UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE) CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

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

Topic (i): Architectures, models and standards

Tentative anatomy of a new generation of IT-architecture to support GSBPM-processes

Invited Paper

Prepared by Jakob Engdahl and Hans Irebäck, Anders Holmberg Statistic Sweden, Sweden

I.   Introduction

1.   Many statistical institutes have recently adopted a process oriented production approach. A consequence of this shift of approach is that a lot of the existing survey-oriented IT-systems are no longer aligned with the business viewpoint. Development is currently on going at Statistic Sweden with the aim to replace survey-oriented IT-systems with process-oriented IT-systems. Since this approach reduces the possibility of creating tailor-made solutions for the survey staff, this shift isn't completely unproblematic.

2.   This development is also taking place during a time when interest for international collaboration, due to requirements of more efficient software development, is increasing.

3.   This paper gives a tentative description of how different parts of a new generation of IT-architecture could be formed to increase the possibility for efficient international collaboration between statistical institutes that have adopted the *GSBPM-model.*
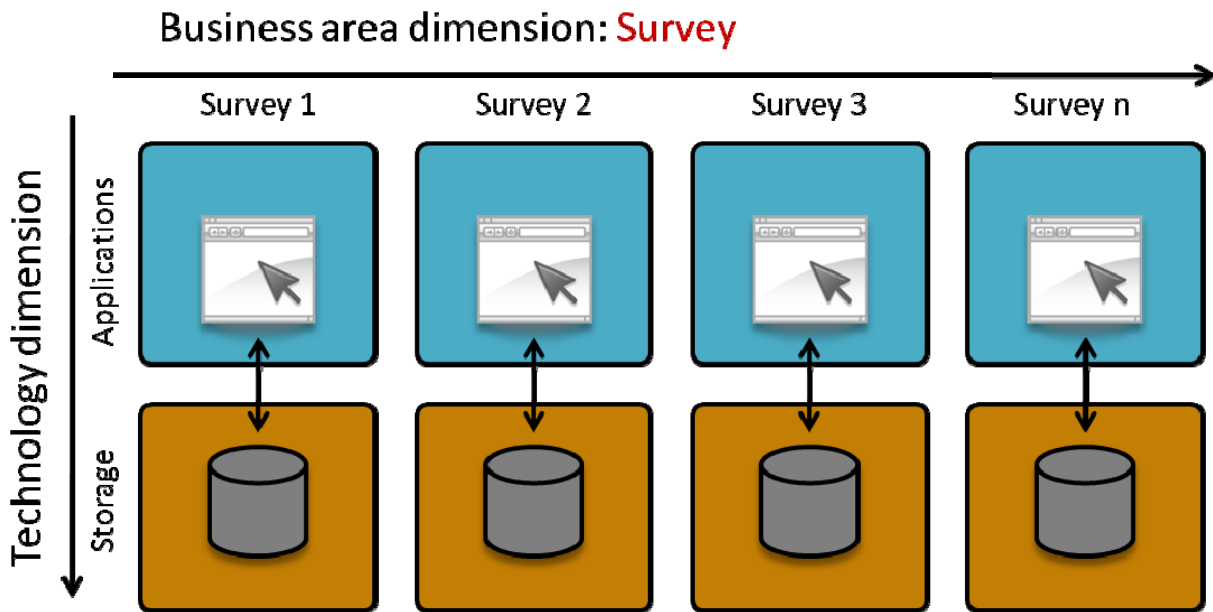
II.   Modularization

A.   Modularization

4.   In early times of software development it was up to the developer to implement all functionality to create a system that gave support to perform a certain business tasks. The developer was free to choose how each feature was implemented but was also required to create unique solutions for everything from graphical

presentation and memory management to data storage, authentication and authorization etc. As the IT-industry matured, reoccurring requirements was identified and solutions for memory management, data storage etc was packaged into modules that the developer could integrate together with the business unique solutions. The idea of modularization is an important feature of any modern IT-system and as the complexity of the business requirements increases, so does the need of using efficient modularization.

5.   The first form of modularization was used to encapsulate different type of technology requirements but as more and more business areas require IT-systems there is also a need to modularize along the different business areas. One commonly used parameter to differentiate business areas within statistical institutes is the different surveys or subject matters areas. Different surveys could be seen as different business areas and a lot of statistical institutes have, together with the technology dimension, used surveys, or in best case used cluster of surveys, as a dimension for creating reasonably sized modules in their IT-platform.

[Picture 1: Modularization based on technology and survey dimension]

## B.   Towards a business oriented approach

6.   Lately, many businesses have shifted their viewpoint of their internal business structure. Focus was shifted away from products towards business processes. A lot of statistical institutes have today, in some form, adopted the *Generic Statistical Business Process Model (GSBPM)* to describe their business.

7.   A shift of approach towards a business oriented approach also means that the approach used for modularization of the IT-systems will also have to be shifted. From a process oriented business viewpoint, the existing IT-systems are seen as stovepipe-systems since these aren't aligned with the new business viewpoint.

8.   This meant that new requirements for the IT-systems are bundled from the process viewpoint instead of the survey viewpoint. An effect of this is that the modularization of the existing software to support statistical production is no longer aligned to the business viewpoint.

9.   Changing a business process, based on new requirements, is today difficult and expensive since a lot of the existing IT-systems are still survey oriented. Many statistical institutes are currently working on creating process oriented IT-systems in order to more efficiently react to new requirements from the business side.

10.	Shifting modularization from survey oriented to process oriented isn't something that happens overnight. Since the business oriented approach also means that the possibility for the survey staff to require tailor-made IT-systems is reduced, the result often means that the survey staff thinks that this is due to IT-related viewpoints when it is really a consequence of a shift of approach from the business side.
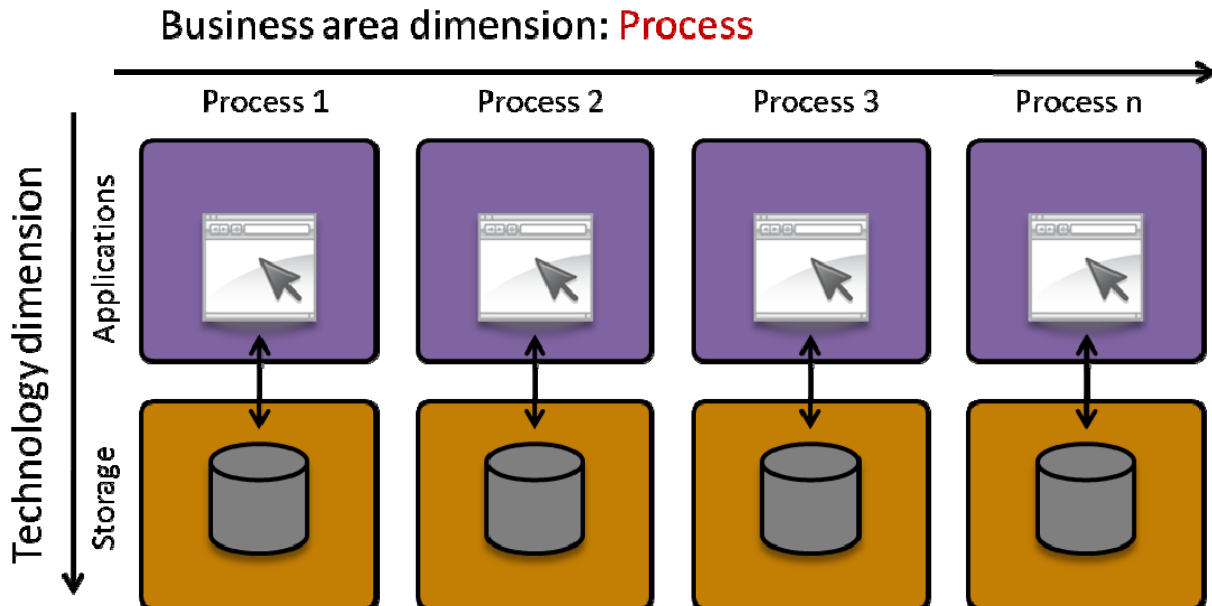
*[Picture 2: Modularization based on technology platform and process]*

## III.	Extended technology dimension

### A.	Increased need for integration – Service layer

11.	The shift from survey oriented modules to process oriented modules increases the need for efficient integration solutions between the modules. When data has been processed in the IT-system for one process, the result must be made available for the next process to be further refined.

12.	Two common approaches to solve this new requirement is to either create data warehouse modules that are shared between many process oriented modules or to transfer the data when it has been processed from one module to the other.

13.	Since the data warehouse approach means that a software module will be shared between many processes, it has the risk of creating big dependencies between the process modules since any change in a shared data warehouse module will affect several other modules. During the 00s some new principles for software design was gathered under the concept of *Service Oriented Architecture (SOA)*. Some of these principles were derived from modularization principles and enabled software encapsulation aligned with business areas such as data collection or technology areas such as e-mail-communication or data storage. These principles could be used to overcome dependency problems. A new layer of modules in the technology platform dimension could be added to create a buffer between the process-oriented modules and the data warehouse-modules. This would allow for some changes in the data warehouse module without affecting the other modules.
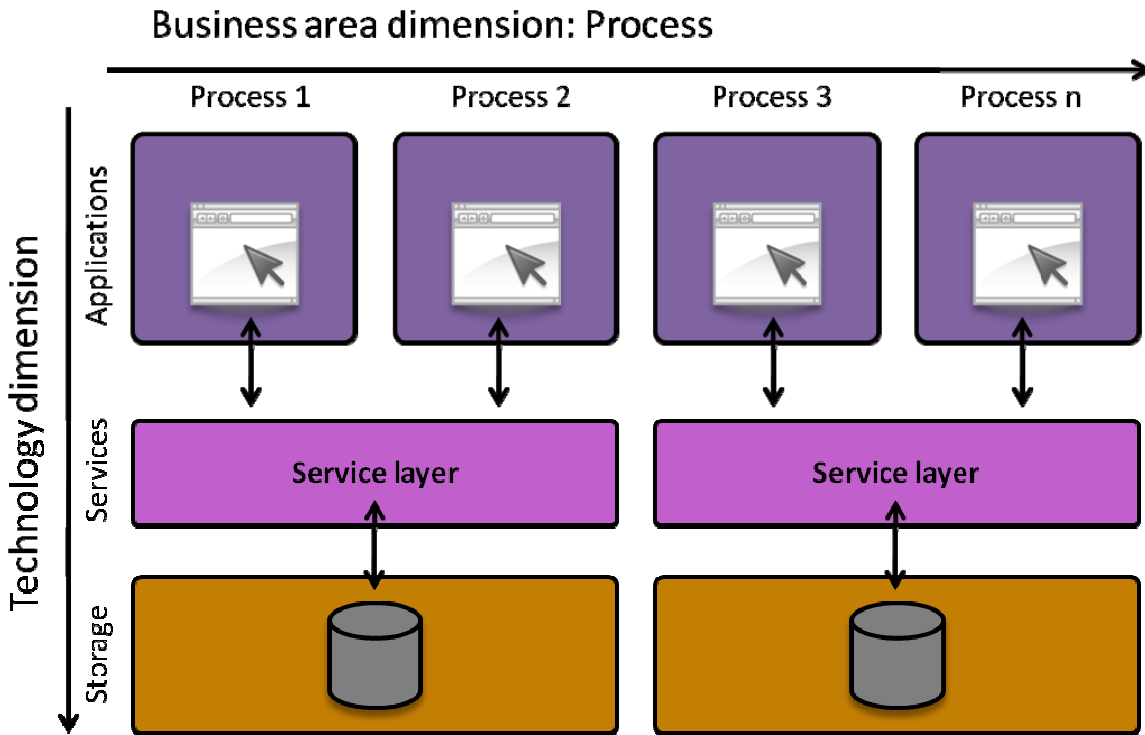
[Picture 3: Protecting data warehouse modules with a service layer]

B.   Keeping state - workflow

14.   A problem with a simple service layer is that it has no state. In other words, a service is only responsible for transferring information from a data warehouse module to a process oriented module. The service layer will therefore not keep track on when it is time for the next process oriented module to begin processing the data. In this case, it will be up to the users to keep track of when to use a specific process oriented module and when to use the next.

15.   Requiring users to keep information about state has a couple of problematic side effects. In many processes and surveys, the work is being done in a group of people. This means that everyone in that group needs to have the same conception about the state of the current process and if someone in the group doesn't have the same the same conception, there is a risk that a process is initialized to soon or is unnecessarily repeated.

16.   Keeping information about the process state at the staff also means that automation of process oriented modules is difficult to establish. Even though a process could be automated it would still require the user to initialize the process since no software module has sufficient information about the state of the process to automatically initialize the process.

17.   To solve this problem, information about the state of the process must be kept inside some part of the IT-system. No specific process oriented module will be able to keep track of the overall process state and therefore a new module must be established with the responsibility of keeping information about the process state.

18.   More advanced service oriented solutions have been created that not only adds a buffer between the process oriented modules and data warehouses, but also keeps information about the state of the overall process. Common names for these solutions are *workflow engines* or *BPMN/BPEL-engines* (based on the notation language). These solutions provide the users with the possibility of predefining a workflow or a set of orchestration rules which tells the module how to transfer information from one module to another.

19.   By keeping information about the current process state inside a workflow module, calls to automated process oriented modules could be made by the workflow engine. The user would only see the result of the process execution and would not have to initialize the call anymore.

Business area dimension: Process
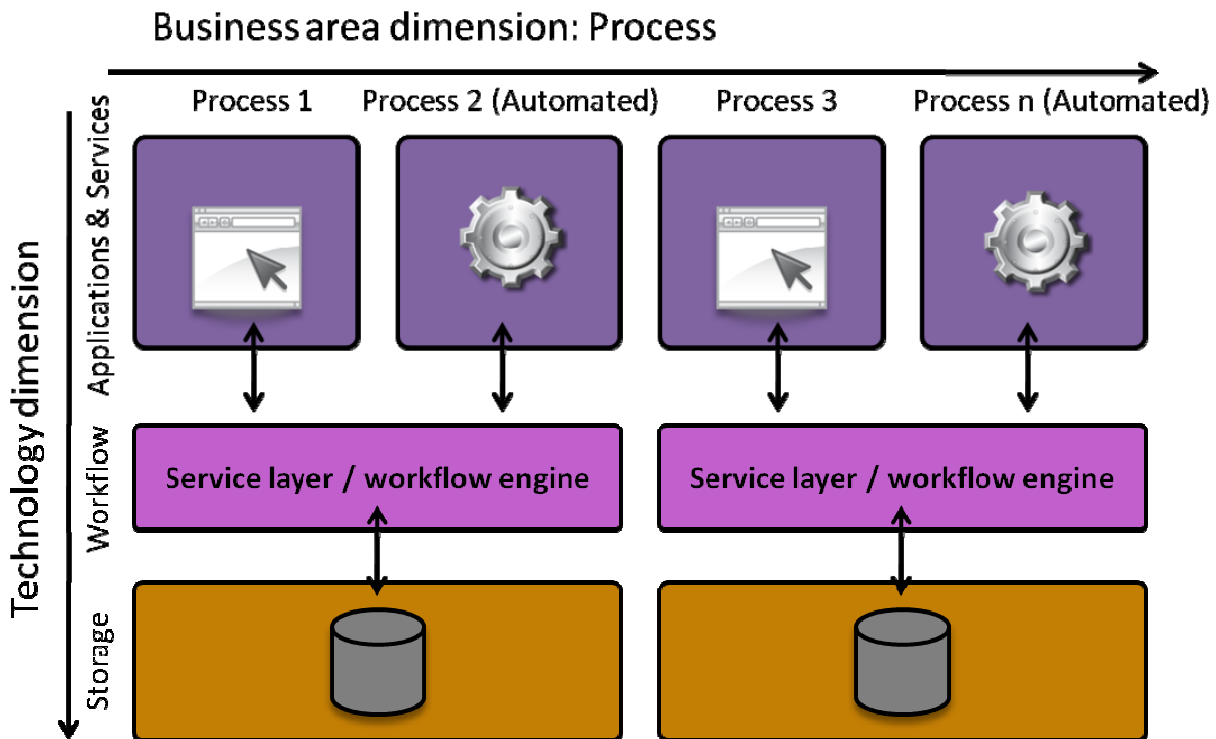
[Picture 4: Overview of modules: process oriented modules, service layer/workflow, data warehouse]

20.   Dependencies to other infrastructure-related solutions such as security, logging etc should also be minimized within the process oriented software modules. Each process oriented software module should not require specific solutions for this but instead use the *workflow-layer* to communicate with these solutions.

## C.   Creating an overview – Process dashboard

21.   When realigning the IT-systems to support processes instead of surveys, a problem occurs for the survey staff. Previously a lot of their work could be performed in the same system, but in a process oriented IT-platform, the same work will instead be performed in several smaller (process oriented) systems. None of the process oriented system will be able to provide an overview of the overall survey progress.

22.   To provide this feedback, a new process dashboard module is needed to provide the users with the overview and guide the users in their work. The dashboard should give a survey oriented view into the process oriented software modules.

23.   The responsibility for this module could include:

(a)   Provide an overview of the progress of the survey process;
(b)   Provide process execution feedback to supervise the process and to support decision making;
(c)   Guide the users to the right process oriented applications based on process state.

24.   To provide an overview of several processes, this module must overarch multiple processes.
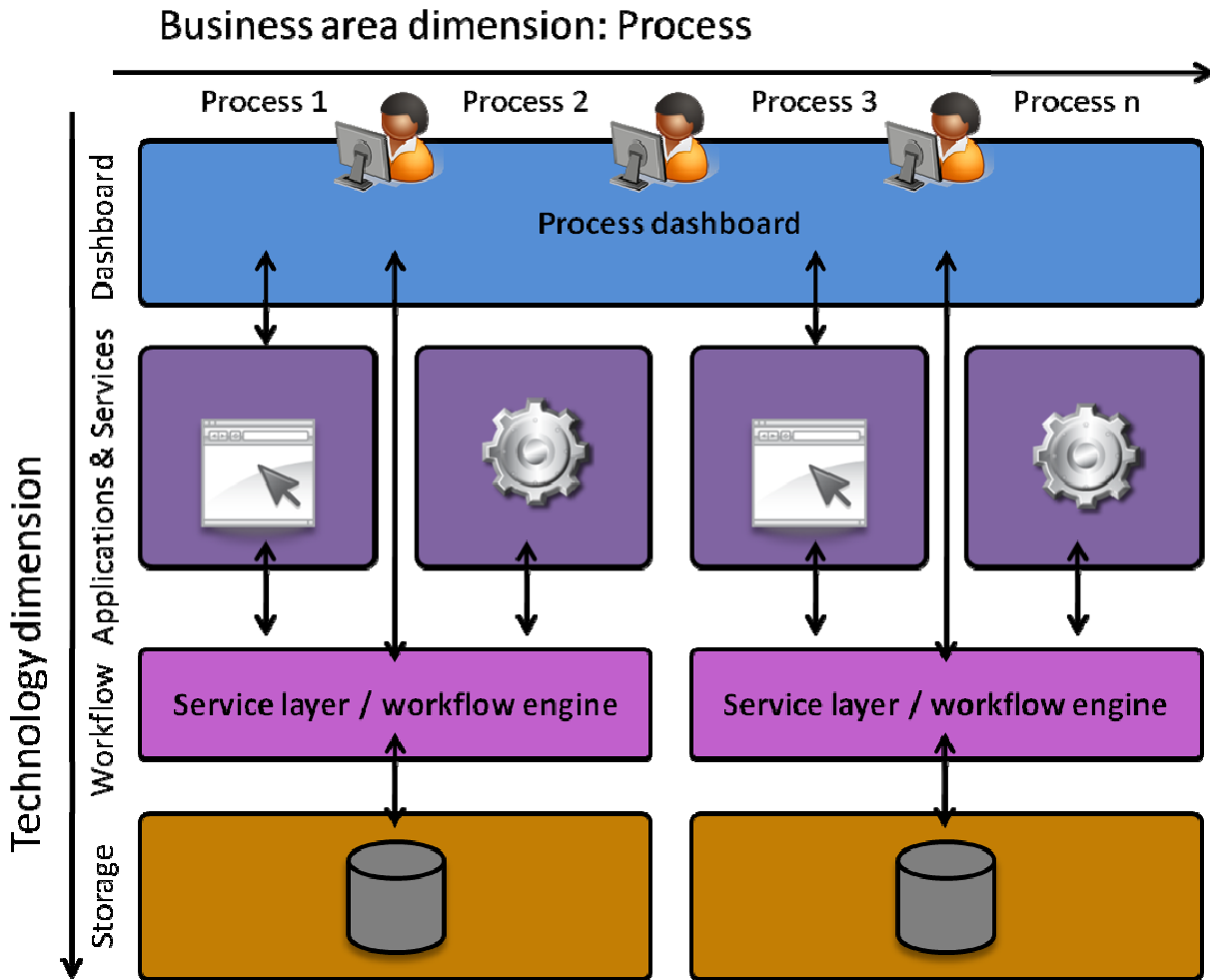
[Picture 5: Overview of modules: Dashboard, process oriented modules, service layer/workflow, data warehouse]

IV. Reusing modules

A. Design-time reuse or Runtime reuse

25.   To provide the business side with a set of process oriented modules two different types of solutions could be considered:
- (a)   Process oriented templates that could be used to generate or manually create survey-specific modules;
- (b)   Reusable modules with built in flexibility so that survey-specific parameters could be applied.

26.   These options could be called design-time reuse and run-time reuse and both these options could be used to provide the necessary process oriented IT-systems for the business side.

27.   One advantage for only using design-time reuse is that building flexibility into the modules might be difficult. Requirements from several surveys needs to be collected and requirements from different surveys might be incompatible.

28.   There are however several advantages in aiming for run-time reuse. The effort of creating flexibility within a module is a onetime thing whereas converting a template module to a survey specific module will have

to be done for every service. The accumulated time and resources for this will be larger than designing a run-time reusable module.

29.   Making changes to a process which will affect the underlying process oriented module will also only have to be done in one system rather than in all modules based on the same template. This means that new requirements are easier to implement in modules that are runtime-reused. This is also expected to lead to lower maintenance costs.

30.   The same principles apply to the other modules as well. Minimizing the amount of run-time modules for workflows, data warehouses etc, will increase the possibility of ensuring the stability and quality of these and changes based on technology requirements or business requirements will be easier to implement.

B.   Extended lifecycle of a process oriented software module

31.   A survey oriented IT-system has a lifecycle which corresponds with the GSBPM-processes and includes:

(a)   Development;
(b)   Test;
(c)   Production.

32.   A consequence of establishing run-time reuse of process oriented software modules is that the lifecycle of a specific module will include its own development, test and production phase, but will also include the survey implementation lifecycle for each survey that makes use of the software module. The new lifecycle of a process oriented software module will therefore include:

(a)   Software module: Development;
(b)   Software module: Test;
(c)   Software module: Production;
     a.   Survey implementation: Development;
     b.   Survey implementation: Test;
     c.   Survey implementation: Production.
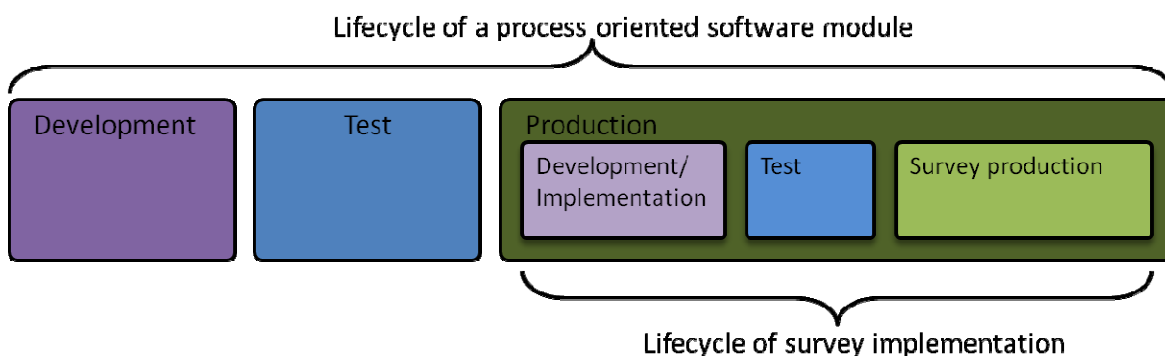
[Picture 6: IT-system: development, test, production; Survey-implementation: development, test, production]

33.   One advantage of this extended lifecycle is that the IT-system will be more thoroughly tested since more than one survey will test both their survey specific implementation but also the IT-system itself. This is an important part of quality assurance in statistics production. Moreover it also decreases the risk of too early releases. A larger effort in ensuring other quality requirements like security, availability and performance can also be better focused when minimizing the amount of run-time software modules.

V.   Business Information Model

A.   Parameters for a process oriented software module

34.   To provide run-time reusable process-oriented modules, survey-specific details cannot be hardcoded into the module but must instead be provided using parameters. Process oriented modules should accept parameters that control how the software module behaves. Just like the design of a specific survey will affect the way a process is performed, the survey design will affect how the software module behaves.

35.   The parameters could of course take the form of a programming language where the survey-specific design is described by using programming code. This however, will create a gap between the business side and the IT-side and will also require IT-resources when a survey should be implemented in a software module. Therefore a more high level form of parameters should be used.

36.   An example of this could be that a *questionnaire* could be sent to a *web questionnaire software module* but instead of using html, a more high level language should be used. This language would probably consist of concepts like question, options, variables etc.

37.   To be able to use the same software module, this language has to be agreed on within all business areas. The same language and vocabulary have to be used. This language will not only be a vocabulary since the relation between for instance a question and an option will have to be the same throughout all business areas. This language could therefore be seen as the process oriented statistical business language and it consists of business information objects that are used to control both the process but also the software module that supports the process.

38.   The input and output of each process oriented software module should therefore be described in business information objects and includes:

(a)   Information controlling the behavior of the software module
(b)   Information objects that the software module should add value to
(c)   Other necessary information resources
(d)   The result of process completion
(e)   Information that further describes the result of the process execution
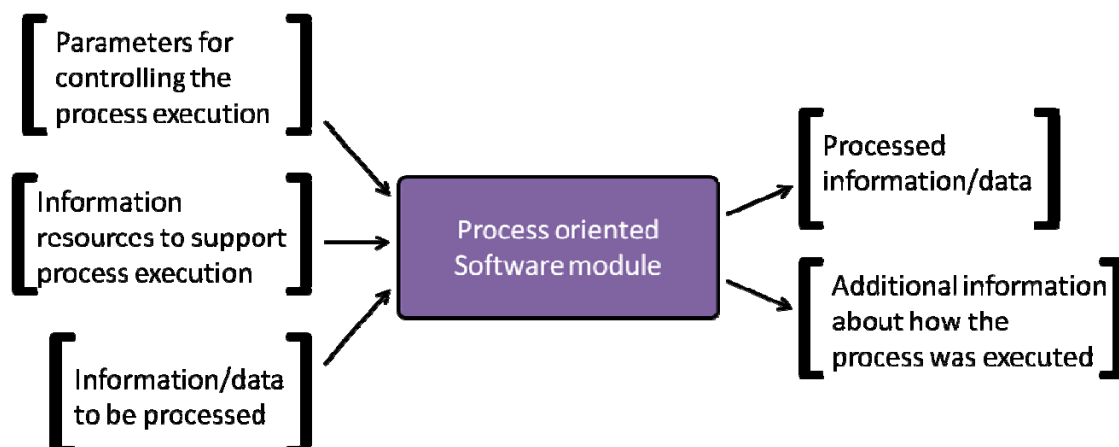
[Picture 7: Input and output of a process/process oriented software module]

B. Common international view - GSIM

39.   If process oriented modules should be used by multiple surveys within the same statistical institute it is important to establish a shared view of the business information model used within the production process. Using the same information model will enable different surveys to communicate better with each other but also to communicate new requirements more efficiently.

40.   The next step in this evolution is to create a commonly shared view of the business information model on an international level. The *GSBPM-model* is today a commonly shared view of the statistical process and an information model to correspond with this process model will enable more efficient international collaboration to take place. Before statistical institutes will be able to share software solutions such as a *web questionnaire tool*, we need a shared view of the information objects (*questionnaire, question, option etc*) that this tool uses.

41.   The work of creating a shared view of a *Generic Statistical Information Model,* which corresponds with the *GSBPM-model,* has been initialized in different forms of international collaboration.

VI. International collaboration

A. Sharing software modules

42.   Due to increased requirements of efficient software development within statistical institutes there is also an increased interest in international collaboration for software development. To be able to share software modules, it is important that we have a common understanding of which type of software modules that could be shared and that new software modules are created in a way that increases the possibility of reuse.

43.   Even though a lot of statistical institutes have adopted the *GSBPM-model* as a way of viewing their process, this does not mean that every statistical institute has exactly the same way of performing this process. The chronological order of processes and activities vary and different statistical institutes put a different amount of effort in different processes. This means that software modules that represent the internal view of chronological order of processes will be difficult to share. Module-layers that contain information about this are the *workflow-layer* and the *process dashboard* and these will therefore be difficult to share.

44.   The infrastructure of each statistical institute will also vary. Statistical institutes will have their own infrastructure in form of operating systems, security platforms, database management system etc. This means that sharing *data warehouse-modules* and other modules that could be categorized as statistical infrastructure will also be difficult.

45.   Sharing software modules that support a specific GSBPM-process or represent an implementation of a specific statistical method will however be an efficient way of collaboration. By minimizing dependencies towards specific data warehouse-solutions, specific infrastructural platforms, requirements of a specific chronological process order etc, these modules will be easier to share between statistical institutes.

46.   When offering a specific process oriented software module for sharing, a good way of explaining what is does is to make a reference to the *GSBPM-model*. To further explain how the software module works, a description of *Business Information Objects* is needed.

B. Preparing the internal architecture

47.   By establishing a *Dashboard-layer, Service- & workflow-layer, Data warehouses* and other statistical infrastructural layers, each statistical institute will be better prepared for efficient international collaboration. By creating good modularization architecture, a statistical institute will be able to make use of shared process

oriented software modules by connecting these to the existing architecture instead of creating a new architecture for each used software module.

48.   The work of creating a common view of our *Business Information Objects* will take a long time and if changes are made to the process, the information model must also be updated to correspond with the new process model. This means that all software modules will have some form of dependency to the information model just as all software modules are in some form dependent on the process model. A way of preparing the architecture for this kind of changes is to implement an *anti-corruption layer.* This should be implemented in the *workflow-layer* and could be used to translate between different views of the same *Business Information Object.* If the interpretation of what a *"Questionnaire"* is differs between two process oriented software modules, the *anti-corruption layer* in the *workflow-engine* could be used to translate between these different interpretations.

C.   Deployment strategies

49.   The development of new deployment solutions has lately created a lot of opportunities for creating more efficient software deployment strategies. Virtualization solutions enable statistical institutes to even further decrease dependencies between software modules since it creates a buffer between the software modules and the infrastructure. The possibilities of using shared software component will to a much lesser extent be limited by which technology it was implemented in.

50.   Another recent development is the possibility of using cloud based software modules. This means that statistical institutes could use software modules offered in private, mixed or public clouds to support different types of processes. Cloud based software modules exist today that could support data collection processes and dissemination processes. Different statistical institutes could however be required to follow legislations that prevent usage of this.

VII.   Conclusion

51.   A common challenge for many statistical institutes will be to shift their current survey-oriented IT-systems to process-oriented IT-systems. By establishing separate architectural layers to handle *Process Dashboards, process oriented modules, services and workflow-engines and data warehouse-modules*, statistical institutes will be better prepared to more efficiently make use of shared process oriented software modules.

52.   By minimizing dependencies towards specific data warehouses, security-platforms etc when developing process oriented software modules, statistical institutes can more easily collaborate and share software modules.

53.   A commonly shared view of the *Generic Statistical Information Model* could be seen as the next cornerstone for international collaboration and it is important that the next generation of architectures, to support *GSBPM-processes,* is designed to incorporate this efficiently.