COOPERATION MODELS FOR SOFTWARE DEVELOPMENT

Valentin Todorov[1]

[1] United Nations Industrial Development Organization (UNIDO)

Meeting on the Management of Statistical Information Systems
(Luxembourg, 23-25 May 2011)

Outline

1. Motivation
2. Statistical Collaboration Models
3. Statistical Collaboration Groups
4. Socio-Technical Issues
5. Summary and Conclusions

Collaboration

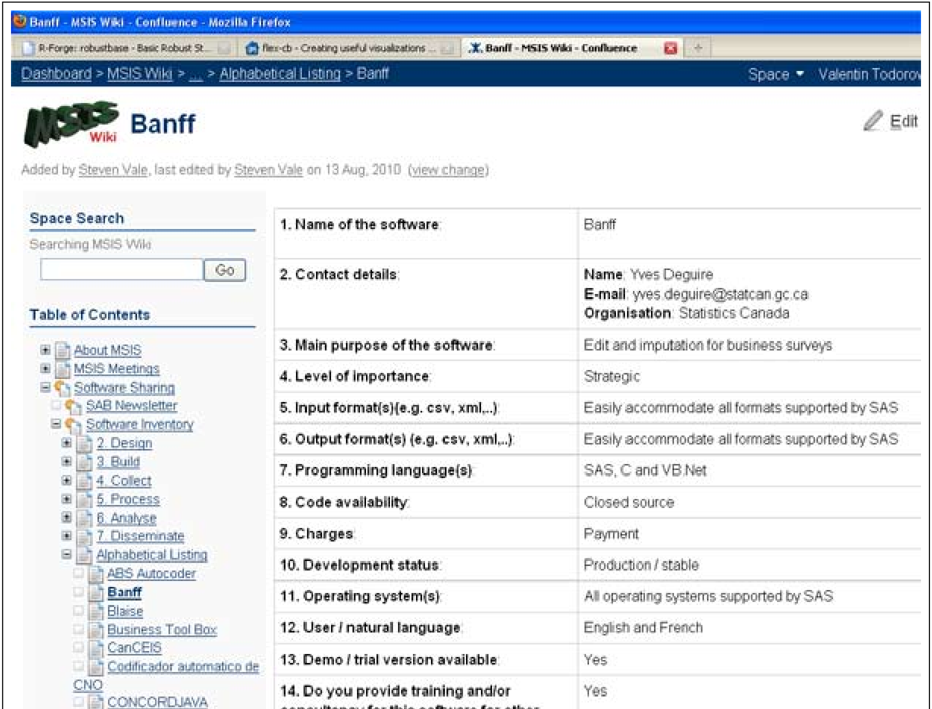
"Symphony of collaboration: multiple-stakeholder collaborations provide a clear path to integrated data exchange and system interoperability "
by **Robin Thomashauer** *on clinical data exchange and eHealth*

Collaboration and common software

- Using common software solutions in national and/or international statistical organizations
  - results in savings in time and money
  - enhance mutually the institutional knowledge
  - promote and enable the implementation of statistical standards
- Using statistical standards like SDMX will in turn
  - facilitate the data exchange mechanisms
  - improve the data quality on both sides

The Software Inventory

- Software Inventory platform launched by the Sharing Advisory Board (SAB)
- `http://www1.unece.org/stat/platform/display/msis/Software+Inventory`
- To collect information about existing shared products, products under development or even products that are undergoing planning
- Compiled in cooperation with the ESSnet project on a Common Reference Architecture (CORA)
- Already more than 50 products
- Several more products which are not yet in the Inventory (EUROSTAT, FLEX-CB)

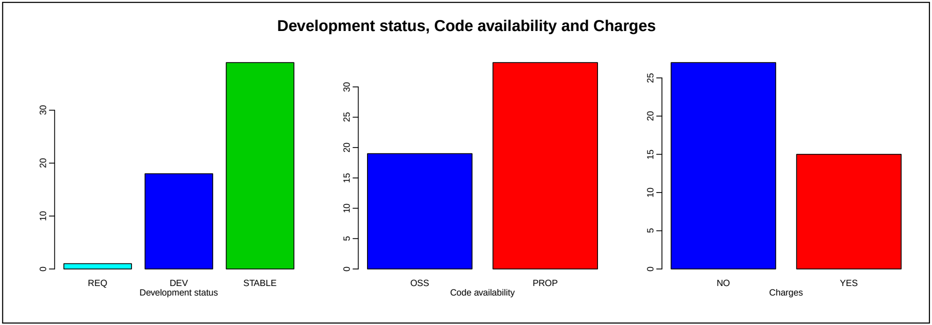


Banff

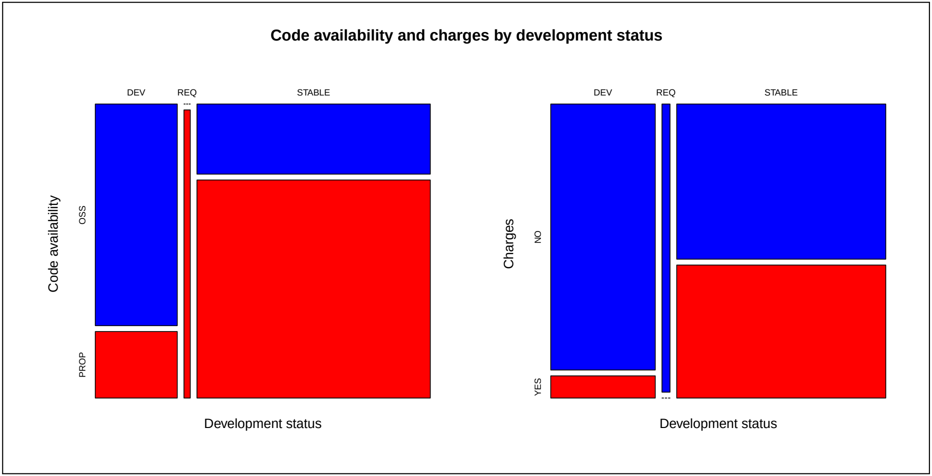
Added by Steven Vale, last edited by Steven Vale on 13 Aug, 2010 (view change)

1. Name of the software:	Banff
2. Contact details:	**Name**: Yves Deguire **E-mail**: yves.deguire@statcan.gc.ca **Organisation**: Statistics Canada
3. Main purpose of the software:	Edit and imputation for business surveys
4. Level of importance:	Strategic
5. Input format(s)(e.g. csv, xml,..):	Easily accommodate all formats supported by SAS
6. Output format(s) (e.g. csv, xml,..):	Easily accommodate all formats supported by SAS
7. Programming language(s):	SAS, C and VB.Net
8. Code availability:	Closed source
9. Charges:	Payment
10. Development status:	Production / stable
11. Operating system(s):	All operating systems supported by SAS
12. User / natural language:	English and French
13. Demo / trial version available:	Yes
14. Do you provide training and/or	Yes

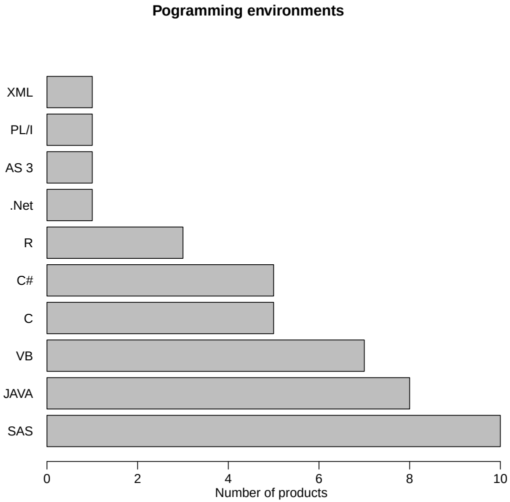
Development status, Code availability and Charges



Code availability and Charges vs Dev. Status

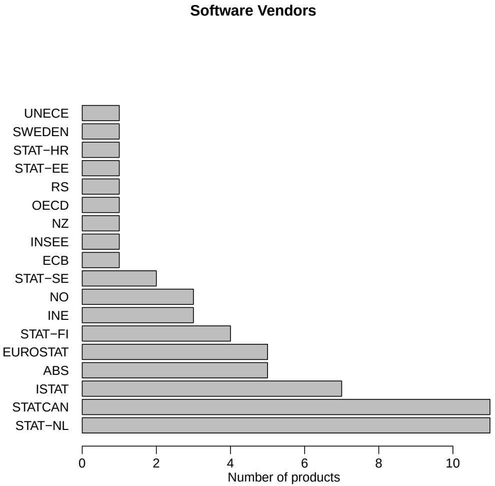


Programming environments used



- Three main groups of products according to the programming environment
  - SAS
  - Java
  - Any other (VB, C, C#, etc.).
- Only three products written in R and only one of them states to have a stable version in production
- Of course there are many more R packages at CRAN

Who is developing shared software?



- Most of the products offered for sharing were developed by StatCan or Netherlands - eleven each
- ISTAT has seven and ABS five
- EUROSTAT has also five, but they are not yet entered in the Inventory

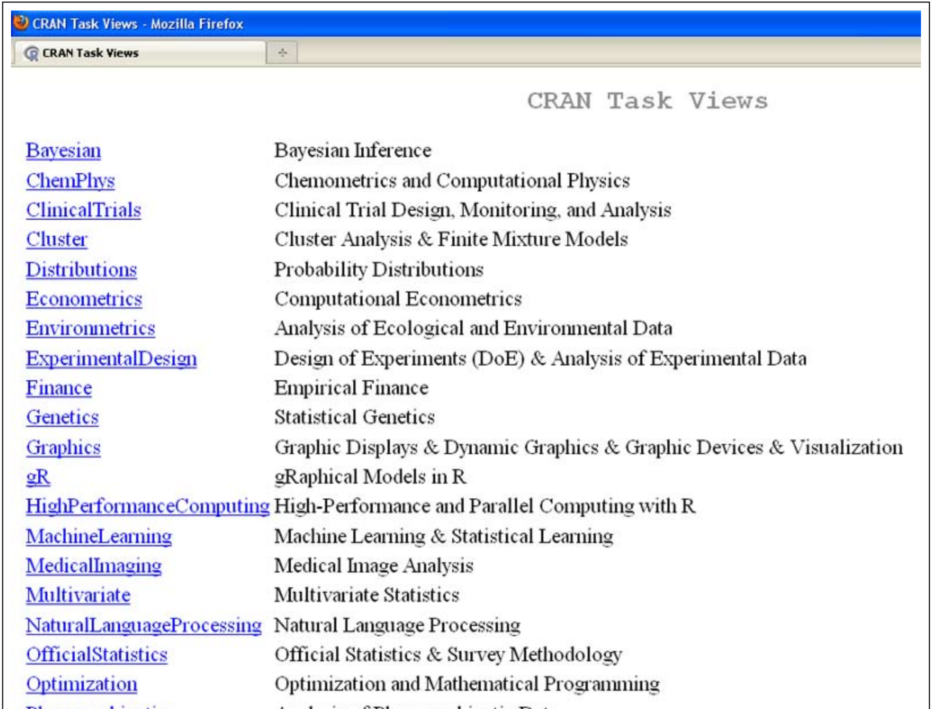
Types of Collaboration Projects

Three main approaches for collaboration

1. Software developed by one organization and released as Open Source or Free Software
   - EUROSTAT: *eDamis*, *SDMX Reference Architecture*
2. Software developed by one organization and offered for collaborative sharing for a fee
   - **PC-Axis**: a suite of software designed for disseminating and visualizing statistical data (GSBPM 7.x)
   - **OECD.Stat**: designed for disseminating and visualizing statistical data (GSBPM 7.x)
   - **Generalized Systems Suite (StatCan)**
3. "True" open source software which was developed as OSS and follows the OSS models.
   - **FLEX-CB**: to create useful visualizations of statistical data from institutions that employ **SDMX**.
   - R packages at CRAN: `http://cran.r-project.org/`

CRAN Task View for Official Statistics

- CRAN Task Views, Zeileis (2005): `http://cran.r-project.org/web/views/`
  - Collections of packages which belong to the same area
  - Relevant descriptions of the packages, further grouped in functional areas
  - Possible to install or update all packages from a given task view
- Recently a Task View dedicated to **Official Statistics** was added
- Methods typically used in official statistics and survey methodology
- Currently more than 40 packages are listed



CRAN Task Views

Bayesian	Bayesian Inference
ChemPhys	Chemometrics and Computational Physics
ClinicalTrials	Clinical Trial Design, Monitoring, and Analysis
Cluster	Cluster Analysis & Finite Mixture Models
Distributions	Probability Distributions
Econometrics	Computational Econometrics
Environmetrics	Analysis of Ecological and Environmental Data
ExperimentalDesign	Design of Experiments (DoE) & Analysis of Experimental Data
Finance	Empirical Finance
Genetics	Statistical Genetics
Graphics	Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
gR	gRaphical Models in R
HighPerformanceComputing	High-Performance and Parallel Computing with R
MachineLearning	Machine Learning & Statistical Learning
MedicalImaging	Medical Image Analysis
Multivariate	Multivariate Statistics
NaturalLanguageProcessing	Natural Language Processing
OfficialStatistics	Official Statistics & Survey Methodology
Optimization	Optimization and Mathematical Programming

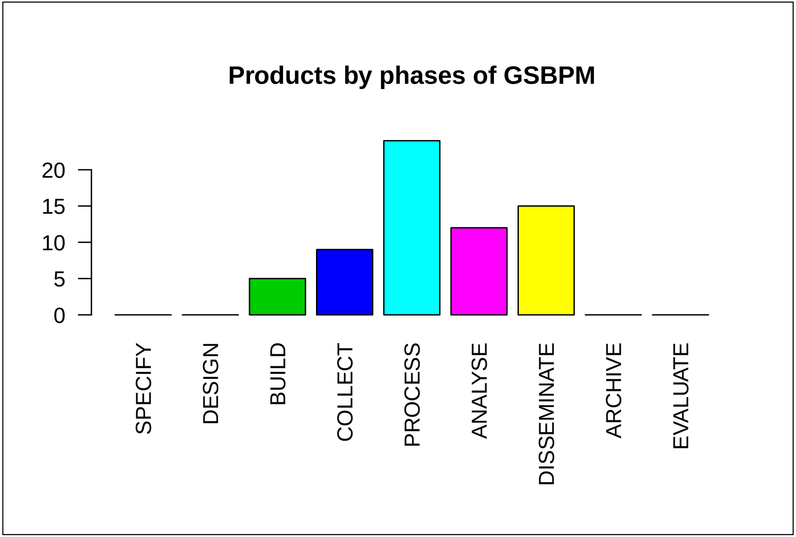
CRAN Task View: Official Statistics and Survey Methodology

- **CRAN Task View: Official Statistics and Survey Methodology**
  - Complex Survey Design: General
  - Complex Survey Design: Details
  - Complex Survey Design: Point and Variance Estimation
  - Complex Survey Design: Calibration
  - Editing and Visual Inspection of Micro data
  - Imputation
  - Statistical Disclosure Control
  - Seasonal Adjustment
  - Statistical Record Matching
  - Indices and Indicators
  - Additional Packages and Functionalities
- Most popular packages: *Amelia, impute, mice, RecordLinkage, robCompositions, rrcovNA, sampfling, sampling, sdcMicro, survey (core), VIM, x12*
- Other task views: TimeSeries, Econometrics and SocialSciences

Generic Statistical Business Process Model (GSBPM)

- Published by METIS in 2009 as a tool for describing and benchmarking statistical production processes
- Rapidly becoming a defacto global standard
- Initially developed to provide a basis for statistical organizations to agree on standard terminology to aid their discussions on developing statistical metadata systems and processes
- However it is apparent that the model can be very useful in other areas like
    - harmonize the statistical computing architectures
    - facilitate the sharing of statistical software
    - provide a basis for explaining the use of SDMX in a statistical organization
    - provide a framework for process quality assessment and improvement
    - and many more ...

The Role of GSBPM



Licensing Issues

- Necessary to define the commercial and legal foundations for the exchange of software
- Open Source Software (OSS) approach is a viable alternative for exchange activities
- However limited by different factors
  - The particular legislation of the country (e.g. Canada)
  - Total Cost of Ownership (TCO)
  - Other issues of "compatibility" between OSS rules and the rules of the organizations
- EUROSTAT COmmon Reference Architecture (CORA) ESSnet - detailed description of the available licensing models
  - GNU type licenses
  - Apache/BSD type licenses
  - European Union Public License (EUPL)
- Details: EUROSTAT (2010). CORA: Models for licensing, support and legal properties, Work package 4.1

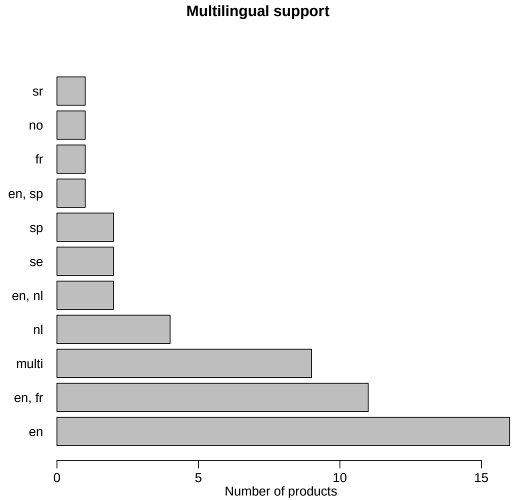
Multilingual Support

- In the national context - either not a requirement or its priority is very low (budget, resources and time constraints)
- Rarely becomes a key part of the software architecture
- The lack of multi-language support: a major barrier to sharing (statistical) software
- Statistics Canada's experience (see Karen Doherty, 2011)
- Three scenarios:
    1. All controls are displayed in the both languages
    2. The user once (by starting or installing the application) selects the language
    3. The user can switch between languages at any time without loosing the context

Multilingual Support

- Must be approached early in the design of the system and become a key part of the architecture
- The preferred way of implementing multi-lingual support is to apply frameworks and relevant tools
- Principles and Guidelines for Developing Multi-lingual Statistical Software prepared by the Sharing Advisory Board (SAB) - available at the MSIS Wiki: `http://www1.unece.org/stat/platform/display/msis/Home+Page`

Multilingual support in the inventory



- Multilingual support of the products from the software inventory
- Reflects the experience of StatCan
- Eleven bilingual (Englis/French) products offered by StatCan
- Several other bilingual products (English/Dutch, English/Spanish)
- Ten products which are declared multi-lingual: Blaise, OECD.Stat and PC-Axis are among them.

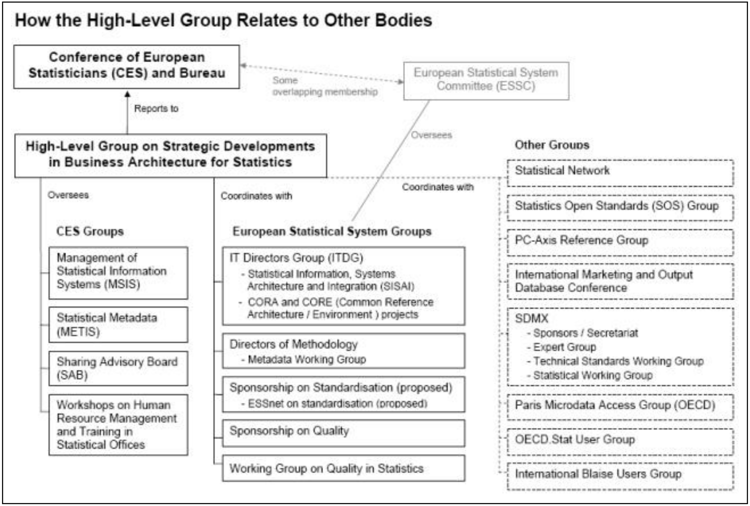
Support, Documentation and Training

- Documentation and training activities: an important component of the total cost of a product or system
- This is what the open source software products are mostly criticized. Example: **R-Help**
- Evaluation version - important if the software product is not free of charge
- **Inventory:** for less than the half of the products support and training are available
  - Support and training are offered for **Blaise**, **Banff**, **PC-Axis**, **OECD.Stat** and **eDamis**
  - StatCan offers support for most of the products comprising the **Generalized Systems Suite**
  - **FLEX-CB** acts in the same way as most of the open source projects: sparse documentation; the source is available to modify and contribute; help forum provides for quick support granted by the few involved developers

Inventory of International Groups

- An overview of the different international groups whose work is related to the enterprise architecture of statistical organisations.
- Available at the **MSIS Wiki** page
- Several of the groups are especially interesting for the topic considered here - their main objective is to foster collaboration approach in striving for improvement in Statistical Information Management and to facilitate better cooperation in the field of development of software and uptake of IT related standards within the respective statistical production systems
    - Statistics Open Standards (SOS) Group
    - Statistical Network
    - User groups of different statistical software products

Relations between the International Statistical Collaboration Groups



How the High-Level Group Relates to Other Bodies

Conference of European Statisticians (CES) and Bureau

European Statistical System Committee (ESSC)

Some overlapping membership

Reports to

Oversees

High-Level Group on Strategic Developments in Business Architecture for Statistics

Oversees

Coordinates with

Coordinates with

CES Groups

- Management of Statistical Information Systems (MSIS)
- Statistical Metadata (METIS)
- Sharing Advisory Board (SAB)
- Workshops on Human Resource Management and Training in Statistical Offices

European Statistical System Groups

- IT Directors Group (ITDG)
  - Statistical Information, Systems Architecture and Integration (SISAI)
  - CORA and CORE (Common Reference Architecture / Environment) projects
- Directors of Methodology
  - Metadata Working Group
- Sponsorship on Standardisation (proposed)
  - ESSnet on standardisation (proposed)
- Sponsorship on Quality
- Working Group on Quality in Statistics

Other Groups

- Statistical Network
- Statistics Open Standards (SOS) Group
- PC-Axis Reference Group
- International Marketing and Output Database Conference
- SDMX
  - Sponsors / Secretariat
  - Expert Group
  - Technical Standards Working Group
  - Statistical Working Group
- Paris Microdata Access Group (OECD)
- OECD.Stat User Group
- International Blaise Users Group

Statistical Network

- Formed at the *Informal CSTAT Workgroup on Stronger Collaboration on Statistical Information Management Systems* in Paris last June
- Main Objective: *"Working together with pace and passion to better meet our societies' information needs while driving down costs"*
- Participants: Australia, New Zealand, Sweden, Norway, UK and Canada
- Areas of common interest and cooperation are
    1. Innovation in Dissemination (New Zealand),
    2. Confidentiality and Disclosure Control (Sweden),
    3. Common Metadata/Information Management Framework (OCMIMF) (Australia),
    4. Editing (Norway) and
    5. Web Data Collection (Canada).
- Details: Value Creation Group (2010), Draft Output Report

Statistics Open Standards (SOS) Group

- A *consortium of NSIs* who have agreed that they wish to and are able to contribute to common development of their statistical production environments
  - homogeneous group in relation to use of IT
  - similar underlying production models
  - they share a set of common visions
- Participants: Denmark, Finland, Iceland, Netherlands, Norway and Sweden
- Areas of common interest and cooperation are
  1. Dissemination databases
  2. Metadata
  3. data collection
  4. Architecture and tools
  5. Standards.
- Details: Rune Gløersen (2008)

User Groups

- Groups organized around a particular software product
- Key objectives: promoting the implementation and use of the software family in national statistical offices and other organizations; serve as forums for discussion and exchange of ideas and experiences
- **PC-Axis Reference Group**
- **OECD.Stat User Group**
- **International Blaise Users Group**

Socio-technical issues of collaborative software development

- Global software development (Noll et al. 2010)
- Main practices associated with collaboration:
  1. Identify common **goals, objectives and rewards**
  2. Collaboratively establish and maintain the **product ownership boundaries**
  3. Collaboratively establish and maintain **interfaces and processes**
  4. Collaboratively develop, communicate and distribute **work plans**.
- Barriers to these practices:
  - geographic distance
  - temporal distance (locations in different time zones)
  - language and culture differences (including corporate culture)
  - infrastructure and product architecture

Collaboration tools for software development

- Solutions to the collaboration barriers
    - **Virtual teams** and **Global teaming model**
    - Beecham et al. (2010), Todorov (2009), Theußl and A. Zeileis (2009)
- Software development infrastructure and the relevant collaboration tools - Lanubile et al. (2010)
    - **R-Forge**: a central platform for the development of R-related software, Theußl and A. Zeileis (2009)
    - **OSOR.EU**: Open Source Observatory and Repository, a platform for exchanging information, experiences and FLOSS-based code for use in public administrations.
    - **Google Code**: Google's site for developer tools, APIs and technical resources

Summary and Conclusions

- Review and analysis of collaboration activities for sharing of software among national and international statistical organizations
- Several main issues and opportunities
  - licensing, position in the statistical business process model, ownership and governance, sustainability, standards utilization, methods for distributed software development and technical communication advances
- Three main approaches for collaboration:
  - software developed by one organization and released as Open Source or Free Software
  - software developed by one organization and offered for collaborative sharing for a fee
  - "true" open source software which was developed as OSS and follows the OSS models.

Summary and Conclusions

- The collaboration adds new dimensions of complexity in the technical, administrative and legal aspects
- But will substantially contribute to
  - the harmonization of the software architectures
  - the adoption of international standards like SDMX
  - will be beneficial for reducing the costs for developing and maintaining of statistical software
  - will improve the user support and capacity building

References I

United Nations Economic Commission for Europe, (2009)
Generic Statistical Business Process Model, Version 4.0,
URL: www.unece.org/stats/gsbpm.

Noll, S. Beecham and I. Richardson (2010)
Global software development and collaboration: barriers and solutions,
ACM Inroads, **1**, 66–78.

EUROSTAT (2010)
CORA: Models for licensing, support and legal properties, Work package 4.1,
URL: http://cora.forge.osor.eu/WP4__Design_of_the_Organizational_Architecture.htm.