Business Management of Data Quality in Information Systems

Assoc. Prof. Valentin Kisimov, Ph.D.
Department "Information technologies and communications"

Resume: Multiple applications run in the information systems lead to creation of different levels of non-quality data – from partially missing data, through doubled data with different content, to non-relevant data to the business service. Today's integrated business partners' system using B2B information flows, also lead to non-relevant information content. Depending of the level of non-quality of the data, the business damages can vary, but there are always there.

A new Method for management of the data quality is offered in the article, working with business rules. ICT solution is offered for supporting the Method, providing either fully automation, or partially automated tasks when a business specialist intervention is required. Business data classification is developed according to principles of storing and according to principles of logical values of the data. Technical solution for quality management of each type of data is provided.

Evaluation Parameters for management of data quality are developed while for each parameter an appropriate technical solution is offered. The entire ICT solution is proposed on a level of Conceptual design, with used Supporting system, and with specially designed Creator of business communications. Technical solution is presented for each Method's task. In the article are presented the used software components, through which the tests have been made.

Key words: Data Quality (DQ), Business data classification, Method for Dynamic management of DQ, Evaluation Parameters for DQ, SOA for DQ, ICT management of DQ, Creator of Business communications.

JEL: C6, C63, C8, C81, D8, D85.

1. Introduction

According Teradata reports (a division of NCR) and Intelligent Solutions Inc, USA, [10][12], the problem of bad data quality cost to the business in the USA more than 600 billion dollars. It is obvious for any type of company – small or big, that the data quality problem is very serious. The analysis of the problem sources in data quality demonstrates their link with data entry, with the integration processes of incoming data into the systems from different distributed sub-systems of the same information system, as well as with incoming data from business partners to the business systems [5]. A main reason of data quality problems is also merging of a few companies and merging of their data [4]. All these symptoms can be seen on the Bulgarian market as well, for which reason the data quality is not only a problem to foreigner companies, but also to Bulgarians.

In general, the companies need to have information systems with accurate strategic data, on which to relay. To manage the data quality, it is necessary to clarify what is "quality of the business data" and how it influences

the business. Also, an appropriate level of data quality is sufficient for one company and not sufficient to another. The resolution of the data quality problem requires multidimensional approach and it is much more than a technical solution [3]. On top of this, the problem solving is not a single act; it consists of many steps needed to run continuously to support the relevant permanent data quality level, on all incoming and outgoing data.

In the literature there are few methods and models for data quality technical evaluation [14] [15], which are in two directions – manual analysis on the stored data, and automatic data analysis after their storing. These methods are not related to the business content of the data and may be applied only for keeping technical level of the data quality. The other problem of the mentioned methods is that they are working on a static data – data already recorded – in files, databases, repositories, and not on data being in transit to the places where they will be treated or stored.

The goal of the current paper is to offer a new Method for creation and maintenance quality of business data, according to formulated business rules, applicable for static data (stored already) and dynamic data (data in transit). The paper has also a goal to offer an Information and Communication Technology (ICT) solution, supporting the Method. The proposed ICT solution includes automatic steps for checking the data quality, as well as a solution helping a person (business data quality specialist) to execute appropriate steps (some of the decisions to improve the data quality cannot be taken automatically – they need human intervention and final business decision). The proposed Method and ICT solution support dynamic data quality, based on dynamic changed business rules. In this way the delivered data quality will correspond to the dynamically changed business needs of the real life.

2. Data quality is a business problem, not technical problem

The term "trusted data" is a variable term and depends on current content of the business rules. The Information system has to have "true data" in dynamic aspect, depending of the dynamics of the business rules [1].

Often, data are created from one application, and used by another application [7]. It is often also two independently created applications from different vendors, to lead to creation of the same data in many copies, but with different level of completeness and accuracy, creating different level of "business truth". The current Business-to-Business integration leads to automatic acceptance of data from a business partner, which data not always correspond to the business goals [6]. The data quality problem is a problem of the management of the entire information system and of the entire business goal of the information system [8][9][13].

From technical point of view, data quality is expressed mainly in *acceptance, storing, protecting, giving access, and accessing* the data, where from technical point of view what is the content of the data is not an issue and is not a point of attention. The value of the data, semantics of its connection, the interrelationships with other data, the completeness of all necessary data for a process, is purely business issue.

From business point of view, the data quality is expressed mainly in *completeness* (whether all necessary data for execution of a process are available), *consistency* (whether the data participating in a single business process are with non-contradiction values), *correspondence* (whether the data participating in a single business process are with values within business reasonable rages), *accuracy* (whether the data value precision and exactitude correspond to the needs of the business process in which

they will participate), *timeliness* (whether the data creation is on time suitable for the time of running the business process using that data).

It is also necessary to know how much cost one percent not-quality data to the company and which are those important data. That is way the data quality is a business issue, and the purpose of an ICT system is to provide that quality.

3. Business data – the base for data quality management

Business processes are based on business data. Business data has to be well identified, with standardized formats and fields, to have relevance with the business processes where they are used, not to have duplication, and to have business relational integration, through which the business quality is provided.

For the purpose of focusing on business data quality, I will make a detail classification of the business data, and based on it I will identify how to provide the necessary data quality.

Data in an Information system can be classified by two principles:

- *According to the persistency;*
- *According to their logical value*.

According to the principle of persistency, data can be defined as:

a) Stored data;
b) Derivative data;
c) Data in transit.

The *stored data* are data already written in files, data bases, or in Repository for structured or unstructured data. Their writing is provided by execution of application or system programs. The *derivative data* are created dynamically from stored data, using some extraction and treatment processes as calculation / aggregation / relation / association / etc., when is necessary their use. *Data in transit* are data in the communication systems and are data in a way to the corresponding data treatment systems, such as files, data bases, or in Repository for structured or unstructured data, or any system or application program. If there is a good data quality system, the data in transit has to be checked for appropriate quality, before to be sent for storing or used for treatment. In this case if the quality of the transient data is not right, this data has to be rejected.

According to the principle of logical value, data can be defined as:

a) Technical atomic data
b) Business atomic data;
c) Technical aggregated data;
d) Business interim data with sub-categories:
- Business aggregated data;
- Meta data.

e) Business end-data.

Technical atomic data is the lowest level of data aggregation, having value without distinguishable business content, and they are usually a field from a file record, a column in a system database table, operating system variable, etc.

Business atomic data are also from the lowest level of aggregation, in which the value has some business content, and as a separate data it has identified business value, for example company name, product price, client's physical address, etc.

Technical aggregated data are set of atomic technical and atomic business data, which do not have business meaning, and their aggregation is mad only for technical meaning and technical purposes, for example the table SYSTEM. SYSCOLSTATS in DB2 DBMS uses indexes in database and consists of many columns,

representing technical and /or business atomic data, but combined in this table only for the purpose of DBMS performance.

Business aggregated data represent also set of atomic data, but as a collection of data has business meaning. For example a database table "Supplier" has technical atomic data (internal sequence number in the table), and business atomic data (supplier's name, supplier's address), but as an entity this table is a business data. Business aggregated data can be organized into hierarchies.

Meta data are data for the data. They represent a number of atomic data; they have a business meaning and abstraction; and serve to manage those consisting data. Meta data are different from the aggregated data that they can contain additional information non-present in the atomic data, such as descriptors, pointers, management directories, semantics and other management data. Data-flow can also be part of the additional information to the meta data. Meta data can participate in a hierarchy of meta data.

Business aggregated data and meta data are combined into a term "*Business interim data*", which means they are not atomic business data, but can be used for creation of next higher level business data – either higher level Business interim data, or Business end-data.

Business end-data are data with business meaning like business reports, business graphics, business trends and etc. Generally the Business end-data are not stored for the further treatment; they can be stored only for backup or archiving principles.

Technical data are only data serving the business processes. When is going to deal with data quality, there is only one focus – the quality of the business data. Having in mind that the Business end-data are only resulting data, we will consider for the data quality management only three types of business data – Business atomic data, Business aggregated data, and Metadata.

4. Method for Dynamic business management of data quality

The author has developed a Method for Dynamic business management of data quality, which consists of 2 phases:

a) Phase 1 – Initial preparation;
b) Phase 2 – Providing dynamic quality.

The Method is presented graphically on Figure 1.

Phase 1 "Initial preparation" executes the tasks before the main tasks for providing dynamic data quality. This phase has the aim to do initial data cleaning in the Information system. In general, there is a set of data in each Information system, which data are in some level not in good quality. The Phase 1 does the initial data cleansing of that data. The phase executes two tasks. The first task (Task 1-1) is Analysis and provides Identification, profiling and creation of meta data 1 (MD1). By analyzing the information system, the data has to be profiled and based on this the MD1 will be created. I call this meta data number 1 (MD1), because they are created before starting the essential part of the Method – Phase 2. The second task (Task 1-2) does initial data cleansing in the system, where MD1 leads the cleansing process.

The Phase 2 – "Providing dynamic quality" is the essence of the Method, and it executes every time when there are changes in the business rules, or new business rules are created. For this reason the entire phase is a cycle running periodically. At the same time the tasks 2-5, 2-6 and 2-7 are running automatically in a permanent cycle and control all the time the data quality in the Information system. In this way the en-

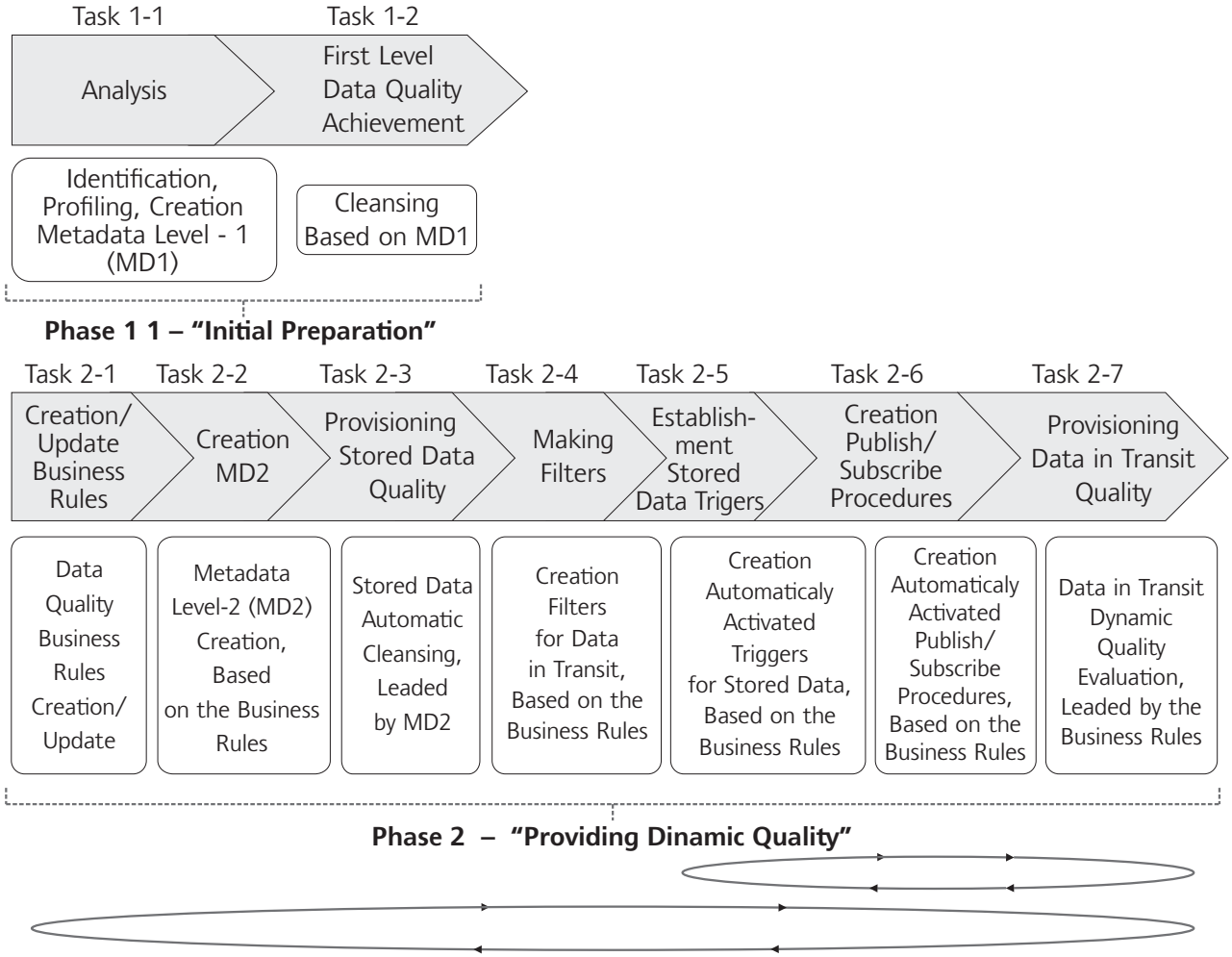

Figure 1

tire Phase 2 is a single cycle running periodically when there is a change in the business rules, in which there is another cycle running all the time checking permanently the data quality.

Task 2-1 is the creation of the business rules for data quality. This task does also business rules update. Business rules are described through the language BPEL (Business Process Execution Language), which a business and not IT language and through which the business processes are presented (modeled). This language is not a programming language for creation of applications. Using BPEL for describing the business rules leads to provide business control to data quality. The offered below ICT system supporting data quality uses BPEL Processor (interpreter), so each change in the business rules – change the way and level of controlling the data quality.

Task 2-2 creates the extended meta data 2 (MD2) – based on MD1 and the business rules for data quality (formed during task 2-1). MD2 is the base of the Phase 2 functionality. Based on MD2 is provided one time Data cleansing in the Information system (task 2-3). Based on the same MD2 are created Filters for selection of the data in transit for future control for data quality (task 2-4). The purpose of the filters are to select those application and business messages between all data moving in the enterprise network, which are target for data quality control. In the proposed below ICT solution supporting the Method, the IP packets are transformed to application messages

(for example HTTP, SAP, FTP, MQSeries and etc.), where one of a few application messages create a business message. In the current Method, business messages are extracted from the moving in the network business data, which are analyzed for the appropriate quality through the BPEL rules. The creation of the mentioned filters serves to extract only the appropriate business messages in which there are business data for controlling, and then to extract the appropriate data for checking their quality. The filters are made based on XML descriptions.

The task 2-5 establishes Triggers in the existing databases, for generation events related to already stored data. The triggers are activated with changes in some data, when the triggers logic corresponds to the business rules from the BPEL code. The triggers can be created with Java or other specific database languages. BPEL description defines for which data to establish triggers. It is feasible to create a trigger for a single data, and after that to use it to manage different levels of data quality for that data. The triggers can inform directly the system supporting task 2-7 – Provisioning Data in transit quality, named "Data in transit Dynamic Quality evaluation leaded by the Business rules", but also they can send information to the Publish / Subscribe procedures defined in task 2-6. The Publish / Subscribe procedures have the goal to aggregate information from different triggers, incoming asynchronously. These procedures have "Subscription" for type of information, which generally means to collect information from a few triggers, executing appropriate rule. The subscription rules are derivatives from the BPEL business rules for data quality. When the information from the triggers arrived in the subscription procedure, it is treated for possible future publishing of event to the corresponding system "Data in transit Dynamic Quality evaluation leaded by the Business rules". That system provides in fact the real checking of the data quality – task 2-7.

5. Key indicators for management of data quality

The current paper offers a set of Key indicators for evaluation of the data quality, which are critical in the functioning of the proposed Method, as well as critical in the operation of the ICT system supporting the data quality. The Key indicators are:

- *Completeness*

This indicator measures the existence or missing of data. For example, in the banking systems, the client's address is collected from special billing reports created by the Municipality or Utility companies, and this data are not always complete. The completeness is usually 95 %, but this 5 % incompleteness can bring big financial damages to the bank. To realize full completeness of data, the data has to be cleansed. The cleansing can lead to three options: correction of the incomplete data with new source of data to make it "true" data; keeping the uncompleted data in that way, but register somewhere the incompleteness; deleting the uncompleted data.

- *Consistency*

Consistent data are those in which having duplicated data, both are with the same content. The consistency can be transactional – during the execution of a single transaction, or stored consistency – a consistency for stored data.

- *Availability*

This indicator requires data to exist, when they needed for treatment (for example in an electronic form of document-order, the field "number" cannot be empty). The indicator is often valid for cases where data coming from external partner do not correspond to the business relationships with the partner.

- *Relevance*

This indicator requires the data values to be in an acceptable range or to be with defined set of ranges.

- *Accuracy / Precision*

The Accuracy presents how much the data value is equal to the "truth" value. The precision is the measure of this indicator. So called "truth" data can be identified only via correct business understanding of the data nature.

- *Timeliness / Freshness*

Data from a Stock Exchange which are published are always old, while data from the historical events are always fresh, timeliness. Data in Business information systems are always old, but for some business processes this "oldness" is acceptable and it is not important. This indicator uses the time for storing data and the time until which the data are correct. The difference between those two times defines the freshness of the data.

6. ICT solution for Business management of data quality

6.1. Conceptual architecture of the ICT solution

The conceptual architecture of the proposed ICT solution, which supports the offered Method, is presented in figure 2.

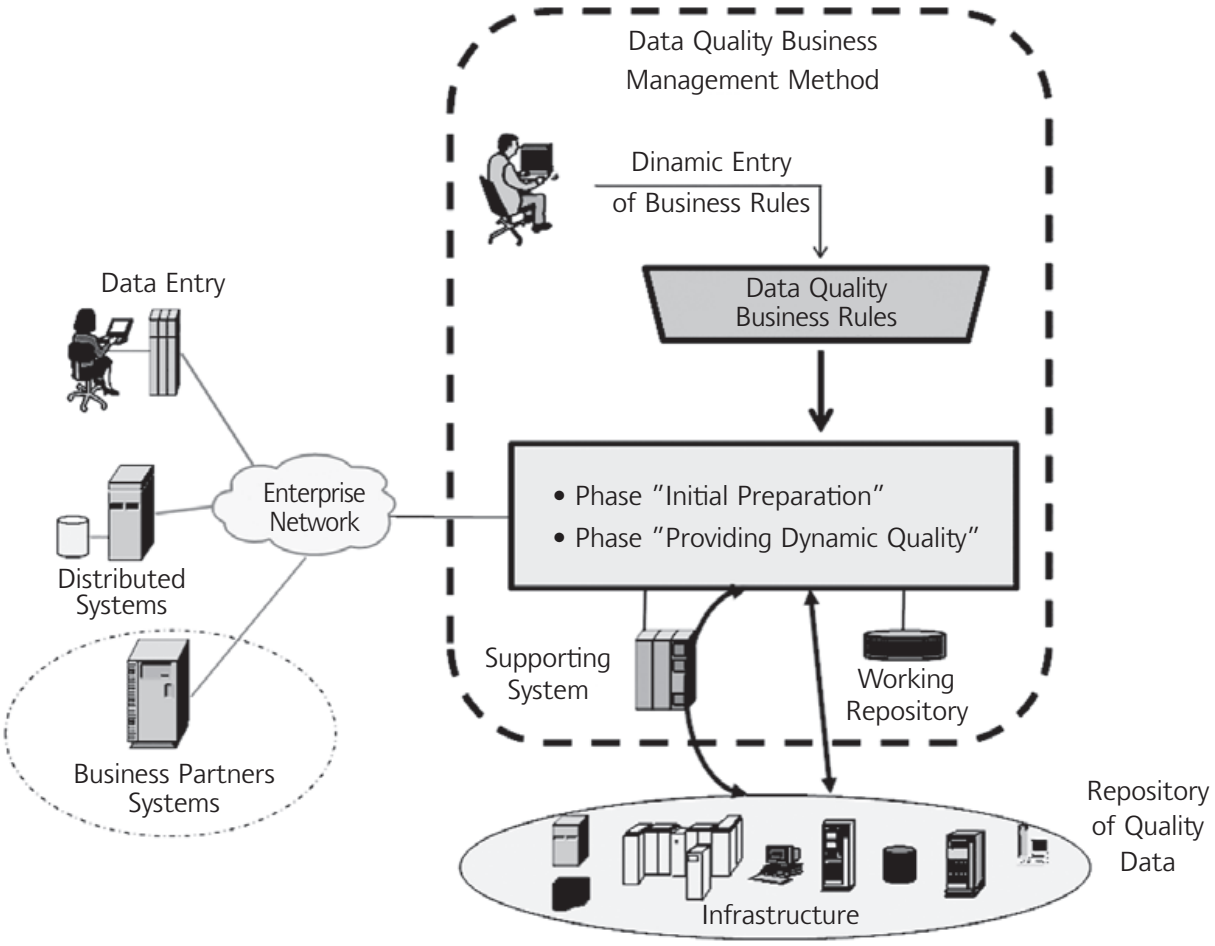

Figure 2

All repositories for data (file systems, DBMS, Repositories for structured and unstructured data) are in the Infrastructure. Data in those repositories and the data in transit have to be controlled for quality via business rules. The data in transit are coming from the enterprise network which means: from the data entry systems; from other systems of the distributed system; and from other partner systems working in Business-to-Business mode. The proposed ICT solution consists of: Component "Data quality business rules", which consists of business rules created via BPEL language; a component which manages the two phases of the Method – Initial preparation and Providing dynamic quality; Supporting system executing main data quality processes; and Working repository keeping temporary data. The proposed ICT solution uses a business specialist to lead and manage the process related to data quality, as well as executing automatic processes for providing dynamic quality. As a result of functioning of the proposed ICT solution, the Working repository will have quality of data required from the business rules.

Table 2

No BPEL ser-vice	Data quality Key Indicator	Description of the BPEL service	Parameters in the BPEL service
1	Completeness	Data cleansing according to "truth" data" with specified ending way	• „Truth data" (constants and variables) • Ending way (correction, rejection, keeping with additional data)
2	Consistency	Checking the value of duplicated data	• Type of consistency (Stored – Stored; Stored – In transit) • Stored data (type of data in MD2) • Data in transit (Partner, Type MD2, type data in MD2)
3	Availability	Checking for data availability – stored or in transit	• Data for checking (stored / in transit) • Condition for availability (! = null; ! = 0; ! = space) • Data in transit (Partner, Type MD2, Type data in MD2)
4	Relevance	Checking for data value to be is appropriate range	• Type of relevance (Stored / In transit) • Stored data (Type data in MD2) • Data in transit (Partner, Type MD2, Type data in MD2)
5	Accuracy	Define Time period (TP): t2-t1; tcurrent-t. For TP: Current Accuracy (CuurAc) = √((InitAc – TruthAc/ TruthAc)↑2) PossAc) – CurrAc < 0	• Truth Accuracy (TruthAc) • Possible Accuracy (PossAc) • TP (t1÷t2) • Initial moment (ti)
6	Timeliness	Recording arrival time (tarr) with Data name Checking tcurr – tarr < ttime	• ttimeliness (ttime) • Data in transit (Partner, Type MD2, Type data in MD2)

Business rules are entered into the solution by a business specialist, using BPEL (task 2-1). Tasks 1-1 and 1-2 are executed with the help of the Supporting system, under the control of the business specialist responsible for data quality. Tasks 1-2 and 2-3 are executed automatically by the Supporting system. Tasks 2-4, 2-5, 2-6 and 2-7 require respective programming. Those 4 tasks are the tasks with longer duration and initially require a few men-months. After that, for changing some business rules, it is required much shorter duration for re-coding of the tasks–few men-hours, few man-days, or few man-weeks. This means when changing BPEL business rules, an interval of a few hours-days-weeks is required to tune the ICT solution.

The used language for data quality business rules – BPEL, is a standard language for modeling business processes, which can use the services of Service Oriented Architecture (SOA) [2]. BPEL is a language for business rules orchestration (having dependencies between different services with a central management), and it is not for choreography (using dependencies between services in pairs). BPEL supports creation of aggregated business services, treated also as a service. BPEL is supported currently from the biggest software providers.

The proposed ICT solution uses specific BPEL services – one per each data quality Key Indicator. I have defined 6 Key indicators and for this reason I have created 6 BPEL services in the ICT solutions. For example, for the indicator Completeness is created a BPEL services "Completeness". The entire data quality is managed with these 6 BPEL services, with using different parameters for each of them. The parameters specify for which data to be applied the appropriate BPEL service, and with which details for functionality. It is possible to create unlimited number of requirements for each of the 6 BPEL services, using unlimited number of parameters. There is a buffer for each of the 6 BPEL service, and the parameters are recorded into that buffer, providing bigger performance. The implemented 6 BPEL services are in a form of Java objects – POJO (Plain Old Java Object). It is used a SOA in the proposed ICT solution, which executes all the tasks of the Method. This architecture is based on the product ServiceMix (an open source SOA), which supports JSR181 standard for EJB/POJO.

In the table 2 are shown the parameters for each of the 6 BPEL services, together with the description of the service.

6.2. IBM Information Server as Supporting system for Business management data quality

The proposed ICT solution use IBM Information Server [11] as Supporting system. IBM Information Server provides profiling, cleansing and does data integration from heterogeneous systems. It provides data integration leaded by meta data. IBM Information Server offers many functionalities in the form of web services, such as Improving data quality (using mainly data cleansing); Supporting data transformation (using mainly ETL processes); providing common view of all data – structured and unstructured.

IBM Information Server components are: WebSphere Information Analyzer for defining relationships between data in different physical media; WebSphere Business Glossary – for strong meta data descriptions; WebSphere Quality Stage – for data cleansing based on rules; WebSphere Data Stage – for execution of ETL procedures; WebSphere Federation Server – for linking different data management systems independently of the data forms and the type of management; WebSphere Information Server Directory – serving for publishing of the existing web services; WebSphere Metadata Server – serving as reposi-

tory for meta data, providing also services to meta data such as Access, Integration, Import-Export, Analysis, Research.

6.3. Business Message Builder

The proposed ICT solution use "Business message builder" to treat the incoming messages. This Builder consists of devices and procedures, leading to creation of the business messages. The IP packets from the network are transformed into application messages (using the Convertor), and after that they are transformed into business messages (using a devices based on SOA). Placing the SOA in the middle level of the architecture, e.g. in a middle layer of 7 layered ISO model, this SOA I have called Middleware SOA (MSOA) – figure 3. Again ServiceMix product is used for the designed solution, in which BPEL Processor is incorporated.

The device "Convertor" consists of 2 sub-devices: Cisco Switch or Router with included Cisco AON blade, and Network management server. AON blade is a specialized hardware card with loaded AON software for extraction of IP packets and conversion them into defined requirements. AON is programmed as J2EE application server and in the currently proposed ICT solution it works a specialized proxy server, for redirection of the IP packets to the Network management server – figure 4.

MSOA uses Enterprise Service Bus (ESB), working with the principle of JBI (Java Business Integration) –JSR208 standard. The way in which it transforms a message into a service is presented in figure 5.

Each request from external user of services, such as business messages, is converted into

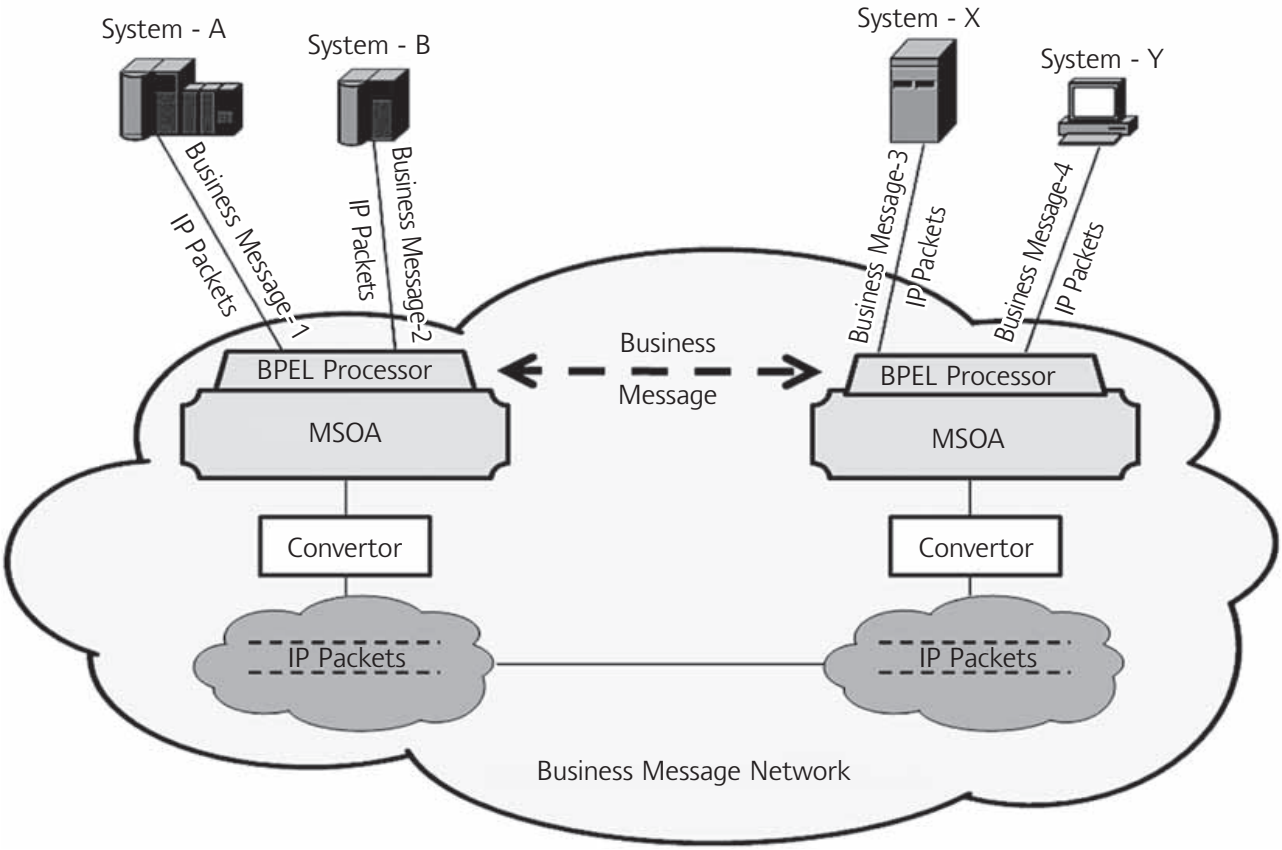

Figure 3

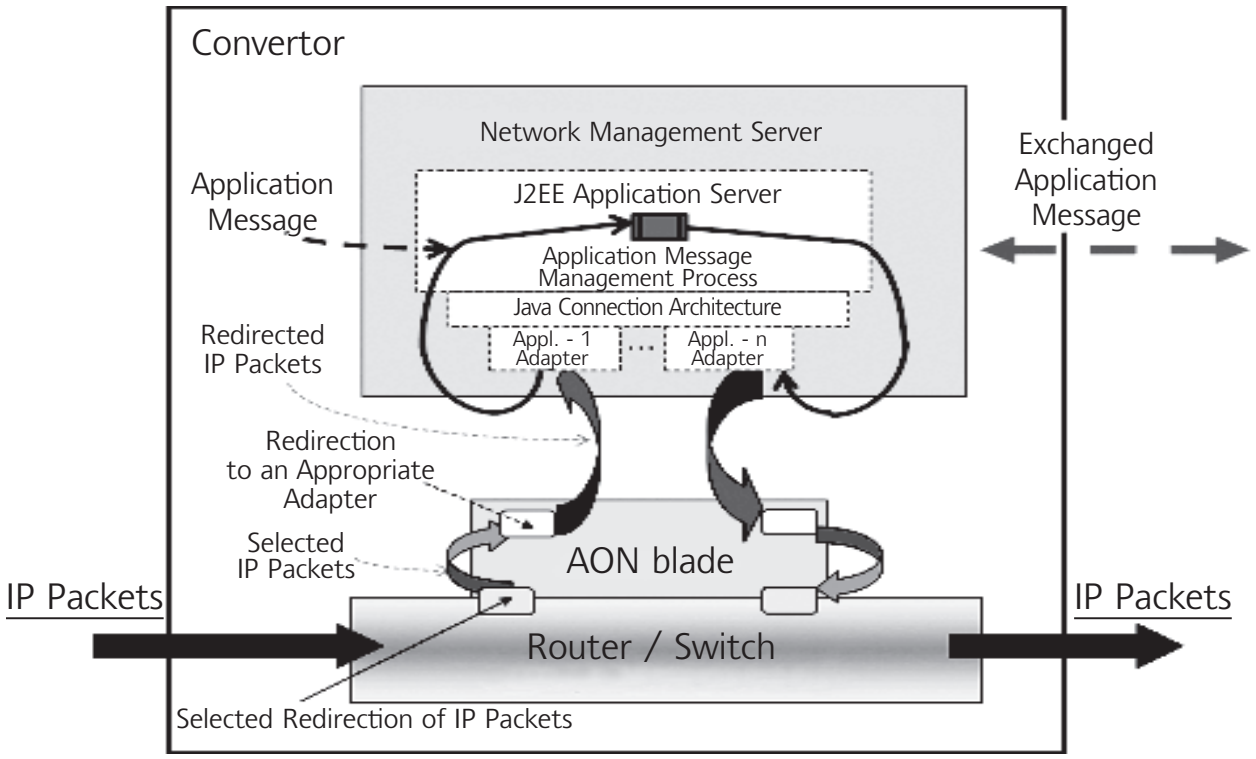

Figure 4

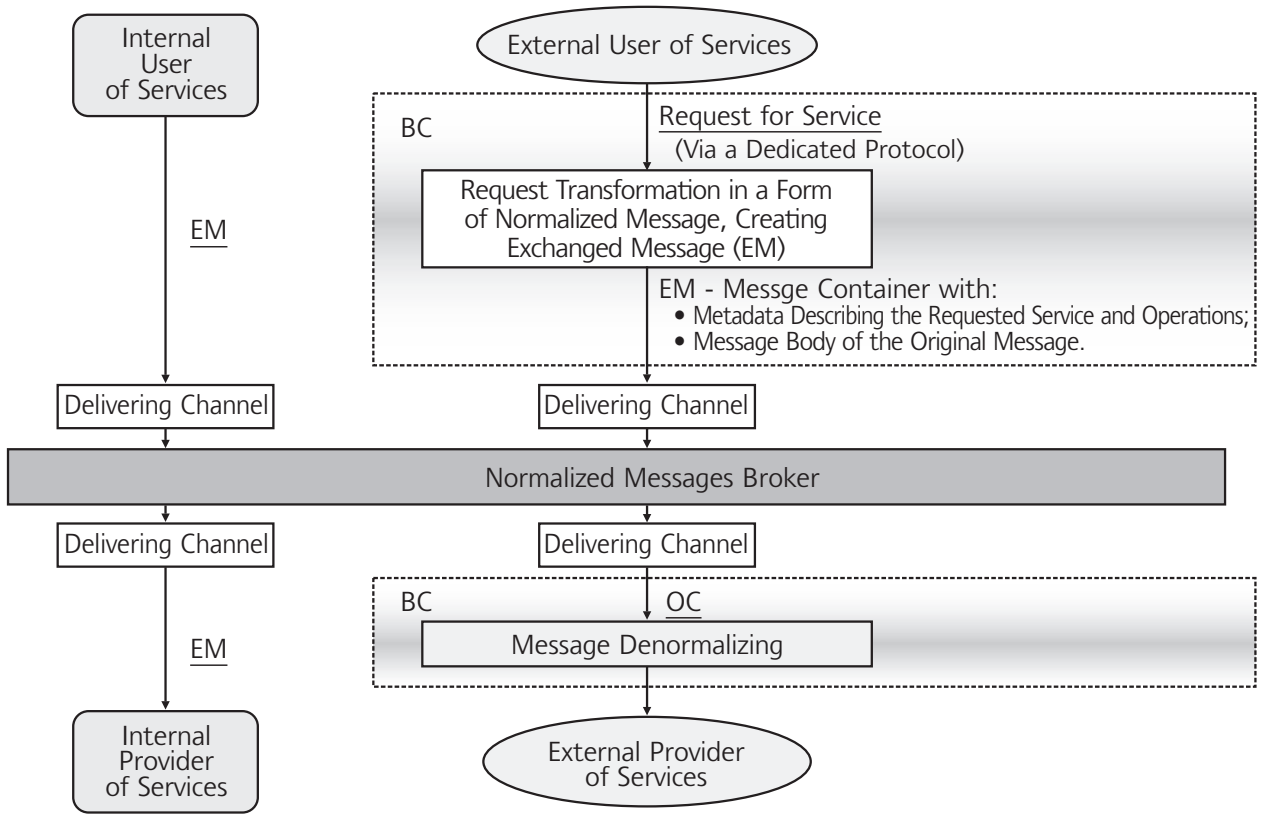

Figure 5

Binding Components (BC), via conversion of the message to Exchanged Message (EM) – adding a metadata on top of the original message. The received in this way Exchanged Message is sent to the Delivering channel, which send it as a Normalized message for the Normalized messages Broker. The last one plays a role as a router for Normalized Messages.

6.4. Task Technology execution of the proposed Method

It is required to use IBM Information Server for the execution of tasks 2-1, 2-2, and 2-3. Business data quality specialist has to run the execution of those tasks – figure 6.

To run the first two tasks, business data quality specialist needs to have appropriate business vision. After the manually supported establishment of MD2 (end of task 2-2), the rest of the tasks run automatically. The business data quality specialist just starts the tasks, identifying additionally correlation between the business data, if IBM Information server cannot find all of them automatically via its analysis, profiling and semantic mechanisms. The execution of the rest tasks from Phase 2 is shown in figure 7.

The tasks 2-4, 2-5, and 2-6 require some programming, when business rules are changed, but the execution of these tasks is run automatically. The task 2-7 also requires programming, but not related to any change of

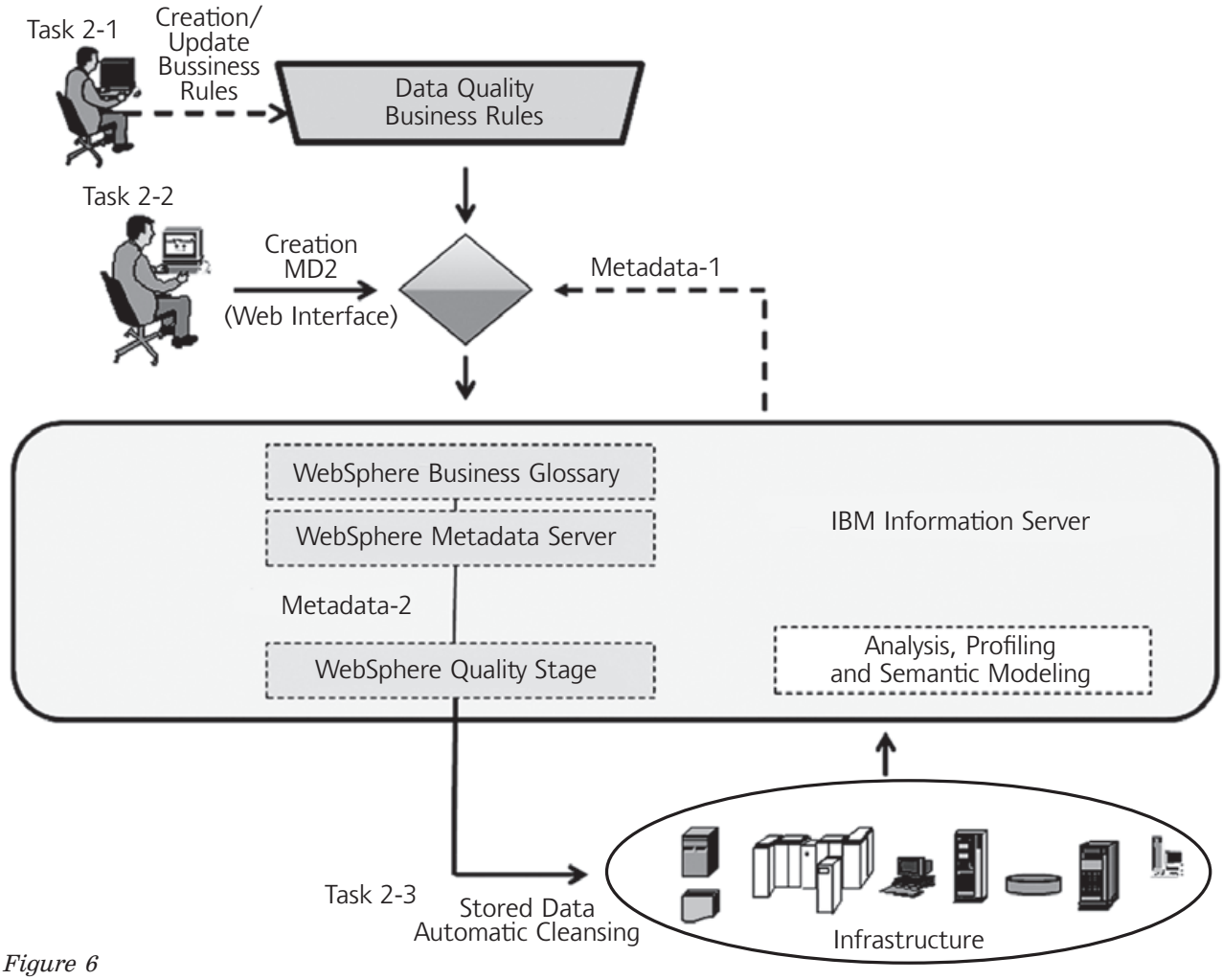

Figure 6

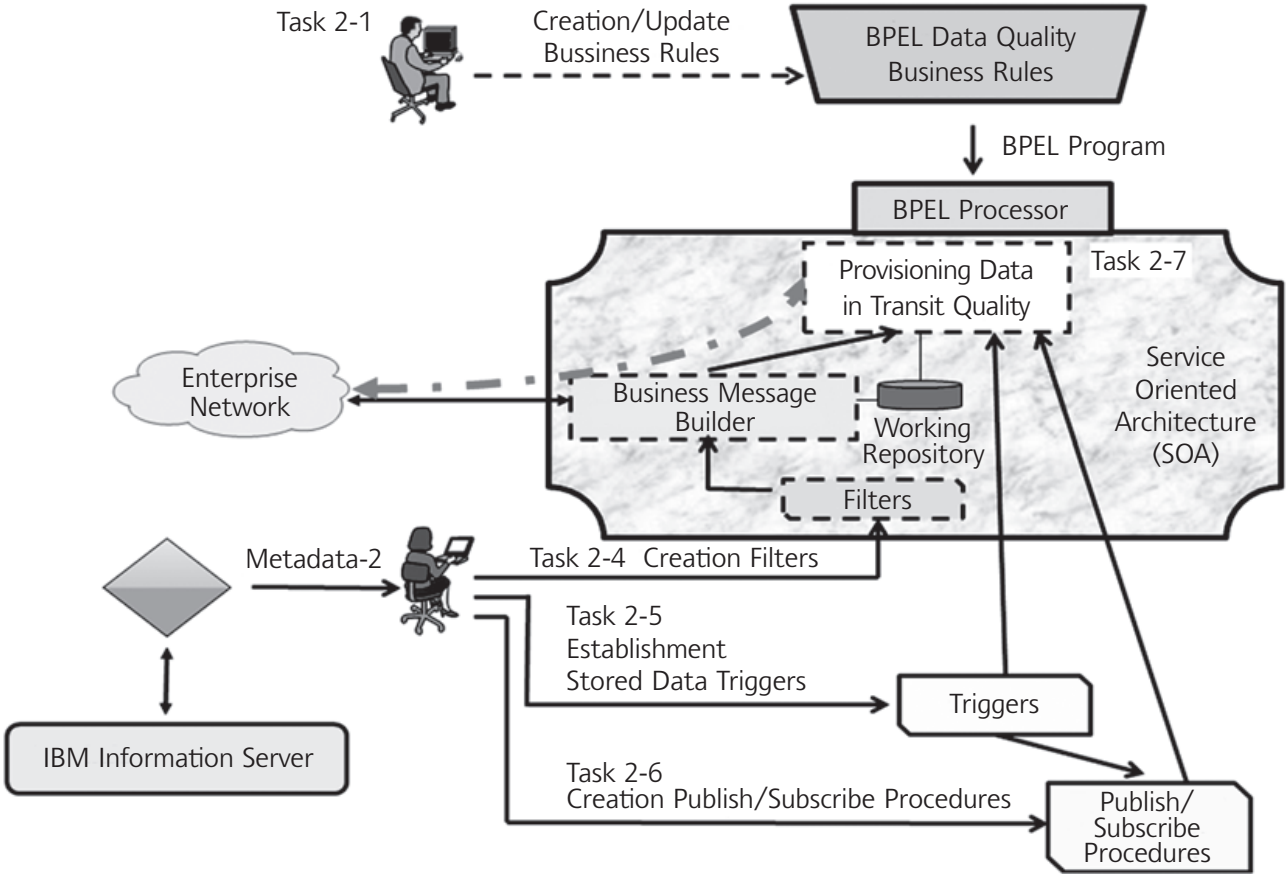

Figure 7

the business rule – the task requires programming on the initial moment – when the ICT solution is created. This programming of task 2-7 can be accepted as part of the creation of the ICT solution.

References

1. Kisimov, V., R. Nikilov, Business intelligent infrastructure – base for Business Intelligent systems in real time, Proceedings of the International Conference organized for celebration of the 40 anniversary of university course "Informatics", under scientific edition of proof. A.Batchvarov, Sofia, 2007.

2. Velev, D., E. Dentchev, K. Stefanova, V. Lazarova, M. Zaneva, A. Murdjeva, SOA – Main trend in software technology evolution, Proceedings of the International Conference organized for celebration of the 40 anniversary of university course "Informatics", under scientific edition of proof. A. Batchvarov, Sofia, 2007.

3. Stefanova, K., V. Kisimov, M. Zaneva, V. Lazarova, A. Murdjeva, D. Velev, E. Dentchev, D. Kabaktchieva, Business Intelligence Competence Centre design, Proceedings of the International Conference organized for celebration of the 40 anniversary of university course "Informatics", under scientific edition of proof. A.Batchvarov, Sofia, 2007.

4. Dentchev, E., K. Stefanova, M. Zaneva, D. Velev, V. Kisimov, A. Murdjeva, V. Lazarova, Problems and solutions in selection of ERP systems, Proceedings of the International Conference organized for celebration of the 40 anniversary of university course "Informatics", under

scientific edition of proof. A. Batchvarov, Sofia, 2007.

5. Zaneva, M., V. Kisimov, A. Murdjeva, E. Dentchev, D. Velev, V. Lazarova, K. Stefanova, Business applications integration – main challenge in system software design, Proceedings of the International Conference organized for celebration of the 40 anniversary of university course "Informatics", under scientific edition of proof. A.Batchvarov, Sofia, 2007.

6. Lazarova, V., E. Dentchev, D. Velev, A. Murdjeva, V. Kisimov, K. Stefanova, M. Zaneva, Criteria for user efficiency defined for Internet based Information systems, Proceedings of the International Conference organized for celebration of the 40 anniversary of university course "Informatics", under scientific edition of proof. A.Batchvarov, Sofia, 2007.

7. Murdjeva, A. K. Stefanova, M. Zaneva, V. Lazarova, E. Dentchev, D. Velev, Distribution of the business logic in multi-layered applications using current technologies, Proceedings of the International Conference organized for celebration of the 40 anniversary of university course "Informatics", under scientific edition of proof. A.Batchvarov, Sofia, 2007.

8. Poor Quality Data: The sure way to lose business and attract auditors, Gartner conference, SA, 2006.

9. Data Integration, The Nucleus of the Date Delivery Evolution, Gartner Conference, SA, 2007.

10. Data Quality Management: Oft-Overlooked Key to Affordable, High Quality Patient Care, Darryl McDonald, Teradata, NCR.

11. IBM Information Server, IBM, http://www-304.ibm.com/ jct03002c/software/data/integration/info_server/overview.html

12. US Public Company Accounting reform and corporate responsibility, http://www4.law.cornell.edu/ uscode/15/usc_sup_01_15_10_98.html

13. Basel II: Revised international capital framework, http://www.bis.org/ publ/bcbsca.htm

14. Batini, C., M. Scannapieca, DataQuality, SpringerLink, 2006.

15. Pilar, A., M. Lachlan, Quality Measurement and Assessment models including Data Provenance to grade Data sources, http://www.macs.hw.ac.uk/~pilar/research/ATINERv3.pdf ИА