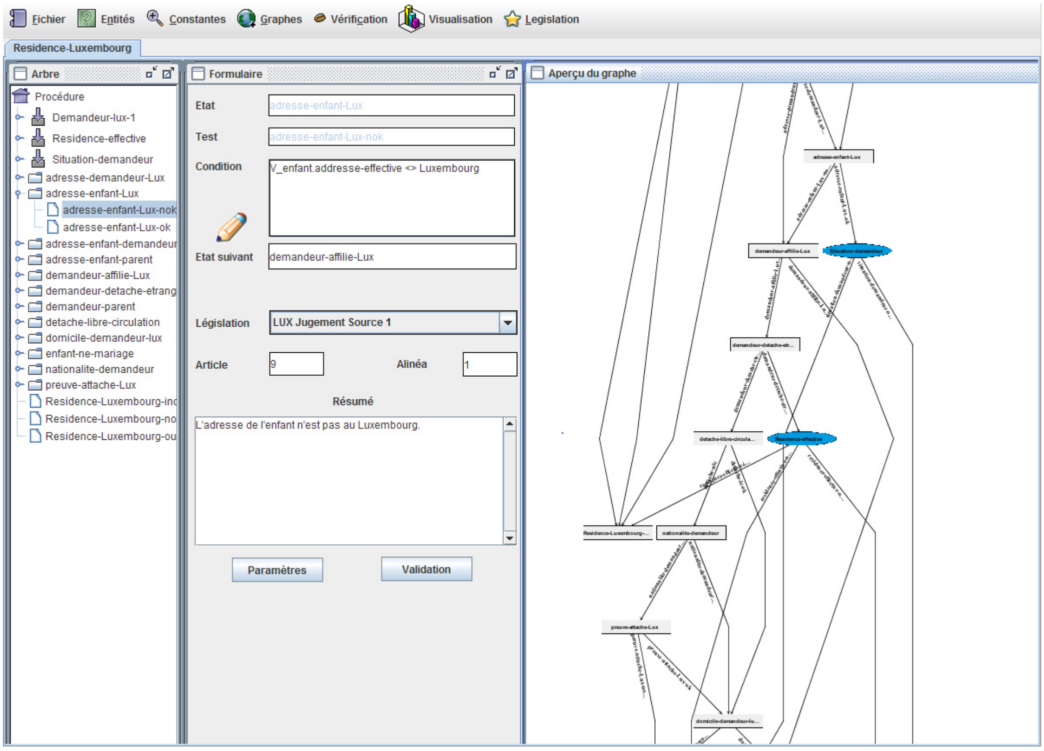
MEASURING THE IMPACT OF DATA QUALITY IN A CORPORATE DECISION-SUPPORT SYSTEM

J. Wax, B. Otjacques, T. Tamisier, O. Parisot, Y. Didry, F. Feltz

Centre de Recherche Public - Gabriel Lippmann
41, rue du Brill, L-4422 Belvaux, Luxembourg
{wax,otjacque,tamisier,parisot,didry,feltz}@lippmann.lu

The benefit of decision support systems (DSS) is in practice frequently impaired by the processing of unsure and erratic data [1]. We present an experience of using Cadral, a business DSS developed at our department, for processing applications received by the administration of Family Benefits in Luxembourg. To better adapt to the operational context [2], Cadral includes informative tools that help the user take into account the quality of the data when modeling the knowledge of the system.

Cadral is a collaborative DSS: it consists of a Knowledge Editor (KE) and a Knowledge Simulator (KS), both supporting teamwork [3]. Knowledge in Cadral is a set of interconnected administrative procedures organized in a hierarchy of decision trees as follows: we start from a unique root node; every non-terminal node denotes a test on specific data; according to the result of the test one arc leaving the node is univocally selected; terminal nodes denote an intermediate (for sub-trees) or final decision state. A display algorithm inspired from Sugiyama's [4] allows visualizing the trees.



Etat: adresse-enfant-Lux
Test: adresse-enfant-Lux-nok
Condition: V_enfant.addresse-effective <> Luxembourg
Etat suivant: demandeur-affilie-Lux
Législation: LUX Jugement Source 1
Article: 9
Alinéa: 1
Résumé: L'adresse de l'enfant n'est pas au Luxembourg.

Figure 1: Knowledge Editor with a decision tree

From operational feedback we know that the correctness of the decisions suggested by Cadral is directly correlated with the place of the tests within the tree and the quality of tested data: in other words, there is a room for improvement provided we reorder the trees according to some strategy of data invocation, as presented in [5]. As a matter of fact data originate from diversified sources that do not have the same level of accuracy, and in particular are not all in digital format. Optimal decision trees should therefore make the most of reliable data, localize hazardous items as well as limit and postpone manual handling.

Following such guidelines is nonetheless by no means obvious to business experts, in particular due to their lack of awareness about the quality of the data, which are all retrieved in the same transparent way from an Enterprise Service Bus. Hence, we have defined some metrics for evaluating data with respect to business requirements and supplemented the graphical representation of the trees by attaching to the nodes and to the arcs information based on these metrics. We use 4 metrics that evaluate an independent data item as follows.

- **Availability**: one must minimize the impact of a possibly missing or manually handled item.
- **Credibility**: an item loses trust due to different factors, in particular if not regularly updated. An untrustworthy item (e.g. and address) can directly lead to a false answer of the DSS.
- **Accuracy**: if the item is not accurate enough for a test (e.g. threshold detection), all subsequent processing can be irrelevant.
- **Cost**: cost concerns the resources (CPU, time, human handling…) needed to retrieve an item. Cost impacts considerably mass-processing and batch execution.

For the sake of simplicity, we do not detail the process of evaluating each data item retrieved by Cadral with respect to the metrics, which is left to IT specialists in charge of data integration. In addition, credibility only is set through numeric values; the other metrics use a two-levels scale (low-high). Some validation routines have also been implemented to help the characterization concerning the accuracy metric.

Visual indications are helpful means to ease the legibility of graphs of significant size [6]. Moreover, specific patterns should be dedicated to data quality [7]. In Cadral, we use 3 patterns to show details on data quality measured thanks to the 4 metrics.

- specific color on nodes with low-quality data
- specific icons attached to the nodes for reporting quality data problems
- specific color on arcs to follow the propagation in the decision trees of the impact of low-quality data.

In addition, relevant messages on data quality are displayed in help and modeling tools (e.g. when selecting variable in a test).

Figure 2 illustrates different informative patterns on decision trees. Depending on the quality of retrieved data, errors or uncertainties about the results of the decision tree are pointed out (see left part). Warnings for potentially inaccurate or unsure results appear in orange. Parts of the tree impossible to process (e.g. due to missing data) appear in red. Credibility is notified in a red gauge, with complementary numeric percentage available in a tool tip (see right part).

Before going into production, the modified Knowledge Editor has been proofed with a group of 10 pilot users. We have designed a simplified test problem to be programmed in Cadral, that does not require any domain-related expertise: the problem is to decide whether a golf game must be played, according to weather conditions. We have defined a set of 10 decision rules that imply problematic or missing data items. The goal was to find a procedural model that delays as long as possible checks on unsure data, what we call here an *optimal model*. Details of the test and the results will be discussed during the presentation. They highlight the following trends:

- Most users are able (with some help) to find an optimal model.
- Visual information has been correctly interpreted, and was necessary: the majority of the users mentioned that they should not have been able to optimize the model without
- The main difficulty was, as far as unsure data is concerned, to take into account the importance of the item in the model.

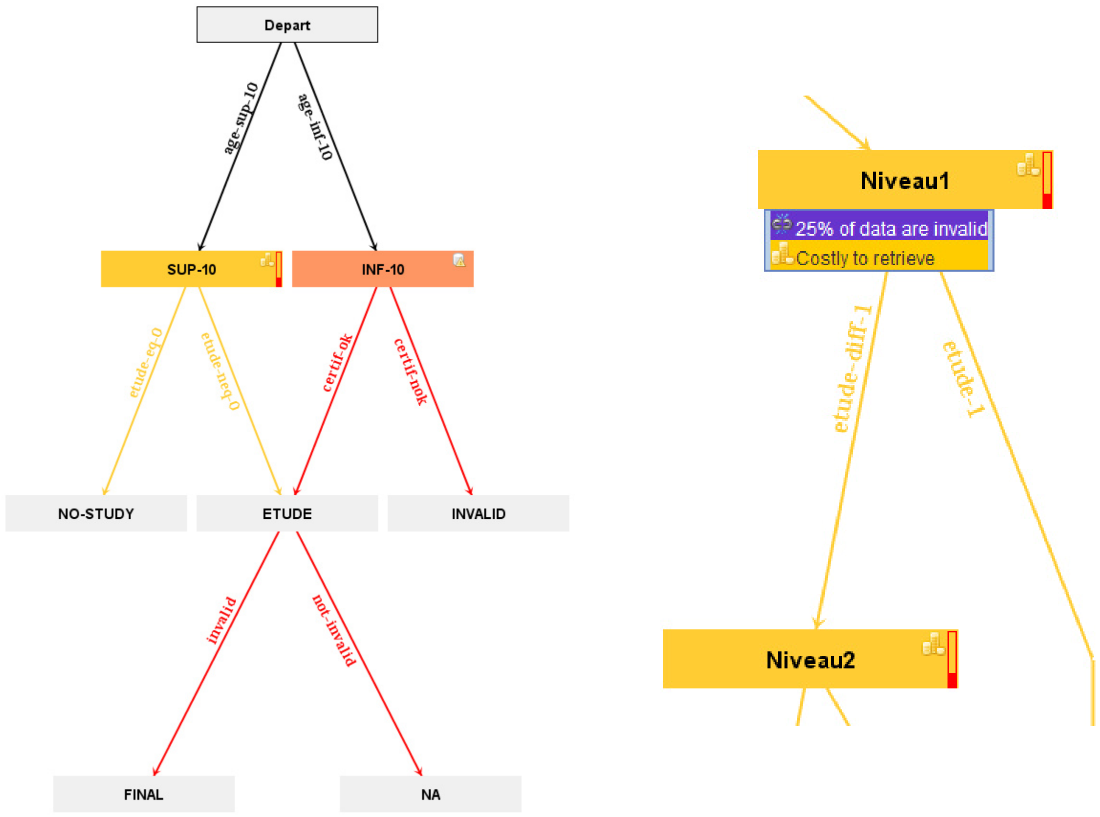


Figure 2: details and impact of data quality

This practical study opens the way to 2 directions of improvements. First, the automatic reordering of the decision trees to minimize the impact of low-quality data. Second, in addition to processing application files, Cadral is also used for socio-economic prediction with respect to projected demographic data or administrative and legal framework: the second research direction is to evaluate the model on real data, in order to speed up prediction results by discarding as much as possible unsure data.

References

[1] T. Redman, "The impact of poor data quality on the typical enterprise", *Communications of the ACM*, 1998

[2] V. Asproth, "Visualisation of data quality in decision-support systems", *International Journal of Applied Systemic Studies*, v. 1, n.3, p. 280-289, 2007.

[3] T. Tamisier, and al, "A Collaborative Reasoning Maintenance System for a Reliable Application of Legislations", *Proc. of the 6th Int'l CDVE Conference*, LNCS, Springer, 2009.

[4] K. Sugiyama, S. Tagawa, and M. Toda, "Methods for Visual Understandings of Hierarchical System Structures", *IEEE Transactions in Systems, Man, and Cybernetics*, v. smc-11, n.2, p. 109-125, 1981.

[5] A. Even, and G. Shankaranarayanan, "Utility-Driven Assessment of Data Quality", *ACM SIGMIS Database*, 2007

[6] R. Wang, H. Kon, and S. Madnick, "Data Quality Requirements Analysis and Modeling", *Journal of Management Information Systems*, 1993.

[7] C. Hao, D. Keim, U. Dayal, and J. Shneidewing, "Business process impact visualization and anomaly detection", *Information Visualization*, 2006