A Multi-perspective Framework for the Analysis of Legacy Information Systems

Silvana Castano[1] and Valeria De Antonellis[2]

[1] Università di Milano, Dip. Scienze dell'Informazione
via Comelico 39 - 20135 MILANO - ITALY
e-mail: castano@dsi.unimi.it
[2] Università di Brescia and Politecnico di Milano,
P.za L. da Vinci 32 - 20133 MILANO - ITALY
e-mail: deantone@elet.polimi.it

Abstract. This paper proposes a comprehensive framework for the analysis of sets of data, processes, and applications of legacy information systems in view of restructuring interventions, according to different perspectives. A "data asset structure perspective" supports the information system analysis in terms of semantic correspondences between data assets shared by different organization units and applications of the system. An "operational structure perspective" supports the information system analysis in terms of the semantic correspondences between processes, with respect to the similarity of the involved data and operations. Finally, an "organizational structure perspective" supports the information system analysis in terms of the organization unit cohesion and coupling properties with respect to the exchanged information flows. The perspectives are discussed with reference to the analysis of italian Public Administration information systems.

1 Introduction

Complex organizations (e.g., Public Administrations, international corporations) are characterized by the presence of a high number of subordinate-level organization units, that are distributed and have their local information systems, with specific application contexts and support technologies. For these organizations, reengineering activities are generally considered challenging and critical activities to be performed [1, 11]. In fact, due to the evolution of organizational requirements as well as to the availability of advanced information technology, complex organizations promote activities for restructuring and integrating their legacy information systems, pursuing different objectives, such as: (i) enhancement of the efficiency of business processes and improvement of the quality of the services and products; (ii) cross-functional data standardization and definition of sharable data structures; (iii) greater availability and accessibility of information.

Reengineering of autonomous, heterogeneous information systems requires a deep understanding of data and processes of the organization, to determine

restructuring and migration requirements [2, 7, 5, 10]. In the paper, we analyze conceptual descriptions of data and processes according to the following *perspectives*:

- *Data asset structure*: according to this perspective, we analyze conceptual descriptions for the aspects related to data similarity to construct standardized data assets describing the key concepts of the system, common to several organization units (e.g., an "Employee" data asset describing in an integrated and unified way all employee data manipulated by processes of different organization units).
- *Operational structure*: according to this perspective, we analyze conceptual descriptions for the aspects related to data and operation similarity, to establish semantic correspondences between processes of the same organization unit, or strictly related units, to identify possible replications or redundancies.
- *Organizational structure*: according to this perspective, we analyze conceptual descriptions for the aspects related to interaction/cooperation between processes, by means of internal / external information flows to evaluate the level of internal cohesion and external coupling.

The proposed framework is developed and experimented within the PROGRESS (PROcess Guided REengineering Support System) project promoted by CINI and the National Research Council, aiming at developing techniques and tools for reengineering the Public Administration information systems.

The paper is organized as follows. In Section 2, we introduce the basic elements of our framework. In Sections 3, 4, and 5, we describe the data asset structure, the operational structure, and the organizational structure perspectives, respectively, together with their associated analysis parameters. Finally, in Section 6, we give our concluding remarks.

2 Basic elements of the approach

In the following, we describe the basic elements of the analysis framework, referring to our experience in the framework of the PROGRESS project.

2.1 Conceptual descriptions of data and processes

In PROGRESS, we have at disposal Entity-Relationship (ER) conceptual schemas describing the data manipulated by processes and a textual description of process functionality and related operations. Following an information processing viewpoint, processes are described in terms of the data that are required by the process for its execution (input and output data) and the operations performed on these data [12]. To support a semi-automatic analysis of processes in our framework, information available on processes is formally represented using *process descriptors*. A descriptor D_i of a process P_i is a data-oriented representation of P_i as a 4-tuple of *features*, $D_i = \langle f_1, f_2, f_3, f_4 \rangle$. Feature f_1

contains the name n_{OU_k} of the organization unit OU_k to which P_i pertains. Feature f_2 contains the set $IN(P_i)$ of names of the input entities of P_i derived from its corresponding schema S_i. Feature f_3 contains the set $OUT(P_i)$ of names of the output entities of P_i derived from its corresponding schema S_i. Feature f_4 contains the set $FUN(P_i)$ of operations performed by P_i, that is, $FUN(P_i) = \{O_{1i}, O_{2i}, \ldots, O_{qi}\}$. Following a disciplined approach [9], an operation $O_{ki} \in FUN(P_i)$ is described as a triplet, $O_{ki} = \langle n_{op_{ki}}, CST_{ki}, CMS_{ki} \rangle$, where $n_{op_{ki}}$ is a verb (i.e., the operation name); $CST_{ki} \subseteq IN(P_i)$ is a set of entity names corresponding to the mandatory constitutive entities required for the execution of the operation; CMS_{ki} is a set of entity names corresponding to optional circumstantial entities involved in the operation execution.

An example of process descriptor is shown in Fig. 1, for process **Personnel Absence Recording** (in the following referred to as P_1) of the **General Management of Cooperation** organization unit of the Labour Ministry. P_1 is responsible for absence recording within this unit, and its associated ER schema (S_1) is shown in Fig. 2(a).

Personnel Absence Recording

f_1: General Management of Cooperation
f_2: {Employee of General Management of Cooperation }
f_3: { Absent Employee, Employee of General Management of Cooperation }
f_4: {⟨ Record, Employee of General Management of Cooperation, Presence Card ⟩}

Fig. 1. An example of process descriptor (D_1)

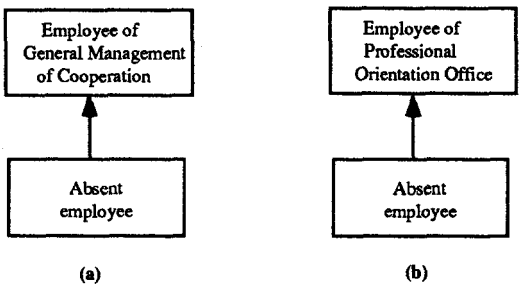

Fig. 2. Examples of ER conceptual schemas (S_1 and S_2)

ER schemas and process descriptors have been obtained through reverse engineering (based on the analysis of logical schemas of databases and on questionnaires to the information systems users) and constitute the basis for information system analysis in our framework [3]. Similarity criteria play an important role in our analysis perspectives. For similarity purposes, we need to compare entity

and operation names appearing in different process descriptors and entity names appearing in conceptual schemas, with the aim of identifying names denoting the same entity/operation or semantically similar entities/operations. To this end, we refer to a semantic dictionary. A detailed description of the construction process for the semantic dictionary is given in [8].

2.2 Semantic dictionary

The semantic dictionary contains concepts organized into *concept hierarchies*. We maintain separate hierarchies for entities and operations, to facilitate the different types of analysis in our framework. Concepts at the bottom-level of the hierarchy have an associated cluster of names, corresponding to entity names or operation names that are semantically similar in ER schemas and in process descriptors. Higher-level concepts are defined by means of the generalization and aggregation abstraction mechanisms.

For names in the semantic dictionary, we introduce the notion of *affinity*. Two names have affinity if they denote the same entity/operation or semantically similar entities/operations. To operationally evaluate affinity, we assign an affinity strength σ to hierarchical links between concepts in the semantic dictionary. Two names have affinity if they refer to a common concept in the semantic dictionary hierarchies, that is, if a path of length l, with $l \geq 1$ (denoted by the symbol "$\rightarrow^l$") exists between them in the semantic dictionary. For names in the same cluster, a path of length 1 is defined.

Definition 1 Name Affinity Coefficient. The Name Affinity Coefficient of two names n_i and n_j, denoted by $NA(n_i, n_j)$, is the measure of their affinity in the semantic dictionary computed as follows.

$$NA(n_i, n_j) = \begin{cases} 1 \text{ if } (n_i = n_j) \\ \sigma^l \text{ if } (n_i \neq n_j) \wedge (n_i \rightarrow^l n_j), l \geq 1 \\ 0 \text{ otherwise} \end{cases}$$

Definition 2 Name Affinity. Two names n_i and n_j have affinity, denoted by $n_i \sim n_j$, if their affinity coefficient is greater than or equal to a given threshold $\alpha > 0$, that is, $n_i \sim n_j \leftrightarrow NA(n_i, n_j) > \alpha$.

High values of the threshold guarantee the selection of names referring to the same concept or to concepts that are very close in the semantic dictionary.

3 Data asset structure perspective

According to this perspective, conceptual schemas are grouped into similarity families to identify a set of standardized data assets, that is, a set of "core" schemas describing the key concepts common to several organization units of the system. This core knowledge can constitute the basis for the analysis of data exchanges and show a potential for data integration in a common shared

database. The *inter-schema semantic correspondences* parameter can be considered, in relation to the number of entities that are similar in the considered schemas. Depending upon the insights the analyst is seeking, inter-schema semantic correspondences can be evaluated for:

1. *all schemas of an organization unit* to identify possible data replication/redundancy within the organization unit;
2. *all schemas of tightly coupled units* to identify schemas candidate for unification/integration, for example, for the construction of a database shared by several organization units;
3. *all schemas of all organization units* to construct reference data assets common to all the units describing key concepts for the whole organization (e.g., "Employee", "Enterprise", "Location" data assets).

3.1 Schema similarity

In order to evaluate inter-schema semantic correspondences, we introduce a schema similarity coefficient, which computes the degree of similarity of two schemas on the basis of the number of similar entities they contain.

Definition 3 Schema similarity coefficient. The *Schema similarity coefficient* of two schemas S_i and S_j, denoted by $Sim(S_i, S_j)$, is the measure of the similarity of their entities, computed as follows.

$$Sim(S_i, S_j) = \frac{2 \cdot \sum NA(n_{e_{hi}}, n_{e_{kj}})}{N_{S_i} + N_{S_j}}$$

where $n_{e_{hi}}$ (respectively, $n_{e_{kj}}$) is the name of e_{hi} (respectively, e_{kj}) with $e_{hi} \in S_i$ and $e_{kj} \in S_j$, and N_{S_i} (respectively, N_{S_j}) indicates the total number of entities in S_i (respectively, S_j). Each entity of S_i and S_j participates at most in one affinity pair (e_{hi}, e_{kj}). The schema similarity coefficient is proportional to the number of entities that are similar in the considered schemas, with $Sim \in [0, 1]$.

For example, similarity of schemas S_1 and S_2 of Fig. 2 is $Sim(S_1, S_2) = 2 \cdot (0.8 + 1)/(2 + 2) = 0.9$. In fact, entity `Employee of General Management of Cooperation` in S_1 has affinity 0.8 with entity `Employee of Professional Orientation Office` of S_2, because they belong to the same cluster associated with concept `Employee of the Labour Ministry` in the dictionary, while entities `Absent employee` have affinity 1 because they are equally defined in both schemas.

3.2 Schema families

Families of similar schemas constitute the basis for identifying conceptual elements which are present in several organizational units. This can be useful for evaluating data replication/redundancy at the intensional level between processes of the same organization unit or different organization units, to recognize

possible inadequate or undesired duplications of data at the extensional level in the corresponding databases and show a potential for data integration in a common shared database. Furthermore, similar schemas can be properly integrated / unified into standardized data assets describing the core knowledge of the organization.

Schema families can be dynamically created by the analyst starting from a selected "reference schema" S_i. In this case, the similarity coefficient(s) between S_i and the other schemas of interest are evaluated. Alternatively, schema families can be automatically constructed using hierarchical clustering techniques, as described in [3].

Given a schema family $\Im$, we can establish *semantic correspondences* between the schemas therein contained, on the basis of their similarity coefficients. Given two schemas $S_i, S_j \in \Im$, the following cases occur:

1. *Semantic equivalence*, $S_i \equiv S_j$, that holds if $Sim(S_i, S_j) = 1$. Semantic equivalent schemas contain identical entities, and in this case, we can recognize schema replication.
2. *Semantic relationship*, $S_i \sim S_j$, that holds if $T_1 \leq Sim(S_i, S_j) < 1$, where T_1 is a defined similarity threshold. Schemas with a semantic relationship contain entities describing semantically similar concepts. In this case, we envisage schema overlapping, such as for S_1 and S_2, which describe employee data for different organization units.

4 Operational structure perspective

Following this perspective, we group processes into similarity families to evaluate possible redundancy, replication, and inconsistencies with respect to data and operations. The *inter-process semantic correspondences* parameter can be considered for evaluation, in relation to the number and the type of similar entities and operations featuring each process.

Inter-process semantic correspondences can be evaluated in different ways:

1. *all processes pertaining to a unit* for a comprehensive analysis of a given organization unit;
2. *all processes of tightly coupled units* to focus on interdependency relationships due to data/process sharing or overlapping, to consider the possibility of work reassignment (e.g., in presence of duplicate exchanges).

4.1 Inter-process semantic correspondences

In order to evaluate inter-process semantic correspondences parameter, the following coefficients are introduced:

- *Entity-based similarity coefficient*, to analyze process relationships due to data commonalities. It can be a point of reference to evaluate the adequacy of data structures in the operational structure.

- *Functionality-based similarity coefficient*, to analyze process relationships due to functionality commonality. It can be a point of reference to check the adequacy of production/manipulation of information in the operational structure.

To evaluate the affinity between sets of names contained in process descriptors (e.g., $IN()$, $OUT()$), we introduce the *name set affinity coefficient* based on the name affinity coefficient. Let $X = \{n_1, \ldots, n_h\}$ and $X' = \{n_1, \ldots, n_k\}$ be two sets of entity names, with $h \neq k$, respectively.

Definition 4 Name set affinity coefficient. The *name set affinity coefficient* of two sets of names X and X', denoted by $A(X, X')$, is the measure of the affinity of the pair of names of X and X' that have affinity in the semantic dictionary, computed as follows.

$$A(X, X') = \begin{cases} \frac{2 \cdot \sum NA(n_i, n_j)}{n_X + n_{X'}} & \text{if } n_X + n_{X'} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

where $n_i \in X, n_j \in X', n_i \sim n_j$ and n_X (respectively, $n_{X'}$) indicates the number of names contained in set X (respectively, X'). Each name of X and X' can participate at most in one affinity pair. $A(X, X') \in [0, 1]$ is proportional to the number of pairs of names with affinity and to their affinity value in the dictionary. If both X and X' contain only one name, we have that $A(X, X) = NA(n_i, n_j)$. Two sets of names X and X' have affinity, denoted by $X \sim X'$, if their affinity coefficient is greater than or equal to a given threshold $\beta > 0$, that is, $X \sim X' \leftrightarrow A(X, X') \geq \beta$.

Entity-based similarity

To evaluate the entity-based similarity of two processes we evaluate the affinity of the input and output entities in process descriptors as follows.

Definition 5 Entity-based similarity coefficient. The *Entity-based similarity coefficient* of two processes P_i and P_j, denoted by $ESim(P_i, P_j)$, is the measure of the affinity of the names contained in their features related to input and output entities, computed as follows.

$$ESim(P_i, P_j) = \sum_{f \in \{f_2, f_3\}} A(D_i.f, D_j.f)$$

The entity-based similarity of two processes is the sum of the affinity of their input and output entity names, with $ESim(P_i, P_j) \in [0, 2]$.

For example, consider process P_2 **Absence Recording** of the **Central Office for Professional Education and Orientation** organization unit, whose descriptor D_2 is shown in Fig. 3 and the associated ER schemas is illustrated in Fig. 2(b). $ESim(P_1, P_2) = [2 \cdot 0.8/(1 + 1)] + [2 \cdot (0.8 + 1)/(2 +$

2)] = 1.7, since entity **Employee of General Management of Cooperation** in both $D_1.f_2$ and $D_1.f_3$ has affinity 0.8 with entity **Employee of Professional Orientation Office** in both $D_2.f_2$ and $D_2.f_3$, while entity **Absent employee** in $D_1.f_3$ has affinity 1 with **Absent employee** in $D_2.f_3$.

Absence Recording

f_1: Central Office for Professional Education and Orientation
f_2: { Employee of Professional Orientation Office }
f_3: { Absent Employee, Employee of Professional Orientation Office }
f_4: {⟨ Record, Employee of Professional Orientation Office, Presence Card ⟩}

Fig. 3. An example of process descriptor (P_2)

Functionality-based similarity

Two processes are similar with respect to their functionality if they perform the same or similar operations on the same or similar constitutive entities. In fact, even if circumstantial entities vary between processes, the semantics of process functionality remains unaltered, because it is mainly determined by process operation and associated constitutive entities. Given two operations $O_{hi} = \langle n_{op_{hi}}, CST_{hi}, CMS_{hi} \rangle$ and $O_{kj} = \langle n_{op_{kj}}, CST_{kj}, CMS_{kj} \rangle$, their similarity coefficient, denoted by $OSim(O_{hi}, O_{kj})$, is the measure of the affinity of their names, constitutive entities, and circumstantial entities, respectively, that is, $OSim(O_{hi}, O_{kj}) = NA(n_{op_{hi}}, n_{op_{kj}}) + A(CST_{hi}, CST_{kj}) + A(CMS_{hi}, CMS_{kj})$.

Two operations are similar, denoted by $O_{hi} \sim O_{kj}$, if their similarity coefficient is greater than or equal to an established threshold (i.e., $OSim(O_{hi}, O_{kj}) \geq \gamma$), and if at least their operation names and constitutive entity names have affinity (i.e., $NA(n_{op_{hi}}, n_{op_{kj}}) \neq 0$ and $A(CST_{hi}, CST_{kj}) \neq 0$). High values of γ are recommended to filter out similar operations.

Definition 6 Functionality-based similarity coefficient. The *Functionality-based similarity coefficient* of two processes P_i and P_j, denoted by $FSim(P_i, P_j)$, is the measure of the similarity of their performed operations, computed as follows.

$$FSim(P_i, P_j) = \frac{2 \cdot \sum OSim(O_{hi}, O_{kj})}{N_{hi} + N_{kj}}$$

where $O_{hi} \in D_i.f_4, O_{kj} \in D_j.f_4$, with $O_{hi} \sim O_{kj}$ and N_{hi} (respectively, N_{kj}) indicates the total number of operations in $D_i.f_4$ (respectively, $D_j.f_4$). Each operation in $D_i.f_4$ and $D_j.f_4$ participates at most in one affinity pair $\langle O_{hi}, O_{kj} \rangle$.

The functionality similarity of two processes is proportional to the number of their similar operations and to their similarity coefficients, with $FSim \in [0, 3]$.

For example, for P_1 and P_2 we have $FSim(P_1, P_2) = 2 \cdot (1+0.8+1)/(1+1) = 2.8$, because P_1 and P_2 perform the same operation (**record**) on constitutive and circumstantial entities with affinity 0.8 and 1, respectively.

4.2 Process families

Families of similar processes constitute the basis for evaluating possible replications / redundancies of processes, in the same or different organization units. For example, one could recognize that an inadequate assignment or duplications affect process execution. This case suggests to look for a possible optimization to improve flexibility in the execution. In the case of many similar processes, it is suggested to investigate on the semantic correspondences between them to evaluate restructuring activities to obtain a better specialization of a process.

Families are defined using the entity-based similarity coefficient or the functionality-based coefficient or both. Similarity-based families can be dynamically created by the analyst, every time a "reference process" P_i is available. In this case, we evaluate the similarity coefficient(s) for P_i and the other processes of interest. Alternatively, processes can be automatically grouped into families using clustering techniques, analogously to data schemas.

For processes in family $\Im$, we can establish *semantic correspondences*, on the basis of their corresponding similarity coefficients. Given two processes $P_i, P_j \in \Im$, two main cases can occur:

1. *Semantic equivalence*, $P_i \equiv P_j$, that holds if $ESim(P_i, P_j) = 2$ and $FSim(P_i, P_j) = 3$. Semantic equivalent processes represent the same real-world activity, that is, they operate on the same involved data and perform the same functionality. In this case, we can recognize process replication.
2. *Semantic relationship*, $P_i \sim P_j$, that holds if $T_2 \leq ESim(P_i, P_j) < 2$ and $T_3 \leq FSim(P_i, P_j) < 3$, where T_2 and T_3 are similarity thresholds. Processes with a semantic relationship represent partially overlapping real-world activities, that is, activities performing similar functionalities on similar data. In this case, we envisage process overlapping, such as in the case of P_1 and P_2, which pertain to different organization units. The semantic relationship suggest a possible unification of the activities for employee absence recording in the Labour Ministry units.

5 Organizational structure perspective

Processes are grouped by organization unit, to understand the information flow network of the unit, and its implications. This requires the analysis of input and output entities to determine, at an aggregated level, the data flow among the separate involved processes. On the basis of the analysis of the volumes and type of the exchanges, it is possible to envisage the possible regrouping of processes to simplify or expedite this flow. In particular, the physical distribution of work has to be carefully considered in order to reflect the demand for a physical network. The *organization unit autonomy / dependency degree* parameter can be considered, in relation to the the number and type of internal and external information flows.

5.1 Organization unit autonomy / dependency degree

In order to evaluate autonomy / dependency parameters, the following coefficients are introduced:

- *Cohesion coefficient*, to analyze unit's internal coordination in process execution. The cohesion coefficient measures the relationships between processes within a unit by identifying the main routes of information flow between them. It can be a point of reference to evaluate coherence of production and utilization of information in the organizational structure.
- *Coupling coefficient*, to analyze the interactions between different units. The coupling coefficient analysis measures the relationships between a given organization unit and other units of the organization, by identifying the main external routes of information flow between them in view of communication and cooperation. It can be a point of reference to check effectiveness of the interaction with the outside environment.

To formally define cohesion and coupling coefficients, we introduce a *closeness function* $Cl()$ that takes in input two sets of entity names and returns 1 if they have affinity in the semantic dictionary, according to definition ??, and 0 otherwise.

Organization unit cohesion
Cohesion of an organization unit OU_k depends on the number of information flows exchanged between the processes of OU_k. The greater the number of exchanged information flows, the higher the organization unit cohesion. We compute the total number of information flows $IF(OU_k)$ exchanged by processes of OU_k using function $Cl()$ as follows:

$$IF(OU_k) = \sum_{i=1}^{N}\sum_{j=1}^{N} Cl(D_i.f_3, D_j.f_2), j \neq i + \sum_{i=1}^{N}\sum_{j=1}^{N} Cl(D_i.f_2, D_j.f_3), j \neq i$$

where N is the total number of processes of OU_k and D_i denotes the descriptor associated with a process $P_i \in OU_k$.

Definition 7 Cohesion coefficient. The *cohesion coefficient* of an organization unit OU_k, denoted by $Cohesion(OU_k)$, is the measure of the tightness of its processes, computed as follows.

$$Cohesion(OU_k) = \frac{IF(OU_k)}{N \cdot (N-1)}$$

The Cohesion coefficient of an organization unit is computed by considering the total number of information flows exchanged by the organization unit's processes with respect to the total number of possible information flows between the same processes. The higher the number of OU_k's inter-process information flows,

the greater its corresponding cohesion coefficient. Such a cohesion coefficient is a nonnegative, normalized coefficient for a given organization unit OU_k (i.e., it is independent of the number of processes in OU_k), as required for cohesion measures [4]. Cohesion coefficient is used to assess the coherence of information flows. High values of $Cohesion(OU_k)$ (e.g., ≥ 0.5) indicate that several routes of information flows exist between OU_k's processes. This can be used as an indicator of utilization of information within the organization unit.

For example, with reference to our experimentation on the Labour Ministry, in Table 1, we show the cohesion coefficients we have computed for 7 organization units of this Ministry. As we can note from the table, we have organization units with rather high cohesion coefficients but also organization units characterized by very low cohesion values. For these units, a further analysis based on their workload can be applied, to evaluate activity restructuring within the Labour Ministry.

OU_k	N	$Cohesion(OU_k)$
General Management of General Affairs and Personnel (OU_1)	47	$274 + 243/2162 = 0.24$
General Management of Cooperation (OU_2)	22	$123 + 199/462 = 0.52$
General Management of Employment (OU_3)	17	$9 + 11/272 = 0.073$
General Management of Observatory of Labour Market (OU_4)	14	$32 + 38/182 = 0.38$
General Management of National Insurance and Social Welfare (OU_5)	34	$9 + 11/272 = 0.14$
General Management of Labour Relations (OU_6)	25	$22 + 24/600 = 0.076$
Central Office for Employee Orientation and Training (OU_7)	34	$34 + 38/210 = 0.34$

Table 1. Cohesion coefficients of the organization units of the Labour Ministry

Organization unit coupling

The effectiveness of an organization unit depends on its level of coupling with outside (i.e., other organization units). Remember that, for a given process P_i, $IN(P_i)$ ($OUT(P_i)$) denotes the set of names of the input (output) entities of P_i. To formally define a coupling measure, we define the following sets. Let $IN(OU_k)$ be the set of the input entities of OU_k, that is, $IN(OU_k) = \bigcup IN(P_i), i = 1, \ldots, N$. Let $OUT(OU_k)$ be the set of the output entities of OU_k, that is, $OUT(OU_k) = \bigcup OUT(P_i), i = 1, \ldots, N$, where N is the total number of processes of OU_k. Let IN_k (respectively, OUT_k) be the complement of $IN(OU_k)$ (respectively, $OUT(OU_k)$), that is, $IN_k = \bigcup IN(OU_p), p = 1, \ldots, M, p \neq k$ (respectively, $OUT_k = \bigcup OUT(OU_p), p = 1, \ldots, M, p \neq k$), where M is the total number of considered organization units. The coupling coefficient of an organization unit OU_k is defined as follows.

Definition 8 Coupling coefficient. The *coupling coefficient* of an organization unit OU_k, denoted by $Coupling(OU_k)$, measures the amount of relationships with outside due to incoming and outgoing information flows as follows.

$$Coupling(OU_k) = \sum_{i=1}^{T}\sum_{j=1}^{S} Cl(\{n_i\},\{n_j\}) + \sum_{h=1}^{P}\sum_{q=1}^{Q} Cl(\{n_h\},\{n_q\})$$

where $n_i \in IN(OU_k), n_j \in OUT_k, n_h \in OUT(OU_k), n_q \in IN_k$ and T, S, P, Q indicate the cardinality of sets $IN(OU_k), OUT_k, OUT(OU_k)$, and IN_k, respectively.

The coupling coefficient of an organization unit is evaluated by considering both the total number of incoming information flows from external organization units and the total number of outgoing information flows to external organization units, using function $Cl()$. These flows are identified by comparing the input (respectively, output) entities of the involved unit OU_k with the output (respectively, input) entities of remaining units of the organization. High values of coupling for a given unit OU_k indicate that its processes need several information exchanges with outside in order to perform their work. For a more precise assessment of the effectiveness of an organization unit, we compute a measure of the relevance of the external information flows with respect to the internal ones, denoted by $Coupling'(OU_k)$, as follows:

$$Coupling'(OU_k) = \frac{Coupling(OU_k)}{IF(OU_k)}$$

High values of this measure indicate a strong inter-dependency with critical distribution of processes among organization units, that requires a more complex inter-unit information flow network. Understanding the information flow network and its implications requires the analysis of input and output entities to determine, at an aggregated level, the data flow among the separate involved processes. This analysis can be performed with the help of the semantic dictionary, by exploiting the entity concept hierarchy therein contained. On the basis of the analysis of the volumes and type of the exchanges, it is possible to envisage the possible regrouping of processes to simplify or expedite this flow improving global effectiveness. In particular, the physical distribution of work has to be carefully considered in order to reflect the demand for a physical network.

For example, the coupling coefficients we have computed for the Labour Ministry organization units are shown in Table 2. As we can note from the table, coupling coefficients for 3 organization units is greater than 1. This can be considered an indication of a critical distribution of work among processes of these units which leads to several inter-unit information flow exchanges. By analyzing cohesion and coupling coefficients in a combined way, we can draw some considerations. For example, in the case of OU_2, the high internal cohesion is justified by the presence of several processes devoted to processing and

checking activities, performed within the unit according to established regulations. These activities are triggered by requests coming from outside, which constitute the only information flows OU_2 has with other external units. As another example, we can observe organization units characterized by high values of coupling and, at the same time, by very low values of internal cohesion, such as the **General Management of Employment** and the **General Management of Labour Relations** organization units. As a possible restructuring intervention, it should be evaluated the possibility of merging these two units into a unique unit, to reduce external coupling and augment their internal cohesion, to improve the global effectiveness of the workload.

OU_k	$Coupling'(OU_k)$
General Management of General Affairs and Personnel (OU_1)	$60 + 55/517 = 0.22$
General Management of Cooperation (OU_2)	$35 + 38/242 = 0.3$
General Management of Employment (OU_3)	$31 + 31/20 = 3.1$
General Management of Observatory of Labour Market (OU_4)	$39 + 41/70 = 1.14$
General Management of National Insurance and Social Welfare (OU_5)	$46 + 48/161 = 0.58$
General Management of Labour Relations (OU_6)	$44 + 34/46 = 1.69$
Central Office for Employee Orientation and Training (OU_7)	$31 + 31/72 = 0.86$

Table 2. Coupling' coefficients of the organization units of the Labour Ministry

6 Concluding remarks

In this paper, we have presented a approach to support a multi-perspective analysis of legacy information systems. According to a perspective related to the information system data asset structure, we can evaluate inter-schema properties looking for standardization of data assets shared by different organization units. According to a perspective related to the information system operational structure, we can evaluate inter-process semantic correspondences, in terms of the level of similarity of the involved data and operations. According to a perspective related to the information system organizational structure, we can evaluate parameters related to the degree of autonomy / dependency of organization units. To evaluate data and operation similarity independently of possible semantic heterogeneities that can arise between descriptions of different processes, a semantic dictionary organized into concept hierarchies has been defined.

Following an information processing view, the proposed approach focuses on the information structures and the information flows between processes / organization units to discover possible inconsistencies, redundancies, or replications

which can affect the workload and the autonomy of each organization unit. In the PROGRESS project, extensions are planned to analyze also execution modalities of single processes and communication protocols between them, by employing workflow specifications [6]. With these extensions, work structure and workflows can be analyzed in detail to discover possible inefficiency and failures.

Acknowledgments
The authors thank prof. B. Pernici for initial discussions on the analysis perspectives.

References

1. J.D. Ahrens, N.S. Prywes, "Transition to a Legacy- and Reuse- Based Software Life Cycle", *IEEE Computer*, October 1995.
2. P. Aiken, A. Muntz, R. Richards, "DoD Legacy Systems - Reverse Engineering Data Requirements", *Communications of the ACM*, Vol.37, No.5, May 1994.
3. C. Batini, S. Castano, V. De Antonellis, M.G. Fugini, B. Pernici, "Analysis of an Inventory of Information Systems in the Public Administration", *Requirements Engineering Journal*, Vol.1, N.1, 1996, Springer.
4. L.C. Briand, S. Morasca, V.R. Basili, "Property-Based Software Engineering Measurement", *IEEE Trans. on Software Engineering*, Vol.22, No.1, January 1996, pp.68-86.
5. G. Canfora, A. Cimitile and U. de Carlini, "A Logic-Based Approach to Reverse Engineering Tools Production", *IEEE Trans. on Software Engineering*, Vol. 18, No.12, December 1992, pp. 1053-1064.
6. F. Casati, S. Ceri, B. Pernici, G. Pozzi, "Conceptual Modeling of Workflows", in *Proc. of OO-ER'95, Int. Conf. on the Object-Oriented and Entity-Relationship Modelling*, Gold Coast, Australia, December 1995, Springer Verlag, pp.341-354.
7. S. Castano, V. De Antonellis, "Reference Conceptual Architectures for Reengineering Information Systems", *International Journal of Cooperative Information Systems*, Vol.4, Nos.2&3, 1995, pp.213-235.
8. S. Castano, V. De Antonellis, "Techniques for Process Analysis and Unification", in Proc. of *CAiSE'96*, Crete, Greece, May 1996, Springer Verlag, pp. 234-254.
9. V. De Antonellis, B. Zonta, "A disciplined Approach to Office Analysis", *IEEE Trans. on Software Engineering*, Vol.16, No.8, August 1990, pp.822-828.
10. T. Kudrass, M. Lehmbach, A. Buchmann, "Tool-Based Re-Engineering of a Legacy MIS: An Experience Report", in Proc. of *CAiSE'96*, Crete, Greece, May 1996, Springer Verlag, pp. 116-135.
11. D. Karagiannis, (Ed.), *Special Issue on Business Process Reengineering*, SIGOIS Bulletin, Vol.16, No.1, August 1995.
12. E. Yourdon, *Modern Structured Analysis*. Englewood Cliffs: Prentice-Hall Intl., 1989.