistant Partner	NSI	Germany

Scope and Objectives

The main goal of the project was to analyze the accuracy of estimates while taking into consideration different aspects of practical needs, such as non-response rates and response behaviour, imputation, rotation schemes and applicability of the methods to large scale universes. An additional task was to develop efficient methods for combining data from surveys and registers. These methods are normally useful in reducing response burdens and may help to improve data quality, especially when dealing with rare events in small areas. An issue in this case is variance estimation. The specific objectives were:

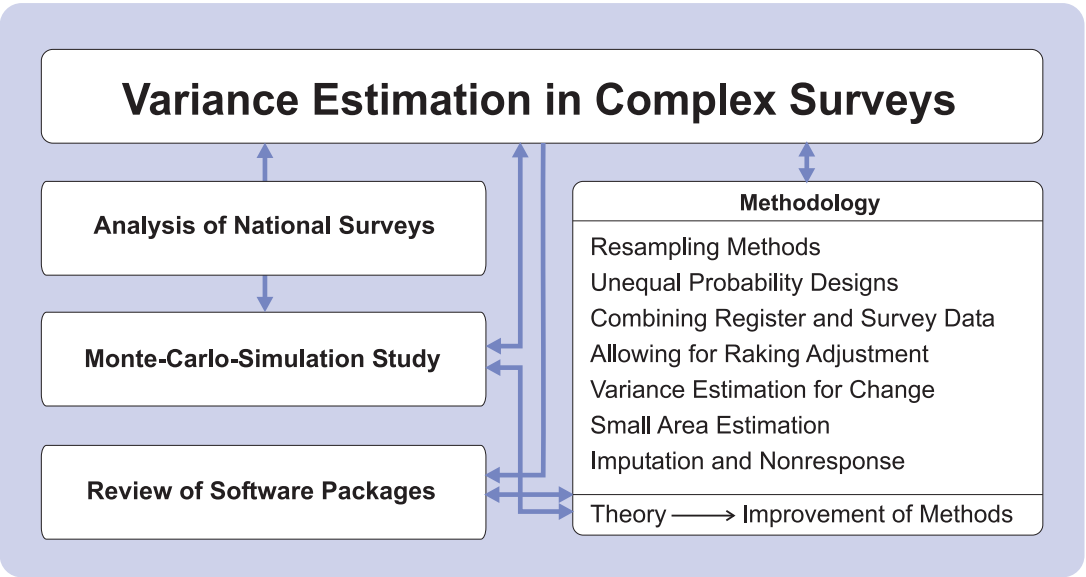
- Standardization and harmonization of variance estimation methods used to calculate sampling errors. This would be achieved with special emphasis on different national surveys within the European statistical system;
- A catalogue of instructions and criteria to be compiled, allowing the user to choose the most effective procedures for variance estimation for complex sample designs;
- Inspection of relevant methods, analyzed and evaluated with respect to their applicability to complex surveys. If the methodology is limited, it would be improved or further researched;
- The scrutiny and evaluation of standard software packages for survey sampling, with special emphasis on the implementation of the above mentioned variance estimation methods for complex surveys;
- Dissemination of the results and the estimation methods to NSIs and other users; this would include the delivery of a recommended practice manual as well as the source and pseudo codes of all relevant estimation procedures.

Outline of methodology

The project was divided into two major parts that would converge in the main workpackage, WP1, on Variance Estimation in Complex Surveys, where all the research results would be merged. One part would be based on the analysis of national surveys that would play a major role in the Monte-Carlo simulation study as a basis for realistic data sets and their practical characteristics. The other part would consist of fundamental theoretical research on different variance estimation methods. These would be evaluated within the Monte-Carlo simulation study regarding their practical applicability to complex survey designs. In an extra activity, standard software packages would be tested with respect to the implementation of different variance estimation methods in complex surveys and the possibility of extending these packages to include new methods.

The methodology would focus on variance estimation of sampling errors for cross-sectional and to some extent also for longitudinal data. Additional work would be done through the examination of variance estimation methods for non-sampling errors with respect to missing values and their imputation.

The methodological framework of variance estimation techniques can be inferred from the following diagram:



Main results achieved

DACSEIS accomplished its major task of investigating data quality, reliability and estimation of precision of values gained from surveys with complex designs. The main focus was on elaborating and improving the accuracy of adequate point and variance estimators in theory and application, and this was done. The research fields covered state-of-the-art methods for re-sampling techniques, unequal probability methods, repeated weighting, raking and calibration, variance estimation for change, and small area estimation methods. The research was conducted under full response and non-response with weighting as well as under single and multiple imputation.

Various software packages were evaluated for their availability of relevant point and variance estimation methods including their applicability. Special emphasis was put on European household and individual surveys such as the Labour Force Surveys and the Austrian and German Microcensus as well as on selected Household and Budget Surveys. In order to gain best practice recommendations for the applicability of the methodology, a large Monte-Carlo simulation study was conducted on the basis of the surveys mentioned above. The aim was to investigate the relevant point and variance methodology in a close-to-reality environment to allow for recommendations on their practical use.

The recommendations derived from the simulation study played a central role in the final project report and in the recommended practice manual.

All results were documented and available in paper form as well as electronically and as scientific books. The voluminous simulation results were illustrated in graphs and tables. Specific outcomes of the research were:

- Set-up of the recommended practice manual database for the implementation of variance estimators;
- Investigation of European individual and household surveys with respect to design, including non- response rates and behaviour, variance estimation and statistical software packages in use;
- Investigation of national household and labour force surveys as follows:
    - Labour Force Surveys from Finland, Netherlands and United Kingdom;
    - Microcensus from Austria and Germany;
    - Household Budget Survey from Switzerland;
    - Income and consumption survey from Germany;
    - Establishment of the criteria for the evaluation of the statistical software packages;
    - Overview of variance estimation methodology including source and pseudo codes;
    - Completion of the universes and sampling procedures for the Monte-Carlo simulation study;
    - List of quality measures for estimators and variance estimators;
    - Investigation of the use of register data for improving labour force surveys

Possible impact on ESS and/or the end-user

DACSEIS offered a methodology for survey sampling to NSIs in Europe to improve the quality of their data and to develop harmonized standards for data quality. Special emphasis was put on the practical needs and the applicability of the methodology to the needs of the ESS. A pervasive and detailed Monte-Carlo study enabled an empirical evaluation of variance estimation methods in realistic European settings for household and individual surveys. Recommended practices emerged to facilitate the end-user from NSIs and other statistical institutes. DACSEIS kept in close touch with NSIs; indeed six out of the ten DACSEIS partners were NSIs, selected to ensure the practicability of an innovative methodology. In fact, the project was designed to facilitate the implementation of the recommended variance estimation methods in Eurostat and in NSIs. DACSEIS has in turn benefited from cooperation with NSIs and also with universities in several European countries. This, together with other projects such as EUREDIT and EURAREA, did lead to a strengthening of the skills infrastructure in Europe.

Specifically, the ESS should find the following extremely useful:

- The recommended practice manual of variance estimation methods for complex surveys which was issued as a database to all potential users;
- Recommendations as a basis for possible standardization and harmonization of variance estimation methods and standards in Europe;
- Analysis of selected national household and labour force surveys to obtain a broader knowledge of their main characteristics in the context of DACSEIS;
- Exchange of information with Eurostat and other users to improve the practical relevance of the research results.

Dissemination and exploitation prospects

A website and a project brochure were established. All results were disseminated in many articles in scientific journals as well as in the recommended practice manual, which was itself accompanied by an electronic version. Theoretical research and simulation results were made available to officials, non-officials and other interested users.

The final conference was embedded in the European Conference on Quality and Methodology in Official Statistics in Mainz, May 2004. Other dissemination activities included presentations at the ISI conference in Berlin; JPSM meeting in Calcutta; and a lecture day from DACSEIS partners. Many papers were accepted in international journals and conference proceedings. Additional papers were published as DACSEIS research papers.

The web-based platform will continue at least two years. An overview of the project is given in deliverable 12.1 (cf. http://www.dacseis.de).

Future research challenges/possibilities

The DACSEIS research has played an important role in future applications of variance estimation methods in official statistics. CBS intends to combine small area techniques with repeated weighting. Further, the use of registers and the alignment of tables estimated from different surveys are under investigation (ISI 2007 satellite meeting in Southampton). DESTATIS / Münnich (now University of Trier) developed variance estimation applications for the German access panel and DE-SILC. Additionally, Hulliger and Münnich are conducting research on the influence of outliers on variance estimation methods (cf. JSM 2006 invited session 177). Finally, University of Trier (Münnich) and ZUMA Mannheim (Gabler) started research into the estimation methodology for the German register-based census, in which DESTATIS variance estimation methods have played an important role.

Bibliography

- Berger, Y.G. and Skinner, C.J. (2005). A Jackknife Variance Estimator for Unequal Probability Sampling. Journal of the Royal Statistical Society Series B, 67, 79-89.
- Davison, A. C.,Münnich, R., Skinner, C. J., Knottnerus, P. und Ollila, P. (2004): The DACSEIS Recommended Practice Manual. DACSEIS deliverable D12.3, http://www.dacseis.de.
- Davison, A. C. and Sardy, S. (2007) : Resamping variance estimation in surveys with missing data. Journal of Official Statistics, to appear.
- Knottnerus, P. and C. van Duin (2006). Variances in Repeated Weighting with an Application to the Dutch Labour Force Survey, Journal of Official Statistics, 22, pp. 565-584.

DIASTASIS
Digital Era Statistical Indicators

Timetable	1/12/2002 - 30/11/2004 (24 months)
Website	http://www.eurodyn.com/diastasis/

The Consortium

Member	Role	Institutional type	Country
European Dynamics SA (ED)	Coordinator	Company (IT)	Greece
Fraunhofer Gesellschaft – Institut für Autonome Intelligente Systeme (FhG/AiS)	Partner	Research	Germany
Institut d'Estadística de Catalunya (IDESCAT)	Partner	Public sector	Spain
Universitat Politècnica de Catalunya (UPC)	Partner	Academia	Spain
Serveis d'Acces a Internet de la Universitat Politècnica de Catalunya (UPCnet)	Assistant Partner	Academia	Spain

Scope and Objectives

One of the main aspects of the new economy has been the continuously extending use of the Internet. Thus, the measurement and assessment of this use, as well as the precise definition of the characteristics and the profile of Internet users, is a very specific research issue. Although significant efforts were made towards the measurement of the socio-economic impact of the Internet, existing methodologies and tools were insufficient and needed to be extended. Moreover, there was the need to develop the required capabilities and means to measure, assess and comprehend the changes taking place and their effects. Thus, the collection, classification and detailed analysis of the associated information have become a primary goal. The new framework should have a multinational (and especially a pan-European) dimension. Experts should be able to access high-quality, accurate data at an international level. Therefore, it was of high importance to produce and measure new indicators that would provide a clear view of these issues.

The main objectives of the DIASTASIS project were:

- To develop a new methodology for correlating statistical data on SMEs and household research data (panels) with statistical data on Web usage (visitor/Internet demographics and preferences);
- To define, measure and exploit new socio-economic indicators regarding the use of the Web;
- To design and develop computer tools to gather data on Web usage; implement the new data collection and correlation methodology; and use the new indicators;
- To Investigate how new indicators could be used for supporting decisions on the new economy in Europe;
- To evaluate the results of the new methodology and the impact of the use of the new indicators on perceptions of, and decision-making on, the new economy.

Outline of methodology

The methods that were used as a starting point for the project belonged to the family of techniques called “Data Fusion” and “Data Pruning”. This type of technique, where fusion is more often implemented than pruning, had originated from market studies in which it was not possible to ask all the questions to the same individual because of the extensive length of the questionnaire. Therefore, in order to acquire extensive additional information about the market and the customer base without commissioning a new survey, analysts opted for the solution of fusing two different surveys.

This broad methodology spawns many different specific techniques that range over a broad spectrum of complexity. One of the simpler techniques is that of profile matching, which researchers have been using for many years, in order to link two surveys. For example, if from one survey the analyst observes that people of a certain characteristic (e.g. age group) tend to use a certain product, then it is logical to choose the TV program that is mostly watched by audience of the same age-group in order to advertise it. The techniques of matching two different data sets that were explored in the project are a refinement of this simple example.

For specific activities were to:

- Examine the adequacy of existing statistical methods used to produce indicators;
- Identify suitable information sources at national and pan-European levels and define all the possible legal and access rights problems;
- Study the content and format of the available data;
- Design and develop innovative methodologies for the analysis of the data;
- Define, propose and measure new indicators based on the new methodologies;
- Design, develop and test computer-assisted applications that would implement and support those methodologies;
- Verify the usefulness of the new methodologies and the system by running a pilot application and suggesting improvements.

Main results achieved

DIASTASIS developed efficient and innovative methodologies and tools for the analysis of the existing information relating to the use of the Internet and its socio-economic impact at a pan-European level. This was done on the basis of data available from household surveys and in the systems of Internet Service Providers (ISPs). In this respect, new statistical indicators were produced and calculated on a regular basis, thereby providing reliable information on the issues concerned. In outline, the accomplishments were:

- Development of enhanced methodologies for the analysis and processing of the available data;
- Development of a reliable and accurate way to measure the impact of the Internet and its use and to produce indicators on a regular basis;
- Development of methods which enabled users to gain valuable information regarding Internet usage;
- Development of the DIASTASIS Statistical Collaboration platform (SCoP) which:
    - Supported the collection of high quality information on web usage;
    - Supported the generic innovative statistical framework combining statistical data from different sources (detailed web data and socio–economic data of Internet users);
    - Provided a state-of-the-art environment where application and validation of DIASTASIS methodologies were performed to a wider population that did not display the homogeneity of the sample used during the Pilot operation;
    - Provided a state-of-the-art environment that should allow experimentation on DIASTASIS methodologies and the extraction of meaningful results regarding statistical indicators on ICT usage.
- Improvement of the reliability and the timeliness of the information and the indicators produced by the project.

Possible impact on ESS and/or the end-user

The prospects for extensive exploitation by the ESS should be bright. DIASTASIS had initiated a marketing campaign to disseminate project results and investigate NSIs' interest in common development exploitation projects. ED provided NSI representatives with user accounts for logging into SCoP platform and has established close communication with these representatives.

There would be integration with modern statistical processing environment, which would allow statistical users to access to large open source developed statistical processes. Its modular design would allow for easy integration of NSIs' statistical tools.

DIASTASIS also created a questionnaire which NSI representatives were requested to use for typifying their feedback and it had also established a communication channel with the interested NSIs for common development projects. Feedback was collected from interested NSIs and DIASTASIS planned to pursue opportunities with them as well as EUROSTAT for common development projects.

Dissemination and exploitation prospects

There were the project website, CD-Rom with the project outputs, project brochures, workshops, commercial fairs, participation in special interest groups and in the EU concentration activities, production of scientific papers, publication in scientific journals and participation in conferences.

The user groups targeted were NSIs, data providers/publishers, data analysts and statistical tool developers, specifically private sector companies, market research firms, special Common Interest Groups, RTD institutes and universities.

Conferences DIASTASIS actually participated in and the presentations made to them were:

- Presentation in the conference, New Research for New Media, Tarragona, October 2004;
- V. Meléndez the 1st Panel of Experts Workshop, Barcelona, July, 2003;
- Tomas Aluja, Gerhard Paaâ, Albert Prat, Ingo Schwab, "Enhancing socioeconomic surveys by data about internet usage" KDNet Symposium, Bonn, Germany, June 2004.

The consortium had intended to attend:

- IEEE International Conference on Data Mining;
- ACM/SIGKDD International Conference on Knowledge Discovery and Data Mining;

One of the main goals of the DIASTASIS project was to build up new excellence in web-mining and statistical technologies and to transfer that knowledge to industry and research. The DIASTASIS project actually encouraged further exploitation of its socio-economic indicators; the data analysis and statistical methodologies; the platform; and the consultancy services. In order to deliver a set of methodologies adapted to the needs of users conducting research in the statistical sector, and particularly on statistics related to the use of Internet, the consortium would, on various terms, deliver consultancy services, including customisation, development of extension modules, adjustments and fine-tuning.

BIBLIOGRAPHY

- Statistical Methods and Applications, Springer – ISSN: 1618-2510
- Computational Statistics, Springer - ISSN: 0943-4062
- Statistical Papers, Springer - ISSN: 0932-5026
- Computational Statistics & Data Analysis, Elsevier - ISSN: 0167-9473
- Statistics & Computing, Kluwer - ISSN: 0960-3174
- Journal of the American Society for Information Science and Technology, Wiley - ISSN: 0002-8231
- International Journal of Scientometrics, Informetrics and Bibliometrics, Cindoc - ISSN 1137-5019
- Knowledge & Information Systems, Springer - ISSN: 0219-1377
- Applied Soft Computing, Elsevier - ISSN: 1568-4946
- Information Fusion, Elsevier - ISSN: 1566-2535
- Data Mining and Knowledge Discovery, Kluwer - ISSN: 1384-5810
- Information Retrieval, Kluwer – ISSN: 1386-4564
- Transactions on Information Systems, ACM - ISSN 1046-8188
- Knowledge and Data Engineering, IEEE Trans., Product No.: 016-147-TBR

ECOSTAT
Environmental Consolidated Statistical Tools

Timetable	1/1/2001-31/12/2003 (36 months)

The Consortium

Member	Role	Institutional type	Country
Panteion University of Social and Political Sciences, Athens	Coordinator	Academia	Greece
Turku School of Economics and Business Administration, Turku	Partner	Academia	Finland
University of Aegean, Mytilini	Partner	Academia	Greece
World Systems (Europe) Limited	Partner	Company	Luxembourg
Statistics Finland (StatFi)	Partner	NSI	Finland

Scope and Objectives

There seemed to be a growing demand for environmental statistics and specialized statistical methodologies that could cope with the particular needs of environmental scientists. ECOSTAT was intended to produce statistical methodologies and tools that would meet that demand. It would engage in research along the most important issues associated with the statistics of environmental analysis:

- The use of data mining software tools and advanced specialized statistical modeling to identify significant relationships;
- The identification of user needs concerning statistical methodologies and software tools required for the analysis of environmental issues;
- The effective statistical data selection process from multi-source data;
- The elaboration of statistical indicators to measure environmental phenomena;
- The creation of a database/inventory of sources, data and metadata in the field of water quality for a specific geographical coverage i.e. at least 4 countries spanning both northern and southern EU countries;
- The development of a specialized prototype software to be used by environmental scientists;
- The dissemination of the final S/W among others through the website, leaflet etc.

Outline of methodology

The models already developed and surveyed by Eurostat in a recent project on 'Current activities on Environmental Pressure Indicators Modelling' together with later environmental models would be analysed and evaluated in terms of their prediction potential. Data mining software tools would be used to identify latent relationships and to establish new environmental models. The results would be incorporated into the ECOSTAT prototype software.

Research would also be undertaken in the area of Material flow accounting (MFA) because existing methods failed almost completely to analyze the interaction between economic activities and natural resources i.e. material flow. A scientific calculation model would be developed suitable for forecasting total consumption of natural resources and economic development, the so-called decomposition model. This model could be utilized in the analysis of sustainability to compare the functioning of national economies with sectoral performance. This model would be tested in Finland where relevant data were available. It would create new possibilities for the development of advanced empirical evaluation methods of sustainability.

Existing theory would be better used and further developed. Pioneering work by the OECD on a conceptual framework, namely the PSR model, (later extended to DPSIR framework by the European Environmental Agency), has influenced the indicators activities of a number of countries and of various international organizations. The OECD programme on environmental indicators was path-breaking. It aimed at the integration of environmental concerns into sectoral policies, environmental and natural resource accounting and the development of indicators for use in environmental performance reviews. The DPSIR framework was used to present the information related to each section in a structured manner. ECOSTAT's target was to connect policy with data from the model and provide information from other components, such as the database/inventory.

The use of environmental indicators was increasing worldwide and many countries had developed a system of indicators. Indicators were used to increase awareness of environmental issues, to influence and guide policy development and to stimulate planning efforts to reduce environmental pressures.

Latest technology was used as UML and RUP in order to develop the software prototype. This would be achieved through well-defined requirements and specifications.

Main results achieved

The most important innovations were:

- The use of data mining software tools combined with environmental pressure-state modeling and MFM in the field of environmental science, with improvements and adaptation of the MFM to the case of water and agriculture;
- The statistical research and its applications to environmental statistics, including the creation and use of multi-source data;
- The development of a user-friendly software prototype incorporating statistical methods and tools to be used for solving multiple problems in the water quality environment field. The conceptual architecture of the software prototype was established.

Specifically, a public survey on users needs was conducted. The results shaped the profile of the software requirements, including the Graphical User Interfaces (GUI). A number of use-cases were also identified that imparted structure to environmental policy.

There was an extensive analysis of the database specifications required for the statistical analysis and the implementation of the environmental modeling and the associated support of data mining tools. Instead of relying on conventional client/server platforms of relational databases (Oracle, Microsoft SQL server, etc.), a combination of Microsoft Excel and Access was chosen.

A data mining software was developed including state-of-the-art components and it was tested.

The multisource software prototype developed in the project could be adapted to water quality and agriculture-environment statistics such as the GIS-framework. It was flexible and could be adapted to many different platforms.

The software prototype was constructed according to a three-tier model. Decisions concerning users' requirements were incorporated in the three layers. These layers were the GUI layer, the Business Logic layer and the Data layer. The prototype software was used to form policy on environmental issues and, more specifically, to propose actions according to the water balance algorithm.

The software was tested by various bodies of users in Finland and Greece. These bodies were the Environmental Planning Division of the Greek Ministry for Environment, the Statistical Service of Finland (Environmental Department) and the University of Turku in Finland – Department of Economic Statistics – Finland Future Research Center. The software in its final form was tested technically by World Systems in Luxembourg,

ECOSTAT research had shown that the plethora of existing indicators sets were not based on a common theoretical rationale, neither were they serving the same purpose. Furthermore, solutions to the aggregation problem were mostly simplistic and scaling varied greatly between indicators.

The research on indicators took account of the outcome of the user' survey. The selection of indicators reflected a mixed ecological/economic methodology. The environmental indicator system for agriculture-water system was developed using the so-called DPSIR-framework.

Finland's Futures Research Centre (FFRC) had deployed the decomposition approach in sustainability analysis of national economies and published the results in various articles. They had also developed the environmental indicator system for the agriculture-water system using the so-called DPSIR-framework; and they had worked extensively on the analysis of different Multi-Criteria Decision Analysis methods.

Possible impact on ESS and/or the end-user

The statistical methods that were developed and applied in this project should contribute to the advancement of the ESS in general. Statistical methods dealing with the construction and properties of statistical indicators and decisions concerning multi-source data could advance the state-of-the-art in statistical research, further improving the ESS.

This project should affect positively the economic development of the European countries since a better management of natural resources, especially water, could create additional prosperity. Improvements of the environment should translate into better population health and less expenditure on the health services. Moreover, the project should offer a considerable scientific contribution to European technological progress because it contains innovative elements in software development and related products.

Dissemination and exploitation prospects

There was the website, which gives access to all results of the project together with a time-limited demonstration copy of the ECOSTAT software for free download by any interested internet user. There was a project leaflet that was disseminated widely, and a dissemination/exploitation plan, which aimed at increasing market acceptance and boosting commercial penetration for the software developed in the project, together with its user manual. ECOSTAT partners participated in conferences, seminars and meetings of thematic networks and they published many articles and related publications, see Bibliography below.

Intended as demonstrators, the pilot applications have the potential to be exploited. Both pilot applications could provide services to a broad customer base. It is in the interest of all partners to make the component architecture available to the broad public as seed ware. The partners agreed to license the results of ECOSTAT to each other beyond the end of the project but not to develop a detailed joint exploitation plan. Rather, each partner would exploit his results individually.

Future research

Future research might be oriented to global models, especially to evolutionary interactive ones combining environment with society and economy and, seeking valid forecasts for the future state of the world, oriented also to the environment. Through such models, it should be possible to predict the impact of several policies on the future state of the environment, to evaluate them and to take timely corrective actions.

BIBLIOGRAPHY

- Hoffrèn, J. (2003) Use of Multisource Data in Environmetal Statistics, Suomen Tilastoseuran vuosikirja
- Kaivo-oja, J.K., Jyrki Luukkanen, J.L and.Pentti M.P. (2002) Methodology for the Analysis of Critical Industrial Ecology Trends: An Advanced Sustainabiltiy Analysis of the Finnish Economy, Indicators of Sustainable Development (Ed. Olli Hietanen) Futura, Vol 21,
- Kaivo-oja, J.K., (2002) Social and Ecological Destruction in the First Class: A Plausible Social Development Scenario, Sustainable Development, Vol 10
- Katko, T.K, Seppälä, J.K. and Kaivo-oja J.K (2001) Management of Water, Wastewater and Solid Waste Services in Comparative Historical and Futures Perspective. A Nordic Research Workshop supported by NorFa: TUT, IEEB
- Hoffrèn, J.H., Luukkanen J.L. and Kaivo-oja J.K. (2001) Decomposition Analysis of Finnish Material Flows: 1960-1996, Journal of Industrial Ecology, Vol 4
- Kaivo-oja, J.K., Luukkanen J.L. and.Malaska P.M. (2001) Advanced sustainability analysis. In Tolba, M.K. (Ed.) Our Fragile World. Challenges and opportunities for sustainable development, Encyclopedia of Life Support Systems and Sustainable Development
- Kaivo-oja, J.K., Luukkanen J.L. and.Malaska P.M. (2001) Sustainability Evaluation Frameworks and Alternative Scenarios of National Economies, Population and Environment, Vol 21
- Tassopoulos A.T. and Papaioannou D.P. (2003) Design of Information Systems for Environmental Managers: An example using interface prototyping, Neural, Parallel & Scientific Computation
- Papaioannou D.P. and Tassopoulos A.T. (2003) Metadata as a knowledge management tool: Supporting Environmental Information Systems and end user access to spatial data, Implementing the Learner-Centered Design Paradigm for Web-Based Training Curricula, Special Issue
- Tassopoulos A.T. and Papaioannou D.P. (2003) The Politics of Participation in Watershed Modeling, Working Papers of Panteion University, Special Issue
- Hoffrèn H.J. (2003) Use of Multisource Data in Environmetal Statistics, Suomen Tilastoseuran vuosikirja,
- Kaivo-oja, J.K, Jyrki Luukkanen, J.L.and Pentti, M.P.(2002) Methodology for the Analysis of Critical Industrial Ecology Trends: An Advanced Sustainabiltiy Analysis of the Finnish Economy, Indicators of Sustainable Development (Ed. Olli Hietanen) Futura, Vol 21, Issue 2
- Kaivo-oja, J.K, (2002) Social and Ecological Destruction in the First Class: A Plausible Social Development ScenarioSustainable Development, Vol 10
- Katko, T.K.,.Seppälä, O.S. and Kaivo-oja, J.K. (2001) Management of Water, Wastewater and Solid Waste Services in Comparative Historical and Futures Perspective. A Nordic Research Workshop supported by NorFa: TUT, IEEB
- Kaivo-oja, J.K, Luukkanen, J.L. and Malaska P.M. (2001) Advanced sustainability analysis. In M.K. Tolba (Ed.) Our Fragile World. Challenges and opportunities for sustainable development, Encyclopedia of Life Support Systems and Sustainable Development, Vol 2
- Kaivo-oja, J.K, Luukkanen, J.L. and Malaska P.M. (2001) Sustainability Evaluation Frameworks and Alternative Scenarios of National Economies, Population and Environment, Vol 21, Issue 1
- Tassopoulos, A.T. and.Papaioannou, D.P. (2003) Design of Information Systems for Environmental Managers: An example using interface prototyping, Neural, Parallel & Scientific Computation,
- Tassopoulos, A.T. and.Papaioannou, D.P. (2003) Metadata as a knowledge management tool: Supporting Environmental Information Systems and end user access to spatial data Implementing the Learner-Centered Design Paradigm for Web-Based Training Curricula, Special Issue
- Tassopoulos, A.T. and.Papaioannou, D.P. (2003) The Politics of Participation in Watershed ModelingWorking Papers of Panteion University, Special Issue

EICSTES
European Indicators, Cyberspace and the Science-Technology-Economy System

Timetable	1/12/2000 -31/3/2004 (40 months)
Website	http://www.eicstes.org

The Consortium

Member	Role	Institutional type	Country
Consejo Superior de Investigaciones Científicas (CSIC)	Coordinator	Public sector (research)	Spain
Austrian Research Center Seibersdorf (ARCS)	Partner	Public sector (research)	Austria
Centre National de la Recherche Scientifique (CNRS)	Partner	Public sector (research)	France
Computer Technology Institute, University of Patras (CTI)	Partner	Public sector	Greece
University of Amsterdam (UvA)	Partner	Academia	Netherlands
University of Surrey (UNIS)	Partner	Academia	United Kingdom
Institut d'Estadística de Catalunya (IDESCAT)	Assistant partner	Public sector	Spain
Koninklijke Nederlandse Academie van Wetenschappen (KNAW/NIWI)	Partner	Public sector (research)	The Netherlands
Danish Technological Institute (DTI)	Partner	Public sector	Denmark

Scope and Objectives

The project scope was to offer statistics and to derive indicators about the European Science-Technology-Economy System in Internet. The objective was to test and use agents to recover data from the web in an automatic way and to apply new models and concepts to discover relationships between the various actors of the new economy. A series of case studies involving different and complementary aspects, especially relevant to the European scenario, were analysed. Because of the special nature of the Internet, the intended coverage of the proposal was global, with strong emphasis on EUROPE. The specific objectives were:

- To develop automatic web data collection procedures using agent technology;
- To obtain data about the Internet economy from sources other than the Internet itself;
- To build statistical databases with the complex, large and multidimensional information obtained in the project;
- To derive indicators from those databases by applying powerful methodological tools for quantitative processing such as graph theory, complexity and chaos theories and social network analysis;
- To understand and generate models about the dynamics of the new economy, uncovering relationships among R&D sector, industry, economic sectors, culture and society;
- To apply quantitative methodologies to describe empirical case studies such as the network of University-Industry-Government relations, the new role of intermediation, the balkanisation of some disciplines, trans-border co-operation and the way electronic information was consumed by end-users;

- To analyse the role of intermediaries in cyberspace, the new emerging groups and the tasks they were facing in a self-organising environment;
- To show the indicators of the new economy in a universal, user-friendly environment based on new techniques of visualisation.

Outline of methodology

The chosen approach was global, involving not only the 15 EU members but indirectly the rest of European countries and others linked from the academic webspace. Indicators were developed in those areas where the specific needs of European citizens were strongest, a priority that was reflected in the selection of the different case studies.

A first step was to review the state of art of the different topics to be targeted in the project producing several public deliverables that are still strongly cited in current papers. The most important are those related to comparative analysis of commercial crawlers and the one devoted to visualizations tools.

The project created 74 indicators designed to answer a wide variety of research questions, with a view to map the development of the science system in interaction with economic, technological and political developments, and thus to be able to promote understanding and analysis of the knowledge-based economy.

The EICSTES partners built a large database containing information on the internet web sites of 791 European Universities from the 15 EU countries. This information included (a) how many pages and what kinds of files were present in each site; and (b) where or whom the sites were linked to. Furthermore, the data on the web sites could be classified by country, university, department (scientific field) and region because each individual page stored in the database was labelled with this valuable additional information.

Main results achieved

EICSTES results could be classified as methodological or technical and statistical or assessments of the European R&D System. A model, called Amsterdam Model, provided an overall strategy for implementing a large and complex project, including linkages between its activities. The Amsterdam Model might serve as a reference in future Cybermetrics projects.

One of the most interesting results was the development of Miri@d, an interactive server for statistical work. It generated descriptive statistical data on Web users' searching behaviour and what use was done from a digital bibliographical database.

EICSTES developed a tool to visualize and map its results. Appropriate displays of cluster points could give an insight that would not be possible from tables or simple summary statistics. For some tasks, appropriate visualization was the only tool needed to solve a problem or confirm a hypothesis. The prototype of a visualization tool developed in the project could map complex objects in a two-dimensional space.

EICSTES tested the tool BibTechMonTM for structuring, visualizing and analyzing data it collected. BibTechMonTM was a software based on bibliometric methods. It was constructed in such a way as to calculate the data for the network topology on its own, using data from the literature databases.

EICSTES described the Multi-Maps SOM prototype and its visualization-based analysis functions. Multi-Maps SOM prototype was based on a Kohonen self-organizing map for clustering and mapping according to a multi-maps extension. The prototype was called Multi-SOM. It deployed a mix of technologies on internet servers and clients using the capabilities of current Servers and Browsers.

The EICSTES database contained information on the internet sites of 791 European Universities from the 15 EU countries. This information included (a) how many pages and what kinds of files were present in each site; and (b) where or whom the sites were linked to. Furthermore, the data on the web sites could be classified by country, university, department (scientific field) and region because each individual page stored in the database was labeled with this valuable additional information.

The project established empirical knowledge of the possible relations between web data and the knowledge-based economy, and it performed analyses of links and onsite pages at a cross-country and cross-scientific field levels to determine the quality and intensity of such relations. One result was that the relations amongst websites of academic bodies located in the EU-15 reflected the relations that prevailed off line.

Additional substantive results were:

- Central and North European EU countries performed better in terms of knowledge indicators than the countries in the South;
- Different types of scientific areas applied the internet to different extents. Technical areas, such as Mathematics and Computer Science, used the internet to a larger extent than other, less technical areas to make information available;
- Within two selected scientific areas, Computer Science and Biology-Biochemistry, the departments concerned attracted more in-links from universities in the 15 EU countries than "average" university departments.

Possible impact on ESS and/or the end-user

The consortium intended to satisfy both short-term and long-term needs of users, for which purpose users would be involved from the outset not only to disseminate results but also to give feed-back. In the short run, users would mainly be such colleagues as researchers, technical experts and statisticians.

In the long run, the Webometrics Ranking of World Universities (www.webometrics.info) published from 2004, with semester updates, is the most visited secondary result of the project, with more than 1 million visitors per year.

Dissemination and exploitation prospects

The target users were the academic community, R&D and innovation-related communities, statisticians, enterprises, the citizen and other end-users. There was a project website with advanced visualization techniques and a widely-disseminated project presentation leaflet.

The central website (www.eicstes.org) is still operative with details of the project, most of the empirical web data collected and a repository of relevant documents compiled for/from the project.

Several papers were published in general or popular journals when data became available in order to disseminate results and basic conclusions to a wider audience. Over 15 conferences were attended with contributions from partners.

Software developed included French (INIST-CNRS) Miri@d, not available for open use because the technical and computer requirements were very high and also because it was customised to a specific database, Austrian BibTechMon, that is now public domain and it can be obtained upon request from ARCS and the English (UNIS) web log analysis tool (shareware).

BIBLIOGRAPHY

- Widhalm, C.; M. Topolnik, A. Kopcsa, E. Schiebel and M. Weber (2001): Evaluating Co-operation Patterns - Application of a bibliometric visualisation tool to the 4th Framework Programme and the Transport Research Programme; Research Evaluation, 10 (2001), 2: 129-140.
- Polanco, X.; Francois, C. & Lamirel, J. C. Using artificial neural networks for mapping of science and technology: A multi-self-organizing-maps approach. Scientometrics, 51(2001), 1: 267-292.
- Besselaar, P. The cognitive and the social structure of Science and Technology Studies. Scientometrics, 51 (2001): 441-460.
- Boudourides, M. & Antypas, G. A Simulation of the Structure of the World-Wide Web. Sociological Research Online, 7 (2002), 1. http://www.math.upatras.gr/~mboudour/articles/ssweb.pdf
- Heimeriks, G.; Hörlesberger, M.; Besselaar, P. Mapping communication and collaboration in heterogeneous research networks. Scientometrics 58 (2003): 391-413.
- Lamirel, J.-C.; Shehabi, S. ; Francois, C. & Polanco, X. Using a compound approach based on elaborated neural network for Webometrics: An example issued from the EICSTES project. Scientometrics, 61 (2004), 3: 427-441.
- Lamirel , J.-C.; Francois, C.; Shehabi, S. & Hoffmann, M. New classification quality estimators for analysis of documentary information: Application to patent analysis and web mapping. Scientometrics, 60 (2004), 3: 445-562.

EPSILON
Environmental Policy via Sustainability Indicators on a European-wide NUTS-III Level

Timetable	1/12/2002-30/11/2005 (36 months)

The Consortium

Member	**Role**	**Institutional type**	**Country**
EPSILON International SA (EPSILON)	Financial Coordinator	Company	Greece
National Technical University of Athens (NTUA)	Scientific Coordinator	Academia	Greece
Planungsbüro Prof. Dr. Joerg Schaller (PbS)	Partner	Company	Germany
National Statistical Service of Greece (NSSG)	Partner	NSI	Greece
Department of Statistics and Research, Cyprus (DSR)	Partner	NSI	Cyprus
University of Minho (UM)	Partner	Academia	Portugal
Swiss Federal Institute of Technology Lausanne (EPFL)	Partner	Public sector	Switzerland
Statistical Office of Stadt Köln (Cologne)	Partner	Public sector	Germany
Province of Cosenza (COSENZA)	Partner	Public sector	Italy
MiCE Engineering (MICE)	Partner	Company	Italy

Scope and Objective

The objective was:

- To deliver the theory and the mathematical model (statistical s/w "tool) consisting of: (a) a set of methods and a mathematical model, and (b) a mathematical clustering mechanism & s/w technology (model, GIS interface, web-technology). Both s/w packages would deliver and cluster environmental sustainability indicators by exploiting Internet based IST on a NUTS-III level;
- To demonstrate and disseminate the use of the "tool" while serving the needs of official statistics within the ESS at NUTS-I, II and III levels;
- To assist decision makers in the dynamic allocation of environmental funds and the prioritisation of actions and to provide an early warning system, as well as to in planning for environmental components (e.g., water resources management).

To achieve this aim, EPSILON would advance the statistical and IST technologies in:

- On-line linkage to databases e.g. in Eurostat, NSIs and the EEA;
- Mathematical algorithms & modelling in sustainability indexing;
- Sustainability clustering at NUTS-III level;
- GIS web-based technologies;
- Web-based dissemination and training procedures, including setting up of a Centre of Excellence.

Outline of methodology

EPSILON would be produced, tested, validated and disseminated in five steps:

1. Sustainability Model: adaptation of concepts, development of algorithms and mathematical expressions and development of an integrated computerised model for environmental sustainability indexing (indicators model) at EU-wide NUTS-III level and at river basin level. EPSILON would advance the state-of-art technology with a "tool" aimed at generating "primary", "secondary" and "tertiary" sustainability indicators grouped into 6 categories: physical environment indicators, biotic indicators, social indicators, economic indicators, aggregated indicators and clustered indicators. The proposed approach would be based on a first set of key pressure or "practice indicators" and a second set of "performance indicators". These two levels would be linked through the sustainability model.
2. Clustering Model: the development, application and validation of clustering (e.g., neural network) model geared to the clustering of the sustainability indices produced at 1 above and arranged into user specified clusters at EU-wide NUTS-III and river basin levels.
3. Tool & Reliability: conducting model testing, application, verification and validation in six countries or selected regions of these countries, with possible expandability to cover all EU-15 countries.
4. Web Technologies: web-based operation and visualisation of the sustainability predictions and the clustered data via Internet based GIS technologies at EU-wide NUTS-III level by following the JRC/Ispra Dashboard Indicators & other approaches.
5. Centre of Excellence & Dissemination: establishment of an EU Centre of Excellence on "Environmental Sustainability Indicators" and related activities, see dissemination and exploitation below.

Main results achieved

Data were collected from European and international databases for NUTS 0, I, II and III and a comprehensive database was created. It included a large number of sustainability indicators and sub-indicators. Readily available data were gleaned from several European and international sources. Data manipulation was unavoidable in some cases including those where:

- Data was found for the desired indicator at a specific NUTS level but no data was available for other NUTS levels;
- Data was found for a number of NUTS-codes but not for others;
- Data was found from different sources with incompatible units or years.

The EPSILON sustainability model relied on the establishment of a first coherent sustainability structure for addressing quality of life issues across four dimensions, viz the environmental, the social, the economic and the institutional. Enhancing the DPSIR model was rooted in solid rationale:

- The UNCSD structure was adapted keeping 4 pillars/dimensions but defined across 16 themes and around 50 sub-themes resulting in a 4 by 4 framework structuring more than 150 indicators;
- The distinction between state indicators and response indicators allowed the definition of a dynamic sustainable model;
- Regional sustainability assessments illustrated the interest and desirability to move down from a national assessment to a more local level.

The investigation of the analysis of the interrelations between regions was performed with the definition of an ecological-economic framework integrating trade and multimedia pollutant transfer, with a first illustration across European Member States.

The definition of composite indicators was a key issue of the EPSILON project. Since the sustainability model relied on more than a hundred indicators, a new aggregation methodology based on life cycle and comparative risk assessment was developed for the Environmental pillar to calculate the so called Concerted Factors (CDF), a new metric linking environmental parameters via various methods as concentrations to human health damages via a human exposure and an effect modelling. This metric was calculated also as a demonstration using the IMPACT2002 model, a European multimedia fate and exposure model for the assessment of toxic chemicals.

Space Models were created to emphasize the particularities of the geo-referenced data being analysed. A Space Model integrated groups of regions that exhibited similar behaviour. Each group represented a cluster aggregating regions that were similar in some way, with maximum differences between different regional clusters. The GeoSpace Clustering Technique developed was intended for the creation of Space Models. This technique was implemented in the form of a software tool (API) and was freely available for download.

Existing GIS spatial databases usable for the model application and the definition of user requirements were analysed. The main focus was on the basic NUTS geometry and data for the environmental pillar. The model further developed the web application prototype based on ESRI WebGIS and ESRI modeller that enabled the user to access and visualise the data collected. The integration of the clustering technology led to a second prototype. A detailed documentation including a set-up guide, a user manual, an online-help and a video-presentation explaining the functionality of the web application were created.

The overall set of applications consisted of basic and extended applications, the former dealing with benchmarking analyses and the latter dealing with sensitivity analyses, analyses of temporal trends; with the transferability of the model philosophy to other sustainability assessment contexts. The applications were created using/adapting the model and the GIS tool developed within the project. Most of them had required the insertion of new data/indicators in the database, new normalisation/aggregation operations and also, in one case, the development of a new sustainability model. All of them evinced high capability, versatility and flexibility of the model and of the GIS tool to cope with different sustainability scenarios.

Possible impact on ESS and/or the end-user

EPSILON is operational via Internet based GIS technologies (ESRI technology), compatible with those used by Eurostat, NSIs and the EEA. The techniques and software for clustering are very relevant to the activities of the ESS. Thus, at the least the outputs of EPSILON should be a kind of background reference for practising statisticians.

Dissemination and exploitation prospects

There were the project website, a widely circulated project brochure, the establishment of an EU "Environmental Sustainability Centre" of Excellence, targeted bulk mail, participation in conferences, publications in RTD journals, presentations in newsletters, formation of an EEIG, distribution of s/w and information via CD-ROM, and a licensing agreement with international (non-EU) and US organisations, and more than 20 major publications.

The partners were committed to continue their efforts for dissemination and future use of the tool beyond the formal end of the project. An exploitation plan was produced outlining the dissemination and use intentions of the partners for the project tool. The exploitation options adopted by the partners were the following:

- The Web version of the tool would be used for marketing and for demonstration, without charge;
- Users would be able to contact the project consortium to buy a CD-ROM standalone version of the tool that was based on ARC GIS;
- The project team would provide consultancy and assistance to enter the users' data into the standalone desktop version and to run the tool to obtain results;

The above options did not exclude the provision of additional services, for example:

- Sale of the product and database;
- Producing sustainability maps upon request.

ESIS
Development of a European Satisfaction Index System for the new economy

Timetable	1/12/2001-30/11/2004 (36 months)
Website	http://www.esisproject.com

The Consortium

Member	Role	Institutional type	Country
Spad	Coordinator	Company (IT)	France
Agilis S.A.	Partner	Company	Greece
Università degli Studi di Napoli Federico II (DMS)	Partner	Academia	Italy
Universitat Politècnica de Catalunya (UPC)	Partner	Academia	Spain
Kepler Prodimpex	Partner	Company (IT)	Romania
Yahoo! France	Partner	Company (IT)	France
Finmatica s.p.a.	Partner	Company	Italy
Électricité de France (EdF)	Partner	Company	France

Scope and Objectives

ESIS aimed to provide a new economic indicator that would measure customer satisfaction annually. It was based on customer evaluations of the quality of goods and services that were purchased in Europe and produced by both EU and non-EU companies that had a substantial European market share. Thus ESIS would develop and implement regular measurements of customer satisfaction in Europe supported by an ad hoc software tool and a data warehouse system able to collect and manage the questionnaires involved in the project. The system was intended to deliver index values for individual companies, industries and for sectors as well as the entire economy. ESIS would also calculate a global index and explanatory indices of customer perceived quality, expectations, company image, perceived value and, as the main performance indicator, a loyalty/retention index.

Outline of methodology

The basic structure of activities was:

- User requirements: this would cover users' needs in the different parts of ESIS such as the design of the questionnaires adapted to the new economy; collection and storage of survey data in a data warehouse; writing of the model linking users' answers to the different facets of satisfaction; statistical estimation of the coefficients of the model and satisfaction indexes; datamining of the database; and dissemination of resulting statistical reports;
- Specification, methodology and tools: this would be dedicated to the writing of the technical specification sheets of the principal stages of the project;
- ESIS computational engine would concern the development of the heart of the calculation system;

- ESIS Information Delivery System would seek the development of the software tools of the computational engine needed to create a complete management system for the input data (the questionnaires data warehouse) and the output results (the datamining of the satisfaction indexes in the database);
- Pilot and validation: would include the installation of the pilot of the system, the statistical calculations, validation and tests on the integrated set;
- Dissemination, as elaborated below.

Main results achieved

The main output was the project's final software product called SPAD ESIS. The ESIS developments integrated in this software were the PLS Path modeling developed by SPAD, DMS and KEPLER; the survey grafting program developed by UPC; OLAP Data Mining developed by UPC; and OLAP Graphical processing developed by KEPLER. The specific, detailed outputs were:

- The tasks on end-user requirements were completed. Thus, the technical choice was mapped on to user needs. This consisted of the definition of the user context, user requirements, the scope of pilot applications and benchmarks for all levels of application. Users and experts together defined the principles for the questionnaires, structural equations and the methodology for data treatment;
- Each part of the project was defined to ensure the coherence of the method used in each development phase, to verify that the successive stages were correctly linked together and to define precisely the tools used in each part. This included tackling the problem of the statistical estimation of satisfaction from the data collected through surveys;
- The operational system defining all the necessary steps for the calculation and the validation of a satisfaction index was designed. A set of software tools along with the appropriate interface were developed to do this. Special attention was given to the functional design of the interface and the efficiency of the statistical algorithms;
- The prototype of a tool for managing the inputs/outputs of the ESIS software was developed. Thus, developed also was general access data warehouse of survey data for secondary use and the set-up of a resulting satisfaction index database for datamining and regular information dissemination;
- Users evaluated the results, specifically the computational engine, the software prototype, the user manuals and help facility.

Possible impact on ESS and/or the end-user

ESIS should attract the attention of many stakeholders, such as methodologists, politicians, managers, quality associations, societal groups, the media, researchers, universities and companies. The possible impact is as follows:

- The methodologies used in the construction of satisfaction indices should be of interest to the ESS;
- The methodology behind ESIS should make it possible to link the satisfaction index to economic returns. This should be important if the future economic situation requires connecting company actions to customers' reactions;
- ESIS should benefit consumers by giving voice to their evaluations of the products and services they buy and use. That should stimulate quality improvements;
- ESIS should benefit employment and society in general.

Dissemination and exploitation prospects

ESIS dissemination strategy targeted three groups:

- Private companies in both the traditional and the new economy;
- Public administrations and other semi-public organizations;
- Academics.

There was a project website. The core dissemination instrument was the ESIS User Group, which comprised invited representatives from NSIs and academia from major European countries. The User Group acted through workshops, press conferences and participation in information days.

There were the following specific dissemination activities:

- Seminars and conferences where the scientific achievements of ESIS were presented. The following were the latest conferences mentioned by the consortium:
  - The XLI Scientific Meeting of the Italian Statistical Society , Bari, June, 2004;
  - Meeting of the International Federation of Classification Societies, Chicago, July 2004;
  - XVI Symposium on Computational Statistics, COMPSTAT, Prague, CZ, August 2004;
  - 55th Session of the International Statistical Institute, Sydney, April 2005;
  - 4th International Symposium on PLS and Related Methods – Focus on Marketing, Barcelona, September 2005.

- The (non confidential) results of the project were submitted for publication to scientific journals in the field of official statistics;
- Methodological aspects developed during the project were published in international publications and disseminated through well-known international conferences, such as the "Annual conference of SFDS" (Société Française de Statistique) or the KDD annual International Congress (Knowledge Discovery and Datamining) held in 2002.

Suggested Further Work

The consortium participants were committed to continue their cooperation beyond the project duration in order to develop ESIS fully. The goal was to achieve product maturity; in this respect, a consortium agreement was planned that would formally establish the required collaboration.

Bibliography

All the documentation on the ESIS project could be found in www.esisproject.com.

- SEM state of art
- PLS State of Art
- ESIS methodological specifications
- PLS methodological tutorial
- PLS Satisfaction presentation
- Demonstration of the system functionality
- Result precision Program
- Survey Grafting module
- OLAP DataMining Module
- Others documents about conferences or dissemination activities

EURAREA
Enhancing Small Area Estimation Techniques to Meet European Needs

Timetable	1/1/2001-30/06/2004 (42 months)
Website	http://www.statistics.gov.uk/methods_quality/eurarea

The Consortium

Member	Role	Institutional type	Country
Office for National Statistics (ONS)	Coordinator	NSI	United Kingdom
University of Southampton (U.Soton)	Partner	Academia	United Kingdom
Statistics Finland (StatFi)	Partner	NSI	Finland
University of Jyvaeskylae (JyU)	Assistant partner	Academia	Finland
Instituto Nacional de Estadística (INE)	Partner	NSI	Spain
Statistics Norway (SSB)	Partner	NSI	Norway
Statistics Sweden (SCB)	Partner	NSI	Sweden
University of Economics, Poznan (AE)	Partner	Academia	Poland
Istituto Nazionale di Statistica (ISTAT)	Partner	NSI	Italy
University of Rome III (R3)	Assistant partner	Academia	Italy

Scope and Objectives

The overall objective of the project was to improve small area estimation methods currently used by NSIs.

The first part of the project would consist of assessing the effectiveness of 'standard' small area estimation techniques. By 'standard' techniques was meant the techniques of domain estimation (synthetic estimators, GREGs, and composite estimators) which had entered into use in the United States and Canada in the 1980s and were the subject of steady theoretical refinement since. The focus would be on up-to-date but relatively straight forward linear and logistic versions of these estimators. Seven 'standard' estimators would be implemented:

- Direct Estimator;
- GREG with Standard Linear Model;
- Synthetic Estimator, Model A (Linear Normal model with unit level covariates and model-based estimate of within-area variance);
- Synthetic Estimator, Model B (Linear Normal model with area level covariates and pooled sample estimate of within area variance);
- Synthetic Estimator, Model C (Logistic model with area level covariates);
- EBLUP (Composite / Model A above);
- EBLUP (Composite / Model B above).

The second part of the project would be theoretical innovations to enhance the 'standard' techniques in four major ways which would reflect the requirements and strengths of the ESS, as follows:

- Borrowing strength over time - using time series data;
- Borrowing strength over space - taking account of spatial correlations;
- Investigating the effect of complex sample designs and developing sample design criteria that would be optimal for small area estimation;
- Providing improved estimates of cross-classifications, using a modified version of the SPREE approach.

In order to assess the 'standard' and innovative methods, the project would also provide an extensive external validation of the estimators. The validation exercise would be conducted using real data from various European NSIs and simulation exercises would also be performed. The project would target three variables: household composition, ILO unemployment and income.

A further objective was that the procedures investigated within this project should be easily implemented by third parties. To ensure this, pieces of program code would be written in SAS language, along with the algorithm descriptions, to accompany the theory and results.

OUTLINE OF METHODOLOGY

The main part of the project concentrated on developing the 'standard' small area estimators by investigating four major technical themes.

Theme 1: borrowing strength over time: investigating ways of using survey data from earlier years to increase the precision of estimates made in the current year. The existence of annual surveys (particularly the EU-wide LFS) and the use of panel structures in the EU-sponsored ECHP provided substantial scope for methods of this kind;

Theme 2: borrowing strength over space: this had two aspects: (a) taking advantage of the spatial correlation of residuals; and (b) coping with the Modifiable Area Unit Problem (MAUP). The MAUP refers to the fact that, when domain estimation techniques are applied to spatial units of different sizes, they tend to result in different and inconsistent estimators;

Theme 3: adapting the standard methods for complex sample designs: the standard theory is based on the assumption of simple random sampling but this hardly ever applies in practice. Clustered samples tended to be the norm. There were two sub-themes: (a) choice of estimators and investigation of their precision and robustness; and (b) the choice of sample designs that are optimal for various kinds of small area estimation;

Theme 4: estimating cross-classifications: for policy purposes, it is often important to estimate cross-classifications – for instance the unemployment rate by age and sex, or income by family type – as well as the overall means and proportions. A middle way was needed between (a) domain estimation techniques, which are relatively flexible and have relatively well understood error-properties but only deal with one variable at a time; and (b) calibration-like techniques such as Structure Preserving Estimation (SPREE), which make fuller use of covariate data but which have relatively poorly understood error properties and may constrain estimates too tightly to the structure of the covariate data.

Within each theme the work consisted of a review of the existing literature, theoretical development, software development, setting up the databases and analysis.

Main results achieved

The "Freeware" SAS program for all of the "standard" and enhanced estimators produced in the project together with appropriate documentation were tested and hence completed.

The key substantive findings of EURAREA were:

- The results of the performance of model-based estimation methods in the context of official statistics were unambiguous. At NUTS3 level, model-based estimators achieved comparable or slightly better levels of precision than design-based estimators. At NUTS5 level, model-based estimators substantially outperformed design-based methods. This finding does not extend to the performance of confidence intervals calculated using model-based methods. Though in some instances they performed well, in others they achieved coverage rates substantially below face value. Model misspecification was a potential source of error. Once various problems were taken into account, theoretical expectations regarding the performance of model-based estimators under sampling from fixed populations seemed to be broadly supported;
- "Borrowing strength over time", that is, making use of data from earlier time periods for the area concerned substantially enhanced the precision of estimates for individual small areas. "Borrowing strength over space", that is, allowing for spatial auto-correlation of random area effects was less effective. It was possible that improvements might be achieved with different autocorrelation structures or distance metrics, but a more pronounced gain could come from incorporating time series data. The enhanced log-linear methodology proved effective in estimating change-since-last-census for cross-classified data. A generalized linear structural mixed model achieved the best results in most cases. The associated confidence intervals tended to be underestimated for SPREE and GLSM estimators and were usually too conservative in the case of the GLSMM estimator;
- Concerning the distribution of area means, the standard deviation of the set of estimated area means from a single sample tended either to underestimate or to overestimate the standard deviation of the set of actual area means. For design-based estimates, the standard deviation of the estimates was too high, while model-based estimators, particularly synthetic estimators, produced the opposite effect. Composite estimators within the model-based framework whose standard deviation approximately matched that of the underlying area means were possible;
- Regarding data set-up and sample design, effective model-based estimation required that sample data were matched to area-level covariates with high explanatory power. If possible, unclustered sample designs were also favorable and increased the success of estimation models.

The recommendations were that:

- Eurostat should encourage member states to adopt model-based methods of estimation for area sizes below (possibly including) NUTS3. However, more research would be required into:
    - Methods of estimation that reproduce the underlying distribution of area means;
    - Robust estimation of variance-covariance parameters to improve the coverage properties of confidence intervals and optimize the performance of composite estimators.

Possible impact on ESS and/or the end-user

The results of the validation exercise should enable statisticians from countries all over Europe to make an informed judgment about the practical benefits of adopting the methods investigated in EURAREA; and the rest of the project should provide them with the theory and software needed to apply the chosen methods. In addition, the enhancements made to the standard methods as a result of this project should place a greater range of small area estimators at the disposal of the statistician, which should result in improved estimates using more appropriate estimators.
Dissemination and exploitation prospects

Dissemination has been through the project website, from which all the main project deliverables (including computer programs and manuals) are freely downloadable. Papers were published in journals and presented at various conferences. The main findings of the project were gathered together into a single reference volume designed to serve as a guide to practicing statisticians throughout the EU. The user focus was on NSIs methodologists.

The end-of-project conference, which was held in University of Jyväskylä, Finland, in August 2005 and attended by over 50 participants, included presentations of the project results by the participants and workshops illustrating the use of the software. An important aspect of the conference organization was the establishment of an effective advertising strategy aimed at worldwide government organizations and academic institutions.

All the members of the consortium planned to apply the research results and deliverables of this project directly in the development and enhancement of small area estimates. The consortium expressed the belief that NSIs generally and international institutions would be keen to apply the techniques validated by EURAREA.
Suggested Further Work

EURAREA's recommendations were that:

- Eurostat should encourage member states to adopt model-based methods of estimation for area sizes below (possibly including) NUTS3. However, more research would be required into:
    - Methods of estimation that reproduced the underlying distribution of area means;
    - Robust estimation of variance-covariance parameters to improve the coverage properties of confidence intervals and optimize the performance of composite estimators.

A number of issues affecting the performance of small area estimation went beyond the research remit of EURAREA and are therefore candidates for further research. These included:

- Impact of non response;
- Methods of covariate selection and alternative approaches to model building;
- Estimation of within-area distributions of continuous variables;
- Ensuring consistency of estimates at different geographic levels;
- Estimating local change over time.

BIBLIOGRAPHY

1. Heady, P, Hennell, S (2001) Enhancing small area estimation techniques to meet European needs, Statistics in Transition, Vol. 5, No. 2.
2. Kordos J and Paradysz J, (2000) Some Experiments in Small Area Estimation in Poland, Statistics in Transition, Vol. 4, Number 4, P963-977.

Main Publications from Project

- EURAREA Consortium, 2004, Enhancing Small Area Estimation Techniques to meet European Needs: Project Reference Volumes 1-3, http://www.statistics.gov.uk/eurarea/download.asp
- Zhang, L. and Chambers, R.L., 2004, Small area estimates for cross-classifications, Journal of the Royal Statistical Society, Series B, Vol. 66, Issue 2, p. 479 – 496.
- Dehnel, G., Golata, E. and Klimanek, T., 2004, Consideration on optimal sample design for small area estimation, Statistics in Transition, Vol. 6, No. 5, p. 725 – 754.
- Heady, P.J. and Ralphs, M.P., 2005, EURAREA : An overview of the project and its findings, Statistics in Transition, Vol. 7, No. 3, p. 557-570.

EUREDIT
The Development and Evaluation of New Methods for Editing and Imputation

Timetable	1/3/2000 - 28/2/2003 (36 months)
Website	http://www.cs.york.ac.uk/euredit/

The Consortium

Member	Role	Institutional type	Country
Office for National Statistics (ONS)	Coordinator	NSI	United Kingdom
Royal Holloway College	Partner	Academia	United Kingdom
University of Southampton (U. Soton)	Partner	Academia	United Kingdom
University of York	Partner	Academia	United Kingdom
NAG Ltd	Partner	Company	United Kingdom
Statistics Netherlands (CBS)	Partner	NSI	Netherlands
Statistics Finland (StatFi)	Partner	NSI	Finland
University of Jyväskylä	Partner	Academia	Finland
Swiss Federal Statistical Office (SFSO)	Partner	NSI	Switzerland
Qantaris	Partner	Company (IT)	Germany
Istituto Nazionale di Statistica (ISTAT)	Partner	NSI	Italy
Statistics Danmark	Partner	NSI	Denmark

Scope and Objectives

The main objective of the EUREDIT project was to investigate and evaluate methods for automatic editing and imputation. Specifically, EUREDIT aimed:

- To establish a standard collection of datasets for the project;
- To construct a methodological evaluation framework and develop evaluation criteria;
- To establish a baseline by evaluating currently used methods;
- To develop and evaluate a selected range of new techniques;
- To evaluate different methods and establish best methods for different data types;
- To disseminate the best methods via a single computer package and publications.

Outline of methodology

The EUREDIT basic technical project structure was:

- The provision of a standard collection of datasets of different types, presented both as “clean” data and with a broad range of error types, for the purpose of evaluating the selected methods;

- The definition of quality and evaluation criteria by which to judge each technique, and the provision of a methodological framework within which the evaluation might take place;
- The adaptation and application of a diverse range of new methods (multi-layer perceptron, correlation matrix memory, self-organizing maps, support vector machines) to data editing and imputation;
- The development of new statistical techniques for multivariate edit and imputation based on the application of outlier robust methodology to detection and modification of representative outliers in survey data;
- The investigation of editing techniques that could handle mixed data types;
- Development of non-parametric regression techniques for edit and imputation, particularly in the context of temporal (panel) data series;
- An overall comparison of all methods evaluated in EUREDIT identifying the weaknesses and strengths of each, with particular reference to error attributes;
- The development of an overall framework which identified recommended strategies for data editing and imputation according to known or expected error attributes of the data set in question.

Almost all NSIs employ edit and imputation techniques, as do those involved in academic and business research. However, these methods are typically based on simple statistical ideas and little is known about the comparative performance of various methods in given situations. The EUREDIT project combined recent developments in statistical and computer science to develop and evaluate novel edit and imputation methodologies. This should provide a significant enhancement to the methods used by NSIs, commercial and academic analysts. In particular, EUREDIT focused on the use of new and innovative technologies based on statistical, neural network and related methods for edit and imputation in large-scale statistical data sets. The project has developed and evaluated edit and imputation applications for a number of state-of-the-art neural architectures, including multilayer perceptron, correlation matrix memory, self-organizing map and support vector machines. It has also developed non-parametric regression-based edit and imputation applications.

In order to avoid excessive “tuning” of methods to particular situations, EUREDIT developed a new standard suite of trial datasets and generic statistical evaluation criteria that were used across all work packages in the project.

Main results achieved

In broad terms, the main outputs of EUREDIT were:

- A CD-ROM containing 6 datasets and the associated data dictionary was produced and distributed;
- Operational evaluation criteria were developed and agreed. An ‘Evaluation Handbook' was also developed;
- Evaluation software, developed by NAG Ltd to implement the statistical criteria, was tested and distributed to all partners. CBS developed software for automatic error localisation and it completed research on the use of regression trees for selective editing. The actual evaluation of the methods was completed by the partners;
- The following were developed and evaluated:
    - New methods for error localisation based on multi-layer perceptron (MLP) type neural networks;
    - New methods for data editing and imputation based on CMM (Correlation Matrix Memory) neural networks;
    - New methods for error localisation based on self-organising maps (SOM's) type neural networks;
    - Currently used methods for data imputation to establish best practice methods;
    - New methods for imputation based on MLP type neural networks;
    - New imputation methods based on SOMs;
    - New methods for data imputation based on support vector machine (SVM) techniques;
    - New imputation methods for panel and time series data;
- The evaluation and validation of these methods to determine the "best" methods in different circumstances;

- The integration of the individual edit and imputation methods into a prototype software system. NAG completed the final system that included the multivariate normal EM, BACON-EM, DIS and WAID algorithms, as well as utility functions.

Possible Impact on ESS and/or the End-user

Almost all NSIs as well as Eurostat have to use editing and imputation in their work, and it is not surprising that the consortium included 6 NSIs. Thus, the findings and software of EUREDIT should be indispensable to them. The project has provided a significant advance in the quality of the methodologies and techniques that underpin data gathering and analysis. In turn, this should help to improve the quality of decision-making based on socio-economic research data at all levels and ultimately better services for the European citizen.

Dissemination and exploitation prospects

There was a project website. NAG had produced a poster and a flyer on EUREDIT that was displayed at Compstat 2000 and RSS 2000 conferences. The technical report on measuring the quality of edit/imputation procedures was published in the ONS Methodology Series. The results of EUREDIT research were presented in a number of conferences, in particular a conference organized by EUREDIT (Data-Clean) in May 2002 immediately after the UN/ECE Workshop on Statistical Data Editing. Papers from the Data-Clean conference were to be published in a special edition of the Journal of the Royal Statistical Society, series A. Partners have also presented their work at various seminars, such as a half-day meeting about EUREDIT that took place at the Royal Statistical Society where John Charlton, Ray Chambers, and Pasi Koikkalainen gave presentations.

The commercial exploitation of the EUREDIT project software was dependent on the existence of a demonstrable market opportunity and competitive advantage. The approach considered was NAG Data Mining Components that was launched in November 2001. Release 2 of the Components contained many of the underlying algorithms used in the methods considered in EUREDIT.

Suggested Further Work

EUREDIT extensive market research had indicated that data cleaning and mining were quickly becoming an essential functionality for applications in bio-informatics, CRM, e-business, financial services, fraud detection, web analytics and many other areas. Furthermore, there was a commercially viable market opportunity for good quality data cleaning and data mining components that could be integrated into existing or new applications. NAG was well placed to pursue these opportunities.

Bibliography

- Beguin, C. (2004) “Multivariate outlier detection in incomplete survey data: the epidemic algorithm and transformed rank correlations” Journal of the Royal Statistical Society Series A, 167 Part 2, pp.275-294.
- Charlton, J. (2004) “Editorial: Evaluating automatic edit and imputation methods, and the EUREDIT project” Journal of the Royal Statistical Society Series A, 167 Part 2, pp.199-207.
- Chambers R. et al (2004) “Robust automatic methods for outlier and error detection” Journal of the Royal Statistical Society Series A, 167 Part 2, pp.323-339
- Eurostat (2004) “Methods and Experimental Results from the EUREDIT Project” Deliverable 6.1 EUREDIT Project IST-1999-10226
- Eurostat (2004) “Towards Effective Statistical Editing and Imputation Strategies - Findings of the EUREDIT Project” EUREDIT Project Deliverable D6.2, IST-1999-10226
- Manzari, A. (2004) “Combining editing and imputation methods: an experimental application on population census data” Journal of the Royal Statistical Society Series A, 167 Part 2, pp.295-307.

EUROKY-PIA
Developing European Knowledge for Policy Impact Analysis

Timetable	1/1/2003-30/4/2005 (28 months)
Website	http://www.informer.gr/eurokypia/index.htm

The Consortium

Member	Role	Institutional type	Country
Istituto Nazionale di Statistica (ISTAT)	Coordinator	NSI	Italy
Board of Inland Revenue	Partner	Public sector	United Kingdom
CEIS – University of Rome "Tor Vergata"	Partner	Academia	Italy
Informer S.A	Partner	Company (IT)	Greece
Global Insight	Partner	Company	Italy
Mantos Consulting	Partner	Company	United Kingdom

Scope and Objectives

The original objectives of this project were the following: (i) serving the Lisbon goals and gathering a critical research mass of centres of excellence around a model framework research environment interlinking policy areas, skills, officials, administrators and academia; (ii) supporting EU Governance according to the five principles of openness, participation, accountability, effectiveness and coherence; (iii) developing needed PIA (Policy Impact Analysis) tools and addressing key EU and IST 2002 Work Programme concerns, by bringing together technology developments and EU policy areas; (iv) developing a 'vision' and a 'roadmap' and prepare the ground for a future FP6 project. Preparing a project for FP6 was dropped from the original objectives as a result of changed circumstances.

Outline of methodology

The methodology revolved around the four groups of technical expertise which participated in the project, viz. NSIs, Academia, Research Institutions, and IT and consulting firms. Correspondingly:

- The first component related to the information that was needed to support the Lisbon Objectives and New Governance in the social, business cycle and economic-structural/market areas, which were carried out by NSIs;
- The second component concerned the support that could come from New Technologies. The IT group in the network took responsibility for this activity. The focus was on how technology could help not in doing the same but in doing it faster;
- The third component bore on the broader scientific, analytical and general aspects of policy and socio-economic analysis. The perspective was on the soundness and strength of the analytical tools that had been used for PIA. This group would study how changes in the new economy had influenced the concept, design and role of indicators. Universities in the network took responsibility for this activity;
- The fourth component envisaged the use of existing tools and methods in an applied perspective focused on what needed to be done to support policy making. Under this activity, the members in the Research group of the network would explore existing leading-edge tools and methods and how they could be developed in the medium term. Opportunities would be charted, including the obstacles (which might be associated with technology or data) that might exist and might hinder their development and EU-wide use.

MAIN RESULTS ACHIEVED

In a nutshell, the project confirmed:

- That neither the EU nor its members, generally, have acquired a sufficient stock of PIA experience and knowledge; evidence-based policy in the EU was definitely underdeveloped;
- The value of PIA; PIA could make a difference;
- That PIA was required for "best" regulatory, distributive, environmental, sector and territorial policies;
- The importance of and need to invest in PIA data and tools;
- The increased sophistication of PIA;
- That better policies do not just happen but need to be designed availing of PIA knowledge (data, methods and tools);
- The importance of implanting PIA capacity in Governments;
- The importance of a networked, synergistic approach involving PIAs at both the EU and national levels;
- That the key PIA input was integrated and systematized information systems (ISIS);
- That the key PIA tools were:
    - IT ways to access ISIS for research purposes;
    - Microsimulation models;
- That the key PIA outputs were:
    - Counterfactuals;
    - Multivariate composite and decomposable indicators that satisfied certain properties.

In more detail, the results of questionnaire responses confirmed a priori expectations that there was little development of PIA and of its underlying tools. They provided some broad indications on broad areas for improvement, as follows:

a) All respondents (including those in the US) agreed that there was still substantial room for improvement in the usage of administrative data;

b) In half of the cases (excluding the US), there was no feed-back from evaluation studies into data production;

c) Linkage of microdata was in most cases deemed possible but only in a few cases this was carried out in a major way;

d) Access to microdata was restricted in most cases, mainly on grounds of privacy and confidentiality.

A number of respondents stated that there was a need to include the dynamic dimension (demography) into the more static (structural) dimension.

Concerning the gaps in information, Euroky-Pia identified the needs by working on a set of themes as follows:

- First, the general framework for PIA information, as was deduced from the experience of the UK ONS;
- Secondly, the best way to track the new economy and the information needed for that purpose;
- Thirdly, sustainable growth and ways to monitor its development in a multi-dimensional and multi-level framework.

The project drew the following general conclusions from its assessment of the differences between the EU and the USA concerning the application of modern ICT to PIA:

- There were no significant differences in available technology;
- The current technology was well-known in both the EU and in the USA;
- EU started research on PIA earlier than the USA;
- The USA implemented current state-of-the-art ICT as soon as is was available;
- EU dealt with the changes in technology by continuing research projects;
- Most of the results of the EU research projects have not been carried through to production.

The project found that reducing the gap between EU countries and the USA required (i) investing in research; (ii) developing and applying tools and methods of policy analysis; and (iii) expanding access to administrative and survey microdata. There were important areas where EU countries lagged behind the USA. These areas concerned the difficulties encountered in accessing, merging and systematizing large microdata bases.

The remedies where EU lags behind USA in implementing the results of ICT research included tighter and more respectful cooperation between IT-groups and the NSI- groups in the EU. Complex IT solutions should be designed by IT-groups, leaving NSIs to define the requirements. Multi-disciplinary task forces could play a major role in the implementation of the necessary ICT infrastructure needed for an effective PIA.

There were many areas of PIA which required quantitative tools. The need appeared more urgent in the following:

a) Capital and labour markets, productivity and competitiveness both of individual industries and of the industrial system as a whole;
b) Relatively new policy issues, such as the shift from national to a regional focus and the environment;
c) New policy needs, such as more timely responses to economic and monetary developments, as well as the need to incorporate the impact of micro policies into the assessment.

EUROKY-PIA suggested a list of questions for investigation to fill specific gaps in existing macro-economic models:

a) How to structure the formal links between micro-models and macro-models?
b) What is the right level of geographical aggregation: regional, country, Eurozone or EU as a whole?
c) What are the data issues relative to each of the geographical aggregations?
d) Should an inter-industry approach be considered?
e) What time horizon should be considered as optimal, taking account of the increasing interdependence between monetary/financial conditions and consumer behaviour?
f) How can the impact of environmental policy be fed through to a macro-model?.

In order to answer these questions, EUROKY-PIA presented a comparative assessment of twelve multi-country macro-econometric models.

Possible impact on ESS and/or the end-user

Generally speaking, NSIs do not themselves undertake PIAs. That tends to be left to individual policy Ministries. However, NSIs have a critical role to play in making PIAs possible.

All respondents to the questionnaires studied in Euroky-Pia agreed on the desirability of a stronger effort by NSIs on both the generation and the coordination of data series in support of policies, including statistical activities to improve the quality and relevance of data for policy analysis. Data quality evaluation should be more structured to cover all stages of data collection and processing, and the suggestions and the needs of users should be more extensively taken into account in the design of the evaluation process. A majority of respondents thought that coordination should be assigned to NSIs, while others declared that other Agencies also should be involved.

Dissemination and exploitation prospects

As an Accompanying Measure, EUROKY-PIA was essentially a dissemination project. There were a web site, a project presentation leaflet, two workshops to calibrate progress, two conferences, publication of scientific papers and the publication of a final report.

Suggested Further Work

The description of the project's findings in the above section on Main results achieved contains many suggestions concerning the way forward in establishing PIA as a routine supporting activity in NSIs and as an activity in its own right in Government generally. The central finding that PIA awareness, PIA modeling tools and microdata support need to be greatly enhanced at all levels of Government cannot be over-emphasized. Calls for proposals for research on, inter alia, policy impact indicators in FP7, Cooperation, Theme 8 on Socio-economic Sciences and Humanities is a welcome recognition of the importance of PIA.

FLASH
Flash Estimates of Quarterly National Accounts Main Aggregates

Timetable	1/1/2001-30/6/2003 (30 months)
Website	http://eemc.jrc.ec.europa.eu/softwareFLASH.htm

The Consortium

Member	Role	Institutional type	Country
National Institute of Economic and Social Research (NIESR)	Coordinator	Independent	United Kingdom
Banque Nationale de Belgique	Partner	Central bank	Belgium
Deutsches Institut für Wirtschaftsforschung (DIW)	Partner	Independent	Germany
Istituto Nazionale di Statistica (ISTAT)	Partner	NSI	Italy
Joint Research Centre (JRC, Ispra), European Commission	Partner	Public sector (research)	Italy

Scope and Objectives

Quarterly National Accounts (QNA) are an integral part of the national accounts system. Their increasing role in economic analysis and in the evaluation of economic policy measures justifies the growing interest of the ESS in them. For monetary institutions such as the European Central Bank (ECB) and for short-term analysts or policy-makers, the earlier the main economic indicators are available the better. The indicators of most interest are the Gross Domestic Product (GDP) and its main components.

Flash estimates in the system of QNA are at present released by only few NSIs, with far too long time-lags. The objective of the FLASH project is to produce prompt (and it is hoped reliable) estimates of the main QNA constant price aggregates for the European Economic and Monetary Union (EEMU), for the EU and for member states. The target delay is 40-45 days after the end of the reference period. The project would supply a coherent system able to help in short-term economic analysis and monetary policy decisions, while the shortcomings of the delay in the availability of the official quarterly figures would be avoided. The specific aims of FLASH are:

- To analyze and assess the main relevant sources of information available for each country of the EU;
- To develop methodologies for flash estimates of volume aggregates;
- To implement and assess the overall scheme in computer programs compiled in C++;
- To apply flash estimates methodologies to the EU-15 data;
- To compare the aggregate with the total derived by adding up separate countries using both EEMU-11 and EU-15 data;
- To disseminate user-friendly procedures for flash estimates to all European countries.

Outline of methodology

There are two methodologies in hand. Both of them are statistical. They rely on the use of related data to indicate movements to accounting aggregates rather than requiring the collection of additional data. The first method is regression-based. Accounting aggregates (such as GDP and personal consumption) are related to indicator variables using regression equations.

Appropriate indicator variables are either quarterly variables published ahead of accounting aggregates (and within 45 days of the end of the quarter) or monthly variables for which two months' data were published and the third month can be forecast using statistical forecasting methods.

The second method involves the use of principal components. The principal components are extracted from the covariance matrix of current and lagged indicator variables. Regression methods are then used to explore the ability of the leading components to explain movements in the aggregate of interest. The combination of the regression coefficients and the factor loadings then provides weights linking current period indicator movements to movements in the aggregate of interest.

In both cases the methods were tested out of sample. The results were compared with each other and with those of an autoregressive forecasting model.

Models were estimated both for individual countries and for EMU-12 and EU-15 aggregates. Early country-specific as well as area-wide information was used in the estimation of the figures for the EMU-12 and EU-15 aggregates. It was intended that the out of sample performance of area-wide figures calculated by aggregating country results should be compared with that obtained by direct modeling.

Main results achieved

A survey paper was produced setting out the statistical theory behind the production of FLASH estimates.

Software for the production of FLASH estimates was produced. The program allowed users to input both monthly and quarterly variables. It assessed the statistical properties of each individual series and allowed users to study relations between pairs of series. It then produced flash estimates of quarterly aggregates based on appropriate estimating methods. Traditional econometric models remained one important means of producing forecasts, though recently non-structural models were preferred. Regression techniques were used to characterize the historical relationship between the variable to be forecasted, often quarterly constant price GDP, and a set of indicator variables. Since the indicator variables were available more promptly than GDP, unknown GDP could be forecasted from the regression equation. Little generalization was possible about which were the good indicators and, in any case, each indicator explained only a relatively small part of the variation in GDP or consumption growth in each country. Indicators of the monthly variables, industrial production and retail sales generally had less explanatory power than those of the quarterly variables. There was a general problem in that there were often a large number of potential indicators which could not simultaneously be used in a regression equation because there were too few degrees of freedom.

Two different ways of using indicator variables were explored. The first, regression method, relied on variable selection on the basis of contemporaneous correlation, with only the most correlated variables being used. The second method used factor extraction to reduce a larger number of indicator variables to a small number of factors which could be used in regression equations. Most data series were relatively short, which made it difficult both to have satisfactory data periods to which to fit models and to have sufficient data to provide a good indication of out-of-sample performance. These methods were applied to the Euro Area and the EU as single entities and also to those member states which produced regular quarterly national accounts: Austria, Belgium, Denmark, France, Finland, Germany, Italy, the Netherlands, Spain, Sweden and the United Kingdom. Brief reports were produced on the statistical situation in the remaining four countries: Greece, Luxembourg, Portugal and the Irish Republic.

An important part of the approach was that estimates calculated using indicator variables were compared with those generated using autoregressive forecasting. The main practical conclusion reached was that it was possible to produce flash estimates of GDP growth for the Euro Area with a RMSE of around 0.1%. The consortium took the view that an error of this magnitude was likely to be acceptable to data users. The main reason for this good performance was that early GDP estimates were available for Germany and Italy, which together accounted for about 50% of the Euro Area. A flash estimate could also be produced for the EU with similar reliability, using the early estimate for the United Kingdom in addition to those for Germany and Italy. FLASH addressed in some detail the performance of the FLASH indicators for the pan-European aggregates with those derived by aggregating the indicators for individual members. The consortium found that the factor models generally did not perform well.

When individual countries were considered, the indicator models performed better than the autoregressive ones. The RMSEs were, however, considerably larger than 0.1%.

Possible impact on ESS and/or the end-user

Timeliness is a key quality objective which the ESS must observe. FLASH should help official statisticians and economists to monitor the evolution of the EU national economies and of the EU economy as a whole. With more timely information, policy-makers should be in a better position to make informed choices for short-term macro-economic policy making.

More specifically, FLASH did make advanced statistical techniques available to key players in the analysis of the EU economy, such as Eurostat, NSIs and the central banks, including ECB. It has offered a coherent and harmonised approach to the analysis of trends of the EU economy that is crucial to the functioning of the Union.

FLASH has contributed to the design at EU level of a uniform methodology for providing prompt estimates of GDP that could be adopted by NSIs and, in due course, become a standard for the European Member States.

Dissemination and exploitation prospects

A web page for FLASH is maintained. This web page, on the JRC-ISIS site, can be reached at http://eemc.jrc.ec.europa.eu/softwareFLASH.htm.

An interest group was set up consisting of statisticians from NSIs interested in early estimates. An advisory board capable of advising the consortium and reviewing its work was also established.

Another action was direct support to the dissemination of the software FLASH, as follows:

- A help-desk on Web, within the site "Time Series Analysis for Official Statisticians";
- Implementation on the premises of a list of test-users, which received prioritized support.

A further line of dissemination included:

- Presentation to conferences and special events, for example, the ETK-NTTK and the AMRADS conference held in Crete in June 2001 and the AMRADS workshop held in Greece in May 2002;
- Articles on specific theoretical issues which emerged in specialized literature;
- Use of web technologies, as the web space and help desk already mentioned;
- A logo for the FLASH project.

The FLASH consortium was very active with users. For example, staff from Eurostat and the Director of Research of the ECB attended the meeting of the group in September 2001 and participated in discussions about progress. Again, the consortium took steps to ensure the exploitation of results obtained by the FLASH project for accession countries. Thus, the JRC-IPSC had planned to convene a workshop to be held in Ispra (Italy) where there would be about 50 participants with 15 practitioners invited from Candidate Countries.

Suggested Further Work

The FLASH software has resulted in a tool which allowed users to build on, and improve, the work accomplished in the project as new data sources became available or as better access is gained to data which FLASH was not able to exploit. The project itself showed that indicator variables had important roles to play in the construction of early estimates of national accounting aggregates. Since the work was completed, EUROSTAT has focused on encouraging member states to accelerate their data production and the software is available to help them in this. There is further work to be done on methods of short-term forecasting, and the project has developed some methods which may be of use. However, the experience of the project was that appropriate methods and indicators are country-specific. While general techniques can be applied, it is less clear that general models are very useful.

Bibliography

- Camba-Mendez, G., G. Kapetanios, R.J. Smith and M.R. Weale. (2003). "Tests of Rank in Reduced Rank Regression Models". Journal of Business and Economic Statistics. Vol 21. pp 145-155.
- Blake, A.P and G. Kapetanios. (2003) A radial basis function artificial neural network test for neglected nonlinearity. Econometrics Journal Vol 6 pp 357-373.

INSPECTOR
Quality in the Statistical Information Life-Cycle: A Distributed System for Data Validation

Timetable	1/7/2001-31/12/2003 (30 months)
Website	http://www.agilis-sa.gr/projects/inspector/

The Consortium

Member	Role	Institutional type	Country
Intrasoft International S.A.	Coordinator	Company (IT)	Belgium/Luxembourg
Universita degli Studi di Napoli Federico II (DMS)	Partner	Academia	Italy
University of Vienna	Partner	Academia	Austria
Instituto Nacional de Estatística (INE)	Partner	NSI	Portugal
National Statistical Service of Greece (NSSG)	Partner	NSI	Greece
Agilis S.A.	Partner	Company	Greece

Scope and Objectives

The main objective of the project was the design and development of a generic, distributed and flexible data validation system, which would be able to be seamlessly integrated in the current (or future) processes of statistical data collection, in order to ensure and monitor data quality. The system would be accessible in a distributed way in order to validate large statistical data sets before their transmission or even throughout their production, ensuring homogeneity and consistency which are critical quality parameters of the validation process. Among the critical project objectives were: the development of a formal framework for the classification and semantic notation of validation rules; the design, development and population of the rules repository; the development of the validation engine and a Java application implemented; and Validation Client.

Outline of methodology

INSPECTOR'S approach to automated Data Validation was based on a declarative concept: rules are in fact information metadata and, if declared upon a semantically rich underlying data structure, can be stored in a suitable metadata inventory, such as a relational database. A suitable validation engine can then reconstruct the rules out of metadata and apply them to actual data. Such an inventory can be distributed and be broadly accessible, thus ensuring homogeneity in the validation process.
In order to implement such a metadata repository-based data validation system, the following actions would be required:

1) Develop a sound abstract formalization (i.e. classification and notation) of data validation rules, which would uncover the logical structure and methodological patterns underlying the seemingly ad hoc nature of the applied validation processes. In order to group the validation rules into more comprehensive and meaningful categories, the targets of the validation process had to be identified. These were the data elements (variables), which were elementary data building blocks and had values, the entities, which were groups of related logical elements that possessed instances and the data schema, which was a group of related entities.

2) Based on this classification and notation, lay out an object-oriented (or extended ER) data model that would be implemented on a relational rules repository in a straightforward way. The model was characterized by a hierarchy of data objects, namely (from top to bottom): data schema consisting of different data sets, which in turn included different data record types. Each data record was an ordered vector of several generic data elements. Recursive Data Element Hierarchies allowed the definitions of unlimited-level hierarchies of simple and composite data

elements. This way, the structure of complex data sets could be semantically declared in the form of metadata. A similar hierarchical structure existed for objects that represented domains of definitions, which were independent objects.

3) Implement a set of application modules (the validation engine) that would be able to reconstruct the (procedural) validation logic out of the rules declarations at run time as well as the necessary client. The system architecture would consist of (a) a distributed (and replicated) repository of validation rules, (b) a validation engine, implemented as a set of platform independent Java modules and (c) a user client, using a local database as a repository for transient data sets under inspection and including various modules.

Main results achieved

The actual outcome of the project consisted of an innovative domain-centered declarative validation process, and a generic, distributed and flexible data validation software system, which implemented the innovative data validation process. The INSPECTOR software system:

- Seamlessly integrates in the current & future processes of data collection;
- Provides consistent and harmonized metrics for data quality monitoring;
- Improves data quality evaluation, by providing the means for quantitative evaluation of data accuracy;
- Provides a formal framework (template) for the definition of data set structure and expression of validation rules semantics;
- Optimizes the data validation process by enabling: (a) harmonization of the process, (b) efficient formulation and tuning of data validation rules and (c) definition and computation of a wide range of data quality metrics.

Possible impact on ESS and/or the end-user

The consortium stated that the primary target group for the exploitation of INSPECTOR results consisted of the ESS. A secondary group included all Public Information Systems (PISs), i.e. governmental and other authorities involved in the exchange of statistical information, such as central banks, social security, health and local administrations. A third extended target group included private sector companies operating data warehouses, collecting and exchanging large volumes of information such as in the financial and insurance industry.

At the stage when the prototype product was being developed, it was mainly the primary target group, the ESS, whose specific requirements were taken into account and who would be the focus of the project's exploitation activities. However, on the whole, the objective of the exploitation plan was to be able to initiate specific application / implementation projects in NSIs immediately after the end of INSPECTOR. These projects would enable further development and would allow subsequent market penetration to the other two target groups, namely PISs and Private Sector systems.

Dissemination and exploitation prospects

The project website was established. The core dissemination instrument was the User Group consisting of representatives from NSIs, academia and selected other disciplines from major European countries. The intention was that the User Group would extend beyond the project duration. Other dissemination activities were:

- Direct mailing: bulletins on the project's aims, progress and outcomes were sent to all members of the e-mailing list;
- Design and dissemination of INSPECTOR flyers and/or brochures; press releases;
- Project clustering;

- Scientific publications such as in the "Research in Official Statistics" (ROS), the "Journal of Official Statistics (JOS). A paper presented in the Q2001 conference was submitted to JOS and the same was planned for a paper presented at the NTTS-ETK conference in 2001. Further publications were planned.
- Workshops / information days and conferences, such as the following international conferences:
    - A conference on Data Clean organized by Statistics Finland in May 2002;
    - UN/ECE Work Session on Statistical Data Editing of the Conference of European Statisticians of the UN, Finland, May 2002 and October 2003;
    - MCS 2002, International Workshop on Multiple Classifier Systems, Cagliari, Italy, June 2002;
    - IWSM 2002, International Workshop on Statistical Modeling, Crete, Greece, July 2002;
    - IFCS 2002, 8th Conference of the International Federation of Classification Societies, Cracow, Poland, July 2002;
    - 26th Annual Conference of the Gesellschaft für Klassifikation, University of Mannheim, Germany, July 2002;
    - The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Alberta, Canada, July 2002;
    - Compstat 2002, International Conference on computational Statistics, Berlin, Germany, August 2002;
    - The International Conference on Improving Surveys, Copenhagen, Denmark, August 2002;
    - IAOS 2002, 'Official Statistics and the New Economy', UK Office for National Statistics, London, August 2002.

Suggested Further Work

There has been a clear potential for offering value-added services based on the project's results. These included:

- Consulting and expert services to NSIs in the field of data validation and data quality in general, ranging from training of NSI personnel and support for the implementation of best practices to the development of customized methodological solutions for specific data validation problems;
- Novel data validation methodology developed within INSPECTOR could be "packaged" and marketed as an autonomous "soft" product, possibly independent from the software product;
- The software architecture of INSPECTOR would allow the offer of data validation services via the Internet, following the ASP (Application Service Provider) paradigm.

Further research possibilities and challenges

Data intensive domains, other than statistics, also involve 'data validation'. For example, Fraud detection in health insurance claim forms or in customs declaration can be achieved with the application of rules very similar to data validation rules. The difference is that, unlike official statistics, rules need to evolve continuously and rapidly in response to new forms of fraud. INSPECTOR explored an approach where historical data are analyzed with data mining algorithms in order to obtain associations between variables which can be transformed into validation rules. This approach offers promising prospects for application in fraud detection.

Particular future research challenges are:

- The development of efficient methods for detecting changing patterns, which is beyond the scope of manually crafted rules based on previously detected fraud cases;
- The specification of statistical tests to assess the ability of detected patterns efficiently to distinguish between fraud and acceptable behavior; if the ability is high the patterns concerned are transformed into rules.

BIBLIOGRAPHY

- Petrakos, G., Conversano, C., Farmakis, G., Mola, F., Siciliano, R. and Stavropoulos, P. (2004) New ways of specifying data edits. Journal of the Royal Statistical Society (series A), 167, 2, 249-274.
- Farmakis, G., Figueiredo, J., Mota, D., Petrakos, G., Santos, D. and Stavropoulos, P. (2004) An alternative approach for the implementation of data editing: the INSPECTOR approach. Proceedings of the European Conference on Quality and Methodology in Official Statistics. Mainz, May 24-26.
- Conversano, C. and Siciliano, R. (2003) Incremental Tree-Based Missing Data Imputation with Lexicographic Ordering. Computing Science and Statistics, 35.
- Petrakos, G., Farmakis, G. and Stavropoulos, P. (2003) Data validation through measurements on conditional distributions of a logically related group of variables. United Nations Statistical Commission and Economic Commission for Europe, Work session on statistical data editing, working paper n. 20. Madrid, October 20-22.

IPIS
Integration of Public Information Systems and Statistical Services

Timetable	1/2/2000-31/1/2003 (36 months)
Website	http://www.instore.gr/ipis/

The Consortium

Member	Role	Institutional type	Country
Quality and Reliability S.A. (Q&R)	Financial Coordinator	Company (IT)	Greece
University of Athens (UoA)	Scientific Coordinator	Academia	Greece
Centre d'Etudes de Populations, de Pauvreté et de Politique Socioéconomiques (CEPS/INSTEAD)	Partner	Research Center	Luxembourg
General Secretariat for Information Systems, Ministry of Finance (GSIS)	Partner	Public sector	Greece
Universite des Sciences et Technologies de Lille (USTL)	Partner	Academia	France
Organization for Vocational Education and Training (OEEK)	Partner	Public sector	Greece
Centro de Formação Profissional para o Comércio e Afins (CECOA)	Partner	Public sector	Portugal
National Statistical Service of Greece (NSSG)	Partner	NSI	Greece

Scope and Objectives

The general objective of the IPIS project was to develop new tools and services to enable Public Administrations to design, organize, develop and disseminate Public Information Systems (PIS) in a pre-harmonized and standardized way. The IPIS system addressed the urgent needs of Public Administrations for reliable, timely and accurate statistics, aiming also in harmonization of incompatible sources' merging. Consequently, IPIS would help to lighten the burden on data providers, reduce costs and enhance the competitiveness of their productive systems in the global information society. The specific objectives of the IPIS system were the following:

- To integrate statistics produced by different data providers through Europe in order to enhance data accessibility in a homogeneous way;
- To harmonize data and related metadata from various sources, produced in different time periods, in order to enhance statistical data comparability;
- To widen available statistical information through the use of existing public administration archives;
- To assist policy makers in the preparation and monitoring of policy plans;
- To provide experts and policy makers with new facilities for widening the exploitation of existing statistical information.

OUTLINE OF METHODOLOGY

The projects objectives were fulfilled through the division of work in the following technical axes (milestones):

a) Identification of Users Needs/User Requirements Models in three main areas namely: Vocational Education and Training (VET), Cross-Border Statistics (CBS) and Household Budget Surveys (HBS) consisting of a collection of standards, indicators and methodologies with the use of case studies from IPIS partners

b) Harmonisation procedures and development of related transformations;

c) Development of the IPIS statistical metadata model integrating these transformations and quality issues;

d) Architectural design of the IPIS prototype;

e) Development of the IPIS system; continuous assessment and evaluation by the IPIS user community and feedback;

f) Continuous dissemination and exploitation of the results of the project.

Some limited details on these milestones are provided:

a) Market Research in the areas of VET, CBS and HBS:

Classification of user needs.

For each policy area user categories and their needs were identified. In general, the user categories consisted of policy makers, end-users and experts.

Identification of Users Needs/User Requirements Models:

The Decision Making Process was considered as a control process acting through specific control variables to a given input flow in order to transform the characteristics of the output flow.

A specialised methodology was developed using a systemic approach to identify the functions and tasks of the policy making process and then testing them in the project's application areas. Specific Requirements models according to functions and tasks were developed for each of the three policy areas, namely VET, CBS and HBS, based on the variables considered by the corresponding responsible IPIS partner.

Proposal of a User Requirements model for the Software System.

Then, based on the common variables and classifications schema, for each function and task, a common user requirements model was selected and sets of relevant indicators were identified to assist decision-making and policy monitoring.

b) Harmonisation procedures and development of related transformations

The IPIS system was designed to be able to handle data from different sources using different frames for collecting, compiling and releasing data and, where possible, to alleviate the effects of breaks in time series that appeared between sources or within the same source. These functions were referred to as transformations. Two main categories of transformations were examined: i) Mapping transformations (alleviate breaks in time series due to errors in classifications conversion) and ii) Methodology-correcting transformations (correct errors due to methodological inconsistencies and restore resulted breaks in time series).

c) Metadata Model Development

The IPIS statistical metadata model includes Semantic Metadata, Logistic, Documentation and Process Metadata and followed the concept of OECD's list for Main Economic Indicators. An object-oriented paradigm was adopted as a designing methodology. In general, the classes of the IPIS metadata model hold information on statistical populations, survey variables, indicators, classifications and other standards, data quality issues, source agencies and data collection information, quality, logistic metadata (which reveal how and where the data are stored), and process metadata (which are used during statistical processing). The IPIS metadata model was considered at the COSMOS cluster Conference in Athens in May 2002 as the best-defined model of all 5 projects of the cluster and was, hence, selected as the guiding metadata model for COSMOS.

d) Architectural design of the IPIS prototype.

The architectural design provides a lot of useful functionalities. The roles of the different actors of the system are customisable as well as their access rights. It is possible to use access lists to implement user policies when required. A number of high- level design entities can be redefined or created and after their characteristics are discussed in detail. Also, the required databases on which these entities would interact could be customised.

e) System Development

Test data are loaded into the database tables. Modules are developed to interface between the database and the application. The functionalities to view MORs and to perform an initial set of manipulations can be customised and implemented. In addition, modules in the application to handle the administration, login and register functionalities are developed. The application modules would be web-based and the whole system would have a multi-tier architecture.

Main results achieved

The scientific and technical achievements of the IPIS project could be summarized as follows:

a) The design, specifically for the IPIS requirements, of a solid statistical metadata model, which is flexible enough to embrace further requirements and technological progress.

b) The development a user-oriented information system, designed according to non-IT expert user requirements, allowing them to access homogenized and centralized data from various sources. The system generates innovative and value-added services in terms of:

- Harmonization of information collected in various sources, in different countries and for various time periods. Additionally, the system uses process metadata for automating the statistical processing of information so that users need only to describe the statistical tables they were interested in and not how to construct them;
- Handling of huge datasets. This was achieved by (i) the use of a Relational Database system, which was designed to support virtually unlimited volumes of data; (ii) the use of a data abstraction layer, which enhanced the performance of the system; and (iii) the mapping of Metadata model to a relational model using a master/detail/sub-detail structure;
- Establishing a common, user-friendly data/metadata interchange format based on OPEN standards which, if followed by data providers, would improve the quality of the collected data/metadata;
- Enabling easy access to comparable and pre-harmonized data at reduced cost as well as access to corresponding statistical metadata produced by separate producers throughout Europe;
- Allowing the creation of new indicators as well as the development of specialized collections of indicators for specific domains;
- Introducing the IPIS system as a common Data repository / Expert collaboration area, a service that required minimal computing requirements, where the people involved in the field might store, combine and process their data.

Possible impact on ESS and/or the end-user

NSIs and other organizations that attended the COSMOS cluster Conference in Athens in May 2002 had expressed appreciable interest in the IPIS system and its underlying metadata model. Public Administrations and in general policy-making bodies, which have been the main target user-group of IPIS, collect, store and disseminate large amounts of statistical information and meta-information for both management and policy-making. By delivering harmonised, comparable, up-to-date statistical data of high quality to them, IPIS would facilitate data analysis and decision making by EU member states.

Regarding, for example, the CBS area of application, the integration of IPIS with certain European custom operations will contribute to the facilitation of intra-European and extend trade. Increased intra-European mobility of goods, capital and labour, plays a significant role in reducing labour market rigidities and promoting employment generation.

Furthermore, IPIS can also assist the private sector with coherent information on their competitive environment but also benefit citizens who need reliable information on their educational and training choices, employment, mobility, etc.

Dissemination and exploitation prospects

In outline, these were:

- Creation and continuous update of the project website;
- Publication of research papers in scientific journals;
- Conferences - presentations of project results at various workshops and seminars;
- Creation of Special Interest Groups, such as regional administrations, professional bodies, private businesses, researchers;
- The IPIS software was made available to NSIs and other Public Administrations with the permission of Eurostat;
- Invitation to representatives of various European NSIs to participate in IPIS meetings;
- Demonstration activities of IPIS to Eastern and Balkan NSIs;
- Participation in clustering projects;
- The IPIS metadata model was selected as the COSMOS cluster core metadata model.
- The IPIS metadata model was presented and included in the 'Documents Library' of the website of the METANET project.

To elaborate, the dissemination of the results of the IPIS system included the wide acceptance of its fundamental scientific results metadata model and the harmonization transformations. The basic research results, mainly the metadata model and related operations, were published in journals oriented towards Official Statistics, Computational Statistics, and Information Systems, such as 'Research in Official Statistics', 'Computational Statistics' and 'Journal of Intelligent Information Systems'. The IPIS metadata model was presented at the Joint UNECE/Eurostat 'Seminar on Integrated Statistical Information Systems and Related Matters', Geneva, Switzerland, 2002. It was also presented in the 'Thirteenth International Conference on Scientific and Statistical Database Management (SSDBM)', Virginia, USA, 2001.

Suggested Further Work

a) Regarding the IPIS system: most IPIS partners had intended to exploit the system both for their internal needs and in further activities, such as application in new areas and even commercialization. Indeed, the administrative coordinator had initiated discussions with commercial software firms about the possible future exploitation of the IPIS final product. QRI had made contact with several of his clients who had shown particular interest in using IPIS data to make statistical comparisons between public statistical data and commercial corporate data. GSIS had contacted SOYA MILLS S.A, which had expressed strong interest in the IPIS system.

b) The IPIS metadata model was used as baseline and extended to satisfy the requirements of the CODACMOS project. It also formed part of the classical data in the ASSO project. Currently, it is being enriched and extended so that it could be considered for use in health statistics, specifically for the purpose of managing and monitoring multi-center clinical trials' statistical process.

Bibliography

- Haiman G. and Preda C. (2001). "Harmonisation of Statistical Data and Metadata. The French case study" to the NTTS-ETK 2001 Conference, June 18-22, 2001, Crete, Greece, Pre-Proceedings, (2), 925-926.
- Papageorgiou, H., Vardaki, M. & Pentaris, F. (2000), 'Recent advances on metadata', Computational Statistics, 15(1), 89-97.
- Papageorgiou, H., Pentaris, F., Theodorou E., Vardaki M. and Petrakos M. (2001), "Modelling Statistical Metadata", Proceedings of the Thirteenth International Conference on Scientific and Statistical Database Management (SSDBM), Virginia, USA. 25-35, IEEE Computer Society.
- Papageorgiou, H., Pentaris, F., Theodorou E., Vardaki M. and Petrakos M. (2001), "A statistical metadata model for simultaneous manipulation of data and metadata", Journal of Intelligent Information Systems (JIIS), 17(2/3), 169-192.
- Papageorgiou H., Petrakos M., Vardaki M., Theodorou E. and Pentaris F. (2001), "Metadata based Assessment of the level of fragmentation of Data Series and Multisource Statistical Tables", presented to the NTTS-ETK 2001 Conference, June 18-22, 2001, Crete, Greece Pre-Proceedings, (1), pp.263-272.
- Papageorgiou, H. Vardaki, M. & Pentaris, F. (2001), "Data and Metadata Transformations", Research in Official Statistics, 3(2), 27-43.
- Papageorgiou H., Vardaki M., Petrakos M., Theodorou E. and Pentaris F. (2001), "Harmonisation of Economic Classifications and related Transformations", NTTS-ETK 2001 Conference, June 18-22, 2001, Crete, Greece, Pre-Proceedings, (1), 345-354.
- Papageorgiou H., Vardaki M., Theodorou E. and Pentaris F. (2002), 'The use of Statistical Metadata Modelling and related transformations to assess the quality of statistical reports' Invited paper in the Joint UNECE/Eurostat Seminar on Integrated Statistical Information Systems and Related Matters (ISIS 2002), Geneva, Switzerland, jointly organised by the Conference of European Statisticians and Eurostat.

IQML
A Software Suite and Extended Mark-Up Language (XML) Standard for Intelligent Questionnaires

Timetable	1/2/2000-31/4/2003 (39 months)
Website	http://www.epros.ed.ac.uk/iqml

The Consortium

Member	Role	Institutional type	Country
University of Edinburgh	Coordinator	Academia	United Kingdom
Dimension EDI	Partner	Company (IT)	United Kingdom
DESAN Research Solutions (DESAN)	Partner	Company	Netherlands
Comfact AB (Comfact)	Partner	Company (IT)	Sweden
Statistics Norway (SSB)	Partner	NSI	Norway
Central Statistics Office (CSO)	Partner	NSI	Ireland
National Technical University of Athens (NTUA)	Partner	Academia	Greece

Scope and Objectives

The objective of the project was to improve the accuracy and timeliness of statistical data collection from enterprises and individuals whilst at the same time reducing the burden of statistical reporting on enterprises. The aim was to achieve this by following three research strands:

- To research the realities of metadata interchange and object standards in order to facilitate an active contribution to the metadata interchange standards by implementing, in software, chosen aspects of the international standards for metadata interchange, and carrying out trials in the area of intelligent questionnaires;
- To harness the emerging technologies to facilitate the automation, user friendliness, and application integration of raw data collection demands of collection agencies;
- To assist raw data collection agencies to build collection instruments in a variety of forms (eg CATI, CAPI) using a common metadata model which would facilitate the development of, and access to, a common metadata repository.

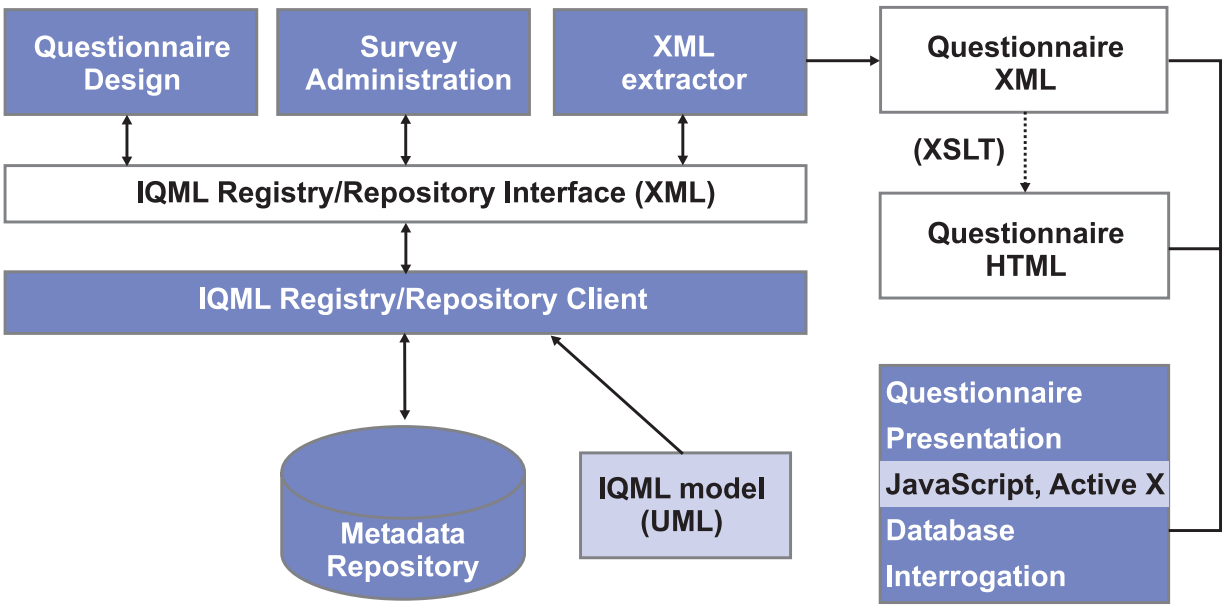
Outline of Methodology

The IQML software suite consists of five separate tools which interact together. Two basic technologies were utilised, and the software tools interacted via these technologies. The first technology was XML and the second was generic metadata repositories. The five software tools are as follows:

- The Metadata registry and repository (MRR): this product supports the definition of metadata objects, and holds the IQML common model. APIs are available which allow the other tools to exchange metadata via the repository. A second function of the repository is to make the metadata available for other software products which support the processing or dissemination of statistical data;

- Questionnaire Design Tool (QDT): this package enables the user to design and manage questionnaires which can be deployed using the other software of the suite. The tool allows the user to define questions and questionnaires. The finished questionnaires are stored in the repository for the other tools to utilise;
- Questionnaire Presentation Tool (QPT): this tool takes an XML version of the questionnaire from the repository and displays it for use with a PC or with a web browser. The XML used was developed by the IQML team in conjunction with EEG6 (the European e- business standards group for the statistical domain). Validation, navigation and calculation are specified in the QDT and will be implemented by this tool. This allows users to fill in the data and or have it validated as appropriate;
- Database Interrogation Tool (DIT): this tool supports the extraction of data from existing databases to fill a questionnaire specified in XML. Once the mapping is configured, it supports the automated loading and extraction of data to and from databases and the electronic questionnaire;
- Survey Administration Tool (SAT): this tool allows the questionnaires to be integrated with registers and sample frames. It interacts with the QDT to allow the specification of fields to be pre-filled prior to dispatch, and it tracks the dispatch and receipt of questionnaires to individuals and organizations.

The following diagram shows the overall architecture of IQML:



Main results achieved

The ambitions and orientation of the partners changed over the life of the project. The OMG and CWM proved to be less useful than had been anticipated, and ebXML proved to be a more fruitful line of enquiry. In some respects the project was too much in advance, since it was unable to take advantage of software solutions (eg the soft forge implementation of an ebXML registry) during the lifetime of the project. The rather ambitious timetable for the trials was successfully kept, with the final trial giving good feedback on the strengths of the software and the paths for future development.

There were three large and innovative outcomes of the IQML project. These were the software suite with 5 modules as listed in the preceding section; contribution to standards; and trials in 3 countries.

The IQML suite of software modules has demonstrated the achievement of some long-held ambitions concerning a number of innovative ideas. First, IQML established an XML-based data capture software (QPT and DIT) that could operate on any questionnaire defined in XML4DR, including calculation and validation. Secondly, IQML established a model independent registry/repository and stored objects conforming to the model in it.

The extraction from the object model to XML4DR was demonstrated. Thirdly, a 'bottom-up' questionnaire design tool was developed which could interface with the Registry, or with other software modules, and which could support the concepts of question banks and re-usable components. Fourthly, IQML developed an administrative tool that would allow for the management of multiple media surveys (eg paper, web and e-mail). Finally, ideas of data extraction were tested in a real-world situation, and some experiences on the use of secondary sources of data were taken on board.

Concerning the trials:

- The first was carried out between months 6 and 9 of the project. Six test users were recruited by the CSO in Ireland and eight in Statistics Norway. The software tested was a precursor of the QPT, and the findings contributed to the development of the later release;
- In trial 2, the objective was not to implement and run large-scale surveys but to test the functionality provided by each of the tested modules and to provide feedback to the developers for further improvements;
- Trial 3 took place from December 2002 to January 2003 at one site: the CSO. Its main purpose was to test the DIT. The latest version of the QDT, QPT and the registry were also used.
- After the three earlier trials which were used to refine and improve the software, the final trial was designed as a validation of what had gone before and only minor improvements could be expected after the trial was completed. In addition, the final trial brought together all five modules, and so integration was a key element. The evaluation methodology of trial 2 was again used for the final trial.

Possible impact on ESS and/or the end-user

Good data collection is the cornerstone of the statistical process. In recent years there have been a number of changes in the way data were collected, with emphasis on the reduction of the response burden, the re-use of metadata within the statistical process, the development of web-oriented collection techniques and the need to support multiple response modes. In addition, the use of administrative data sources has become more widespread. The development of metadata models and registries to support data collection activities should contribute to the goal of a European standard for data collection. Raising the awareness of bodies such as the OMG of the issues concerned with statistical data collection has been important and has brought the OMG's activities to the notice of official statisticians. Finally, IQML has a set of software tools that should aid the design, management, collection and storage of the statistical data and its associated metadata in NSIs and elsewhere.

Dissemination and exploitation prospects

There was the website, a project presentation leaflet and a synthesized description of user needs. A self-registering mailing list had attracted 58 members. IQML partners contributed numerous papers and attended many conferences. These were detailed in the project's final report. There was strong participation by partners in OMG meetings.

The User workshop on the IQML software suite was held at the offices of the Croatian Central Bureau of Statistics, Zagreb, in April 2003. Five presenters from the IQML consortium were in attendance and 14 users took part. All were drawn from NSIs and 10 were from candidate countries. The IQML project, including its software, was demonstrated. The entire system was made available to NSIs at the end of the project.

IQML was part of the COSMOS Cluster. Ongoing work within the COSMOS project was already developing models and software for interoperability between IQML and other related projects and systems. Methods for achieving exchange metadata between different projects were also developed.

The various organizations in the project would develop a variety of products, all of which would be available to any project member for commercial exploitation and/or further development. One effect of the changing nature of the technology base upon which the products would be developed was that the metadata model would be the cornerstone for product integration.

BIBLIOGRAPHY

- Brannen, K. (2001) "Intelligent Use of Metadata in the Questionnaire Design Process" in NTTS2001 – ETK 2001 Pre-proceedings Volume I, pp 155-162, Hersonissos 18-22 June 2001.
- Caarls, N and Hartkamp, J (2003) IQML Deliverable 11b: IQML Survey Administration Tool Functional Specifications and Interfaces.
- Delamere, M (2003) IQML Deliverable 7: Evaluation Report of Trial 3.
- Folkedal, J and Hoel, T (2000) IQML Deliverable D5: Evaluation report of Trial 1.
- Ma, D and Taylor, A (2003) IQML Deliverable D11a: Questionnaire Design Tool: Functional Specification.
- Nelson, C (2003) IQML Deliverable 8: Report on standardisation issues and object model.
- Nelson, C (2003a) IQML Deliverable D11c: IQML Registry and Repository Interface Specification.
- Panagopoulou, G (2002) IQML Deliverable D6: Evaluation Report of Trial 2.
- Taylor A and Hartkamp J, (2003) Report of IQML User Workshop, Zagreb, 24 April 2003.
- Törnqvist, A and Johnsson, M (2003) IQML Deliverable D11d: Questionnaire Presentation Tool: Functional Specification & Interfaces.
- Törnqvist, A and Johnsson, M (2003a) IQML Deliverable D11e: Database Interrogation Tool Functional specification & interfaces.

MANTLE

Mapping Night-time Light Emissions in the EU using satellite observed visible – near-infrared emissions as a policy tool

Timetable	1/11/2000-31/10/2002 (24 months)

The Consortium

Member	Role	Institutional type	Country
HTS Development Ltd (HTS)	Coordinator	Company (IT)	United Kingdom
Imperial College of Science, Technology and Medicine (ICSTM)	Partner	Academia	United Kingdom
Centre for Computer Science in Agriculture, Accademia del Georgofili (CeSIA)	Partner	Academia	Italy
National Observatory Athens (NAO)	Partner	Public sector (research)	Greece
Science Systems Space Ltd (SSSL)	Partner	Company	United Kingdom
Rasmussen and Witthøft (R&W)	Partner	Company	Denmark
Institute of Agrometerology and Environmental Analysis Applied to Agriculture (CNR-IATA)	Partner	Academia	Italy
Regione Liguria (Liguria)	Partner	Public sector	Italy

Scope and Objectives

Information on the total level of night-time light emissions within the EU is lacking, despite the prevalent use and increasing levels of lighting in almost all urban areas in response to a range of societal and economic influences.

The specific objectives of MANTLE did not change significantly throughout the project and can be summarized as follows:

- To assess the potential of using existing satellite data from the Defense Meteorological Satellite Program's Operational Linescan System (DMSP-OLS) sensor to produce maps of light emissions and urban night-time intensity levels in the EU. The accuracy of these maps was to be quantified at different spatial scales and in different countries and areas of the EU in order to assess the potential for routine mapping of urban night-time light emissions using either existing satellite technology or new sensors;
- To investigate the factors that affect urban night-time light emissions, including urban morphology and layout; land use characteristics; lighting configuration; population density and the extent of urban areas;
- Based on this analysis, to assess the ability to use urban light emissions as indicators of population size/density; urban population numbers; total energy consumption by sector of activity; energy waste; gross domestic product; urban typology; rural tranquility and landscape/skyscape quality;
- To evaluate the available capability to enhance the quality of this information either by post-processing of the data or by using higher resolution sensors.

Outline of methodology

Night-time light emissions maps derived from coarse resolution satellite imagery provide a means to readily obtain continent-wide or even global estimates of the area covered by human settlements. Such night-time light emissions maps could also be regarded as a surrogate measure for a range of other socio-economic and environmental indicators, notably population number & density, GDP, energy consumption & wastage, skyglow and tranquil areas. MANTLE built upon a project definition phase incorporating a state-of-the-art review and user-requirements survey carried out across the EU to define, produce and present a range of information products derived from night-time satellite imagery that were of value to national and international users.

Several satellite data sources were considered and used in the project. The most important was the DMSP OLS, as used in previous studies on mapping light emissions. DMSP OLS imagery has the advantage of enabling faint sources of light emission to be detected on the Earth's surface. It also provides continuous coverage of 3000 km during daytime and night-time conditions. The data are initially collected at a spatial resolution of 0.56 km but on-board averaging produces "smoothed" data with a nominal resolution of 2.7 km. Most of the data received by the NOAA National Geophysical Data Centre (NGDC) is in the smoothed resolution mode.

To understand the information contained in the DMSP OLS data, a "composite" image of the data was created, using a procedure to calculate the 'Weighted Sum Intensity' for DMSP OLS images. The basic concept behind the "composite" image is that DMSP low-, medium- and high-gain images are combined into a single 16-bit representative image.

Main results achieved

The MANTLE project has developed a number of important results, including:

- A methodology for developing policy-related indicators (population, GDP, etc.) from night-time light emission data;
- A GIS-based model for modeling light pollution and sky-glow from night-time light emission data;
- A spatially enhanced map of night-time light emissions in the EU;
- A methodology to measure artificial light signatures from airborne-mounted visible and infrared sensors;
- An airborne-based streetlight mapping approach;
- A new layer for the tranquility mapping product based on light pollution measurements;
- User sector requirements for night-time satellite data which would enable new sensor / mission design;
- A methodology for assessing seasonal and multi-annual changes in land use and light emissions;
- An evaluation of the capability to improve these data by enhancing their spatial resolution either post hoc or at source.

Modelling of population, GDP and energy consumption was undertaken at EU, national and regional levels. All referenced data sets were integrated into a GIS in order to allow processing and manipulation, and spatial modelling. Development and testing of these methods were carried out at three linked scales: locally in one or more selected Tran frontier areas, and at EU level.

MANTLE used DMSP OLS data to derive country and EU-wide night-time light emissions mosaics. The study analysed the relationships between DMSP light emissions and CORINE land cover and it assessed socio-economic indicators based on DMSP and CORINE land cover data. The project identified that strong relationships existed between these DMSP OLS data and such parameters as population number and density, GDP, road lighting, energy wastage and total energy consumption.

Analysis within the selected Tran frontier areas was designed to allow issues of data consistency and spatial structure to be investigated. One of the main problems was the Modifiable Areal Unit Problem (MAUP), arising from administrative units varying greatly in size and shape from one country to another or not conforming closely to the actual boundaries of urban areas, thereby producing considerable uncertainty and fuzziness in population maps. An important part of the analysis was therefore to explore the capability of using light emission data to model socio-economic factors across Tran frontier regions, as a basis for deriving spatially more consistent data. Equally important was testing the robustness of the models developed to interpret light emissions for one area and other adjacent areas.

Possible impact on ESS and/or the end-user

Many EU policies had provided the backdrop to the need for a project such as MANTLE. For example, Decision (No.2179/98/EC) had looked forward to improving the basis for environmental policy through reliable and comparable statistics and indicators. The policy benefits to be derived from MANTLE are that it should provide:

- Data on the extent and magnitude of environmental light pollution, which can be used in habitat preservation, for example through identifying tranquil zones which form part of habitat corridors, urban spatial planning and energy conservation. These basic data were not available;
- A dataset which can act as surrogate for other indicators, including socio-economic indicators, and thus provide the capability to estimate missing data;
- The classification of types of towns, areas and objects generally useful for spatial planning policy applications, e.g. as method for targeting/stratifying policy action.

It was envisaged that initially the main beneficiaries of MANTLE would be the EU, particularly DG Agriculture and DG Environment including its Economic Unit, the European Environment Agency and Eurostat. This would be because of the homogeneity of the dataset at the European level and the provision of products that should assist in the implementation of Community-wide policies. MANTLE examined the 'Public Interface'. This was broken down into three main parts. The first was to assess the needs of users. The second part involved the dissemination of the results and the third part examined possible means of exploitation. The assessment of user needs consisted of two phases. The first phase was an assessment of data needs within the MANTLE team and the second involved conducting a survey of a targeted user audience in the UK, Italy, Greece, Denmark and Eire.

Overall, a clear pattern emerged from the user survey:

- Users were unlikely to use MANTLE data in isolation. MANTLE data must be compatible with other datasets;
- Users were keen to acquire data more frequently than currently available from their existing data sources;
- Users were keen to obtain data with improved geographic coverage;
- Some users were keen to use night-time data, which were hard to obtain or nonexistent;
- Users were concerned that the resolution might not be good enough for their needs;
- Users were concerned by the accuracy / uncertainty / 'fuzziness'/ ambiguity of the data or the interpretation of the data, especially as they would tend to use derived data products rather than the raw material.

Dissemination and exploitation prospects

There were a website; a colour project leaflet; CD-ROM with maps and main results; an electronic newsletter; and an overview slide presentation of the project. Partners made presentations at conferences for the project in Greece, the UK, the Czech Republic and Turkey, including a newspaper interview related to the project in Greece. MANTLE presentations were made to the following conferences:

- 6th Greek Geographical Conference of the Greek Geographical Agency;
- International Conference on Light Pollution, Chile, Serena, March 2002;
- 22nd EARSeL Symposium: Geo-information for European-wide Integration, Prague, Czech Republic, June 2002;
- 3rd Symposium of Remote Sensing of Urban Areas, Istanbul, Turkey, June 2002;
- Protection and Restoration of the Environment VI, Skiathos Island, Greece, July 2002;
- 6th Greek Conference of Meteorology and Climatology, September 2002;
- A paper was also submitted for the 2002 European Plan for Research in Official Statistics.

Suggested Further Work

Further research is required, and is currently on-going, to model policy-related socio-economic indicators to monitor urban development and population change and to map tranquillity. Scope exists for enhancement and wider applications, such as extending the research to investigate associations between DMSP data and other human-related activities, for example air pollution and noise modeling.

Much of this research has so far been carried out using coarse resolution data, and further studies of DMSP data are required in order to understand and analyse the data fully. Obtaining timely acquisition of DMSP OLS data would improve this process, as well as increase the potential for commercial uptake of the data. The MANTLE project demonstrated that DMSP OLS was a valuable data source that could be used to support policy-related measures. Given their inherent association with human activity, light emissions data could also be a surrogate measure for a range of other socio-economic and environmental indicators.

It is clear that enhanced spatial resolution DMSP OLS data would be extremely useful in tranquil areas mapping. Specifically:

- The spatial resolution of available data is a limitation;
- DMSP fine resolution data were not archived (need for a non military European receiving station);
- Spatial transformation methods (e.g. via CORINE land cover) provided scope for enhancement;
- There was potential for the design and development of a new night-time sensor.

METANET

A NETWORK OF EXCELLENCE FOR HARMONIZING AND SYNTHESISING THE DEVELOPMENT OF STATISTICAL METADATA

Timetable	1/11/2000-30/4/2003 (30 months)
Website	http://www.epros.ed.ac.uk/metanet/index.html

THE CONSORTIUM

Member	Role	Institutional type	Country
University of Edinburgh	Coordinator	Academia	United Kingdom
Statistics Netherlands (CBS)	Partner	NSI	Netherlands
University of Athens (UOA)	Partner	Academia	Greece
University of Vienna (ViU)	Partner	Academia	Austria
Statistics Sweden (SCB)	Partner	NSI	Sweden
Statistics Norway (SSB)	Partner	NSI	Norway
Survey & Statistical Computing (SSC)	Partner	Independent	United Kingdom

SCOPE AND OBJECTIVES

These were:

- To develop proposals for standards in the methodology used for describing statistical metadata and statistical information systems;
- To develop proposals for recommendations on the metadata objects in a common conceptual model of statistical metadata;
- To disseminate these proposed standards to the relevant user communities and standards bodies;
- To interact with relevant FP5 projects on the development and agreement of these proposals, and to advise on methods of achieving coherence of approach in the field of metadata for statistical information systems;
- To integrate the different views of metadata into one model and bring together these different perspectives.

OUTLINE OF METHODOLOGY

The network was based round four different work groups. Members for the work groups were recruited from the Kick-off conference, and they did contribute to the work by submitting papers and by attending meetings. A structure of working meetings was arranged, with a final dissemination conference.

Work Group (WG) 1 had the title “Methodology and tools”, and its deliverable was titled “Overview of technical aids to Metadata representation”. In practice, it was decided at the Kick-off conference that this paper would be a scoping and information collection exercise, which would feed into the remaining work packages.

WG 2 had the title “Harmonization of metadata structures and definitions” and its deliverable was titled “The concept of metadata: A report on the nature of metadata and how these concepts can be used in practice”. It was agreed at the Kick-off conference that this work group would concentrate on analyzing the findings of WG 1.

WG 3 had the title “Best practice for migration” and its deliverable was titled “Reference book for metadata standards and methodology”. This work group would consider the impact of metadata on the working of a statistical organization, and would produce an implementation model to help it understand the implications of introducing a metadata system.

WG 4 had the title “Adoption issues” and its deliverable was titled “A training manual for the adoption of metadata standards and systems”. It would consider the impact of introducing a new system on the workforce in an organization, and it would recommend good practice for helping users to work with the new systems.

Main results achieved

The main outcomes of the MetaNet project were the activities and results from the various WGs and two conferences. The Proceedings of the Kick- off Conference were summarized in the report, deliverable D3. The Conference resulted in a number of outcomes, with over 60 participants spanning various sectors of industry and government across Europe. The key outputs were:

- Network Website and Project Presentation leaflet;
- Kick-off Conference Proceedings;
- Overview of technical aids to Metadata representation;
- The concept of metadata: A report on the nature of metadata and how these concepts could be used in practice;
- Reference book for metadata standards and methodology;
- A training manual for the adoption of metadata standards and systems;
- Proceedings of the second MetaNet Conference.

WG1 provided a broad overview (D4) of technical aids related to metadata and, whilst serving as an input to the activities of the remaining three WGs, D4 can nevertheless serve as an independent reference text. There was an expansion of WG1 objectives to the broader “methodology and tools” coverage and also in the expansion from “technical aids to implementing metadata systems and exchanging metadata descriptions” to include exploration of the alternative metadata schema.

WG2: the main result was D5: The Concept of Statistical Metadata. It covered the following topics: the role of metadata in statistics; the multiple modeling levels in statistical information systems; the initial version of the UMAS-based Statistical Metadata Standards Proposal; and how the developed methodology should be used in practice via a mapping approach. A recommendation for a basic set of terms was also supplied via the UMAS terminology index.

WG3: the main results were summarized in D6, Developing and Implementing Statistical Metadata Systems. It explored how best to develop and implement metadata infrastructures supporting the production and usage of statistics. Supported by 43 illustrations, the content reflected the activities in WG3 and was structured over 6 chapters.

WG4: the deliverable D7, Adoption Issues, comprised ten chapters. The first three chapters respectively considered the background to the MetaNet initiative, overview of the deliverables of the other WGs and discussion of the necessary reassessment of task and contribution of the other WGs. Chapters 4 through to 9 presented the rationale behind the inclusion of each of the five main issues explored in the project. The final chapter 10 considered the dissemination of experiences and good practices in training on metadata in statistics.

An additional WG5 (Terminology): compiled an inventory of all relevant terms from the MetaNet deliverables as well as from some other relevant sources. XML descriptions of the contributions were generated, containing standard bibliographical information, including abstracted key terminology terms used in the document. A simple metadata search tool was also developed, supporting the identification of terms throughout the deliverables and their cross-referencing. After the analysis and evaluation of the collected terms, the MetaNet "Glossary of Terms" was produced.

The Final Conference was hosted by the University of Athens and was held on the Greek island of Samos, May 2003. The Conference aimed at disseminating the conclusions of the WGs, whilst providing an opportunity for non-members to comment and to indicate the relevance of the results to their own work. Additionally, the Conference sought to focus on ways of taking the network forward, concentrating on the integration of all activities that were conducted during the project's lifespan and the possible exploitation of results. The sharing of work in progress and new paths were also encouraged.

The Conference was attended by 45 delegates from 16 NSIs, 6 universities, 8 public national and international organizations and 4 private companies. Although attendance was lower than originally anticipated, the level of active participation from these delegates was high, resulting in fruitful dialogue. There was a report (D8) on the Proceedings of the Final MetaNet Conference.

Possible impact on ESS and/or the end-user

The advent of the Web has had a huge impact on the ESS and its users, as statistical organizations turned more and more to publishing results and data online. This has focused the attention of publishers towards the metadata to support that data. At present much of that data is human readable, resulting in a waste of resources as data is re-typed. The successor to the Web is the Grid, which has been foreseen as being much more powerful, with processing as well as data being distributed. End users would begin to demand single portals to data from a variety of sources and would expect much more automatic data processing, but they would also expect to be able to track that processing in order to understand how it was maintained. MetaNet contributed to the realization of these expectations by bringing together the major players in the field, by concentrating on models for structured metadata and by providing a forum and resource centre for interested parties.

Dissemination and exploitation prospects

Formal dissemination activities were conducted on an ongoing basis via the Web and through a series of presentations. There was a widely disseminated project brochure. Two conferences were also held, one at the start and one near the end of the funded lifetime of the network, each targeted at different audiences and offering delegates yet another level of network involvement. The individuals comprising the network aided continuous informal dissemination of the emergent findings through interactions within their own organizations, with other NSIs and statistical agencies, through interest groups and various other FP5 projects. MetaNet had planned to continue after the end of formal funding in 2003.

Suggested Further Work

Regarding possible areas of future research in the field of statistical metadata, the user survey of the project found that there was a strong priority at the organizational level for developing metadata solutions that targeted users. A greater need to explore the concept of different types of audience in the best practice for implementation of metadata systems and to incorporate this in the metadata model was suggested. However, the consortium felt that, with the exception of the data archives, where satisfying external user needs was a founding tenet, external user contexts, needs and behavior had as yet received insufficient attention from NSIs. The further development of minimum provisions in support of Internet publication, standards, supporting guidelines and indeed portals might assume higher importance in future, requiring collective action.

It was also found that the majority of the respondents to the project's survey gave high priority to a common strategy for handling data/metadata within their own organization, followed by a common 'model' for handling data/metadata. Concerning improvements in the international context, there was clear support for agreed international standards for data/metadata concepts and terms, but also some support for agreed common models for handling data/metadata. Both areas appeared to be strong contenders for the execution of research at a European level.

The type of activity to be supported by emergent technologies such as the Grid would require good metadata to support it. Moreover, future users would want to integrate statistical data with other types of data. Although this would increase the demands on statistical agencies, it would also provide an opportunity to integrate efforts with other disciplines. Inter-disciplinary collaboration is consistent with the current thinking of European research.

BIBLIOGRAPHY

- Papageorgiou, M., Vardaki, M., Petrakos, E., Theodorou, P. and Pentaris, F. (NTTS & ETK 2001) Harmonisation of Economic Classifications and related Transformations, New Techniques and Technologies for Statistics and Exchange of Technology and Know-how.
- Lamb, J.(NTTS & ETK 2001)Sharing best methods and know-how for improving generation and use of metadata, New Techniques and Technologies for Statistics and Exchange of Technology and Know-how
- Lamb, J (2001) The Metanet project, ONS Workshop on Web-publishing and Metadata, London,
- Lamb, J (2001) Metanet:. A network of excellence for harmonising and synthesising the development of statistical metadata, Eurostat metadata workshop, Luxembourg
- Brannen, K. and Lamb, J. (2001) Using agent technoloyg to disseminate statistics via the web, ASC International Conference on Survey Research Methods, Chesham, UK
- Froeschl, K.A., Grossman, W. and Denk, M. (2002) Statistical composites: a transformation-bound representation of statistical databases, SSDBM
- Grossman, W. and Ofner, P. (2002) A self documenting programming environment for weighting, COMPSTAT
- Lamb, J. (2002) Metanet: A network of excellence for harmonising and synthesising the development of statistical metadata: a progress report, Joint UNECE/Eurostat Working Session on Statistical Metadata
- Brannen, K. (2002) Metadata uses, US Bureau of the Census, Washington DC
- Westlake, A. (2002) Metis, UN/ECE Eurostat, Luxembourg, 6-8 Mar 2002 by
- Lamb, J. (2002) Using metadata in official statistics, Conference on Applied Statistics in Ireland, Ballycastle, N Ireland
- Lamb, J. (2002) METANET: Towards an integrated view of statistical metadata, COMPSTAT
- Brannen, K. (2002) Standardising statistical metadata methodology, IASSIST 2002, Connecticut, USA

METAWARE
Statistical Metadata Support for Data Warehouses

Timetable	1/2/2000-31/1/2003 (36months)

The Consortium

Member	Role	Institutional type	Country
Statistics Sweden (SCB)	Coordinator	NSI	Sweden
Statistics Norway (SSB)	Partner	NSI	Norway
World Systems (Europe) Limited	Partner	Company	Luxembourg
Central Statistical Office (KSH)	Partner	NSI	Hungary
Instituto Nacional de Estatística (INE)	Partner	NSI	Portugal
Statistics Danmark	Partner	NSI	Denmark

Scope and Objectives

The objective was the development of a standard metadata repository model for data warehouses (DWH) and related standard interfaces and functions to exchange metadata between the basic statistical production system and data warehouses. That would make statistical data warehouse technologies more user-friendly for user access by the public sector. It would support the application of official statistics in society and broaden the scope of users. The proposed system would be able to operate both in the traditional client/server environment and in the Internet world. It was also the aim to support and enhance standardization at national and international levels.

Outline of methodology

Based on the research results from FP4, such as IMIM, and in coordination with such FP5 projects as IQML, MetaNet and COSMOS, METAWARE would design a generalised metadata repository for a statistical data warehouse. Modern object-oriented methods and tools would be used in all phases of the project. A common interface would be designed to facilitate the connection of the metadata repository to different kinds of statistics production systems. Some important results would be illustrated by means of a prototype in a real statistical environment using commercial data warehouse software.

The work of the project would be based on documented user requirements, both requirements that were long- standing and requirements that were just emerging as the “metadata maturity” in statistical organizations was gathering pace. The user requirements input would come from the statistical organizations that are members of the project as well as from other sources. Towards the end of the project, preliminary results would be disseminated to a number of NSIs for their review and comments.

Main results achieved

At the outset, it was expected to develop a more extensive software solution based on the conceptual work of the project. Later, with reduced financial resources, it was decided to focus only on the conceptual part of the project. The project has delivered an extensive description of a metadata model related to the needs of a statistical data warehouse and a simple software prototype that proved the basics of the proposed conceptual model.

Specifically, the following were delivered:

- Review of the situation prevailing at the time and of the basic metadata object types;
- Principles of a standard metadata exchange;
- METAWARE process model;
- METAWARE object model;
- User requirements for METAWARE tools;
- Software prototype.

The project results included a description of the then current situation in those NSIs which were project participants: INE Portugal, HSCO Hungary and Statistics Sweden. These descriptions, attached as annexes to the final consolidated report, demonstrated the concrete needs and the production environment of data warehouse developments in real statistical information systems.

An important aspect was the feedback from a number of NSIs on the preliminary results of the project. This feedback was positive and provided recommendations for the improvement of the final documents, which were included in the final deliverable.

Different aspects of the metadata model were considered, such as:

- Objectives of a statistical data warehouse (SDW);
- Main processes in a SDW environment;
- Actors:
  - Human actors,
  - Technical actors;
- Necessary competences and resources for the development of a SDW;
- Global Data Schema;
- Processes and operations;
- Object descriptions of all metadata object types included in the model;
- Comments about the relation between METAWARE and the Common Warehouse Metamodel (CWM);
- A structured description of processes;
- Implementation of processes – description of possible implementation using commercial DW software;
- User requirements for tools.

The SDW could not be covered only by a static metadata model but must take into account such aspects as:

- Processes;
- Integration with the remaining statistics production system;
- Needs of users and the corresponding user tools.

Possible impact on ESS and/or the end-user

One of the main intentions of the project was to contribute to the harmonisation of public statistical information systems across EU member states, pre-accession nations and EFTA countries. The European reference environment for statistics, developed by Eurostat, could be seen as an important step towards a European statistical data warehouse. The results of the METAWARE project would contribute to a European reference framework for metadata and would define some possible tools for implementing a harmonised European metadata approach.

Such a metadata-harmonised statistical information system would make it easier and more efficient to use statistics from different countries and would facilitate the achievement of comparable statistical results at the European level. The tools envisaged should support multiple languages and be capable of storing time-series oriented views.

Multi-lingual support would make it possible to store and exchange standard classifications in different languages, and it would also enable a user to switch from one language to another while searching for, and accessing, statistical information in the data warehouse.

Dissemination and exploitation prospects

The results of the project were disseminated through a website and on a CDROM consisting of:

- The final outputs of the METAWARE project;
- A software prototype that demonstrated the implementation of some main results of the METAWARE project.

The results were also demonstrated at the following events:

- EPROS meetings in Luxembourg (2000, 2001, 2002);
- Meetings of the METANET Group;
- COSMOS Project meetings;
- Data Warehouse Meeting in Australia, November 2003.

The METAWARE outputs would be used as inputs into several future activities in a number of NSIs and in cases of cooperation amongst them. All deliverables from the project were public and could be used freely. The results would also be an input into the development of commercial software products.

Suggested Further Work

Because all results from the project were public, commercial software companies were invited to use the results for the development of metadata systems applicable to statistical offices. At least one German company, which was cooperating with NSIs, was using these results in their future development work. Thus major results from the METAWARE project should be implemented in future commercial software packages.

The METAWARE deliverables were also used in some Phare and Tacis projects to support the development of statistical offices in candidate countries and countries of the former Soviet Union.

MISSION
Multi-Agent Integration Of Shared Statistical Information Over The (Inter)Net

Timetable	1/1/2000-31/3/2003 (39 months)
Website	http://www.epros.ed.ac.uk/mission/index.html

The Consortium

Member	Role	Institutional type	Country
University of Edinburgh	Coordinator	Academia	United Kingdom
Office for National Statistics (ONS)	Partner	NSI	United Kingdom
Central Statistics Office (CSO)	Partner	NSI	Ireland
Statistics Finland (StatFi)	Partner	NSI	Finland
University of Athens (UOA)	Partner	Academia	Greece
University of Ulster (UU)	Partner	Academia	United Kingdom
DESAN Marktonderzoek BV (DESAN)	Partner	Company	Netherlands

Scope and Objectives

The background vision of MISSION was a number of independent organisations publishing their data within a framework which made comparisons and harmonisation possible. Experts could share their methods, so that the community had a much richer information source for understanding statistical information and its interpretation. Users, be they public sector, business or private, could have access to the published methods and a tool which applied these methods to the data of their choice. Agent technology would permit this free association of data providers and users.

The project aimed at producing a software suite that would allow statistical data providers to publish data on the Web. The software should satisfy a number of criteria that are specified below under Main results achieved.

Outline of methodology

The architecture of the system would comprise three basic logical, or conceptual, units or building blocks, which could be deployed in different scenarios. The components are: The Client, The Library and the Dataserver. The Client is a down-loadable module connecting a user to a home Library. The Library is a repository for statistical metadata, holding different kinds of metadata to support searching, access and explanation. The user formulates his/her queries using a graphical interface supplied by the Client. The Client sends the request to the Library, which analyses it, sending queries to other Libraries if necessary. A series of agents analyze the query, and, based on the metadata in the Library, develop a plan for obtaining an answer. This involves decomposing the request into sub-queries and sending requests for the results of these subqueries to the appropriate Data servers, registered with the Library. Once the results of the sub queries are known, they are combined and the result is displayed by the Client in a graphical interface.

The project would proceed through two major stages of development. The first was to develop a mechanism for homogeneous queries, i.e. queries on datasets whose structure (variables and value sets) were shared. The second was to extend this mechanism to deal with heterogeneous queries, where mapping (correspondence tables) existed for different classifications. To achieve this, there would be a number of releases over the life of the project, each of which would have progressively increased functionality.

The project would develop a metadata model to support the query processing. The model was shown to be compatible with the cristal model, and with data warehousing. In addition, a Mission Classification Server model would be developed to support the heterogeneous queries. Also, a number of different types of agents, with different functions, would be developed and used.

Main results achieved

The software product of the MISSION project was a software suite allowing statistical data providers to publish data on the Web, satisfying the following criteria:

- Suppliers of official statistics can subscribe to an integrated network of datastores via an interface to their existing data;
- Suppliers retain control over all aspects of access to their existing data;
- Users can make requests in a declarative manner, with a minimum of understanding of statistics or of the domain area, and still retrieve meaningful results;
- Users can tailor their working environment, from simple requests to detailed in-depth analysis;
- Methods of data manipulation and analysis can be retained, re-used and published;
- Libraries of metadata can be constructed and made available to other users;
- A flexible architecture allows third parties to act as Independent Metadata Providers, thus encouraging free exchange of knowledge;
- Users can build up individual profiles, accessing data and methods most relevant to their needs.

As outlined under Methodology above, the components were the Client, the Library, the Compute server and the Data server. A server as described here means service.

Godfather, the MISSION Agent Platform, was developed entirely within the MISSION project. However, Godfather was designed and implemented in such a way that was a generic product to be used in any circumstances in which an agent platform was required. It is an autonomous, stateful, active computing entity that operates in an agent environment and interacts with other agents within it. Following the general idea of the FIPA standards (FIPA00001, FIPA00023), the agent platform performs three general tasks:

- Provides the agent environment where agents are created and operate;
- Manages the agent life cycle and;
- Gives external entities access to agents.

As well as the software for the full system, MISSION produced three other results worthy of note. There was considerable theoretical development, as witnessed by the large number of academic papers produced during the life of the project; the 'drag-and drop' user interface of the Client, which was suitable for table construction in other contexts than MISSION; and the Agent Platform, which had wider applications than MISSION.

Possible impact on ESS and/or the end-user

MISSION was an investigation into the sharing of data between different institutions. The ideas attracted interest from a number of organizations. NSIs seemed to focus on disseminating information from a single institution. From that point of view, the data archive world, particularly in the UK, would most likely to be the immediate beneficiaries. However, the existence of the Mission system, which is Open Source software, should raise awareness of what could be achieved. Currently, users must search different organizations' website to download the data they need, and perform the matching themselves. MISSION gives them the benefit of utilizing existing mappings, made available through the Library, as well as defining their own.

Dissemination and exploitation prospects

There were the website and the project leaflet. MISSION has a self subscribing mailing list, which had 43 members in 2003. Both the public and private parts of this site would be maintained in the coming years, resources allowing. During the lifetime of the project several partners published and presented papers at numerous conferences for various audiences. A complete list of papers and conference presentations can be found at the public site.

At the end of the project, an overview paper submitted by the University of Ulster and DESAN was accepted for the Dublin Core 2003 conference. Apart from the paper presentation, the MISSION software itself would be demonstrated at DC2003. Project partners have also made several presentations of the MISSION Software on request from interested organizations. Without being complete, these included:

- Swiss National Statistical Office, Neufchatel;
- UK Data Archive, Essex;
- City Council Educational Department, Rotterdam;
- UK Office for National Statistics, London.

Following these presentations, DESAN was in touch with a number of organizations that showed serious interest in the MISSION approach. DESAN was willing, in cooperation with other consortium partners, to discuss and research adaptation and fine-tuning of the software on request, implementation depending on resources.

After the first version of the second prototype of the MISSION software was developed and installation guides and user manuals were compiled, a User Workshop was held at Schiphol Airport near Amsterdamin in September 2002. The Consortium regarded the User Workshop as very successful.

Following the last MISSION software upgrade, a public MISSION server was set up, hosted by DESAN.

Suggested Further Work

MISSION encountered a number of theoretical questions during the construction of the system, as it tried to move from abstract ideas to practical solutions. MISSION did not solve all of them satisfactorily but tried to find a real-world balance between serving the needs of the 'casual' end-user and the analysis specialist. This is an ongoing problem found in all applications aimed at making statistical data available over the web.

Regarding the institutional organisation suitable for a MISSION system, not many institutions could foresee a scenario in which they would deploy such a system. The exceptions were the Data Archives, particularly the UK Data Archive. More recently interest was shown in the United States, where the need to allow users to combine data from different sources was becoming more apparent. However, the consortium argued that MISSION was a system that should be given more attention in coming years. MISSION has a potential application to GRID technology, and, as the basic systems' developments were completed and more middle-ware became available, MISSION and systems like it would be able to take advantage of that technology. Moreover, in such a climate, the sharing of data, metadata and methods would become more acceptable.

Concerning future dissemination and exploitation, DESAN would continue to develop, promote and disseminate the MISSION tools, both for internal purposes as well as to external parties. The software was successfully extensively tested by the consortium and several 'outsiders'. Exploitation would consist of consultancy, implementation of tailor-made solutions and maintenance of the systems. Project partners would be invited to participate in future exploitation depending on the customer's request and the ability of the project partner at that point in time to deliver the requested input.

MISSION did aim at a number of immediate dissemination activities, with the aim of arousing the interest of third parties in the results of the project and to discuss with them how the MISSION software and/or the MISSION approach might be used in, and by, their organizations. Special attention was paid to SMEs, survey research agencies and other organizations that tended to be outside the traditional EPROS circle.

BIBLIOGRAPHY

- Barrie, K (2003) MISSION Deliverable 10: Evaluation Report for the 2nd MISSION Prototype.
- Karali I. et al (2003) MISSION Deliverable 7: MISSION System Prototype 1".
- Ranta, J (2003) MISSION Deliverable 8: The Evaluation Report of The MISSION System Version 1.1.
- Sakkis, G, Kapos, G-D, Karali I and, Hatzopoulos, M. (2001) "The MISSION macro importing module", MISSION, internal report, 2001.
- Tweedie, G. et al (2003) MISSION Deliverable 13: MISSION User Documentation.
- Tweedie, G. et al (2003a) MISSION Deliverable 14: MISSION System Documentation.
- Wooldridge, M (2002)"An Introduction to Multiagent Systems", John Wiley and Sons Ltd.

NESIS
New Economy Statistical Information System

Timetable	1/12/2001-30/11/2004 (36 months)

The Consortium

Member	Role	Institutional type	Country
Informer S.A	Coordinator	Company (IT)	Greece
Joint Research Centre (JRC, Ispra), European Commission	Partner	Public sector (research)	Italy
University of Bath	Partner	Academia	United Kingdom
Istituto Nazionale di Statistica (ISTAT)	Partner	NSI	Italy
Statistics Finland (StatFi)	Partner	NSI	Finland
Statistics Netherlands (CBS)	Partner	NSI	Netherlands
University of Bocconi	Partner	Academia	Italy
EU-Qualify	Partner	Company	Italy

Scope and Objectives

Economic and social changes have been accelerating under the twin impact of globalization and the new information technologies. But how are these processes interrelated? Are they impelling us towards a common socio-economic future? What can governments do if they want to manage and steer the direction of development? NESIS accompanied EPROS. The specific objectives of NESIS were:

- Through users' surveys, to ascertain more fully the policy needs for indicators on the new information economy and to better understand the methodologies by which the Commission services produced and measured them;
- Conceptually and statistically to contribute to the appraisal of existing EU benchmarking indicators;
- To re-conceptualize indicators for the new information economy;
- To cluster the indicator activities of current SINE research projects amongst themselves and in relation to the indicators emerging from the Commission services, with a focus on eEurope and with a view to achieving greater coherence, parsimony and taxonomy;
- To undertake a limited number of pilot case studies in selected areas to achieve the twin targets of:
    - Increasing understanding (metrics) as a prerequisite to designing appropriate indicators on complex issues;
    - Assisting to delineate the broad contours of a statistical information system.
- To disseminate awareness of the urgent need for the ESS to respond to the challenges posed by the dynamics of the new information economy in its routine production process;
- To promote indicator exploitation and dissemination through publications on the European Digital Economy;
- To serve as a repository and a help-desk on best practice concerning indicator methodology in the field of the new information economy.

OUTLINE OF METHODOLOGY

To achieve its objectives, NESIS had three broad activities:

- Observatory
- Metrics
- Dissemination

First, NESIS acted as an Observatory, taking stock of ongoing efforts by the scientific community, policy-makers, and a variety of social and economic actors, including the Commission services, other statistical indicator projects on the New Economy (SINE), and leading-edge work in the US, Japan and perhaps other non-EU OECD countries. Stocktaking and best-practice identification necessarily entailed the exercise of judgments. In making these judgments, NESIS developed clear appraisal quality criteria.

Secondly, it would be premature to attempt to produce indicators in certain fields without a deeper understanding of the underlying phenomena. This pre-design stage of indicators, termed metrics, involved field-tests on a limited number of issues, such as the impact of ICT on business organization. The expected outcome would be an array of statistical indicators which were, on the one hand, informed by careful scanning of leading-edge work across the developed world in particular, and which were, on the other, validated through field-testing in specific and carefully controlled pilots.

Thirdly, NESIS would disseminate these fruits of its work through its website, through its workshops, the IDWG, its publications, and through targeted and customized advice through its conferences.

Given these three broad activities:

- The NESIS approach was organized around four policy and conceptual pillars that had emerged from the Lisbon strategy. These pillars were (a) the New Information Economy and E-Europe, (b) Productivity and Competitiveness in the New Information Economy, (c) Human Investment in the New Information Economy and (c) Social Inclusion in the New Information Economy;
- Each of these pillars constituted an effective conceptual node for the workpackages in its subject-matter domain and therefore would be one source of guidance within the project for the pilot case studies designed to test concepts and to configure an ESS-based statistical information system;
- The standard format of the work in each workpackage with a measurement orientation was (a) dissemination of a stock-taking report based on the partners' experience and consultations with the Commission, (b) identification of best-practice through a user-driven, multidisciplinary, multi-national workshops, (c) revision of (a) on the basis of (b), with recommendations on the design of the pilot case study, (d) the execution of the pilot case study, and (d) report and dissemination;
- To network more widely with the various IST stakeholders, an advisory Indicator Development Working Group (IDWG) would be created;
- The exploitation and dissemination of indicators and the involvement of users would be promoted through meetings of the IDWG, through two large conferences, publications on the European Digital Economy and through NESIS websites and help-desks.

MAIN RESULTS ACHIEVED

NESIS came up with its own definition of the new economy centered on ICT as an enabling technology but recognizing the fundamental importance of knowledge in economic growth. NESIS also came up with its own conceptual framework characterized by the innovation process and articulated with the four pillars and "Eurostat's" taxonomy of indicators.

Using this definition and framework as a gateway and inspired by the Lisbon process, NESIS conceptualized and re-conceptualization a very large number of issues in the areas of:

- Macro-economic stability and sustainability;
- Productivity and competitiveness;
- Investment in human capital;
- Social inclusion;
- Enterprise demography and firm dynamics;
- The impact of ICT on the organization of firms;
- Knowledge stocks and flows;
- Spatial networking and competitiveness;
- E-Government;
- Regional disparities;
- Social capital and e-citizen;
- Satellite accounts on information and knowledge.

There were two major publication deliverables of NESIS:

a) A publication, “The European Challenge” by Graham Room and his co-authors based at the University of Bath, UK;

b) A (forthcoming) statistical publication "Measuring the New Economy” by Teun Wolters of Statistics Netherlands, associated with an internal report entitled Remedial Statistical Programme (RSP) by Mikael Akerblom of Statistics Finland.

The publication at (a) was quintessentially the new economy. This book addressed questions with particular reference to the European Union, which had made the development of a socially cohesive, knowledge-based economy its central task for the present decade. It assessed both the challenges and the policy instruments that were being deployed, focusing in particular on:

- The dynamics of the ‘new economy’;
- The new organisational architectures associated with rapid innovation;
- The transformation of education and training;
- The implications for social cohesion and exclusion;
- The role of policy benchmarking in promoting policy learning and enhancing national performance.

The European Challenge presented the most up-to-date research on the development of the knowledge-based economy and its social and policy implications. Its accessible and integrated treatment of the processes of economic, social and technological change makes it an invaluable resource for those studying and researching in the fields of public and social policy, organizational and technological change and innovation. It is also highly relevant to policy-makers who need to understand and manage this change.

The lessons that had emerged from (b) on statistical gaps from the work over the duration of the project are mentioned under Future Research below.

Possible impact on ESS and/or the end-user

NESIS was intended to stimulate the ESS to modernise its concepts, definitions, nomenclatures and methods of data collection so as to adapt them to the rapidly changing economic and social environment triggered by the advent of the new economy. The integration of indicator measurement within the regular official statistical production system, a key objective of NESIS, has to be coordinated at the level of the ESS in order to obtain comparable concepts, definitions and statistics on a sustained basis. Moreover, the ESS does already have a major presence in the collection of the more limited supply side of ICT statistics, and are making further efforts in that direction, and so collection burdens and costs should not be duplicated through isolated, ad hoc efforts at the measurement of benchmarking indicators. Whereas the ICT data developed so far tended to concentrate on industry, the new information economy encompasses a wider range of actors and concerns, covering both penetration to and impacts on citizen and on Government. Action at the level of the ESS is also necessary because concerns about quality have intensified under one of the great challenges facing the information society, namely the demands by the various actors in the Union for more and more information, of higher and higher quality, faster and faster, more and more comparable, yet at lower and lower cost.

Dissemination and exploitation prospects

Exploitation and dissemination were pursued through the project website, help-desk, project brochure, publication through selected journals such as ROS, presentation of NESIS results at Eurostat working groups such as EPROS, the 8 NESIS international workshops, 3 meetings of the IDWG, 2 major conferences, playing a flagship role at a Commission Workshop involving all 18 SINE projects in April 2002, bilateral concertation with other SINE projects such as participation in the STILE conference in September 2004 and BEEP conference in May 2003; and the publications mentioned under Main Results above.

The IDWG was a key NESIS network of high-powered experts and users from academia, the Commission, industry, Research Institutes and NSIs, including experts from USA. The 3 IDWGs were planned in an inter-related way with the workshops and conferences. The network was a very valuable resource and the NESIS consortium had expressed a strong wish for this network to continue after the project came to an end.

Suggested Further Work

The consortium had expressed the desire for a continuation of the IDWG.

There was a need for further research on Intangibility; Standards in the accounting field, hence links with the SNA/ESA; the local-globalization nexus; Stability; and Environmental sustainability. These new economy topics had not featured squarely and explicitly within NESIS workpackages. Consequently, only issues about them were stated but these issues were not crystallized in NESIS into statistical indicators.

NESIS statistical activities had led to the following observations:

1. Statistical Gaps that could be addressed by ESS without further ado-the short-term

   Existing systems for collecting statistical information lagged behind developments in the new economy. Nevertheless, some of these difficulties could, in principle, be addressed by instituting appropriate data collection systems, such a module piggy-backing on existing surveys: some innovations in terms of indicators would be needed but sufficient consensus probably existed for this to be manageable. Some of the gaps identified by CBS and Statfin in their RSP fell in this short-term category.

2. Areas requiring further research in terms of concepts, modeling and measurement – the longer-term

   Brand new indicators were not always needed but just old ones slightly revamped. Nevertheless, in some areas, further statistical research was needed, involving new forms of data mining and data integration to derive multidimensional indicators and longitudinal analysis at a micro-level, before any new indicators could be specified. This sort of work was, for example, needed in relation to the various factors which promoted dynamism and innovation at the level of the enterprise;

   It might also involve the development of new statistical frameworks such as new satellite accounts, for example, on knowledge or information, bearing in mind that such frameworks might be necessary but not sufficient to meet all the analytic needs. In other areas, major conceptual, modeling and measurement issues needed to be resolved before further progress could be made, such as 'blue skies' research, with no certainty that matters would be quickly and easily resolved. Examples were the conceptualization and measurement of globalized intangibles and the re-assessment of national income accounts.

3. Application to the New Member States

   NESIS gave only limited attention to the policy and statistical challenges posed by enlargement. These would need much more attention. Enlargement raised policy problems of coordination across more diverse members, and of the comparability and completeness of statistics. Enlargement also raised more sharply than hitherto the usefulness of benchmarking indicators that sought to measure national performance against common standards, but said nothing about the unequal relationships between countries at starkly different levels of socio-economic development.

4. A New Institutional Architecture for New Economy Statistics

   The major institutional arrangements for collecting and publishing statistical information remained at a national level. The development of the new economy was however taking place across a global terrain. These global transformations were driven disproportionately by multinational enterprises, which provided only limited reporting of their cross-national activities. There was a strong danger that the information society would become increasingly opaque. A new partnership might be needed between such public and private actors, concerned with the supply and governance of statistical information.

BIBLIOGRAPHY

- (Forthcoming) "Measuring the New Economy" by Teun Wolters of Statistics Netherlands, associated with an internal report entitled Remedial Statistical Programme (RSP) by Mikael Akerblom of Statistics Finland.
- Room, G and co-authors (2005) "The European Challenge: Innovation, policy learning and social cohesion in the new knowledge economy", The Policy Press, University of Bristol, UK
- "The EU-15's New Economy, A Statistical Portrait", Statistics Netherlands, Voorburg/Heerlen, 2005.
- "New economy achievements and lessons for the future", Pre-proceedings, Volume 1, The NESIS Summative Conference, 11-14 October 2004, Athens, Greece.

NEWKIND
New Indicators for the Knowledge-based Economy

Timetable	1/11/2000-31/7/2002 (21 months)
Website	http://www.researchineurope.org/newkind/index.htm

The Consortium

Member	Role	Institutional type	Country
SPRU, University of Sussex	Coordinator	Academia	United Kingdom
MERIT, University of Maastricht	Partner	Academia	Netherlands
Observatoire des sciences et des techniques (OST)	Partner	Public sector	France

Scope and Objectives

Various studies have re-emphasised the specific role of knowledge as a factor influencing the growth of output, productivity, employment and competitiveness in industrialised economies. The NewKInd's principal aims were to identify a group of key issues related to the accumulation, distribution, and use of knowledge that influence the rate and direction of economic change and to gauge their quantitative importance in comparative terms. A principal challenge that will face statistical offices in the near future is the problem of mapping and measuring the growth of the 'new economy'.

Outline of methodology

The main technical methodological axes of this project were :

- The development of macro-level indicators of the accumulation of intangible capital across five main European economies including a representative collection of sectors within each economy. The analysis was performed at the national and sectoral levels;
- The development of indicators of the emergence and extent of the new 'information infrastructure' of electronic commerce that could be applied at both the firm and the meso levels;
- The conceptualisation, development and implementation of micro-level indicators of the changing structure of the knowledge base in two sectors based upon the connection between patents and the scientific literature. The indicators were designed to measure the 'integration' of scientific and technological knowledge and the 'persistence' of patterns of knowledge specialisation in one country over time. The sectors selected for analyses were pharmaceuticals and tyres. The use of information resources in the life insurance industry was also studied;
- The exploration and conceptualization of micro-level indicators measuring the changing structure of the knowledge base in a sector which did not perform formal R&D, but which nonetheless was a major user of knowledge. This task examined the impact of increasing medical, sociological and demographical knowledge on the actuarial practice of life insurance companies;
- The conceptualisation and development of relevant indicators for assessing the relative performance of nations, industries, and firms that might be used to assess the explanatory power of the knowledge-based indicators.

Main results achieved

NewKInd developed an algorithm to identify and structure the citations to scientific articles in the applications forms of the European Patent Office. This algorithm allowed the identification of the journal of publication and the bibliographic notice of the citation in the database used.

The strongest result that emerged from the sectoral studies was that it was not meaningful to refer to a firm's 'knowledge base'. Rather, more attention should be paid to the variety of knowledge bases on which firms relied. NewKInd developed corresponding indicators. For instance, the studies that analysed the knowledge bases of the world's largest pharmaceutical groups, as well as the study that focused on the Spanish firms, highlighted how different were the results when specialisation was measured at the scientific, technological and therapeutic levels. While all firms tended to rely on a relatively common set of bodies of scientific knowledge (as captured by patent citations), significant differences emerged at the technological level (as captured by the IPC codes reported on patents). More importantly, sharp differences emerged at the level of the therapeutic areas in which different firms and groups had specialised. Domestic pharmaceutical firms that had managed to innovate at world class levels obtained a stronger competitive position relative to non-innovative firms; and they displayed a distinctive focus on research activities.

The e-business study demonstrated that a relatively inexpensive and rapidly administered method allowed the 'benchmarking' of industrial sectors and geographic regions. NewKInd research in this area revealed the existence of statistically significant differences between sectors within Europe and other regions and between Europe as a region, and other regions. These differences provided a basis for weights from which results of a more extensive study of a single sector might be extrapolated to an industry or economy-wide estimate of the rate and direction of intangible capital formation in WWW information resources.

The country level study delivered a methodology for creating knowledge stocks of R&D, IT hardware, IT software and telecommunications. It also performed an analysis of direct and indirect contributions of these factors to economic growth and the growth of productivity.

Possible impact on ESS and/or the end-user

The statistical community should be interested in particularly the methodological issues underlying the derivation of the specific indicators of this study.

The consortium's approach to indicator development was geared to bridging analytical gaps between current statistical practices and recent theoretical development in innovation studies. The emphasis was on producing indicators that characterised European patterns in knowledge generation as a form of 'investment' specific of the emergent new economy.

The findings from the examination of the changing structures of knowledge have three implications for statistical practice:

- The first is that they demonstrated that aggregate measures of research intensity were becoming increasingly inadequate as a way of gauging the rate and direction of growth in knowledge-related intangible assets in the new economy;
- Secondly, it is only possible to assess the economy-wide implications of the new economy by detailed examination of each sector;
- Thirdly, future research might overcome the limitations created by the 'industry specific' nature of knowledge evolution. NewKInd pioneered a series of methods and measures that offered a template for more widespread application.

Dissemination and exploitation prospects

Work produced within the project was discussed in workshops internal to the project, and in the course of three meetings organised by Eurostat in Luxembourg. This section focuses on 'external' dissemination, as follows.

- A website, located within a broader electronic hub (http://www.researchineurope.org) that aims to provide a common contact point for a number of EU-funded projects. Two project leaflets were also produced.
- The results were diffused in the French Ministry in charge of research and new technologies. There is also a link to the final report from the ministry's website, and from the website of OST. The results were presented during the meeting organized in Karlsruhe by the Fraunhofer Institute for Systems and Innovation Research (Fraunhofer ISI).

The findings from the tyre study were presented at the project meeting in Maastricht (13 March 2002) as well as at the DRUID Summer Conference on the 'New Economy' held in Copenhagen (6-8 June 2002). Copies of the findings were submitted to the Tire Association of North America (TANA), the industry journal Tire Business as well as to the Director of the Tire Practice of J D Power. Thus far, the findings were disseminated to Michelin and to Pirelli, with plans to send the report to all 10 companies involved for comments.

The findings from the Spanish pharmaceutical industry study were disseminated to firms in the Spanish pharmaceutical industry. Copies of the report on genetics and life insurance were widely circulated in the industry and the industrial association.

Concerning the macro study, the Maastricht team has been implementing a dissemination strategy based upon: scientific publication (forthcoming), book publication (forthcoming), presentation to local statistics office (Statistics Netherlands) and presentation to (regional) policy makers in Italy (Conference in Venice, 2001).

The results of the insurance, tyre and pharmaceuticals study were presented in events in Gothenburg ('The Economics and Business of Bio-Sciences & Bio-Technologies: What can be learnt from the Nordic countries and the UK?' held in Gothenburg, Sweden, Chalmers University, 25-27 September, 2002) and Aix-en-Provence (EAEPE 2002 Conference: 'Complexity and the Economy – Implications for Economic Policy').

There were the following conferences:

- April 2002. V. Acha, S. Brusoni, 'Knowledge on Wheels: new lessons from the tyre industry, SPRU draft, presented at the 2002 DRUID Conference, Copenhagen.
- October 2001. S. Brusoni, R. Cutts and A. Geuna, 'Are Insurers Able to Learn? The Case of Genetic Screening'. SPRU Draft. Presented at the 2002 Meeting of the International Schumpeter Society, University of Florida, Gainsville, Florida, USA.

Suggested Further Work

In a fluid situation of real changes, the existing frameworks and tools for understanding these changes are unsatisfactory and are likely to become increasingly inadequate. These changes cannot be easily captured by simple extensions and revisions of traditional long-term growth accounting methods that would attribute output and productivity to changes in specific 'knowledge-related' inputs. Nor can these changes be easily captured by measures derived from the traditional indicators of shifts in the knowledge base of the economy – the accumulation of patents.

Specifically, traditional R&D-based indicators fail to capture a key characteristic of the innovation process in the knowledge-based economy. Indeed, R&D statistics, innovation counts and similar measures fail to identify the commitment of (some) firms to monitor, acquire, develop and, eventually, exploit commercially new information and knowledge. These findings indicate the need for innovations in methodologies that can capture the relevant developments in knowledge structure in individual industries.

An industry-specific approach may be necessary due to the nature of the changes underway. Knowledge as an input into economic activity has no natural unit of account and the processes of transforming inputs related to knowledge creation, acquisition and use are highly complex and appear to be industry-specific. NewKInd provided ample evidence of the complexity of knowledge accumulation processes and it illustrated a number of methodologies for reducing this complexity to manageable size by considering the specific features of knowledge creation and use in individual industries.

New methods and measures should be devised to trace the changing structure of knowledge as an intangible input into economic activity and to relate knowledge-related investments to measured changes in productivity and output.

BIBLIOGRAPHY

- Brusoni S., Geuna A. and Steinmueller E. (eds.), (2007) Knowledge economies. Understanding and measurements.
- Brusoni, S., Criscuolo, P. and Geuna A., (2005) 'The Knowledge Bases of the World's Largest Pharmaceuticals Groups: What do patent citations to non-patent literature reveal?', Economics of Innovation and New Technology, 14 (5): 395
- Brusoni S. and Geuna A., (2003) 'The Key Characteristics of Sectoral Knowledge Bases: An International Comparison', Research Policy, 32, no. 10, pp. 1897-1912.-415.
- Brusoni S., Cutts R. and Geuna A., (2004) Future Imperfect. Competence-Building in the Insurance Industry in Response to Predictive Genetic Testing. In J. Laage-Hellman; M. McKelvey; and A. Rickne (eds.), The Economic Dynamics of Modern Biotechnologies: Europe in Global Trends, chapter 8, Edward Elgar.
- Brusoni S. and Geuna A., (2004) Specialisation and Integration: Combining patents and publications data to map the 'structure' of specialised knowledge. In U. Schmoch (ed.) Science and Technology Handbook, chapter 33, Kluver.

OPUS
Optimising the Use of Partial Information in Urban and Regions Systems

Timetable	1/5/2003-31/4/2005 (36 months)
Website	http://www.opus-project.org

The Consortium

Member	Role	Institutional type	Country
Imperial College of Science, Technology and Medicine (ICSTM)	Coordinator	Academia	United Kingdom
Transport for London	Partner	Public sector	United Kingdom
Katalysis Limited	Partner	Company	United Kingdom
Swiss Federal Institute of Technology	Partner	Academia	Switzerland
Facultés Universitaires Notre Dame de la Paix	Partner	Academia	Belgium
Systematica s.r.l.	Partner	Company	Italy
PTV Planung Transport Verkehr AG	Partner	Company	Germany
World Health Organisation (WHO)	Partner	International Organisation	Italy

Scope and Objectives

Problems of combining data from different sources to produce consistent estimates of underlying population parameters arise in many fields of study. There did not exist a single coherent statistical framework capable of dealing with the full range of data combination issues, including factors such as: data sources that provided both direct and indirect information on the relevant population parameters; data that were presented at different levels of aggregation; the issues raised by the aging of sample survey data and the consequent need for updating; the effect of sampling and non-sampling errors; and the opportunities presented by new data streams from IST systems themselves. Thus, the key goal of OPUS was to develop and demonstrate new methods for the coherent combination and use of data from disparate, cross-sectoral sources, and so to contribute to improved decision- making in the public and private sectors within Europe. The research was focused on developing an innovative methodology, incorporating statistical and database systems.

The specific objectives of the project were:

- To develop a generic statistical framework for the optimal combination of complex spatial and temporal data from survey, census, real-time and IST sources;
- To apply the framework to the estimation of indicators of urban/regional mobility;
- To develop the necessary metadata, database and estimation software to enable the application of the framework to mobility in London and Zurich;
- To undertake case studies and feasibility studies of the applicability of the methods to a number of related transport and health domains;
- To engage a wide range of academic and practitioner communities and domains in the development of the methods and the exploitation and dissemination of the results.

Outline of methodology

OPUS addressed the combination of data from a variety of different data sources in order to estimate various parameters. Each data source might provide direct observations on the parameters of interest or indirect measurements via intermediate or derived quantities. In addition, data sources might also differ in terms of their spatial and temporal profiles and specificity of their measurements. Moreover, the measurements from each data source would be characterized by a particular configuration of sampling and non-sampling errors. In addition, the measurements from the different data sources would necessarily have certain structural relationships to the parameters of interest and, in general, also amongst themselves. For example, aggregate volumetric count data from road links or public transport systems would be structurally related to individual person-level trip data through the aggregation of individual processes of destination, mode, route and time of day choice, and the flows on different road links would be related one to another through the topology of the network and the routeing decisions of individual travellers. Data from sources of this sort could not be simplistically pooled, since the generating process associated with each source would be different.

Statistical Methodology
OPUS's approach was to view the combination of structural relationships and measurement processes as constituting a complex statistical model linking the observed data to the underlying population parameters of interest. For example, the observed flows on road links would be a function of underlying person trip rates, combined with network topology and a variety of individual level travel behaviour choices. All the structural factors and measurement processes would have distributions and parameters associated with them. In addition there might also be constraints, both stochastic and deterministic. All the parameters would have prior distributions, allowing the formulation of a broadly Bayesian perspective. Domain knowledge would explicitly be encoded in the form of a Bayesian Belief Network. The solution of this BBN in general would result in an empirical posterior distribution, which could be used to characterize any desired property of the underlying stochastic process of interest.

Metadata Methodology
A key element of the OPUS methodology was the use of process metadata to generalize the results. In particular, much could be gained by using emerging XML standards to describe the various data structures. A variety of sources would be drawn in order to provide a process metadata characterization of OPUS's statistical methods. This would assist in the specification and implementation of the estimation software, and in the use of the results of the estimation process in the database. The characterization would also contribute to ongoing initiatives in statistical metadata.

Main results achieved

Although the underlying concepts of OPUS (Bayesian inference linked to comprehensive metadata and process metadata characterization of the inference process) were in themselves straightforward, their implementation in realistic problem settings and with realistic scales of data and model complexity posed formidable challenges. However, the project did deliver a number of case studies and feasibility studies which illustrated the potential of these ideas in various domains.

Specifically, this extensive programme of feasibility and case studies demonstrated the features of the OPUS methodology and tested and extended the OPUS tools and methods. Undertaken in London and Zurich, the case studies revealed how the OPUS methodology could be applied to a range of significant real world transport planning problems. The feasibility studies demonstrated, more modestly than the case studies, how the methods developed by OPUS might in principle be applied to a wider range of problems and domains. For this reason, the feasibility studies involved activity not just in the field of transport but also in the fields of health and planning. In the event, a number of the feasibility studies evolved into more substantial pieces of work, which produced substantive results as well as demonstrating principles of application.

OPUS dealt with two wider trends in the practice of modelling. The first trend was the acknowledgement that point estimates were only the smaller part of the modelling of results of interest. The second trend was the more systematic description and archiving of the model, its structure and estimates, and of the underlying data, which themselves might be model-derived through imputation or external systems, such as transport models. While the methods associated with the first trend were well-established in the health research domain, they were hardly so in transportation research. OPUS has rectified this shortcoming.

In addition to opening the way in general, OPUS advanced two specific methods, the Tebaldi and Dominici samplers, from small scale to large scale application and it solved the problems associated with this enhancement in scale. In the case of the Dominici sampler, the project has also developed a computationally faster Maximum Likelihood interpretation, which should help in popularizing the approach.

With the development of a systematic metadata format, StatModel, for the description of statistical, in particular Bayesian-type, model, OPUS has demonstrated how one can close this important gap. While functioning on its own, StatModel outlined how standard statistical tools could integrate the archiving of data manipulation, imputation and modelling in a comprehensive manner. The prospects for more extensive integration in the near future have become brighter.

In summary, the benefits of the OPUS project include:

- Improved estimation of detailed travel demand, using all available information;
- Avoidance of simplistic combinations of data that could give erroneous estimates;
- Indicators of data quality that could guide new data collection;
- A framework for managing data from rolling survey programmes;
- Better understanding of the role of variability and uncertainty in results and models;
- Avoidance of confusion from different, apparently conflicting, estimates of the same quantity;
- A generalized methodology for other domains of interest.

Possible impact on ESS and/or the end-user

The outputs of OPUS should enable the more effective integration of data from disparate and partial sources into a single consistent framework. Such questions are highly relevant to the challenges confronting ESS, which frequently need to deal with data collected using different means in different national contexts in order to construct indicators for policy development, monitoring and impact evaluation.

The project made a major effort to implement its findings in the context of urban, regional and national transport systems, and it explored the potential of the methodology in other areas, such as health and social policy. These are all preoccupations of the ESS.

There is marked difference across Europe in policies, implementation and available data to support the selected domains (transport, health, social policy). OPUS led the way in demonstrating how international collaboration could address this diversity through shared know-how and common approaches. Indeed, the more generic work on the methodology would be generally applicable across countries and should be of benefit universally.

Dissemination and exploitation prospects

Dissemination was through the project website, including an open discussion forum, through a project brochure, through general publicity material in the form of press releases, information sheets, overview presentations and posters. There were liaison and clustering activities with other relevant Commission/Framework and national projects, participation in relevant professional and academic conferences, workshops, meetings and exhibitions. The work of OPUS was published in suitable technical or academic conferences proceedings and journals. Drawing from its multi-disciplinary User Forum, the OPUS workshop and OPUS conference brought together representatives of relevant academic, technical and user communities from a variety of backgrounds.

In particular, the OPUS Conference was held on the 20th July 2006 in Oxford, as part of the UK Economic and Social Science Research Council's Annual Research Methods Festival. The Conference was successful in bringing the work of the OPUS consortium to the attention of the wider social science community and in providing opportunities for fruitful multi-disciplinary interactions. At least one potential additional application of OPUS methods is currently under development, as a direct result of discussions initiated at the Conference.

Concerning exploitation, the OPUS consortium agreed that:

- OPUS methods (and software code where relevant) would be public;
- There was no collective commercial plan and no restraint on partner use of OPUS;
- Individual partners might build appropriate products;
- One or more partners might pursue ongoing project opportunities;
- Ongoing work should acknowledge OPUS origins.

Suggested Further Work

The project's final report did not explicitly identify future work. However, the conclusions of the OPUS conference furnished important pointers as follows:

- There was widespread agreement that the problems of incomplete and partial data explored by OPUS were relevant to many other domains in the social and behavioural sciences;
- There was widespread support for the pragmatic Bayesian approach adopted by OPUS but the use of these methods inevitably placed more emphasis on the characterization of sources of uncertainty (sampling and non-sampling) affecting existing data sources. That would, in many cases, require new levels of rigour in the treatment of survey non-response and instrument/measurement effects;
- Linked to this, there was a need for concerted and ongoing knowledge transfer activities to ensure that policy makers and practitioners appreciated the opportunities, potential and the requirements associated with Bayesian methods in general and OPUS approaches in particular;
- Further case study applications would be of value, both in refining OPUS methods and in demonstrating their value to practitioners and decision-makers;
- Such metadata tools as StatModel was of fundamental importance, with value not only in the context of OPUS but also potentially much more widely.

Bibliography

- Axhausen, K. W. and M. R. Wigan (2003) Public use of travel surveys: The metadata perspective, in P. Stopher and P. M. Jones (eds.) Transport Survey Quality and Innovation, 605-628, Pergamon, Oxford.
- Chalasani, V.S. and K.W. Axhausen (2005) Conceptual data model for the integrated travel survey and spatial data, in R. Khan, R. Banks, R. Cornelius, S. Evans and T. Manners (eds.) Proceedings of ASC 2005 Maximising Data Value, 123-135, ASC, Chesham.
- Chalasani,V. S., S. Schönfelder and K. W. Axhausen (2002) Archiving travel data: The Mobidrive example, Arbeitsberichte Verkehrs- und Raumplanung, 129, IVT, ETH, Zürich.
- Evers, L. and D. Santapaola (2007) On the use of the IPF algorithm for combining traffic count data with missing dimensions, paper to presented at the paper presented to 86th Meeting of the Transportation Research Board, Washington D.C.
- Lindveld, C., and J.W. Polak (2004) Combining separate datasets in the OPUS project, paper presented to the 6th International Conference on Social Science Methodology, Amsterdam.

- Lindveld, C., M. Logie and J.W. Polak (2006) A Bayesian framework for the integration of multiple large scale incomplete transport data source, paper presented at the 30th Annual Universities Transport Studies Group Conference, Dublin.
- Lindveld, C., M. Logie and J.W. Polak (2006) Using Bayesian belief networks and process metadata to address large scale data integration problems, paper presented at the 23rd European Transport Conference, Strasbourg.
- Lindveld, C., M. Logie and J.W. Polak (2007) The use of Bayesian Belief Networks for the integration heterogeneous transport data sources, paper to be presented at the 11th World Conference on Transport Research, Berkeley.
- Tebaldi, C. and M. West (1998) Bayesian inference on network traffic using link count data, Journal of the American Statistical Association, 93 (442) 557-576.
- Westlake, A. (2004) Data integration through statistical modelling: The OPUS project, paper presented at the CompStat conference, Prague
- Westlake, A. (2005) Combining data and knowledge in models, in R. Khan, R. Banks, R. Cornelius, S. Evans and T. Manners (eds.) Proceedings of ASC 2005 Maximising Data Value, ASC, Chesham.
- Westlake, A. (2006) Managing metadata for statistical models, paper presented at the Statistical and Scientific Database Management Workshop, Vienna.
- Westlake, A. and M. Wigan (2006) Integrating information about complex systems: The role of meta-data in the acceptability of results from models, paper presented to 85th Meeting of the Transportation Research Board, Washington D.C.

SPIN!
Spatial Mining for Data of Public Interest

Timetable	1/1/2000 -30/06/2003 (42 months)
Website	http://www.ais.fraunhofer.de/KD/SPIN/

The Consortium

Participant Name	Role	Institutional type	Country
Fraunhofer Gesellschaft – Institut für Autonome Intelligente Systeme (FhG/AiS)	Coordinator	Research	Germany
Department of Informatics, University of Bari	Assistant partner	Academia	Italy
Institute for Information Transmission Problems, Russian Academy of Sciences (IITP RAS)	Assistant partner	Academia	Russia
School of Geography, University of Leeds	Partner	Academia	United Kingdom
Dialogis Software & Services GmbH (Dialogis)	Partner	Company (IT)	Germany
Professional GEO Systems B.V. (PGS)	Partner	Company	Holland
GeoForschungsZentrum, Potsdam (GFZ)	Partner	Public sector (research)	Germany
Manchester Metropolitan University (MMU)	Partner	Academia	United Kingdom
MIMAS, University of Manchester	Partner	Academia	United Kingdom

Scope and Objectives

The main objective of the SPIN! project was to offer new possibilities for the analysis of georeferenced data. To this end a Spatial Data Mining system was developed. It integrated state-of-the-art Geographic Information System (GIS) and Data Mining functionalities in an open, highly extensible, internet-enabled plug-in architecture. The state-of-the-art in Data Mining was advanced by adapting methods from Machine Learning and Bayesian Statistics to spatial data analysis. The state of the art in GIS was advanced by developing new methods for the visualization of spatial and temporal information. The SPIN! spatial mining system was tested and evaluated in applications to seismic and volcano data analysis and to the web-based dissemination of census data.

Outline of methodology

The SPIN! project integrated research on new data mining and visualization approaches with system development and demonstrator applications. Its main approach was that of a software project. It investigated and developed new algorithms and combined them in an integrated software environment. The system was tested on several application areas.

One activity was on technology integration. The task in this activity was on integrating the existing GIS and Data Mining modules and further incorporating the modules developed in the other activities of the project. The main output was an integrated system, which was deployed in three broad stages:

a) Research: methods were developed for Spatial Data Mining that could be added as a plug-in to the base system. Methods were selected depending on previous experiences and results of the partners:
   - Machine Learning: methods adapted were Inductive Logic Programming algorithms for the discovery of subgroups and spatial association rules;
   - Bayesian Spatial Statistics;
   - Spatial cluster detection;

   These activities all included state-of-the-art review; communication of theoretical advances; implementation and validation; application to real-world tasks; and documentation and final report.

b) The development of methods for visualization of spatial and temporal information, and for the visualization of Data Mining methods;

c) Applications: the technology developed from the above work was applied to the analysis of earthquake and volcano data. There was a second application, which was the analysis and web-based dissemination of census data from NSIs, including census data from North West England together with neighborhood statistics and deprivation data. Analysis and visualization methods developed in the project were applied to the analysis tasks relevant to official statistics. The aim was to develop a demonstrator for dissemination of census data via the internet.

MAIN RESULTS ACHIEVED

The main deliverable of the project was an integrated software environment for spatial data mining. On the methodological side, SPIN! advanced the state of the art in spatial analysis methodology along several lines:

- Exploiting the more powerful representation mechanism of first-order logic for spatial analysis;
- Exploiting the Bayesian capability of model uncertainty handling for spatial statistics;
- Exploiting advanced heuristic search strategies for spatial cluster detection;
- Exploiting temporal information in the data by new methods for the interactive visualization of spatio-temporal data.

The project's final report, which is accessible on the project website, included chapters on the following:

- The Subgroup Miner;
- SPADA (Spatial Pattern Discovery Algorithm), an algorithm for mining spatial association rules;
- A Bayesian approach to Spatial Data;
- Results for Spatial Pattern Analysis;
- Visualization of temporal data;
- The two types of tools used for the analysis of spatial data: computational tools for grid data and visualization-based interactive exploratory techniques;
- Results of applications to data-mining analysis for the Manchester and North West-England data;
- Results of an interactive analysis of the North West England data sets using the CommonGIS module;
- Results of the analysis of subgroup mining to biological data;

- Results of the second major application concerning natural hazard assessment in the areas of volcano and earthquake research;
- An appendix that included a SPIN! user guide

Possible impact on ESS and/or the end-user

The ESS does invest in GIS and ESDA using particularly census data to produce, for example, poverty maps. It is also involved in data mining and knowledge extraction from large databases. Thus both the methodology and the practical results of SPIN! would definitely be of close interest to NSIs.

The project did not have a NSI as a partner. However, two dissemination workshops were organized with the aim of targeting potential users, especially NSIs. Also, ONS of UK was involved as sub-contractor to help in defining suitable test data sets and applications and to assist in the dissemination of the technology to other NSIs.

The main application areas of the SPIN!-system were:

- Analysis and dissemination of census data;
- Analysis and dissemination of seismic data from earthquake and volcano research;
- Analysis and dissemination of environmental data;
- Geomarketing;
- Site selection.

Accordingly, primary user groups are data analysts in official statistics; urban planners; geographers, biologists; data analysts in geomarketing; data analysis for site-selection problems; data mining researchers; GIS researchers; and citizens interested in urban planning and biodiversity issues.

Among the benefits of the product for end users are:

- Interactive thematic mapping over the internet;
- Seamless integration of GIS and spatial data analysis technology;
- Advanced analysis functionality for spatial data mining;
- Integration of data access, transformation, visualization, and analysis in a single system.

Dissemination and exploitation prospects

Results from the SPIN! project were presented to many conferences, workshops and exhibitions and were documented in more than forty scientific and other publications.

Members of the consortium disseminated the scientific results through the European Knowledge Discovery Network of Excellence (www.kdnet.org). Bari organized a work shop on Mining Official Data at the PKDD'02 in Helsinki, the most important European data mining event.

A workshop on mining census data organized by Manchester Metropolitan and Manchester University and focusing on the application to North West England census data was held in 2002 in Manchester.

AIS had set up a business area "Geo-Intelligence", in which the results were demonstrated to end-users from industry and software companies, with the aim of launching joint commercial projects.

The Centre for Computational Geography (CCG) at the University Leeds produced a variety of research outputs, ranging from government funded basic research to commercially funded near-market analysis and tools.

PGS had focused on the development of Java-based GIS technology for the Internet. It had expected to expand its product offerings with data-mining capabilities as a result of the SPIN! technology.

The software was further developed in a national project called Kogiplan and in the IST project MiningMart. The SPIN! consortium had considered distributing a free version of the SPIN! System.

Suggested Further Work

The consortium had stated that it would submit further applications for research in this area. AIS would offer consultancy services partly based on the technologies developed in SPIN! Licensing agreements with commercial partners were also being contemplated.

Specifically, SPADA, FEATEX (Feature extraction from digital maps) and RUDE (the relative unsupervised discretization algorithm) were all available for further testing. For large datasets, a SPADA version more tightly integrated with a spatial DBMS was envisaged. SPADA could be profitably used in document management systems that handled both the layout and the logical structures of document images. Concerning FEATEX, the need was also for adapting the algorithms to specific spatial DBMS and integrating the software in a specific application. For large datasets, a more efficient version of RUDE, possibly more tightly integrated with a RDBMS, was envisaged.

Starting in 2003, Fraunhofer AIS successfully exploited the project results in various business areas, ranging from telecommunication to retail, utilities and outdoor advertisement. The SPIN! system was a key technology in all these projects done for major German companies. General know-how on analyzing spatial data was exploited in the business area “Geo-Intelligence”, having an estimated turnover of 2 million EUR in 2007. The SPIN! technology has become one of the most important assets for Fraunhofer AIS.

Bibliography

- May, M.: The SPIN! project – Spatial Mining for Data of Public Interest, Working Paper No. 13, UN/ECE Work Session on methodological issues involving the integration of statistics and geography, Tallinn, Estonia, 25–28 September 2001
- May, M.; Savinov, S.: An architecture for the SPIN! Spatial data mining platform, NTTS/ETK 2001
- Klösgen, W.; May, M.: Spatial Subgroup Mining Integrated in an Object-Relational Spatial Database, in: Principles of Data Mining and Knowledge Discovery: 6th European Conference, PKDD 2002, Helsinki, Finland, Lecture Notes in Computer Science, Volume 2431/2002
- Willi Klösgen, Michael May, Jim Petch: Mining census data for spatial effects on mortality. Intell. Data Anal. 7(6): 521-540 (2003)
- Turner, A. Density data generation for spatial data mining applications, Geocomputation 2000, Chatham, UK.
- Andrienko, G. and Andrienko, N. Exploring Spatial Data with Dominant Attribute Map and Parallel Coordinates. Computers, Environment and Urban Systems (Elsevier Science), special issue on GIS Research UK 2000. 2001, v.25
- Andrienko, N., Andrienko, G., and Gatalsky, P. Exploring Changes in Census Time Series with Interactive Dynamic Maps and Graphics Computational , Special Issue on "Data Mining and Statistics" 2001, v.16
- Appice, M. Ceci, A. Lanza, F.A. Lisi, & D. Malerba (2003). Discovery of spatial association rules in geo-referenced census data: A relational mining approach, Intelligent Data Analysis, 7, 6, 541-566.
- D. Malerba, F. Esposito, F.A. Lisi & A. Appice (2002). Mining spatial association rules in census data, Research in Official Statistics, 5, 1, 19-44.
- F.A. Lisi, D. Malerba (2004). Inducing Multi-Level Association Rules from Multiple Relation, Machine Learning Journal, 55, 175-210.

STATLAS
Statistical Atlas of the European Union

Timetable	1/5/2001-30/6/2004 (38 months)
Website	http://www.statlas.org/

The Consortium

Member	Role	Institutional type	Country
National Technical University of Athens (NTUA)	Coordinator	Academia	Greece
Swiss Federal Institute of Technology, Zurich (ETH)	Partner	Academia	Switzerland
Institute of Geography and Regional Science, University of Vienna (UNIVIE)	Partner	Academia	Austria
Institute of Regional Geography, Leipzig (IFL)	Partner	Academia	Germany
Agilis S.A.	Partner	Company	Greece

Scope and Objectives

Region and nation-wide statistical data are a vital information source for supporting political, socio-economic, societal and cultural decision and planning processes. However, access to reliable data is not always assured. Quick access supported by meta-databases, user-friendly browsers and data display systems is crucial for the most efficient use of statistical data sources. Moreover, the visualization of such data in thematic maps and atlases, with the support of basic analytical tools, is the most effective way to represent quantitative and qualitative aspects of spatially distributed phenomena as well as their thematic and geographical relations.

The Statistical Atlas of the European Union was an innovative example of this development. It would enable citizens of Europe to acquire almost any kind of statistical information in a fast and efficient way. The basic project objectives of STATLAS were:

- To develop an integrated electronic atlas of statistical information, covering the EU at country and regional levels;
- To flesh out the atlas with carefully designed maps and 3D displays;
- To be able to call up data and analysis for both statistical and topographical information;
- To devise a custom-made interface and a number of functions which would make it possible to collect, process, portray and compare statistical data in a spatial context;
- To display and tailor information to user needs, generating multiple advantages and offering unlimited capabilities.

STATLAS particularly focused on the cartographic side, making it a cartographic visualization environment for statistical data rather than a statistical toolkit with map output.

Outline of methodology

The STATLAS architecture included the following modules:

- STATLAS Repository: the repository is an ORACLE database (object-relational supporting spatial objects) operated on an internet accessible server and used to support both the production of maps distributed on removable media along with the client (i.e. atlas on CD-ROM's or DVD-ROM's) as well as the delivery of geographic and statistical data over live internet connections. The repository would include two conceptual areas (i.e. sets of database tables logically connected via a common NUTS table). These would be Geo Data, which is a geographical database supporting spatial objects migrated from GIS data; and a Stat Data, which is a statistical data warehouse populated with statistical data migrated from third party sources;
- Data Loader: is the module used to retrieve statistical data from external sources, process and transform these data and finally populating (or updating) the Stat Data tables;
- GIS: is an Arc/Gis installation used for the migration of geographical data to the Oracle Spatial environment. The main task of this installation would be the creation of the databases with the same content but varying resolutions. These would require:
    - The generalization of the existing data;
    - The adoption of the appropriate projection and the transformation of the geo data sets to this projection;
    - The migration to the Oracle spatial environment;
- Map Specifications: refers to both an internal STATLAS standard defining the format and structure of the xml files to be delivered to the client and the actual files of the atlas. For off-line mode, a predefined series of the latter would be distributed;
- Map Client: the atlas application, capable to read xml map specifications, would generate the map and provide user functionality, including GIS functionality for the on-line mode;
- Map Parameters: refers to an internal STATLAS standard defining the structure and format of the xml messages which convey the user request in high-level user terms. Predefined maps for the CD-based atlas would be generated through the same mechanism;
- Query Agent: is a module that is accessible by the client through the internet and capable to parse xml messages with map parameters and translate them into lower level SQL queries.

Main results achieved

- The iMap Module is the mapping engine and visualization environment of the STATLAS project. It provides the functionality that manipulates the cartographic attributes of the data and then draws the data into the graphical user interface. Plug-ins modules control the importing of the geographic as well as the statistical data. It is within the iMap module that the statistical data are attached to their corresponding geographical objects. Plug-ins allow for further functionality to be added without significant changes to the iMap module.

- The Digital Terrain Model (DTM) is a representation of the earth's surface displayed as gray-scale hill shading. The DTM can be displayed as background to the mapped statistical data to help orient the user within the map space. A DTM was compiled that covered the furthest extents of the EU. It could be used to develop a 3D model of the topographic features of the EU and also as a basis for the display of thematic information that is closely related to topography.

- The STATLAS client consists of three interrelated modules, i.e. the GUI, the communication modules and expert tools. The GUI is platform-independent, which allows easy porting to other hardware and operating systems. Communication modules mediate between the GUI and all other modules and enable the communication of the GUI with the expert tools, the statistical toolbox and the web services. Expert tools allow customized data presentation, further exploration of statistical maps in conjunction with 3D topographic models. The expert tools offer functionality as follows: 2D data analysis, regional queries and change analysis. The development would lead to a set of software functions, which would altogether form a toolbox.

- The statistical analysis tools were implemented as a separate software module, which communicated with the GUI through SOAP. The tools expected data in the same GESMES/XML format that was used by all other modules. This architecture enabled the easy extension of the toolbox with additional algorithmic modules, as well as its integration in other third party environments or products, including the possibility to operate as a web service. The statistical analysis functions of the Statistics toolbox fell in three categories: descriptive statistics, data mining and spatial econometrics.

- The statistical service aimed to provide on-line updates of statistical data to the STATLAS client and included a multidimensional data warehouse for statistical indicators and a web service for providing data in GESMES/XML format via SOAP. The web service is able to receive abstract requests for statistical data via SOAP/XML and translate them in SQL through the usage of metadata. The data repository includes a multidimensional data warehouse of statistical indicators and related attributes.

- There were many atlases covering a wide variety of thematic areas but their content varied. Therefore, there was a need for a minimum level of standardization of the atlas content. Color was used to support the identification of symbols and thus the visual (graphic) communication.

- The atlas includes a set of pre-designed statistical maps and 3D displays, as well as tools for the creation of customized statistical maps, based on statistical data retrieved ad hoc from a database accessible on the internet. The atlas also offers tools for analysis of statistical data in a spatial context. Map displays can be tailored to user needs. The project has resulted in two main components: a packaged user application (i.e. the digital atlas application) and a supporting information update service (i.e. the internet accessible repositories of statistical and geographical data). STATLAS provided an interactive information, research, analysis and education tool for specialists and experts. STATLAS contents covered more than 200 topics on the European Union's social and economic structures. The project's software components provided an integrated solution to public organizations for the dissemination and exploitation of spatially related statistical information or as a map production facility.

Possible impact on ESS and/or the end-user

The rapid expansion in information technology and the subsequent demand towards faster and more effective ways of information retrieval and knowledge extraction have generated a significant requirement towards new tools for data dissemination. Statistical data have always been a significant factor in sound decision making. However, the demand for wider and faster publicity of statistical information has not been possible to meet by traditional dissemination tools. Thus, the distinctive contribution of the STATLAS project, with its GIS-based technologies, should markedly improve the visualization and dissemination of geographic/statistical information.

More specifically, STATLAS should enable the integration of GIS-based data and statistical information for EU countries by combining geo-referenced data and spatial analysis functions in a hybrid multimedia tool. This multimedia atlas would offer a set of tools that should allow users to realize their own view on a specific subject, at different levels of interaction, which would go far beyond the limited functionality of static atlases. STATLAS comprises carefully designed maps and 3D displays supported by a custom-made interface and a number of functions. Using STATLAS, data and analyses could be called up for both statistical and topographical information.

Dissemination and exploitation prospects

The dissemination strategy plan called for a continuous stream of information about the project's outputs directed towards the target groups. STATLAS target groups for dissemination were organisations that would use the multimedia statistical atlas as a dissemination and decision-making tool. They included parties who would use the software products of the project in their everyday activities, such as the academic community; and scientists from GIS, Social Sciences, Economics, Statistics and Information Technology.

The STATLAS dissemination channels were the project website, the project brochure, a user survey and related user community, scientific publications, workshops, information days, cooperation with other projects such as SPIN and METANET, press releases and conferences. The following conferences were attended:

- "Symposium für Angewandte Geographische Informationsverarbeitung", AGIT, University of Salzburg, Austria, July 2002;
- "Spatial Information and Social Processes: European and Greek Experience in GIS" University of Thessaloniki, June 2002;
- 21st International Cartographic Conference, ICC, Durban, South Africa, August 2003.

Most of the ideas and routines developed in the framework of the STATLAS project are being utilized by the members of the STATLAS consortium in other projects. An example of this is the Austrian Conference of Austrian Planning [http://www.oerok-atlas.at/]

Bibliography

- Cooper, M., Hanewinkel, C., Specht, S. "Graphical user interfaces on the transition between information systems and interaction systems". Proceedings of the 21st International Cartographic Conference. Durban, S. Africa.2003.
- Tsoulos L. "System Design Considerations for the Development of an Electronic Statistical Atlas". Cartography and Geographic Information Science, Vol. 32, No. 3, 2005, pp. 181 - 194
- Kriz K. et. al. "STATLAS – Statistical Atlas of the European Union". Proceedings of the 21st International Cartographic Conference. Durban, S. Africa. 2003.
- Pucher A. "Open Source Cartography: Status Quo, recent trends and limitations of free cartographic software". Proceedings of the 21st International Cartographic Conference. Durban, S. Africa. 2003.
- Sykora, P.: "iMap – Mapping Engine in a Distributed Mapping Environment". Proceedings of the 21st International Cartographic Conference. Durban, S. Africa, 2003.
- Tsoulos L. , Skopeliti A. and Spanaki M.. "An XML-Based Approach for the Composition of Maps and Charts". Proceedings of the 21st International Cartographic Conference. Durban, S. Africa, 2003.
- Sykora, P. (2004) iMap - Entwicklung eines Web Map-Servers als SOAP Web-Service, Strobl, Blaschke, Griesebner: Angewandte Geoinformatik, Beiträge zum 16. AGIT-Symposium Salzburg
- Sykora, P., Schnabel, O. and Iosifescu-Enescu (2007) Extended Cartographic Interfaces for Open Distributed processing, Cartographica
- Pucherand, A. and Kriz, K. (2003) STATLAS – Web Services als Basis eines modularen, grenzübergreifenden statistischen Atlas der EU Strobl/Blaschke/Griesebner (Hrsg.): Angewandte Geographische Informationsverarbeitung XIV. Wichmann Verlag, S. 364-369, Heidelberg
- Kriz, K., Pucher, A. and Katzlberger, G. (2007) AIS-Austria – An Atlas Information System of Austria, Cartwright, W. E., Peterson, M. P. and Gartner, G. (eds) Multimedia Cartography Edition 2, Heidelberg: Springer-Verlag
- Tsoulos, L. (2005) System Design Considerations for the Development of an Electronic Atlas, Cartography and Geographic Information Science, Vol 32

- A Coruña, A. Cooper, M., Sykora, P. and Hurni, L. (2005) The Role of Cartography within Distributed Software Systems; What can we Contribute? How can we Prosper? Proceedings 22nd International Cartographic Conference
- M. Cooper, M., Hanewinkel, C. and Specht, S. (2003) Graphical User Interfaces on the Transition between Information Systems and Interaction Systems, Proceedings of the 21st International Cartographic Conference, Durban, South Africa, 1439–1448, CD-ROM
- Pucher, A., Kriz, K., Hurni, L., Tsoulos, L. and Hanewinkel, C. (2005) STATLAS – Statistical Atlas of the European Union, Proceedings of the 21st International Cartographic Conference, Durban, South Africa, 1411–1418, CD-ROM
- Sykora, P. (2003) IMAP. Mapping Engine in a Distributed Mapping Environment Proceedings of the 21st International Cartographic Conference, Durban, South Africa, 1114–1124, CD-ROM
- Sykora, P. and Pulcher, A. (2002) STATLAS – Kartographische Kommunikation von Statistik- und Geodaten via XML, Webmapping.02, Karlsruhe: Conference Proceedings, XVIII.1–16
- Tsoulos, L., Skopeliti, A. and Spanaki, M. An XML Based Approach for the Composition of Maps and Charts, Proceedings of the 21st International Cartographic Comference Durban S. Africa CD ROM

STILE

STATISTICS AND INDICATORS ON THE LABOUR MARKET IN THE EECONOMY

Timetable	1/11/2001-31/10/2004 (36 months)
Website	http://www.stile.be

THE CONSORTIUM

Member	Role	Institution type	Country
Hoger Instituut voor de Arbeid (HIVA), Catholic University of Leuven	Coordinator	Academia	Belgium
Cork Telework Centre (CTC)	Partner	Company	Ireland
Institute for Employment Studies (IES)	Partner	Public sector (research)	United Kingdom
Camire S.L.	Partner	Company	Luxembourg
Institut für Arbeidsmarkt- und Berufsforschung (IAB)	Partner	Public sector (research)	Germany
Istituto Ricerche Economiche e Sociali (IRES)	Partner	Public sector (research)	Italy
Stichting Organisatie voor Strategisch Arbeidsmarktonderzoek (OSA)	Partner	Public sector (research)	The Netherlands
Institute of Sociology, Hungarian Academy of Science (ISB)	Partner	Public sector (research)	Hungary
Central Statistics Office (CSO)	Partner	NSI	Ireland
Joanne H. Pratt Associates	Subcontractor	Company	US

SCOPE AND OBJECTIVES

STILE aimed to innovate methodologies for the statistical monitoring of the European labour market in the e-economy. In so doing, the project aimed to contribute to the efficient functioning of the European labour market and to the prevention of social exclusion. These aims included efforts to ensure that the project results were disseminated in a manner that was well targeted, appropriate, user-friendly, reliable, and timely. The project strategy set high store on the involvement of users in a systematic and direct way and on the formulation of strategies for European convergence of the statistical monitoring of the labour market in the e-economy.

OUTLINE OF METHODOLOGY

The dimensions of the methodology were:

- Improvement of existing and development of innovative methodologies and content on the statistical monitoring of the labour market in the e-economy
- Extending the coding of the CLFS for the analysis of e-work.
- Updating NACE rev. 1 to include e-enterprises and ISCO 88 to include e-occupations.
- Methodological benchmarking and a module for existing organizational panel surveys to take into account ICT-related labour market issues.
- A questionnaire module for LFS to monitor the dissemination of telework.

- Analysis of sectoral mobility in ICT, using the CLFS and administrative data.
- Profiles of ICT-related occupations, including required qualifications, training needs, type of contracts and likely future developments; benchmarking according to regional differences in ICT penetration.
- User-friendly dissemination tools.
- A project website (www.stile.be) and Newsletter, a concluding European conference targeted at policymakers, scientists, representatives of statistical bodies and all relevant users.

Main results achieved

STILE resulted in a number of reports that were brought together in one book. The book opened with an introduction by the project leader, Monique Ramioul, with An Bollen, explaining the approach of the project and its achievements. The focus turned to employer-based surveys, with a discussion of the obstacles and opportunities of convergence of such surveys at the European level. Then there were some general reflections on occupational mobility in the New Economy, followed by a comparative assessment of administrative databases and surveys in measuring labour market mobility. This was supplemented by a further comparison between the European LFS and data from the Belgian Data warehouse.

In the next section of the book, the focus shifted to an overview of some of the difficulties in capturing the extent of offshore outsourcing using existing statistics. The results of a STILE exercise designed to test ways in which e-businesses and e-occupations were classified to existing sectoral and occupational codes by NSIs were then presented.

A general critical overview of the process by which occupational profiles and the corresponding qualifications had been constructed was made, followed by a presentation of the project's work on European similarities and differences in the development of new occupations in a new economic environment.

The final section of the book focused on teleworking, specifically on the development of a standard module for measuring eWork in social surveys. This 'piggyback' survey concept was illustrated by comparing the results of surveys in the USA, UK, Hungary, the Netherlands and Ireland. Finally, results were presented of some research on eWork in Hungary intended to demonstrate the nature of transitional economies.

Possible impact on ESS and/or the end-user

The STILE project made three main contributions to a better statistical monitoring of the labour market. First, it contributed to a general consciousness of the challenges posed by the New Economy to the statistical community.

Secondly, STILE did not investigate the substantive dynamics of the New Economy. Instead, the focus was primarily on how to measure the socio-economic trends concerned. In this respect, the project team aimed at improving existing statistics and indicators, most of which were indeed designed in such a way that they could in principle allow for international comparative research. The outcomes of STILE included a critical assessment of several aspects of the ESS in a way that should impart an impetus to the production of more reliable and comparable statistics and indicators.

The third statistical contribution of the project was in a critical reflection on the dynamics underlying the realization and the innovation of statistics. The project team's experience with involving users had offered insights into the forces facilitating and inhibiting an adequate statistical system. It was demonstrated that the collaboration between academics, statisticians, policymakers, organizations and other users of statistics could contribute to more useful, comprehensive and reliable statistics. This broad coordination was essential particularly for the innovation in statistics. The STILE consortium had also found that it was difficult to realize such ambitious goals within the confines of an external project-based approach, with fixed duration. Statistical innovation was only possible if it was supported by a comprehensive and coherent statistical policy arising from within NSIs, but with close collaboration from other stakeholders in the New Economy.

The specific key findings that the ESS should consider for exploitation include the following:

Measuring telework

The STILE project developed a module for collecting reliable and comparable data about telework. The questions can be inserted into existing surveys. Because there was no agreed definition of 'eWork' or 'telework,' STILE recommended asking a series of objective questions that pinpointed the location and frequency of work conducted at a distance from the employer. Telework could be defined at the time of data analysis by selecting criteria that were meaningful to the research objectives. A substantial benefit is that questions can be added through an ad hoc module to an existing survey, such as a labour force survey (LFS), or to other general surveys seeking to address changes at work, at whatever level (sectoral, regional, national). The responses can then be cross-tabulated with all other information in the survey.

Labour market mobility

Through a comparison of the results obtained from administrative registers and the Labour Force Surveys, STILE studied the measurement of labour market mobility in Finland and other Nordic countries. The analysis revealed some unexpectedly large differences between the results from the two sources. They were different in a third of cases. The research recommended the harmonization of occupational coding; improvement of ISCO; and the harmonization of wording of questions in national LFS questionnaires.

STILE also measured labour market mobility specifically in the ICT sector using the CLFS and Belgian administrative data. The study showed that the main advantage of the administrative source was its exhaustive character, enabling detailed breakdowns, while the strength of surveys such as the LFS was that they yielded qualitative information to complement the quantitative data within administrative databases and that they could provide internationally comparable indicators.

Classifications and coding of business and occupations

STILE looked at the ways in which industrial sectors and occupations were classified and coded, with a particular focus on 'knowledge-based' industries and occupations. It found that a much more differentiated classification of business services was required if new developments, such as the growth of outsourced call centres, or offshore software supply, were to be tracked. The challenge is not simply to develop new classification categories but also to change the mindset of those who collected and coded economic data. STILE coding experiments demonstrated dramatic differences in the ways in which 'eEconomy' activities were coded. There was less than the desired unanimity amongst coders in making coding decisions relating to the NACE classification system. The diversity of these results pointed to the business services sector as a 'can of worms' requiring serious analytical attention if the information economy is to be monitored consistently and benchmarked internationally.

STILE confirmed that job titles or short descriptions drawn from job advertisements did not provide a good basis for consistent classification in terms of ISCO. Quite frequently, similar titles denoted very different occupations. Without improvements in the quality and comparability of occupational coding, it was unlikely that progress towards a knowledge-based economy could be monitored adequately.

Concerning occupations and skills specifically, STILE focused on three aspects: the characteristics of new occupations; the opportunities and limitations of existing occupational profiling methods for researching new occupations; and concrete research questions that called for international comparative research. Three key trends were advanced as forming the context for occupational change: increasing penetration of ICTs, knowledge as a production factor and an increasing need for flexibility.

Dissemination and exploitation prospects

- National and European user groups on telework to develop indicators and module questions to measure different forms of telework and their impact on the quality of working life.
- Dissemination of ad hoc modules for questionnaires on telework and on ICT-related HRM aspects for relevant existing organization surveys.
- 6 Workshops
- National conferences with employers and employee organizations on the required skills and profiles of ICT-related occupations.
- Website www.stile.be.
- 6 project Newsletters reporting the progress of work and disseminating the project's outcome.
- European conference on the state of the art on the labour market in the e-Economy.

Suggested Further Work

STILE appealed for:

- Stronger systematic and structural co-operation between academics and statisticians, with Eurostat having a crucial role to play in stimulating user groups in all Member States and at the European level. This was particularly so when it came to the innovation of the existing statistics and the creation of new tools and instruments to take account of current changes in the New Economy.
- Improved communication channels and mechanisms to include users in statistical work, whether these were policymakers, 'the citizen' or enterprises involved in research as survey respondents or as research users. It was stated that the dialogue between policymakers and scientists was generally intermittent and often characterised by mutual distrust.
- Streamlining the ESS, which appeared as an 'over'-organised system, dominated by rigid and time-consuming procedures. Despite the thrust STILE had given to dissemination, an effective uptake of the project's results was not at all obvious in the short-term
- Deepen the insights into changing inter- and intra-organisational divisions of labour (networking and outsourcing), organisational changes (production concepts), changes at the workplace (eWork), changes in labour market behaviour (mobility) and work force composition (new businesses and occupations). Also improved knowledge of the impact of these phenomena.

Future research possibilities

The most important spin-off research activities from the STILE project is that, in two successive projects, there were further contributions to the convergence, harmonization and innovation of European organization surveys.

In the WORKS project, Work Organization and Restructuring in the Knowledge-based Society, which was a four-year Integrated Project in the Citizens and Governance Programme of FP6, WORKS included further research on these surveys in order to enhance the comparability of survey design, questionnaires, indicators and concepts to measure changes at work. In the frame of this project, the on-line Digital Toolkit on organization surveys was updated and enlarged. This digital toolkit (www.worksproject.be) enables researchers to investigate how different concepts relating to changes in work are measured in different existing organization surveys. The project also organized two expert workshops on this issue. The first (February 2006) focused on the comparison and assessment of different concepts and indicators on several issues related to changes in work. The second workshop (March 2007) brought together several experts, including experts from the ESS, to discuss future challenges for organization surveys, e.g. the design of international establishment surveys, surveys using linked employer-employee data and surveys that included inter-organizational relationships and their impact on work and organization.

The MEADOW project builds up on this methodological benchmark, which was developed in WORKS dealing with data available at the European level. That is a coordination action funded by the European Commission under Priority Seven (Citizens and Governance) of FP6. It aims at setting out Guidelines for collecting and interpreting harmonized data at the European level on organizational change and work restructuring and on their economic and social impacts. These Guidelines will constitute a first step towards implementing a harmonized European survey instrument.

Main publications stemming from the project

- Ramioul M., Huws U. & Bollen A. (eds.), Measuring the Information Society, STILE report, HIVA-K.U.Leuven, Leuven (HIVA publication n° 959,).
- Bellmann L. & Promberger M. (eds.) with Ester P., Maenen S., Ramioul M., Román A. & Van Hootegem G. (2004), Towards convergence? Current state and future ways of establishment based ICT- and labour market monitoring in Europe, STILE report, HIVA-K.U.Leuven, Leuven (HIVA publication n° 863
- Bollen A., Huys R. & Ramioul M. with della Ratta F., Ester P., Makó C., Oteri C., Pratt J., Román A., Tamási P., Tót E. & van Poppel H. (2004), Understanding occupations in the Information Age, STILE report, HIVA-K.U.Leuven, Leuven (HIVA publication n° 914,
- Huws U. & van der Hallen P. (eds.), with Bertin I., Koltai T., Promberger M., Tickner N. & Verlinden R. (2004), Opening the black box. Classification and coding of sectors and occupations in the eEconomy, STILE report, HIVA-K.U.Leuven, Leuven (HIVA publication n° 905,
- Oteri C. & della Ratta (eds.) with Altieri G., Bates P., Bertin I., Bollen A., Huws U., Lakatos J., Makó C., Pratt J., Ramioul M., Tamási P. & Tickner N. (2004), Working at a distance. How to know about it?, STILE report, HIVA-K.U.Leuven, Leuven (HIVA publication n° 864
- Stimpson A. & Tielens M. (2004), Mobility in the eEconomy, STILE report, HIVA-K.U.Leuven, Leuven (HIVA publication n° 906)

STING
Evaluation of Scientific and Technological Innovation and Progress in Europe, through Patents

Timetable	1/11/2000-31/10/2002 (24 months)

The Consortium

Member	Role	Institutional type	Country
Computer Technology Institute (CTI)	Coordinator	Public sector (research)	Greece
Quantos SARL	Partner	Company (IT)	France
Ecole Polytechnique Federale di Lausanne (EPFL)	Partner	Academia	Switzerland
National Statistical Service of Greece (NSSG)	Partner	NSI	Greece
Industrial Property Organisation (OBI)	Partner	Public sector	Greece
Istituto Nazionale di Statistica (ISTAT)	Partner	NSI	Italy

Scope and Objectives

The STING project set out to design and to develop efficient and novel methodologies and tools for the analysis of existing information relating to European technological innovation. The basis was patents' data. The main objectives of the project were the:

- Development of enhanced methodologies for the analysis and processing of patents data, stored in databases at national and pan-European levels;
- Development of a reliable methodology for measuring technological innovation, from which indicators would be produced on a regular basis;
- Improvement of the quality and the timeliness of the resulting information and indicators;
- Exploitation of the latest IT developments for gaining fast access to universally available patent databases;
- Development of a computer-assisted system for the analysis of patents data.

Outline of methodology

The main technical activities were:

- User requirements: the capture and study of those users' needs that the system should cover; and the conduct of a detailed survey of the current situation in order to identify the problems in existing statistical methodologies concerning the collection and analysis of patent data. A questionnaire with open and closed questions would be designed in order to specify the requirements for patent data more accurately. The use of such a questionnaire would enable users to express their knowledge on the subject in full detail. Also, software on the market that was specialized in the analysis of patent data would be studied for a better understanding of user needs.

- Statistical analysis and development of the statistical methodology: a new statistical methodology for the analysis of patent data would be devised and the resulting indicators would be used to measure S&T progress. The statistical methodology would be based on two different approaches: (a) several kinds of outputs would be produced such as matrices and ready-made reports or different types of graphs, and (b) the use of textual techniques as well as more complex statistical methods such as factor or cluster analysis;
- Functional specifications: this task would define in detail the functional specifications and the system architecture based on user requirements and on the results of the proposed statistical methodologies;
- Application development: the appropriate system modules and user interfaces would be developed and user guides and manuals prepared;
- Validation and verification: the statistical methodology and the resulting indicators would be tested and the feasibility of the system's services would be validated;
- Dissemination and exploitation as detailed below.

Main results achieved

A particularly innovative result of STING was that it permitted the identification of technological trends and innovation in an easy and comprehensible way. Moreover, the production of innovative visualization based on patent information facilitated a rapid user grasp of technological developments. The stability and accuracy of the results were ensured through the application of selected statistical tests. The system enabled the user to control the stability of the results through such techniques as multiple factor analyses and the boostrap, with the presentation of these techniques in a simplified way not requiring deep statistical knowledge. Thus, the main specific output of the project was indicators for the measurement of technological and scientific progress. Other outputs were:

- A user-friendly software for the analysis of patent data and for the production of technology indicators, with detailed user guides and manuals;
- The development of efficient and innovative methodologies for the exploitation of patent data. In more detail, the methodology could analyze all the information describing a patent (e.g.: IPC codes, titles, abstracts, assignees etc.) while the statistical procedure consisted of a linguistic preprocessing step, correspondence analysis and cluster analysis, which enabled the creation of homogeneous groups of patents involving similar technologies. Bootstrap analysis was used to test the robustness of the results of the analysis;
- A market survey which summarized the results of the comparison between different software tools that specialized in the analysis of patents;
- A user-requirement analysis which presented the answers of a questionnaire addressed to European companies concerning the analysis of patent data;
- A vortal server for the dissemination of the project's results and other information concerning the patent analysis;
- The actual results of the analysis of real patent data;
- Dissemination and exploitation arrangements including a business plan.

Possible impact on ESS and/or the end-user

The ESS should gain insights through STING about the appropriateness and limitations of patent data for the measurement of innovation in the new economy.

The European Commission also should gain access to more accurate data regarding technological trends in Europe, scientific activities in different EU regions, the competition inside Europe and those sectors that faced strong competition from non-EU companies. Based on these results, the Commission should be able to take informed decisions and to plan the future in a way that would lead to an improvement in the position of European citizens.

Dissemination and exploitation prospects

An outline of the dissemination and exploitation activities is:

- Production of a project website, which served as a vortal with information on patent databases, statistical analyses, similar projects and publications;
- The design of information leaflets and posters; project presentation in PowerPoint and CD-ROM;
- Establishment and functioning of a core of interested users; user-focus group meetings were held, with publication of their proceedings;
- Development of an exploitation plan and a draft business plan for the future commercialization of STING software;
- Publication of the results of analyzing real patent data;
- Participation in international seminars and conferences;
- Dissemination of the results of the project in the website of the Artificial Intelligence Lab of EPFL;
- Presentation of STING in all IPR (Intellectual Property Rights) from 2001 to 2006;
- Presentation of the project and the adopted methodology in the postgraduate courses of EPFL, which had attracted interest from industry;
- IPR courses and presentation of the project in the Open University and Greek universities in collaboration with OBI;
- Enlargement of the User Group.

Future Research Challenges

An outline of the future research challenges would include:

- Integration of advanced search techniques for mining innovation in patents;
- Extension of the methods developed to citations analysis;
- Improvement of the system performance with new data structures;
- Extension of the methodologies developed to include forecast mechanisms.

Bibliography

- Penelope Markellou, Angeliki Panayiotaki, Spiros Sirmakessis, Antonis Spinakis, Athanasios Tsakalidis, “Using Hierarchy to Extract Innovation: the Use of Patents in Clarifying Innovation”, New Techniques and Technologies for Statistics, Exchange of Technology and Know-How (NTTS & EKT), pp 881?882, Crete, Greece, June 18-22, 2001.
- Jean-Cédric Chappelier, Vivi Peristera, Martin Rajman, Florian Seydoux, Antonis Spinakis, “Evaluation of Scientific and Technological Innovation Using Statistical Analysis of Patents”, 6th International Conference on the Statistical Analysis of Textual Data (JADT 2002), Saint-Malo, France, March 13-15, 2002.
- Konstantinos Markellos, Penelope Markellou, Giorgos Mayritsakis, Georgia Panagopoulou, Katerina Perdikouri, Spiros Sirmakessis, Athanasios Tsakalidis, “STING: Evaluation of Scientific & Technological Innovation and Progress in Europe Through Patents”, New Frontiers of Statistical Data Mining and Knowledge Discovery and E-Business Conference, Knoxville, Tennessee, USA, 22-25 June 2002.

- Spiros Sirmakessis, Konstantinos Markellos, Penelope Markellou, Giorgos Mayritsakis, Katerina Perdikouri, Athanasios Tsakalidis, Georgia Panagopoulou “STING: Evaluation of Scientific & Technological Innovation and Progress” in “Statistical Data Mining and Knowledge Discovery”, Edited by Hamparsum Bozdogan, pp 549?570, Chapman & Hall/CRC Press, 2003.
- Athanasios Tsakalidis, Konstantinos Markellos, Katerina Perdikuri, Penelope Markellou, Spiros Sirmakessis, George Mayritsakis, “Knowledge Discovery in Patent Databases”, in the proceedings of the 11th ACM Conference on Information and Knowledge Management (ACM-CIKM 2002), pp 672?674, November 4?9, 2002, McLean VA, US.
- Athanasios Tsakalidis, Konstantinos Markellos, Katerina Perdikuri, Penelope Markellou, Spiros Sirmakessis, George Mayritsakis, “A multivariate analysis and visualization system for analysing patent data and producing technological and scientific indicators”, in the proceedings of the International workshop on Computational Management Science, Economics, Finance and Engineering, 28-30 March 2003, Limassol, Cyprus.

VITAMIN S
Visual Data Mining System

Timetable	1/1/2001-31/12/2003 (36 months)

The Consortium

Member	Role	Institutional type	Country
ATKOSoft SA	Coordinator	Company	Greece
Università degli Studi di Napoli Federico II (DMS)	Partner	Academia	Italy
Office for National Statistics (ONS)	Partner	NSI	United Kingdom
Istituto Nazionale di Statistica (ISTAT)	Contributor	NSI	Italy

Scope and Objectives

Data visualization plays an important role in statistics. Statistical visualization tools allow users to extract knowledge from data, using both exploratory and confirmatory approaches. Statistical visualization is not the simple graphical representation of numerical analysis results but is a new way to process and analyze data. The data analysts/user can learn through the use of the available tools by interacting with the data, using for example rotation, scaling, translation and transformation in an attempt to identify data patterns, relationships and anomalies. The VITAMIN S project dealt with the development of visualization tools for analyzing large survey data and time series data with the aim of discovering data patterns, trends, associations, clusters and outliers.

The central practical aim of the project was to develop an innovative software for statistical visualization, combining existing graphical methods with the result of researches conducted by the consortium in order to match the needs of NSIs with respect to the treatment of data from large surveys and from time series. Classical statistical methods involving large computational efforts could be ineffective. Instead, visualization methods, which rely on the eyes-brain system, could be more productive.

The objective of the project was the development of a VIsual daTA MINing System (VITAMIN-S) for statistical visualization with a data mining perspective matching the needs of NSIs as well as those of external users such as Marketing Research Companies. The focus was on:

- Unstructured data (large surveys)
- Structured data (time series):

The system would provide linking between visualization methods so that the full potential of several such methods could be exploited during the analysis process. The specific objectives were:

- Enhancement of classical graphical data analysis methods by incorporating dynamic and interactive techniques;
- Implementation of innovative visualization methods;
- Integration of the implemented graphical and visualization methods into a unique and homogeneous environment for a genuine Visual daTA MINing System by: offering classical and innovative visualization tools; providing an interactive framework to explore multivariate data; linking displays, so that the full potential of several visualization methods could be realized during the analysis process.

Outline of methodology

The project would have two phases:

Phase A:

- Identify the needs of NSIs and other end-users in relation to data exploration using visualisation methods;
- Analyse existing literature related to graphical methodologies in statistics and assess s/w tools in order to discover possible weaknesses in existing visualisation techniques;
- Propose and define the functional specifications of enhanced classical graphical methods as well as propose innovative visualisation methods. This would be done on the basis of such criteria as:
    - User group requirements;
    - Resources available;
    - Suitability for future commercialisation.
- Design a prototype demonstrating one of the selected visualisation methods.

Phase B:

- Evolution of the prototype into the proposed full-blown VITAMIN S system;
- Validation and verification of project objectives;
- Exploitation of results.

Main results achieved

The major actual achievement was the VITAMIN S software system that offered visualization methods for exploring unstructured data (large surveys) and structured data (time series). Specifically, for unstructured data the following visualization methods were offered:

- Factorial planes for units (Principal Component Analysis (PCA)) and Multiple Correspondence Analysis (MCA));
- Factorial planes for variables (PCA and MCA);
- Multiple factorial planes for units and variables (PCA and MCA analysis);
- Dendrogram (Cluster analysis);
- Tree-View Representation (Cluster analysis);
- Clusters in Factorial Planes (Cluster analysis);
- Clusters in Multiple Factorial Planes (Cluster analysis).

For structured data the following visualization methods were offered:

- Time series preliminary analysis;
- Best aggregate time series;
- Time series PCA;
- Exploring ARMA models with multidimensional data analysis;
- Time series clustering and comparisons;
- Time series outliers identification method;
- Dynamic PCA.

Finally several statistics graphs were made available such as the box plot, bar charts, pie charts, histograms, scatter plots, scatter plot matrix, parallel coordinates and correlograms.

The capabilities of the above achievements can be elaborated as follows. Regarding the large surveys end-user requirements, the visualization tools of factorial methods and cluster analysis in VITAMIN S would:

- Aid the interpretation of factorial planes in representing units specially when the number of units is very large;
- Aid the interpretation of factorial planes in representing variables using an innovative representation of the points on the plane graph and allowing the user to explore the characteristics of a particular point on the factorial plane;
- Offer a global view of the results of the factorial analysis through a simultaneous representation of the points of more than one plane graph in multiple visualization;
- Visualize the observed units and/or variables along different times/occasions using a simultaneous representation;
- Plot the results of a cluster analysis on the factorial planes offering interaction capabilities to filter out the points, and to compare and explore clusters.

Regarding the time series end-user requirements, the visualization tools developed in the project would:

- Graphically identify possible outliers in time series;
- Find and visualize the best compromise aggregate series of a set of component series through an innovative approach that would allow the user to evaluate how the component series would explain the salient features of the aggregate series; which component series could cause trend growth in the aggregate series; which component series contributed most to the volatility of the aggregate series; and what was the relative role played by each component in the definition of the aggregate variable;
- Equip the user with a set of tools for visual model identification and for model comparison and validation in a Box and Jenkins classical framework;
- Plot simultaneously a large number of time series as points on a common reference space in order to visualize the differences between series and also to display the corresponding values of ARMA coefficients.

Possible impact on ESS and/or the end-user

By enabling users to conduct multivariate statistical analysis in an efficient and cost effective way, the project results should greatly contribute to the optimal use of information and thus should allow the modernisation of the statistical production system, the reduction of costs and greater timeliness of results. Moreover, the project would enable many groups of users to analyse the abundance of data currently available so as better to provide products and services as well as to detect and adapt quickly to a rapidly changing socio-economic environment. VITAMIN S has developed statistical methodologies and techniques that might enhance the capacity of the individual and organisation to acquire insights into several domains. The final visualisation product could help the non-expert user substantially more than the heavier statistical programmes and tools that are available.

Dissemination and exploitation prospects

The target audience comprises organizations and companies applying statistical analysis to large databases, and placing high store on the validity of the statistical results and on the selection of the appropriate statistical software. Users of such statistical software usually belong to the following categories:

- National and international organizations responsible for the collection, processing and dissemination of statistical data;
- Public and private organizations performing data analysis on, for example, economic and social issues;
- Data producers;

- Large and medium size enterprises seeking to be competitive in the market;
- Universities and R&D organizations using the statistical methodology for education or research.

The Consortium would promote the project's outputs to data producers, data analysts/consumers/users, employers of personnel working in the statistical sector, insurance companies, financial analysts, R&D institutes, universities and other relevant groups.

The channels of communication were the project website, the project brochure, press releases, workshops, training courses, conferences and seminars, and publications of papers in scientific journals such as Computational Statistics and Data Analysis; Computational Statistics (Physica Verlag); Review of Official Statistics (ROS, Eurostat); and the Revue de Statistique Appliquée (SFdS). An exhaustive list was provided in the Dissemination and Use Plan. Some of the conferences attended were:

- Scaling and Cluster Analysis, 31st Spring Seminar, Zentralarchiv, Universität zu Köln, Germany, February 2002;
- The 8th International Vilnius Conference on Probability Theory and Mathematical Statistics, Vilnius, Lithuania, June, 2002;
- Joint Statistical Meetings, Hilton and Sheraton, New York, August, 2002;
- 24th European Meeting of Statisticians, Prague, Czech Republic, August, 2002;
- IAOS Conference, Official Statistics and the New Economy, August 2002;
- International Statistical Institute, 54th Biennial Session, Berlin, Germany, August 2003;
- International Statistical Institute, 55th Biennial Session, Sydney, Australia, 2005.

The consortium planned:

- That the software developer (ATKOSoft) would be ultimately responsible for the development, evolution, customization and marketing of the VITAMIN S Software;
- That the partners would be Value Added Resellers of the software responsible for marketing it to end customers in their designated territories with the support of the software developer;
- That the demo sites / statistical organizations of the consortium would provide test bed sites for the VITAMIN S software and whenever requested or possible would offer consulting services that would aid the software marketing to customers.

Also, there would be attempts (a) to form alliances with other countries so that the modules developed were able to work as add-ons to existing statistical software, (b) to reach agreement with companies that produced statistical software in order to provide to the VITAMIN S modules the channels that these companies had already established in the market; (c) to link up with the secondary market of software suppliers and developers that might show interest in developing plug-in modules for Vitamin-S and promote and sell those plug-ins further to their own customers.

SUGGESTED FURTHER WORK

Future research opportunities will include:

- The identification of particular visualization parameters per problem class. This should speed up the visual identification of patterns, without much experimentation, to the application of various parameters in the methods provided;
- The implementation of more methods in the systems and allowing for parallel interoperation of them.

VL-CATS
Virtual Library for Computer Assisted Training in Statistics

Timetable	1/1/2000-31/5/2003 (41 months)

The Consortium

Member	Role	Institutional type	Country
European Dynamics S.A.	Coordinator	Company (IT)	Greece
Quantos S.a.R.L.	Partner	Company (IT)	France
Training of European Statisticians (TES) Institute A.S.B.L.	Partner	Training	Luxembourg
Conservatoire National des Arts et Métiers (CNAM)	Partner	Academia	France
Universitat Politècnica de Catalunya (UPC)	Partner	Academia	Spain
Università degli Studi di Urbino (ECONURB)	Partner	Academia	Italy
Centre d'Etudes des Sciences Appliquées à la Gestion (CESAG)	Partner	Academia	France

Scope and Objectives

The scope of the VL-CATS project was to set up an Integrated System supporting e-learning of specific educational programmes relating to official statistics and providing many relevant on-line services. The system would offer tools enabling shared access to class resources in a controlled, role-based environment as well as a Virtual Library providing Internet users with reference material and other publicly available statistical resources.

The main goals of the VL-CATS project were:

- Provision of access to teaching and educational material available over the Internet, with special focus on issues relating to Official Statistics;
- Exploitation of existing experiences in order to advance on the features offered by existing tools;
- Provision of the option to individuals to publish their own courses without requiring specialized IT knowledge;
- Creation of a critical mass of experts sharing the same background and knowledge, with special focus on improving the skills of official statisticians of the new member states, candidate and potential candidate states;
- Provision of the possibility for statisticians and institutes from non-EU countries to benefit from the knowledge in EU countries and so to converge to the more advanced statistical systems in EU countries;
- Enlargement of the user community and penetration into new market sectors.

For the attainment of the above-mentioned goals, the following objectives were set for VL-CATS:

- To create a virtual library in official statistics with reference material and other publicly available statistical resources;
- To develop structured electronic courses in areas of interest in official statistics;
- To define and implement standards with respect to the structure of teaching material to enable the assembly of complete courses out of discrete modules;

- To develop a system that would host the virtual library and deliver training courses in a controlled environment, enabling selective access to system resources based on user roles (tutor, student);
- To develop a service for the update and quality assurance of VL-CATS; this service would enable the provision of training courses over the Internet in virtual classes;
- To further develop groupware tools that would support enhanced distance training features, such as a shared library, course templates, private areas for students, tools for preparation of exams and quizzes and tools for class administration.

Outline of methodology

The main axes of the methodology were:

- User requirements analysis: specifying the main system services and providing guidance on the structure and layout of the site contents. This was based on an extensive user requirements information collection performed by the VL-CATS consortium. It would cover the functionalities of the VL-CATS web system as well as the requirements on site contents;
- State-of-the-art analysis and selection of tools: based on user requirements, this activity addressed mainly technological issues and state of the art in virtual libraries and distance training with a twofold objective: first, to provide content that would be useful, functional and attractive; secondly, to identify tools and techniques that might be used for implementation of the system functionality;
- Systems design and definition of service: the service that would be provided by VL-CATS was defined as well as the design of the software that would implement system services and the structure of course material;
- Development, testing and integration: implementation of the software tools for VL-CATS based on the design specifications resulting from the previous activities;
- Production of material and population of VL-CATS site;
- Evaluation and assessment: subjecting the system to extensive testing in a real environment. Various scenarios for testing would be drawn up, including testing of course modules by individual students, organisation of (virtual) professional training seminars and delivery of courses;
- Relations with user community – dissemination and exploitation, as detailed below.

Main results achieved

VL-CATS participated in a wider scenario of systems supporting e-learning and e-working business processes, specifically but not exclusively targeting official statistics. VL-CATS aimed at producing a commercial tool to provide e-Learning and e-Training services in a secure and highly collaborative web-based context.

For the purposes of proving the project concepts, a suite of software modules and background services were created ready to be used for supporting electronic courses in the field of official statistics. In parallel, seven original electronic courses were prepared. In order to maximize the value-added of using the web as a medium of delivery, compared with printed material, five major Simulators as well as a number of smaller Applets were implemented and integrated within the courses.

Even if the idea of the VL-CATS project was born out of the needs of official statistics, a thorough examination of user requirements had resulted in the design of an Integrated System that could fulfill its role of e-Learning, e-Training and e-collaborating S/W Platform in any field of knowledge. The main functionalities of the system are:

- User-friendly interface – navigation tools;
- Easy to handle publishing tools;
- Management of course delivery;
- Access control – support for group project;
- Calendar – notepad;
- Conference system;
- Student tracking;
- Grade maintenance – distribution;
- Auto marking of exams;
- Personal work space for tutor and students.

The project has generated the following services:

- Quality assurance of site resources and continuously updating site contents with new material; appropriate templates were defined for the training modules and other site contents; tools were provided for testing conformity of new material to templates;
- A central node for the support of distance training courses that would enable tutors to compose a course out of existing modules and deliver the course in a virtual distributed classroom;
- Facilities for the administration of courses; follow-up of students progress and (semi-)automatic generation of quizzes;
- Facilities for students visiting classes, such as a personal 'student locker' to store personal notes and upload class assignments;
- Administration of access rights to the various VL-CATS resources based on roles (tutor, student, site administrator).

The end service of VL-CATS was organised around a fully functional service centre that provided the infrastructure and support for the delivery of professional training courses in official statistics. This service could be expanded incrementally to include areas of interest to other professional branches.

Possible impact on ESS and/or the end-user

It is expected that the outputs of the VL-CATS should have an important use to the ESS which is engaged in in-house statistical training almost continuously. Subject to the further development of its contents, VL-CATS should contribute substantially to the enhancement of the expertise of the European statistician, keeping him/her up-to-date with the current best practice in EU official statistics. The completeness and automation degree of the VL-CATS System should permit a significant reduction of training and educational costs, and saving on travel time.

The VL-CATS System should allow the utilisation of multimedia technologies during the hosted learning procedures, thus presenting a significant value-added of the Internet on the quality of the material and the resulting course itself. It is expected that the communication and subsequent interactivity between peers of the VL-CATS System will be enhanced. The system will allow its users to get to know each other (even virtually at the beginning) and reinforce interaction and cooperation amongst them.

DISSEMINATION AND EXPLOITATION PROSPECTS

A website and a project leaflet were established. There was the intention to establish also a user group and a user forum, to enhance visibility of the project and to create awareness within the target user community as well as to participate in concentration and clustering activities.

VL-CATS built a Virtual Library that has evolved to a reference website on official statistics that should stimulate dissemination:

- Multimedia training modules for professional training that might be used to compose courses in subjects of interest to official statisticians;
- Links to journals, universities, free statistical software on the web, universities, NSIs, companies that produce statistical software and/or provide statistical databases, and announcements of conferences and other events;

X-STATIS
Extended Statistical Information System

Timetable	1/1/2000-31/12/2002 (36 months)

The Consortium

Member	Role	Institutional type	Country
ATKOSoft S.A (ATKOSoft)	Coordinator	Company (IT)	Greece
Quantos S.A.R.L. (QUANTOS)	Partner	Company (IT)	France
Office for National Statistics (ONS)	Partner	NSI	United Kingdom
Conservatoire National des Arts et Métiers (CNAM)	Partner	Public sector (research)	France
BVA	Partner	Public sector	France
National Statistical Service of Greece (NSSG)	Partner	NSI	Greece

Scope and Objectives

The aim of the X-STATIS project was to make statistical analysis feasible and easy for small to medium sized enterprises and similar organizations that either could not afford to hire statistical experts or needed more effective, user-friendly methods for quickly and correctly analyzing statistical data. Thus, X-STATIS would address the increasing need for a user-friendly statistical information system that could be used by non-experts to deliver searching data analysis and to assist in the interpretation of the results. Specifically, a software system would to be developed to cover the following objectives:

- Assistance and/or guidance to non-expert users in the selection and application of the most appropriate data analysis method in particular situations;
- Provision of help with the interpretation of the results through better guidance to understanding and evaluating the outcome of the analysis;
- Prevention of misuse and misinterpretation of statistics by non-experts;
- Flexibility in the adaptation of the software to several statistical databases held by data providers;
- Provision of a standard interface allowing additional methods to be integrated into the system;
- Full utilization of the most modern and advanced techniques and state-of-the-art IT methods.

Outline of methodology

The very broad steps in the methodology were:

- Capture of user requirements;
- Functional specifications;
- System development;
- Validation;
- Exploitation.

Having certain attributes, data need special processing. Therefore, by defining rules for decisions and identifying constraints and by taking account of the statistical attributes, the system could be designed to guide the user in the selection of the most appropriate method. The extended statistical information system would consist of the following modules:

- The main module having a standard application programming interface (API) that would allow integration of statistical analysis modules;
- Statistical Advisor Module, which would be an open and parameterised module that interacts with the user and other system components in such a way that it hides the complexity required for data analysis. It guides the non-expert user smoothly towards the selection of the appropriate statistical method and it assists in the interpretation of the results obtained;
- Guidance Tools that would allow expert data analysts to create scenarios to be followed by the Statistical Advisor Module and by non-experts;
- A library of statistical methods; all the statistical methods would use a standard API that would allow classification of the output in order to be used as input into other applications;
- Scripting module for automated analysis of repetitive situations;
- Analysis maps showing the overall structure of data analysis sessions;
- Statistical analysis methods, which could be used by the Statistical Advisor Module for guidance in the selection of data and of methods, and in the manipulation of results;
- A metadata wizard, which would be a module that guides the data providers in the description of the metadata database parameters characterising the dataset variables.

Main results achieved

The project actual outcome was the X-STATIS software system consisting of the 8 modules mentioned in the preceding section. Looked at from another angle, it comprised two modules: the main application utility and the expert utility. The main application utility is capable of:

- Guiding non-expert users in the selection of the most appropriate data analysis method through the statistical wizard module;
- Preventing misuse of statistics;
- Invoking the statistical analysis methods and related graphs;
- Providing some flexibility in the interpretation of the results of the analysis;
- Adding/removing application components;
- Adding/removing statistical methods;
- Loading a dataset to be analyzed along with its characterization;
- Editing the related dataset attributes.

Through the expert utility, the user can perform:

- Attribute management;
- Dataset management;
- Dataset characterization;
- Method selection rules management;
- User Objectives Management;
- Database administration.

Related documentation produced were the X-StatIS Statistical Guide, the X-StatIS Statistical Glossary and the X-StatIS User manual.

Possible impact on ESS and/or the end-user

The ESS is staffed both by experts and by non-experts. If the latter category can contribute more to data analysis and interpretation, that should lead to increased efficiency. Thus, subject to the quality of the project's output, X-StatIS should be seen as an aid to the more efficient and effective functioning of the ESS.

Specifically, X-StatIS allows non-expert users to enhance their productivity by being able to analyse and interpret a wide range of datasets. By using several statistical methodologies and a user-friendly interface that guides users and prevents them from misusing statistics, the X-StatIS software fulfils non-expert user needs in a way that enhances decision making for a large number of private and public institutions.

In summary, the benefits of X-StatIS are the maximization of knowledge extraction from raw data, the reduction of costs involved in statistical data analysis and in the interpretation of the results, and the enhancement of the ability to make the right decisions that could have important consequences for organizations and countries.

Dissemination and exploitation prospects

The dissemination activities included:

- The project website and a project brochure;
- Promotion of the project's results in relevant international fora such as the NTTS/ETK conference in 2001 and presentations to user groups coming from Market Research, Marketing Organisations and the Financial Market;
- Utilization at the consortium user sites. The final software developed within the project was deployed in the user / statistical organizations of the consortium, that is, NSSG, ONS, BVA, QUANTOS and CNAM. The deployment of the software in the consortium's statistical organizations had increased the effectiveness of statistical analysis; as users became more accustomed to the software usage, benefits are expected to extend to the optimization of use by non-experts and the reduction of costs associated with statistical analysis;
- The consortium sites would constitute demo sites for commercial exploitation outside the consortium. The need to analyze large data sets transcends statistical experts. It was shown that X-StatIS possessed unique features which would make application very appealing for commercial exploitation outside the consortium. Thus, the target market of X-StatIS consists of organizations that need to utilize both experts and non-experts in statistical data analysis. Such organizations are NSIs, associations and institutes at a national and international levels, Governmental and / or EU authorities involved in statistical data analysis and market research organisations.
- In order to take advantage of the market opportunity provided by X-StatIS, the Project Consortium had envisaged three main parties in such exploitation:
    - Software developer, who would ultimately be responsible for the development, evolution, customisation and marketing of the X-StatIS Software;
    - Network of Value Added Resellers, who would be responsible for marketing the software to end customers with the support of the Software Developer;
    - Demo sites, which would provide test beds for the X-StatIS software and, whenever requested or possible, would offer consulting services.

ANNEXES

II

eurostat

ANNEX
DOSIS RESEARCH

The 18 DOSIS projects covered the following research (the project acronym is shown in brackets):

- Development of a prototype tool capable of documenting and analyzing the contents and structure of electronic questionnaires (TADEQ);
- Development of a network of Geo-data Access Services for European Administrations and data providers (GEOSERVE);
- Exchange of data between small family enterprises and data collectors in Southern European countries (DATAMED);
- Development of the use of electronic means of communication for enterprise reporting (TELER);
- Development of automatic imputation software for business surveys and population censuses (AUTIMP);
- Development of an environment able to extract Bayesian Belief Networks from databases. The first step was the investigation of the feasibility, relevance and applicability of this approach to official statistics by selecting and formulating ten problems provided by NSIs (BAKE);
- Development of a prototype Knowledge Extraction or Data Mining system for identifying patterns and relationships in statistical datasets (KESO);
- Development of an open computer environment for medium and long-term forecasting for use by industry and researchers (FORCE-4);
- The production of a system for automatic linear seasonal adjustment and forecasting of time series (TESS);
- Design and implementation of an integrated statistical data processing environment making use of meta-information methodology, aimed at data collection and statistical processing (IMIM);
- Design and implementation of an integrated statistical data processing environment making use of meta-information methodology, aimed at the storage, access and dissemination of statistical information (IDERESA);
- Investigation of distributed database techniques and World Wide Web technology to improve access to statistical data by the European research and policy communities in order to assist researchers and policy analysts (ADDSIA);
- Demonstration of the feasibility of linking modules of reusable, standardized software (objects), held in widely separated computers, to carry out specific statistical processing tasks (RAINBOW);
- Development of an information retrieval system from economic databanks using meta-information, with navigation using hypertext links (FAPSY);
- Construction of an advanced statistical system based on a visual programming paradigm, capable of supporting both statistical research and statistical application building for end users (STABLE);
- Development of systems for suppressing the disclosure of confidential data when making tables and other forms of abstraction from statistical databases (SDC);
- Design of a prototype software environment where the modules would be able to plug into it, thus creating an Interactive Visualization Statistical System (IVISS);
- Development of techniques and software to facilitate the use of numerical symbolic data analysis in NSIs and companies (SODAS).

NOMENCLATURE ON RESEARCH IN OFFICIAL STATISTICS (NORIS)

NORIS was developed by Eurostat in 1999 as an integral part of EPROS mainly for internal purposes. Projects are classified by their principal outputs, though if there are significant, secondary outputs, the project will be shown under more than one heading. That a project does not appear at all under a certain heading does not necessarily mean that it did not encounter issues under that heading. For example, only CASC appears under Statistical Disclosure Control because that was the subject of its research, but other projects had to take account of confidentiality issues surrounding their research. Not the full classification right down to the 3-digit level is used in this publication because that would have led to a clutter of over-differentiation. There is no implication that NORIS has to be used outside FP5 EPROS. The full classification is as follows:

1. METHODOLOGICAL ISSUES

- 1.1 Concept formation, classifications
- 1.2 International harmonization
- 1.3 Completing the universe
- 1.4 Other

2. ADVANCED TECHNOLOGY FOR DATA COLLECTION

- 2.1 Coding metadata
- 2.2 Sampling
- 2.3 Automated data capture
  - 2.3.1 "Intelligent questionnaires"
  - 2.3.2 Bar codes
  - 2.3.3 Satellite images
  - 2.3.4 Touchstone Data Entry (TDE), Voice Recognition (VR), Optical Mark Recognition (OMR), Optical Character Recognition (OCR), etc.
- 2.4 Telematics for data capture and interchange (Electronic Data Interchange (EDI), IDA)
- 2.5 Other

3. QUALITY ISSUES

- 3.1 Quality assurance systems in the statistical production process (Standardization, Accreditation, and Certification)
- 3.2 Methods and tools for the measurement and improvement of data quality:
  - 3.2.1 Non-sampling (excluding non-response) errors
  - 3.2.2 Non-response
  - 3.2.3 Editing
  - 3.2.4 Weighting
  - 3.2.5 Imputation
  - 3.2.6 Variance estimation
- 3.3 Other

4. DATA ANALYSIS AND STATISTICAL MODELLING

4.1 Data analysis and knowledge extraction
4.2 Nascent techniques, including neural networks
4.3 Timeseries analysis (including now- and forecasting)
4.4 Modern sampling methods
4.5 Models for measuring risk and uncertainty
4.6 Small area disaggregation/estimation/ Geographical Information Systems (GIS)
4.7 Other

5. MULTI-DATA SOURCES, INTEGRATION AND SYSTEMATISATION

5.1 Administrative data for statistics
5.2 Multisource environments
5.3 Integrated statistical processing via metadata
5.4 Distributed database management and analysis systems
5.5 IT infrastructures
5.6 Other

6 DISSEMINATION, DISCLOSURE CONTROL

6.1 Data dissemination
  6.1.1 Publication in paper form
  6.1.2 Publication in electronic forms
6.2 Computer-assisted techniques for training in statistics
6.3 Statistical disclosure control
6.4 New applications in IT: Innovative user-interface and visualization techniques
6.5 Other

European Commission

European Plan of Research in Official Statistics (EPROS) – Main conclusions from the activities in the 5th Framework Programme

Luxembourg: Office for Official Publications of the European Communities

2007 — 205 pp. — 21 x 29.7 cm

ISBN 978-92-79-04705-3
ISSN 1977-0375